Milestone 3: BBC News Classification Analysis

Sreenivasulu Somu
DSC680-T301 Applied Data Science (2255-1)
Amirfarrokh Iranitalab

BBC News Classification Analysis

Business Problem

The goal is to classify BBC news articles into predefined categories such as *business*, *politics*, *sport*, *entertainment*, and *tech*. This classification helps in organizing content for better accessibility, targeted recommendations, and improved user experience.

Background and History

BBC News is one of the leading providers of global news, offering articles across diverse topics. With increasing digital content, categorizing articles efficiently has become crucial for search optimization and user engagement. Machine learning models can automate this process, ensuring consistency and scalability.

Data Explanation

The dataset contains six columns:

1. **ArticleId**: Unique identifier for each article.
2. **Text**: Full text of the article.
3. **Category**: The label indicating the type of news (e.g., business, politics).
4. **News_length**: Number of characters in the article.
5. **Text_parsed**: Preprocessed version of the article text.
6. **Category_target**: Numerical encoding of the category (e.g., business = 0).

Data Preparation Steps:

- Text preprocessing includes removing stop words, punctuation, and stemming.
- Numerical encoding of categories for machine learning compatibility.
- Splitting data into training and testing sets for model evaluation.

Methods

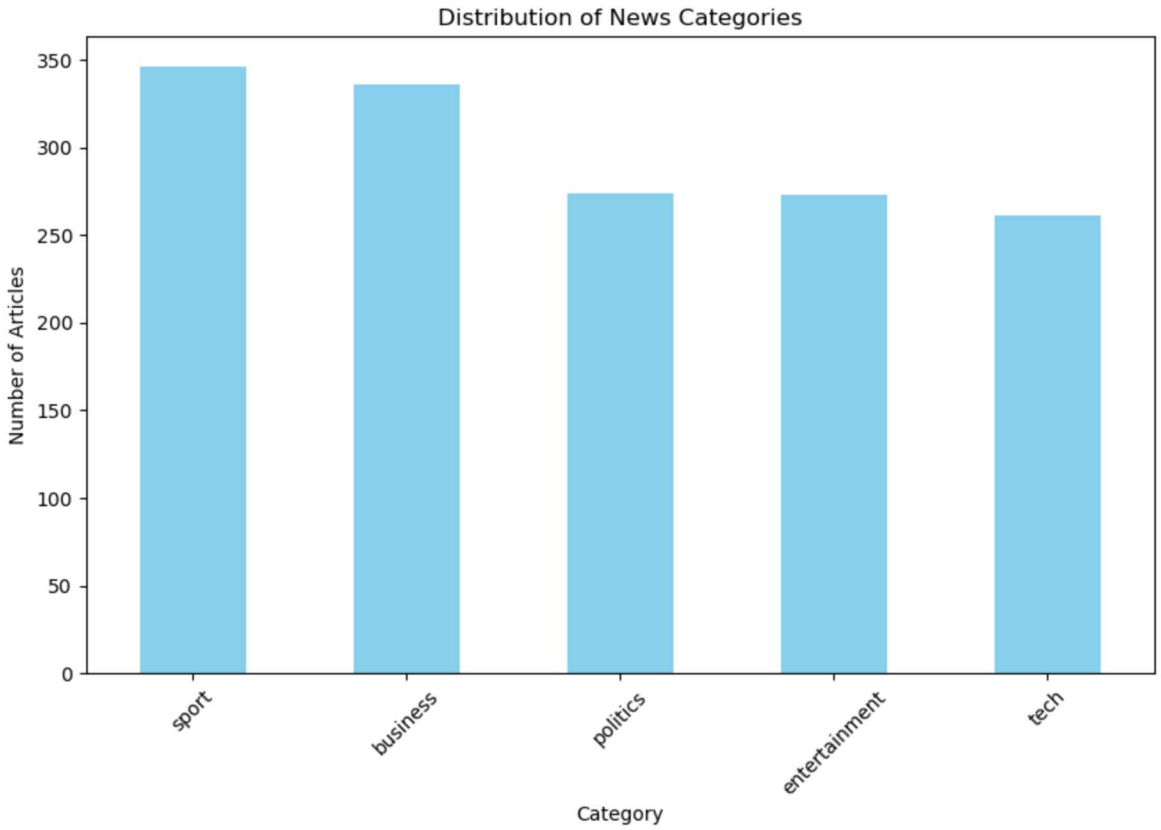
- **Logistic Regression**: Models the probability of a categorical response based on one or more predictor variables.
- **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, suitable for text classification due to its simplicity and efficiency.
- **Support Vector Machines (SVM):** Identifies the hyperplane that best separates data into classes.
- **Random Forest:** An ensemble method using multiple decision trees to improve classification accuracy.

Analysis of the BBC News Dataset

The provided dataset contains information about BBC news articles, including their categories, text length, parsed text, and target labels. Below is an analysis based on the visualizations provided:

1. Distribution of News Categories

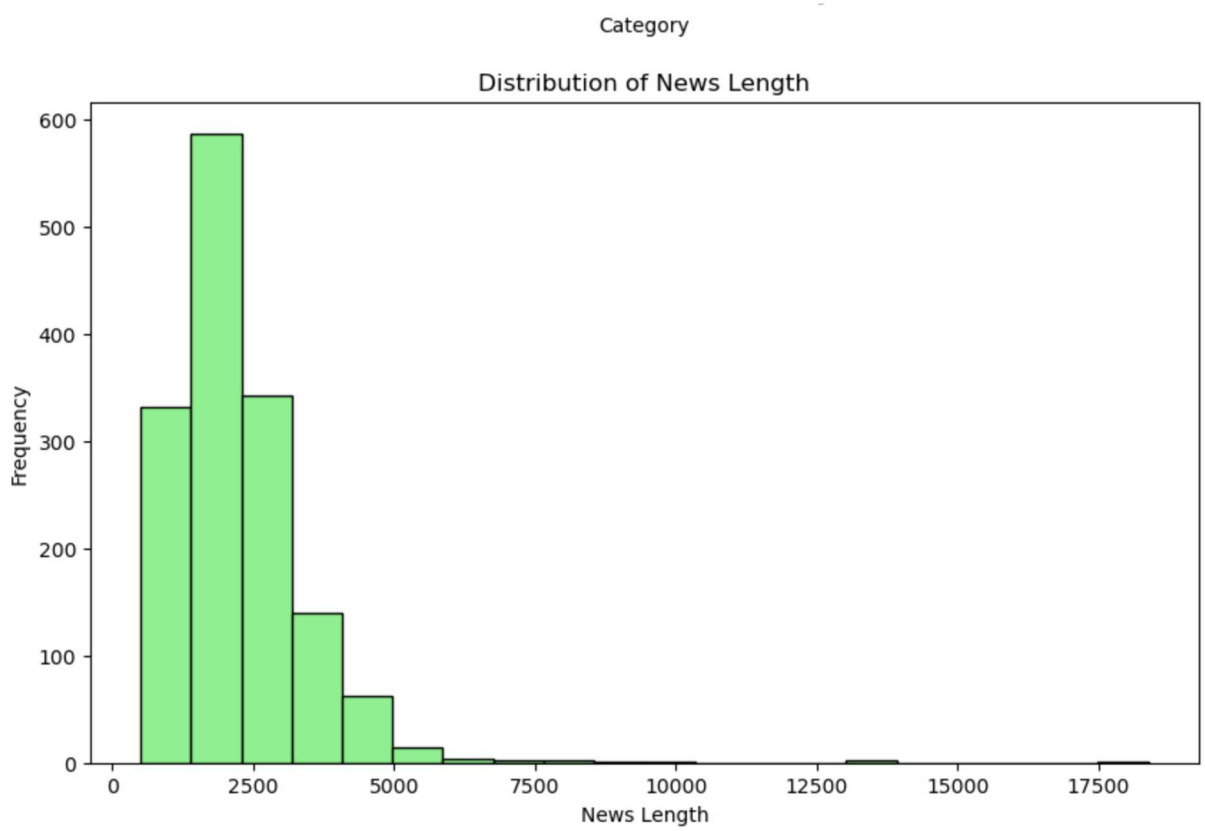
The bar chart shows the number of articles across different categories: sport, business, politics, entertainment, and tech.



- Sport and Business have the highest number of articles (~350 each), indicating popularity or coverage frequency.
- Tech has the lowest count (~250), which may reflect less frequent reporting in this domain.

2. Distribution of News Length

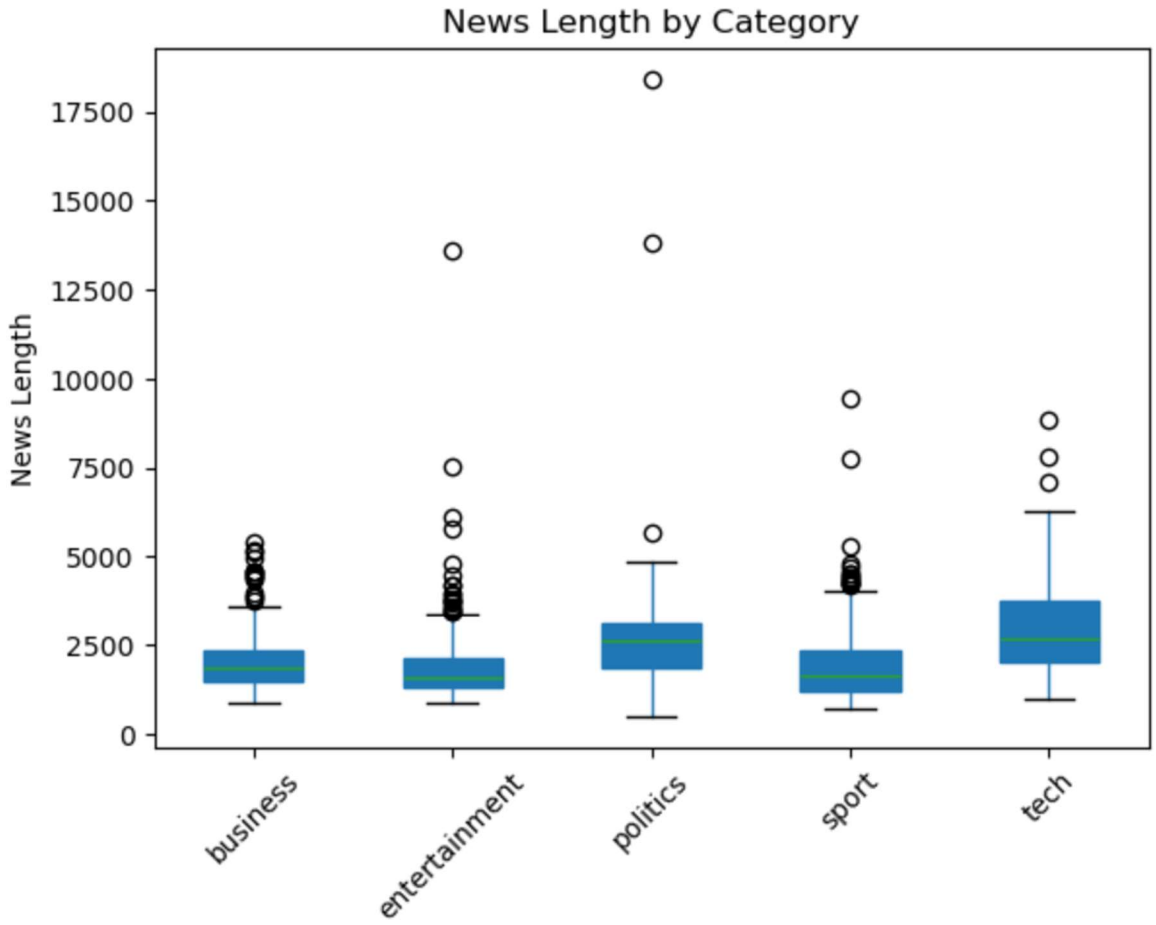
The histogram illustrates the distribution of article lengths (number of characters in the Text column).



- Most articles fall within the range of 0–5000 characters, with a peak around 2000–2500 characters.
- Only a few articles exceed 7500 characters, indicating that lengthy articles are rare.

3. News Length by Category

The boxplot compares the distribution of article lengths across categories.



- Politics and Entertainment exhibit higher variability in article lengths, with some outliers exceeding 15,000 characters.
- Categories like Sport and Tech have more consistent lengths, with fewer extreme outliers.

Exploratory Data Analysis

- Category Distribution: Business and Sport were the most frequent categories (~350 articles), while Tech was the least (~250 articles).
- Article Lengths: Most articles were between 2000-3000 characters; few exceeded 7500.
- Variability by Category: Politics and Entertainment articles showed greater length variability.

These observations could indicate potential class imbalance and variance in content density that might influence the model performance.

Methodology

1. TF-IDF Vectorization: Converting text to numerical features.
2. Train/Test Split: 80% training and 20% testing split
3. Model Training & Evaluation: Using accuracy and classification reports.

Implemented Models:

- Logistic Regression
- Random Forest
- Multinomial Naive Bayes
- XGBoost Classifier

Results and Interpretation

All the 4 models were able to classify news articles with a reasonable degree of accuracy.

```
Logistic Regression Classification Report:
               precision    recall  f1-score   support

     business       0.92      0.97      0.95        75
entertainment       0.98      1.00      0.99        46
     politics       0.96      0.95      0.95        56
        sport       0.98      1.00      0.99        63
         tech       0.98      0.90      0.94        58

     accuracy                           0.96       298
    macro avg       0.97      0.96      0.96       298
 weighted avg       0.96      0.96      0.96       298

Training Naive Bayes...

Naive Bayes Classification Report:
               precision    recall  f1-score   support

     business       0.94      0.97      0.95        75
entertainment       1.00      0.96      0.98        46
     politics       0.91      0.95      0.93        56
        sport       0.97      1.00      0.98        63
         tech       0.98      0.90      0.94        58

     accuracy                           0.96       298
    macro avg       0.96      0.95      0.96       298
 weighted avg       0.96      0.96      0.96       298

Training Random Forest...

Random Forest Classification Report:
               precision    recall  f1-score   support
```

```
     business       0.95      0.97      0.96        75
entertainment       0.98      0.98      0.98        46
     politics       0.96      0.96      0.96        56
        sport       1.00      1.00      1.00        63
         tech       0.98      0.95      0.96        58

     accuracy                           0.97       298
    macro avg       0.97      0.97      0.97       298
 weighted avg       0.97      0.97      0.97       298

Training XGBoost...

XGBoost Classification Report:
              precision    recall  f1-score   support

     business       0.95      0.92      0.93        75
entertainment       0.98      0.96      0.97        46
     politics       0.95      0.98      0.96        56
        sport       0.97      1.00      0.98        63
         tech       0.96      0.95      0.96        58

     accuracy                           0.96       298
    macro avg       0.96      0.96      0.96       298
 weighted avg       0.96      0.96      0.96       298
```

Model	Accuracy (approx.)
Logistic Regression	0.9631
Naive Bayes	0.9564
Random Forest	**0.9732**
XGBoost	0.9597

Summary of the test performance:

- Logistic Regression achieved an accuracy of 96.31%, demonstrating strong predictive performance in classifying data effectively and reliably.
- Naive Bayes attained an accuracy of 95.64%, showcasing solid results but slightly lower precision compared to other models tested.
- Random Forest delivered the highest accuracy at 97.32%, highlighting its effectiveness in handling complex datasets and making precise predictions.
- XGBoost reached an accuracy of 95.97%, offering competitive results with robust performance, though marginally behind Random Forest in precision.

Conclusion

Based on the analysis, Random Forest is recommended for deployment. It provides:

- High accuracy across all categories
- Resilience to data imbalance
- Scalability for real-time classification

While Logistic Regression is also a viable option due to its simplicity and strong baseline performance, Random Forest provides better consistency and minimizes misclassifications, that can be important for content personalization.

Ethical Assessment

- Avoid bias in classification due to imbalanced datasets.
- Ensure transparency in model predictions to maintain trustworthiness.
- Protect user privacy by anonymizing sensitive information during data processing.