Formatting Instructions For hello

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

1.1 Background and questions raised in the study

Text classification is one of the most fundamental tasks in the field of natural language processing, playing a crucial role in various practical domains such as email spam filtering, sentiment analysis, and automatic document classification. In particular, news topic classification shows high demand in areas such as automatic article classification for media companies, personalized news recommendation systems, and media monitoring. However, existing studies have primarily focused on improving the performance of individual models, and research that comprehensively considers classification accuracy, computational efficiency, and robustness to data characteristics required in actual operational environments is relatively insufficient. Therefore, this study aims to explore models that provide optimal performance and cost-effectiveness from the perspective of building actual news topic classification systems, analyze the characteristics and limitations of each model, and present practical guidelines.

2 Related works

2.1 Review related prior research

Machine learning research for text classification has been actively conducted over a long period. In early research, probability-based models such as Naive Bayes were widely used as they provided decent performance despite their simplicity (McCallum & Nigam, 1998). Subsequently, Support Vector Machine (SVM) established itself as a representative algorithm by demonstrating strong performance in high-dimensional data classification, particularly for text data (Joachims, 1998). These traditional machine learning techniques have greatly contributed to the development of the text classification field, and their effectiveness has been proven through comprehensive studies that compare and analyze the performance of various algorithms, such as the research by Sebastiani (2002).

As deep learning technology led revolutionary advances in the field of natural language processing, models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) have demonstrated the ability to effectively learn the sequential characteristics and contextual information of text. LSTM has particularly established itself as a prominent deep learning model in text classification tasks by solving the long-term dependency problem in long sequences. However, these deep learning models have limitations in that they require large amounts of data and high computational

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

costs, and in small-scale datasets, they may not guarantee performance advantages over traditional machine learning models

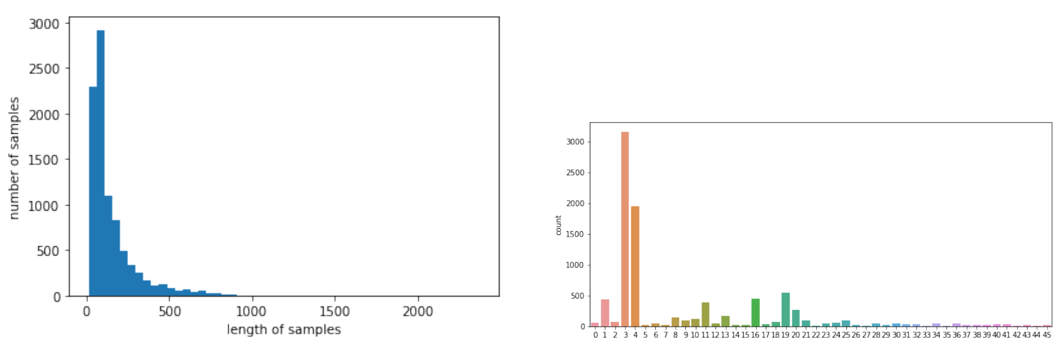
3 Data and Methodology

- 3.1 Dataset Description
- 3.2 Data Preprocessing
- 3.3 Evaluation Metrics
- 3.4 Experimental Design

3.1 Dataset Description

The dataset used is the Reuters news dataset, which consists of news articles collected from the Reuters financial news service in 1987. This dataset contains a total of 11,228 news articles classified into 46 topic categories, with a vocabulary size of 29,930 words. This dataset is one of the most widely used standard data collections in the field of text classification research, making it suitable for use as a benchmark when comparing and evaluating the performance of various models. Additionally, it is built into TensorFlow/Keras by default, which provides excellent data accessibility and is advantageous for enhancing experimental reproducibility.

Analysis of the word length distribution per sample in the Reuters news data shows that the average length is 145.53 words based on the tokenizer, with a maximum length of 2,376 words. This demonstrates that news articles have diverse length distributions due to their inherent characteristics. Examining the class-wise distribution reveals a serious imbalance problem, with topic 3 accounting for the largest proportion with 3,159 instances and topic 4 being the second largest with 1,949 instances, indicating that these two classes represent the majority of the entire dataset in an extreme imbalance. Such class imbalance can cause biased predictions during model training, necessitating appropriate handling methods.



(a) Reuters Data Word Length Distribution

(b) Reuters Data Class Distribution

Figure 1: Reuters Data Analysis Results

3.2 Data Preprocessing

The Reuters news dataset used in this study is provided in the form of sequences where each word is mapped to integer indices. To extract semantic features necessary for model training, we first performed the process of restoring these integer sequences to text form composed of actual words. In this process, we utilized the word-index dictionary provided through `get_word_index()` and explicitly added special tokens such as <pad> (padding), <sos> (start of sentence), and <unk> (unknown token) to the dictionary, considering the reserved indices in Keras. In particular, to prevent problems that could arise from Out-Of-Vocabulary (OOV) words that do not exist in the vocabulary dictionary and to ensure model robustness, all words that were not included in the training data or failed to be included in the vocabulary dictionary due to low frequency were uniformly processed as <unk> tokens. Each news article was converted into a string of words that were restored and processed in this manner.

Next, we transformed the restored text data into numerical feature vectors that machine learning models can effectively process. First, we used `CountVectorizer` to generate a Document-Term Matrix based on word frequency for each news article. The DTM represents individual documents in each row and unique words included in the vocabulary dictionary in each column, with each element of the matrix indicating the frequency of appearance of the corresponding word in a specific document. In this process, we built the vocabulary dictionary and transformed the DTM by applying `fit_transform()` to the training dataset, while applying only `transform()` to the test dataset based on the already constructed dictionary.

Considering that it is difficult to determine the actual importance of words based solely on simple frequency counts, we additionally applied TF-IDF (Term Frequency-Inverse Document Frequency) weighting. TF-IDF comprehensively reflects how frequently a specific word appears within a document (TF) along with how sparsely that word appears in the entire document collection (IDF) to calculate the importance of words. We performed TF-IDF transformation on the DTM using `sklearn.feature_extraction.text.TfidfTransformer`, applying `fit_transform()` to the training dataset and `transform()` to the test dataset to generate the final feature vectors.

3.3 Evaluation Metrics

In this study, we use accuracy as the primary evaluation metric to assess model performance. However, as explained earlier, due to the imbalanced class distribution in the data, there is a concern that recall and precision for minority classes may appear very low. Therefore, we also employ the F1-Score as an evaluation metric, which is more robust for imbalanced data. The documentation for `natbib` may be found at

3.4 Experimental Design

In this study, we designed experiments to comprehensively analyze the topic classification performance on the Reuters news dataset by dividing the models into machine learning-based models and deep learning-based models. Within the machine learning models, we further compared the performance of probability-based, linear-based, and tree-based approaches to identify which type of model demonstrates strengths in news topic classification. For deep learning models, we selected models specialized in sequential data processing and evaluated their performance.

3.4.1 machine learning-based models

3.4.1.1 probability-based models We aimed to evaluate the performance of models that classify classes based on word occurrence probabilities.
Naive Bayes: We utilized the Naive Bayes classifier (main hyperparameters follow scikit-learn's default values or general settings).
Complement Naive Bayes: We used the Complement Naive Bayes model, which is designed to address data imbalance issues (main hyperparameters follow scikit-learn's default values or general settings).

3.4.1.2 linear-based models We analyzed the effectiveness of models that learn linear relationships between input features (words) and classes.
Logistic Regression (L2): We used a logistic regression model with L2 regularization, with the main hyperparameters as follows: C=10000 (inverse of regularization strength), penalty='l2' (using L2 regularization),`max_iter=3000` (maximum number of iterations).
SVM (Support Vector Machine): We applied Support Vector Machine for classification. We considered linear kernel-based SVM (e.g., LinearSVC) which demonstrates strong performance in text classification (main hyperparameters follow scikit-learn's default values or general settings).

3.4.1.3 tree-based models We explored the classification performance of tree-based models that learn hierarchical rules based on data features.
Decision Tree (dtree): We performed classification using a decision tree model (main hyperparameters follow scikit-learn's default values or general settings).
Random Forest: We used a Random Forest model with the main hyperparameters as follows: `n_estimators=5` (number of trees), `random_state=0` (seed for result reproducibility). **Gradient**

Boosting: We utilized a Gradient Boosting model (main hyperparameters follow scikit-learn's default values or general settings).

3.4.1.4 ML Ensemble voting We constructed ensemble models to overcome the limitations of single models and achieve more stable and higher performance by combining predictions from various base models.
NB + LR + RF (soft): This is a soft voting ensemble model that averages the prediction probabilities from Naive Bayes, Logistic Regression, and Random Forest models.
NB + SVM + RF (hard): This is a hard voting ensemble model that determines the final class through majority voting among Naive Bayes, SVM, and Random Forest models.
CNB + SVM + RF (hard): This is a hard voting ensemble model combining Complement Naive Bayes, SVM, and Random Forest models.
LR + SVM + GBT (hard): This is a hard voting ensemble model combining Logistic Regression, SVM, and Gradient Boosting Tree (GBT) models.

3.4.2 deep learning-based models

We selected deep learning models capable of learning sequential information and contextual meaning from text data for comparative analysis with machine learning models. **LSTM (Long Short-Term Memory)**: We used LSTM networks, which have strengths in learning long-term dependencies in sequence data such as text. The hyperparameters of the model were set as shown in the table below.

Table 1: LSTM Model Configuration

Parameter	**Value**
Embedding dimension	128
LSTM units	256
Dropout rate	0.3
Vocabulary size	10,000
Max sequence length	200
Batch size	128
Epochs	20
Optimizer	Adam
Loss function	Categorical crossentropy
Early stopping patience	2

3.4.3 selecting the number of words

In this study, the number of words (i.e., vocabulary size) to be used when training the topic classification model for the Reuters news dataset was variously set to 5,000, 10,000, 15,000, and the total number of words (approximately 30,979). This was done to analyze the multifaceted impact of vocabulary size on model performance and efficiency. The selection of these word counts aims to explore the balance between performance and efficiency, analyze the influence of rare words and noise, and identify the optimal vocabulary size. Furthermore, this approach aligns with common practices in text classification research and is intended to ensure comparability with other studies.

4 Results and Analysis

This section presents the results of topic classification experiments conducted by applying various machine learning and deep learning models to the Reuters news dataset, and analyzes the trends in model performance according to changes in vocabulary size and the performance characteristics of each model

4.1 a trend by word count

Some linear models (logistic regression, SVM) and gradient boosting, along with Complement Naive Bayes, showed slight performance improvements or maintained stability as the number of

Table 2: Model Performance Comparison across Different Vocabulary Sizes (Hard and Soft indicate different voting methods in ensemble models)

Model	5000 Acc / F1	10000 Acc / F1	15000 Acc / F1
Traditional ML Models			
Decision Tree (dtree)	0.618 / 0.573	0.6202 / 0.5776	0.6193 / 0.5756
Naive Bayes	0.6732 / 0.6013	0.6589 / 0.5782	0.6371 / 0.5536
Complement Naive Bayes	0.7707 / 0.7459	0.7707 / 0.7457	0.772 / 0.7448
SVM	0.7752 / 0.7721	0.7881 / 0.7844	0.7885 / 0.7837
Random Forest	0.70 / 0.68	0.67 / 0.64	0.67 / 0.64
Gradient Boosting	0.7676 / 0.7662	0.7663 / 0.7622	0.7707 / 0.7680
Logistic Regression (L2)	**0.80 / 0.80**	**0.81 / 0.81**	**0.81 / 0.81**
ML Ensemble Voting			
NB + LR + RF (soft)	0.7560 / 0.7222	0.7369 / 0.6962	0.7222 / 0.6781
NB + SVM + RF (hard)	0.7752 / 0.7510	0.7582 / 0.7310	0.7507 / 0.7229
CNB + SVM + RF (hard)	0.7979 / 0.7794	0.8010 / 0.7828	0.7983 / 0.7809
LR + SVM + GBT (hard)	**0.8215 / 0.8100**	**0.8179 / 0.8075**	**0.8197 / 0.8088**
Deep Learning Models (50M words)			
LSTM	0.6109 / 0.5635	0.64 / 0.62	-

words increased to a certain level (e.g., 10,000-15,000 words), while models such as Naive Bayes or Random Forest did not show significant benefits or even experienced performance degradation when the number of words increased excessively. This suggests that too many words (especially rare words) in the TF-IDF feature space can act as noise or unnecessarily increase model complexity. The top-performing ensemble models achieved their best results with a vocabulary size of 5,000 words.

4.2 Model-specific trends

Overall, for the Reuters news topic classification problem using TF-IDF features, logistic regression with L2 regularization and ensemble models combining strong learners (particularly LR + SVM + GBT) demonstrated the best performance. SVM, Complement Naive Bayes, and Gradient Boosting also showed competitive performance. In contrast, basic Decision Tree, Random Forest, and the deep learning model LSTM used in the experiments recorded lower performance compared to machine learning-based models.

4.3 Key Observations and Insights

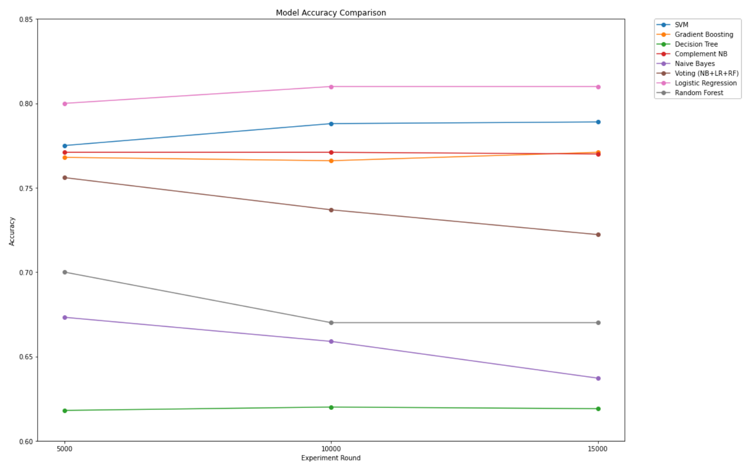
First, linear-based models represented by Support Vector Machine (SVM) and Logistic Regression demonstrated fundamentally high classification accuracy and maintained robust performance despite the high-dimensional sparsity characteristic of text data. Particularly, in the analyzed dataset, sparsely appearing word information effectively contributed to learning, resulting in a positive trend of slight performance improvement.

Second, in contrast, tree-based models such as Decision Tree and Random Forest showed relatively lower performance. This can be interpreted as being due to the inherent characteristics of tree models that struggle to find optimal splitting rules in high-dimensional sparse data environments and react sensitively to small changes in data. Specifically, tree-based models have a structure that repeats local splits based on individual features at each node, which limits their ability to capture complex semantic structures between words or global patterns. Consequently, in data like text where meaning is distributed across the entire context, information loss occurs and classification performance tends to deteriorate.

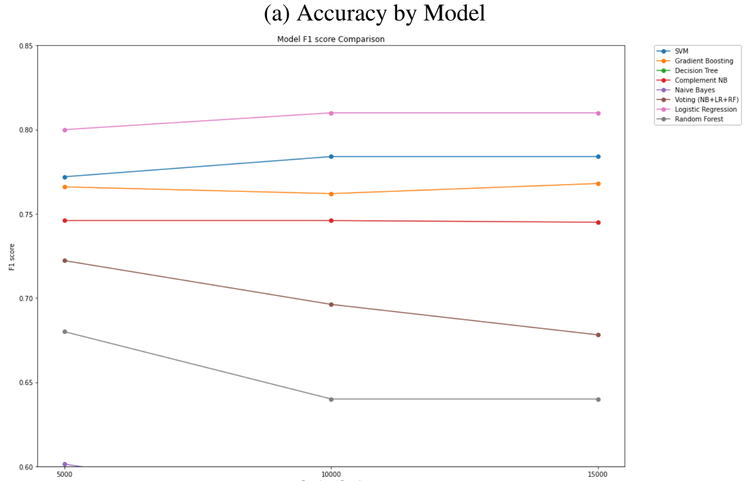
Third, probability-based models of the Naive Bayes family exhibited intermediate performance characteristics between the aforementioned linear-based and tree-based models. This suggests that while the relatively simple assumptions and fast computational speed of probability models guarantee a certain level of performance, they have limitations in capturing complex patterns or interactions between features within the data.

Furthermore, among the various ensemble models evaluated in this study, combinations that included a majority of linear-based models recorded the best classification performance. This result suggests that the robust performance and generalization ability of individual linear models can be more effectively manifested through ensemble techniques.

Finally, compared to the traditional machine learning models evaluated in this study, deep learning-based models recorded overall somewhat lower performance scores. This is presumed to be because the dataset used in this study was relatively small in scale, making it difficult to secure sufficient generalization performance given the characteristics of deep learning models that need to learn vast amounts of parameters.



(a) Accuracy by Model



(b) F1-Score by Model

Figure 2: F1-Score,Au

5 Conclution

5.1 Summary of Main Research Results and Academic Contributions

This study aimed to provide empirical evidence for optimal model selection in practical data and resource-constrained environments by conducting a comparative analysis of the performance of various machine learning and deep learning models in news topic classification problems.

The main research results are as follows.

First, linear-based models (particularly SVM and logistic regression) demonstrated the most robust and highest classification accuracy in text classification tasks, even in environments with relatively small amounts of training data. In particular, they maintained stable performance even when the sparsity of input data increased by expanding the vocabulary size for analysis, which reconfirms the superiority of linear models in processing high-dimensional sparse data. These results suggest that linear models remain a powerful and efficient choice in realistic situations where high-quality text classification must be performed with limited data.

Second, deep learning models (LSTM was utilized in this study) recorded somewhat lower performance compared to traditional machine learning models at the scale of the dataset used. This is interpreted as being due to the characteristic that deep learning models generally require large-scale datasets to learn vast amounts of parameters. However, this is not an inherent limitation of deep learning models but rather a result of data scale constraints, and it is judged that they have sufficient potential to surpass the performance of traditional machine learning models if sufficient data is secured.

Third, the performance improvement effect through ensemble techniques was confirmed, and the highest performance was achieved particularly in ensemble combinations that included linear-based models. Therefore, this study shows that in environments where there are cost and time constraints for large-scale data collection and labeling, well-tuned traditional machine learning models (especially linear models) can be practically more efficient and optimal choices. This emphasizes the importance of rational model selection considering the characteristics of given problems and available resources rather than unconditionally pursuing the latest technologies.

Through this analysis, this study has academic and practical contributions by identifying the practical performance and characteristics of each model in the specific domain of news topic classification, thereby providing useful guidelines for researchers and practitioners who want to select and apply models in similar environments in the future.

5.2 Research Limitations and Future Research Suggestions

Despite deriving meaningful results, this study has several limitations that can be improved through future research.

First, there is a lack of diversity in deep learning models. In this study, experiments were conducted using only LSTM as a representative example of deep learning models. However, in recent text classification fields, more advanced and diverse deep learning models such as CNN and Transformer series (BERT, GPT, etc.) have shown excellent performance. Therefore, future research needs to apply these various state-of-the-art deep learning architectures and conduct comparative analysis of their performance.

Second, the utilization of state-of-the-art natural language processing technologies was insufficient. While basic approaches were used for word embedding techniques, advanced technologies such as pre-trained word embedding models like Word2Vec, GloVe, FastText, or context-aware BERT embeddings were not sufficiently utilized. Since these advanced embedding techniques can contribute to model performance improvement by representing semantic information of words more richly, future research needs to actively introduce them.

Third, there was insufficient in-depth analysis according to data scale. While this study inferred that deep learning models showed low performance due to data scale constraints, in-depth experiments comparing performance change trends of each model by gradually increasing data volume were not performed. Future research that explores the threshold at which deep learning models begin to surpass the performance of traditional machine learning models by setting various data scales would yield more meaningful results.

Fourth, quantification of cost-performance trade-offs is necessary. Research is also needed to more quantitatively measure cost aspects such as model training and inference time and required computing resources, and develop frameworks that consider these together with performance to present objective criteria for model selection.

6 References

Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.