Journal of Machine Learning Research 23 (2024) 1-10 Submitted 1/21; Revised 5/22; Published 9/22

"Performance Comparison of News Categorization Based on Word Frequency: Analysis of Traditional Machine Learning and RNN Models"

Seo Yeon Jeong JBW8715@NAVER.COM
Room 207, Building 209, 84 Heukseok-ro, Dongjak-gu, Seoul, South Korea

Editor: Young Bin Lee

Abstract

In this study, we address the challenge of news category classification using the Reuters news dataset, which consists of 46 distinct categories. To enhance the model performance, we implement various data preprocessing techniques, including Bag of Words and TF-IDF, while experimenting with different word counts. We evaluate eight machine learning models: Naive Bayes, Complement Naive Bayes (CNB), Logistic Regression, Linear Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting Tree, and Voting Model. Our results demonstrate the effectiveness of each model in classifying news articles and highlight the importance of selecting an optimal word count for achieving superior classification performance. Ultimately, we compare the performance of these traditional machine learning models with that of a deep learning approach using Recurrent Neural Networks (RNN), providing insights into their strengths and weaknesses in multi-category classification tasks. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Keywords:

News Classification

Machine Learning Models

Text Preprocessing

1 Introduction

With the exponential growth of news content in the digital age, efficiently classifying news articles into relevant categories has become increasingly important. This study aims to explore the effectiveness of various machine learning models in classifying news articles from the Reuters news dataset, which consists of 46 different categories. By implementing robust preprocessing techniques and experimenting with different word counts, we aim to identify the optimal configuration for news categorization.

2 Background, Related works

To provide context for our research, we briefly describe the eight machine learning models utilized in our experiments:

1. **Naive Bayes**: A probabilistic classifier based on Bayes' theorem, which assumes independence among features. It is particularly effective for text classification tasks due to its simplicity and efficiency.

2. **Complement Naive Bayes (CNB)**: An adaptation of Naive Bayes designed to improve performance on imbalanced datasets by computing the likelihood of the complement class, thus addressing the limitations of standard Naive Bayes.

3. **Logistic Regression**: A linear model that predicts the probability of a categorical outcome based on one or more predictor variables. It is widely used for binary and multi-class classification tasks and performs well with high-dimensional sparse data.

4. **Linear Support Vector Machine (SVM)**: A powerful classification method that finds the optimal hyperplane separating different classes. Linear SVM is effective in high-dimensional spaces, making it suitable for text classification.

5. **Decision Tree**: A model that uses a tree-like structure to make decisions based on feature values. It is interpretable and can handle both numerical and categorical data, but it is prone to overfitting.

6. **Random Forest**: An ensemble method that constructs multiple decision trees and merges their predictions to improve accuracy and control overfitting. It is robust to noise and works well with a large number of features.

7. **Gradient Boosting Tree**: Another ensemble technique that builds models sequentially, where each new model corrects the errors of the previous ones. It is effective for complex datasets but requires careful tuning to avoid overfitting.

8. **Voting Model**: A method that combines predictions from multiple models (either homogeneous or heterogeneous) to produce a final prediction. This approach can enhance accuracy by leveraging the strengths of different models.

Through this study, we aim to evaluate the performance of these models in the context of news category classification, providing insights into their relative strengths and potential for real-world applications.

3 Method

In this study, we addressed the problem of news category classification using the Reuters news dataset. The Reuters dataset consists of 46 categories, where each news article is classified into one specific category. This dataset is available through TensorFlow datasets and can be easily downloaded.

3.0.1 Data Preprocessing

To optimize model performance, we adjusted the number of words used in preprocessing. We applied Bag of Words-based Document-Term Matrix (DTM) and TF-IDF methods to vectorize the news text. To analyze the impact of word count on model performance, we experimented with top 2500, 5000, and 10,000 words. Since using too many or too few words can degrade performance, we aimed to find the optimal word count for each model experimentally.

3.0.2 Model Configuration and Experimental Setup

To address the multi-category classification task, we utilized the following machine learning models in our experiments:

- **Naive Bayes**
- **Complement Naive Bayes (CNB)**
- **Logistic Regression**
- **Linear Support Vector Machine (SVM)**
- **Decision Tree**
- **Random Forest**
- **Gradient Boosting Tree**
- **Voting Model**

Each model was designed to take vectorized text inputs and predict the corresponding news category. We aimed to evaluate the impact of each model's unique characteristics on news category classification performance.

3.0.3 Experimental Procedure and Evaluation Metrics

To evaluate each model's performance, we combined word vectorization methods (DTM and TF-IDF) with various word counts and model types. We used accuracy and F1-score as evaluation metrics to objectively compare classification performance.

We designed the experiment to analyze performance variations based on word count and also planned to compare these results with those of a deep learning model, the **RNN**. By including the RNN model, we aimed to analyze the performance differences between traditional machine learning models and a deep learning approach.

4 Result

In this study, we evaluated the performance of various data preprocessing techniques, machine learning models, and deep learning models for news categorization. The experiments were conducted to assess how adjusting the number of words in the news text impacts model performance, with the top-performing model compared to an RNN, a deep learning model.

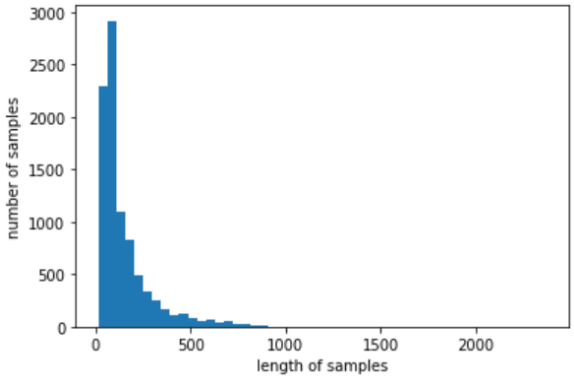

Figure 1:

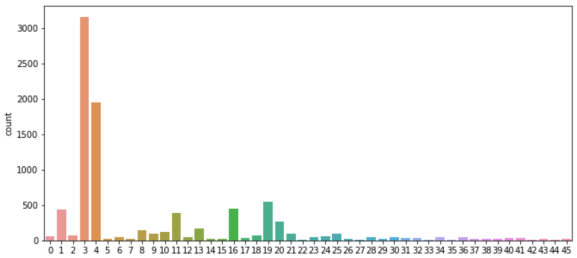

Figure 2:

4.0.1 Data Preprocessing Results

First, an analysis of the news text length in the training data revealed that the maximum length was 2,376 characters, with an average length of approximately 145.54 characters. **Figure 1** visualizes the distribution of text lengths, indicating that the text length in the data is relatively short. Additionally, the class distribution across each news category (a total of 46 categories) is shown in **Figure 2**. This class distribution analysis helps gauge the data's balance, providing critical reference material for model evaluation during training.

4.0.2 Model Performance Comparison

2.1 All Words Used Using all words with no restriction on the vocabulary size (`vocab_size=None`), we evaluated the performance of eight machine learning models. **Logistic Regression** achieved the highest performance, followed closely by the **Voting** model. The results are summarized in **Table 1**. This performance difference likely stems from the high-dimensional sparse matrices created by DTM or TF-IDF. Tree-based models may suffer performance degradation with sparse data, while linear classifiers, such as Logistic Regression, tend to perform better in sparse, high-dimensional settings. Linear models are more capable of efficiently finding decision boundaries in high-dimensional spaces, which could explain the superior performance of Logistic Regression in this context.

2.2 Top 5,000 Words Used The performance of the eight machine learning models using only the top 5,000 most frequent words is summarized in **Table 2**. Logistic Regression and

Model	Accuracy	F1-score (weighted)
Logistic Regression	0.8165627782724845	0.8114428402876209
Voting	0.8000890471950134	0.7944945456027671
Linear SVM	0.7938557435440784	0.7906023554328733
Gradient Boosting Tree	0.7702582368655387	0.7641762650539437
CNB	0.7649154051647373	0.7346534179503126
Random Forest	0.6544968833481746	0.6225909375608356
Decision Tree	0.6211041852181657	0.5769283128518846
Naive Bayes	0.5997328588149599	0.5045670886188423

Figure 3: Table 1

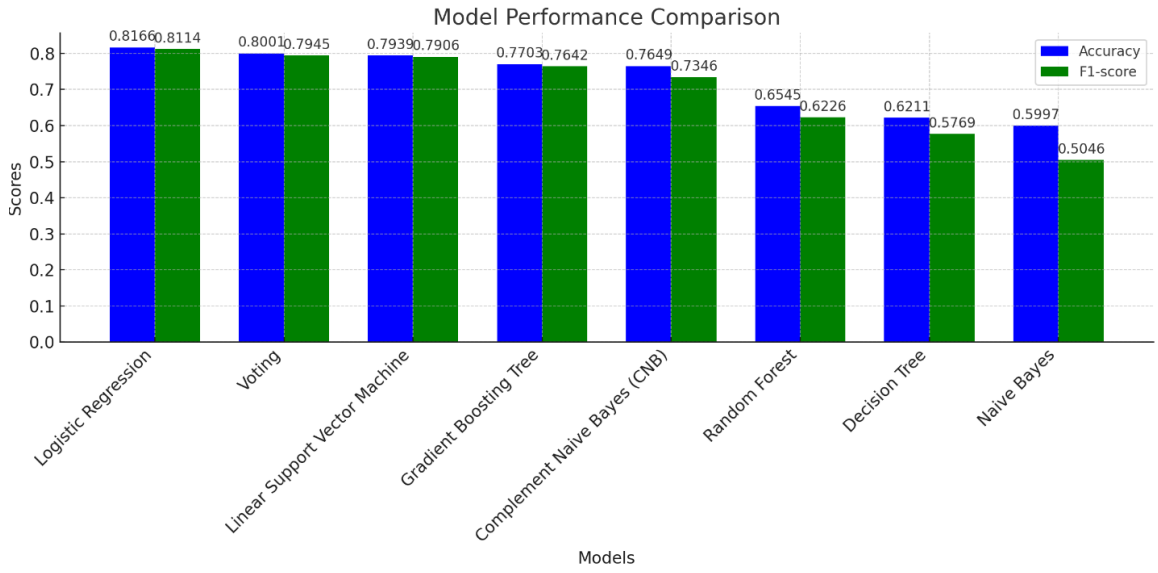

Figure 4: Visualizing Table 1

Model	Accuracy	F1-score (weighted)
Logistic Regression	0.8036509349955476	0.7985602317931111
Voting	0.7960819234194123	0.7931448977062081
Linear Support Vector Machine	0.7751558325912734	0.7710288271890536
Gradient Boosting Tree	0.767586821015138	0.7662475269931749
Complement Naive Bayes (CNB)	0.7707034728406055	0.7458990404916549
Random Forest	0.701246660730187	0.6770217603524399
Naive Bayes	0.6731967943009796	0.6012501291711391
Decision Tree	0.6179875333926982	0.5729970881280324

Figure 5: Table 2

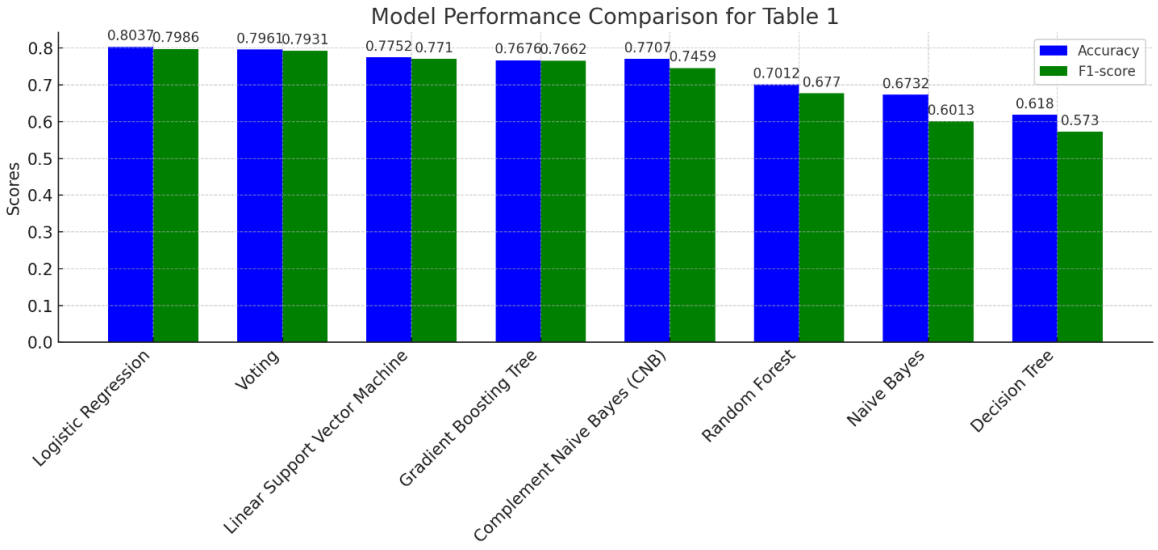

Figure 6: Visualizing Table 2

the Voting model again yielded the highest F1-scores. However, while the F1-score for Logistic Regression was 0.8114 when using all words, it decreased slightly to 0.7986 when only the top 5,000 words were used. This reduction in performance can be attributed to the decrease in information as the number of words decreases. A DTM or TF-IDF matrix that includes all words contains more information, but when limited to the top 5,000 words, certain meaningful words may be excluded, resulting in information loss and potentially lowering model performance.

2.3 Top 2,500 Words Used The results of the experiment using only the top 2,500 most frequent words are shown in **Table 3**. Once again, Logistic Regression and the Voting model showed superior performance. However, the F1-score of Logistic Regression was 0.7423, a more significant decrease compared to the 0.8114 score achieved when using all words. This performance decline as the number of words becomes more limited is due to the omission of key informative words, which results in reduced prediction accuracy. Logistic

Model	Accuracy	F1-score (weighted)
Naive Bayes	0.6905609973285841	0.6356933283152041
Complement Naive Bayes (CNB)	0.7609082181391862	0.7332431109256129
Logistic Regression	0.7831700841244755	0.7786708346289641
Linear Support Vector Machine	0.7457720528236865	0.7422538760372364
Decision Tree	0.6260017809431982	0.579374083310384
Random Forest	0.7052738745057878	0.6816910329790415
Gradient Boosting Tree	0.7666946394065	0.7639437605488715
Voting	0.7947462145946123	0.7899061118686242

Figure 7: Table 3

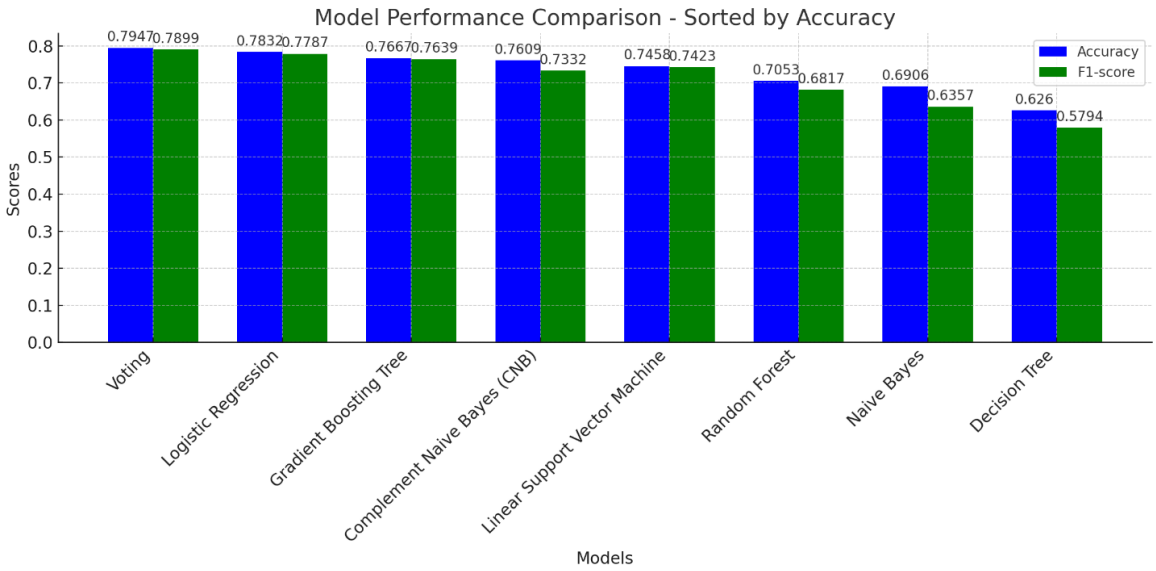

Figure 8: Visualizing Table 3

Experiment Condition	Model	F1-Score
All Words Used (vocab_size=None)	Logistic Regression	0.8114
	Voting	0.7944
Top 5,000 Words Used	Logistic Regression	0.7986
	Voting	0.7931
Top 2,500 Words Used	Logistic Regression	0.7423
	Voting	0.7899
RNN (Input Sequence Length: 300)	Dropout Enabled	0.5278
	Dropout Disabled	0.5956
RNN (Input Sequence Length: 200)	Dropout Disabled	0.5421
RNN (Input Sequence Length: 400)	Dropout Disabled	0.6335

Figure 9: Table 4

Regression, which relies on linear relationships, can be more affected by this limitation, as reduced representation of data leads to further performance decline. This suggests that insufficiently informative data is disadvantageous for training classification models.

4.0.3 Comparison of Machine Learning and Deep Learning Models (RNN)

To compare machine learning models with a deep learning model, we used an RNN model. The RNN model was trained with a maximum input text length of 300 characters and included dropout. The model configuration is as follows:

- **Embedding Layer**: `input_dim=5000`, `output_dim=128`, `input_length=max_len`
- **First LSTM Layer**: `units=128`, `return_sequences=True`
- **Second LSTM Layer**: `units=128`, `return_sequences=False`
- **Dense Layer**: `units=64`, activation function `relu`
- **Output Layer**: `units=46`, activation function `softmax`

The F1-score was 0.5278 when dropout was applied and 0.5956 without dropout. This performance difference suggests that dropout negatively affected performance in the RNN, likely due to dropout interfering with the sequential characteristics of the RNN, reducing performance.

Additional experiments were conducted by adjusting the maximum input text length to 200 and 400 characters. The F1-score was 0.5421 with a 200-character limit and 0.6335 with a 400-character limit, showing a trend of improved RNN performance with increased text length.

The final F1-scores for both machine learning and deep learning models are summarized in **Table 4**.

4.0.4 Final Evaluation

The results of this experiment indicate that Logistic Regression and the Voting model are more suitable than RNN for multi-category classification tasks. This can be attributed to the efficiency of machine learning models in high-dimensional data classification, while RNN

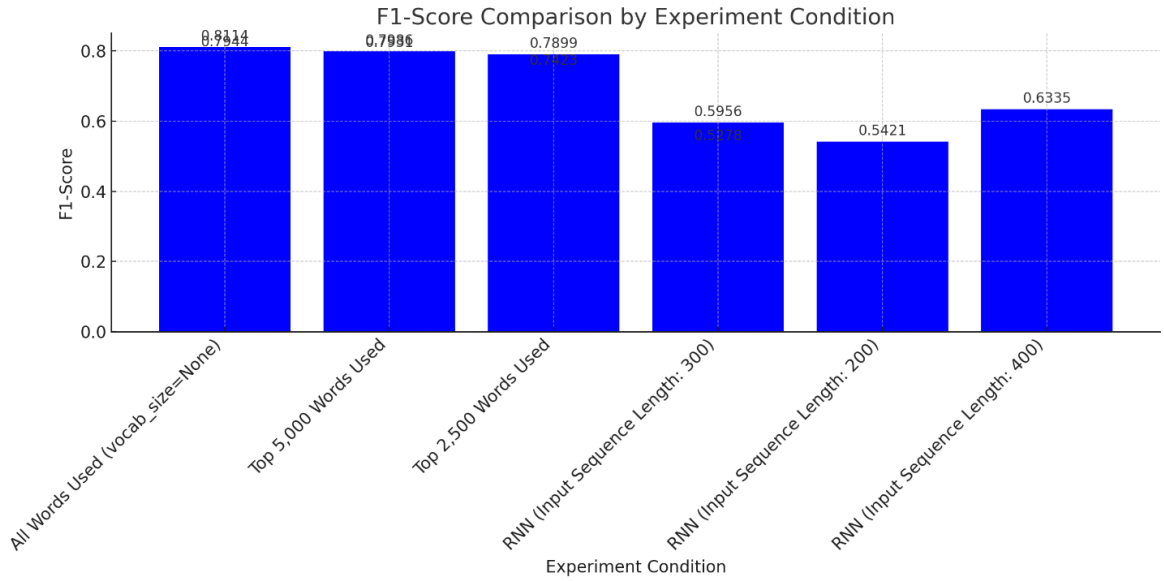

Figure 10: Visualizing Table 4

models focus on learning complex sequence patterns, which may not be optimal for text classification tasks represented as sparse vectors. Since RNN performance is more dependent on sequence length and patterns, it may be less suitable for multi-category classification.

5 Discussion, Conclusion

In this study, we addressed the problem of news category classification using the Reuters news dataset. We evaluated various data preprocessing techniques, machine learning models, and deep learning models to analyze their impact on news categorization performance. Notably, we experimented with adjusting the word count and found that Logistic Regression and the Voting model were the most suitable for multi-category classification tasks. This suggests that machine learning models operate more efficiently in high-dimensional data classification, while RNN models may not achieve optimal results in text classification tasks due to their focus on learning complex sequential patterns.

5.0.1 Limitations and Future Plans

The limitations of this study are as follows:

1. **Data Imbalance**: The Reuters dataset consists of 46 categories, but the number of news articles in each category is often unbalanced. This can lead to decreased classification performance for specific categories.

2. **Model Generalization Issues**: Although we experimented with various models, there is a tendency for certain models to overfit the data. This means that the models may perform well on the training data but poorly in real-world scenarios.

3. **Insufficient Hyperparameter Tuning**: While we compared several models, there was not enough hyperparameter tuning, which could have contributed to suboptimal performance.

Future research will focus on the following directions:

- **Data Augmentation and Balancing**: To address the data imbalance issue, we plan to apply data augmentation techniques or sampling methods to create a more balanced dataset across categories.
- **Hyperparameter Optimization**: We will optimize hyperparameters to enhance model performance and strengthen the model's generalization ability through cross-validation.
- **Exploration of Various Deep Learning Models**: In addition to RNNs, we will explore different deep learning architectures (e.g., CNNs, Transformers) to tackle the news classification problem and analyze the performance differences compared to machine learning models.
- **Transfer Learning Application**: We aim to utilize pre-trained models to improve the dataset's adaptability and enhance overall performance.

Through these plans, we intend to develop a more effective news category classification model and increase its applicability in real-world scenarios.

6 Acknowledgment

This research benefited from the contributions of the Aiffel project from Everyone's Lab. We would like to express our gratitude to the project creator, Yoo won june (@ukairia777), for their invaluable support and resources that greatly enhanced the quality of our study. Their dedication to advancing research and education in the field has been instrumental in shaping this work. Thank you for your commitment to fostering innovation and collaboration.