Fine-Tuning BERT for Automated News Classification

Mohammed I. Salih
Computer Information System Department, Technical College of Zakho, Duhok Polytechnic University, Duhok, KRG, Iraq
mohammed.salih@dpu.edu.krd

Salim M. Mohammed
Computer Science Department, College of Science, University of Zakho, Duhok, KRG, Iraq
salim.mohammed@uoz.edu.krd

Asaad Kh. Ibrahim
Computer Information System Department, Technical College of Zakho, Duhok Polytechnic University, Duhok, KRG, Iraq
asaad.khaleel@dpu.edu.krd

Omar M. Ahmed
Computer Information System Department, Technical College of Zakho, Duhok Polytechnic University, Duhok, KRG, Iraq
omar.alzakholi@dpu.edu.krd (corresponding author)

Lailan M. Haji
Computer Science Department, College of Science, University of Zakho, Duhok, KRG, Iraq
lailan.haji@uoz.edu.krd

Received: 18 February 2025 | Revised: 8 March 2025, 22 March 2025, and 2 April 2025 | Accepted: 5 April 2025

 DOI: https://doi.org/10.48084/etasr.10625

ABSTRACT

Text classification is a fundamental task in Natural Language Processing (NLP) with a wide range of applications such as sentiment analysis, document classification and content recommendation. Traditional approaches like Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF) relied on feature engineering but lacked contextual understanding. Deep learning came into the picture for text classification with transformer models such as Bidirectional Encoder Representations from Transformers (BERT), which could understand contextual words bidirectionally. In this article, we utilize a pre-trained BERT model fine-tuned on the Reuters-21578 dataset to classify news articles. We aim to measure the performance of transfer learning against common machine learning models and non-fine-tuned BERT. The fine-tuned model achieves 91.77% accuracy, which significantly outperforms the non-fine-tuned BERT, and performs better than classical classifiers such as NB, SVM and RF. The results show that fine-tuning allows BERT to contextualize domain-specific intricacies, resulting in improved classification performance. We also address the computational trade-offs associated with transformer models, highlighting the need for optimal methods for deployment. Thus, this study further enables the use of fine-tuned BERT in automatic news classification and is of significant value for information retrieval and content personalization.

Keywords-fine-tuning; BERT; news classification; natural language processing; text classification

I. INTRODUCTION

Text classification is a fundamental task in Natural Language Processing (NLP), which forms the foundation for many applications such as document classification, sentiment analysis, spam detection, and content recommendation [1]. As the amount of textual data generated daily is massive, the automatic classification of text into given categories has become essential. Before the advent of deep learning, text classification followed traditional methods that were largely

based on feature engineering methods such as Term Frequency-Inverse Document Frequency (TF-IDF), n-grams, and Bag-of-Words (BoW) models [2]. Although these methods worked reasonably well, they often couldn't effectively encode the contextual associations between words and phrases. Through its auto-encoding architecture, deep learning has revolutionized text classification. These models provided the necessary progress for certain text classification tasks by preserving sequential information and tracking local dependencies in text [3]. However, these models were still limited because they either did not account for long-range dependencies due to the vanishing gradient problem in Recurrent Neural Networks (RNNs) [4], or their predictions were fixed to a window due to Convolutional Neural Networks (CNNs) [5].

The introduction of Bidirectional Encoder Representations from Transformers (BERT) represents a major revolution in the ability of NLP models to understand meaning in a robust, contextual way [6]. Unlike traditional models that analyze text word by word, BERT uses self-attention that considers every word at once, enabling bidirectional context. This means that a model can understand the meaning of a word means from the words around it, which has led BERT to achieve state-of-the-art results on several NLP benchmarks [7]. We therefore use the Reuters dataset to illustrate the ability of a fine-tuned, pre-trained BERT model for news classification. The Reuters dataset is commonly used in text classification research and consists of thousands of news articles spread across several categories. Using the pre-trained BERT model, we aim to determine whether transfer learning would yield significantly better classification performance than machine learning-focused techniques.

A. News Classification: Why is it Important?

News classification is an essential component of modern information retrieval and content filtering [8]. Manually categorizing articles is becoming increasingly impractical as the number of digital news sources grows. Automated news classification enables media organizations, search engines, and social media to efficiently structure and recommend relevant articles to their audiences [9]. Personalized news recommendation is a popular application of text classification; for example, platforms such as Google News and Flipboard recommend articles by classifying them based on user preferences [10]. Fake news detection is another process of categorizing news into trustworthy and untrustworthy sources [11], which in turn can help combat misinformation. In addition, topic-based filtering allows readers to receive news that is relevant to their interests, while filtering out content that is not relevant to them [12]. Finally, sentiment analysis allows financial institutions to use news sentiment to anticipate stock market trends [13].

B. Drawbacks of Traditional Text Classification Methods

Traditional text classification techniques include Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) [14]. These models use hand-crafted features such as:

- BoW: Representation of each text as a pool of independent words without considering their order [15].
- TF-IDF: It considers the importance of words in the document versus the full representation of documents [16].
- Word embeddings (Word2Vec, GloVe): They provide the meaning of words in a dense vector space, but do not have a deep contextual understanding [17].

Efficient in some scenarios, these methods have several key shortcomings [18]:

- Loss of context: Traditional approaches consider each word in isolation, and therefore miss sentence structure.
- Vocabulary sensitivity: New words or errors with slight changes in spelling degrade performance.
- Manual feature engineering: Requires manual creation of features that have meaningful context, which necessitates domain knowledge.
- Scalability issues: Large text corpora require a lot of preprocessing and computational resources.

C. Revolutionizing Natural Language Processing with Transformers

The advent of transformers brought a paradigm shift in NLP with self-attention mechanisms that processed whole sentences at once rather than sequentially. The BERT model [6] was a milestone because it was pre-trained on large text corpora and learned bidirectional representations. Some key innovations of BERT are [19]:

- Masked Language Modeling (MLM): BERT is bidirectional, unlike traditional left-to-right (or right-to-left) models that rely on the left (or right) context to predict the next word.
- Training objective: Next Sentence Prediction (NSP) allows capturing interdependence in a task such as answering questions.
- Fine-tuning ability: The pre-trained BERT model is capable of fine-tuning on task-specific small datasets.

Text classification is a well-researched task in the field of NLP, and conventional methodologies have used statistical and machine learning methods such as NB, SVM, and decision trees, which can work well with structured text features [20]. TF-IDF, arguably the most popular of these, has been used along with NB or SVM for document categorization [21]. However, these approaches had limitations in semantic understanding and were often unable to capture long-range dependencies in text. As deep learning became popular, RNNs and CNNs were introduced to NLP tasks. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) [22] were major leaps in sequential text modeling because they were able to learn dependencies within long sequences. Although CNNs are more popular for computer vision applications, n-gram feature extraction made CNNs viable for text classification [23]. Although these models expanded the range of possible input data, they were limited in their ability to

effectively handle long-range dependencies, leading to the formulation of attention-based models that form the basis of machine translation [24].

A popular strategy is to fine-tune pre-trained BERT models for text classification. Authors in [25] demonstrate an improvement of at least 10% in the accuracy of BERT-based classifiers over traditional machine learning classifiers on a few datasets such as Reuters, 20 Newsgroups, and AG News. In addition, research shows that domain-specific fine-tuning of pre-trained models improves classification performance on domain-specific corpora, including medical records, legal documents, and financial reports [26]. Recent work by the authors in [27] addressed the problem of Multi-Task Learning (MTL) with BERT by training a single transformer on multiple datasets simultaneously to classify the same piece of text. This method led to higher generalization and robustness, especially for low-resource languages. However, BERT faces challenges and limitations that make it less than perfect for fine-tuning. Research shows that large-scale models of BERT consume too many computational resources, making them unusable for low-power devices [28]. Work on pruning, quantization, and knowledge distillation aims to reduce the model size of the transformer architecture while maintaining the performance [27]. In addition, interpretability is a major issue. Traditional decision-making models, such as SVMs, provide comparatively little information but are transparent, whereas transformers are considered "black boxes" due to complicated multi-head attention mechanisms [28]. Research is still active to develop Explainable Artificial Intelligence (XAI) techniques for BERT-like models.

II. METHODOLOGY

We use the Reuters-21578 dataset [29] to train this model, and preprocess it by trimming vocabulary, specifying maximum sequence length, and performing a train/test split. We use BERT's pre-trained WordPiece tokenizer to tokenize the text data and configure it with token IDs, attention masks, and segment IDs. Since we are fine-tuning the representation of small BERT, our model will have transformer layers, dropout for regularization, and a dense classification layer with softmax activation. We use the AdamW optimizer and sparse categorical cross-entropy loss for training with transfer learning to achieve better classification results.

A. Dataset and Preprocessing

In this study, we utilize the Reuters-21578 dataset, which is commonly used in the information retrieval research field, as a standard benchmark corpus for text classification. It is a news dataset consisting of 11,228 news articles organized into 46 categories. For each document, there is a label that can be one or more topic categories, so it is a multi-class classification task. We preprocess the dataset by truncating the vocabulary for the top 10,000 frequent words. The sequence length was set to 300 tokens based on the distribution of article lengths in this dataset, where most samples contain less than 300 tokens after tokenization, ensuring minimal truncation. Rare classes with too few samples were dropped, and the dataset was divided into training and testing sets (80/20). Neural networks need structured data as input for text classification, so we reverse back to text sequences with word indices and tokenize them with BERT's pre-trained tokenizer. As BERT has a specific input format, we start with the following steps:

1. Text decoding: The dataset provides sequences of word indices instead of raw text. To reconstruct the original sentences, we use a word-index mapping. This splits each word into sub-word segments using BERT's WordPiece tokenizer to allow the model to effectively represent out-of-vocabulary words.
2. Padding and truncation: Since BERT has a fixed input size, sequences must be padded or truncated to 300 tokens.
3. BERT input representation: Each sequence passes through a process before being input to BERT:
    - Token IDs: Describe a financial transaction using a token ID.
    - Attention masks: Binary vectors that are true when the token should be attended.
    - Segment IDs: Used to identify multiple sequences within the same input (e.g., in question-answer input, it allows to separate the question-and-answer parts).

Although certain classes are underrepresented in the Reuters-21578 dataset, this study focused on evaluating the capabilities of BERT without applying class balancing techniques. However, future work could explore methods such as oversampling, class weighting, or data augmentation to mitigate class imbalance.

B. Input Representation and Encoding

The architecture of our BERT-based news classification model is designed to leverage pre-trained transformer embeddings, while incorporating task-specific layers to improve classification performance. Below, we provide a detailed breakdown of the architecture, including the BERT encoder, intermediate layers, and final classification layers.

Data preparation involves transforming input sequences, where each input text is tokenized using WordPiece tokenization, a sub-word-based vocabulary of sequences. The resulting input tokens are then converted into fixed-length vectors, where each token is represented by an integer, a number that reflects a token's place within the BERT vocabulary and is referred to as a token ID. An attention mask is also used, which is a binary vector that is 1 for real words and 0 for padding tokens. Since this is a single sentence classification task, the token type IDs are all set to 0, as they are only needed for sentence pair tasks. In addition, we add a [CLS] token at the beginning of each sentence as a summary representation of the classification, and a [SEP] token at the end of the sequence. We then convert the input sentences to input IDs and pad or truncate them to a fixed sequence length (300 tokens), which guarantees that the input is uniform for all samples.

C. Encoding with BERT and Transformer Layers

The BERT encoder at the heart of our model consists of multi-layer bidirectional transformer blocks. Specifically, in

this study, small BERT (L-4_H-512_A-8) was selected due to its lower computational cost while maintaining effective contextual understanding. It allows faster fine-tuning and inference, making it suitable for real-time applications or resource-constrained environments. We use 4 transformer layers with 512 hidden units and 8 self-attention heads. Each transformer layer consists of two subcomponents, as follows:

- Multi-Head Self-Attention (MHSA): BERT's self-attention mechanism allows it to learn the contextual relationships between words by calculating attention scores across all tokens for each token in the input sequence. Notably, unlike RNNs, which only allow for local dependencies, the self-attention mechanism allows each token to pay attention to all other tokens. In addition, each multiple attention head focuses on different contextual components. We compute the attention function as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

  where Q, K, V represent the query, key, and value matrices, and d_k is the scaling factor for numerical stability.

- Feed-Forward Network (FFN): The token representations are then passed to the attention mechanism, after which they are sent through a fully connected FFN consisting of two dense layers with ReLU activation. This part of the process enables the model to learn complex feature representations within each layer. Such layering with normalization and residual connections allows for stability and better gradient flow throughout training in each individual transformer block. The final pooled output of the [CLS] token is extracted, passing through the classification layers.

D. Fully Connected Layers

After the BERT encoder, task-specific fully connected layers are added on top for news classification. Dropout is applied at a rate of 0.3 immediately after the first post-BERT layer to reduce overfitting by preventing the network from becoming overly reliant on specific neurons (30% of neurons are not used during training). This helps the model generalize better to new data.

Specifically, a dense layer of 256 neurons with ReLU activation is used. This layer takes the feature representation from BERT and processes it at a higher text pattern level. We choose the ReLU activation function to introduce nonlinearity between the two layers to avoid the vanishing gradient problem. The next layer is a dropout layer (0.3), which is used to regularize the model to reduce overfitting. The last layer is the output layer, which has 46 neurons with softmax activation functions equal to the number of classes in the Reuters dataset. The softmax function normalizes the output of each class so that the range of outputs is from 0 to 1 and the sum of outputs across all classes is 1 (i.e., a distribution).

E. Model Compilation and Optimization

The AdamW optimizer, an enhanced version of the Adam optimizer that incorporates weight decay regularization, is used to efficiently train the model. Weight decay is used to restrict significant parameter updates to enhance a model with respect to unseen data. Using the magnitudes of the gradients, the optimizer dynamically adapts the learning rate to help the model to converge faster. Since our dataset labels are integer-encoded, we use sparse categorical cross-entropy as our loss function.

III. RESULTS AND ANALYSIS

In this section, we provide a comprehensive evaluation of our fine-tuned BERT model on the Reuters-21578 dataset. We compare its performance against several simple classifiers as well as a non-fine-tuned BERT model to showcase the benefits of fine-tuning for text classification. The model was trained using an NVIDIA RTX 4060 GPU with a batch size of 32 for efficient memory usage. We trained for 10 epochs, as performance gains plateaued beyond this point, to avoid overfitting.

A. Performance Metrics and Comparative Analysis

The evaluation of the model performance is mainly done using the classification accuracy, which is defined as the number of correctly predicted instances. Besides accuracy, we also evaluated the F1-score, which considers precision and recall, making it more appropriate for imbalanced datasets. Table I presents the comparison of the proposed model with other related models in terms of accuracy, precision, recall, and F1-score.

TABLE I. MODEL PERFORMANCE COMPARISON

Model	Accuracy (%)	Precision	Recall	F1-score
Fine-tuned BERT (proposed model)	91.77	0.92	0.91	0.915
Non-fine-tuned BERT	83.45	0.84	0.83	0.835
NB	76.32	0.78	0.75	0.765
SVM	80.19	0.81	0.79	0.80
RF	78.41	0.79	0.77	0.78
CNN-LSTM hybrid	85.67	0.86	0.85	0.856

Our fine-tuned BERT model outperforms all other models with an accuracy of 91.77%, as can be seen from the results. The only model that comes close is the CNN-LSTM hybrid with 85.67%, whereas all traditional classifiers such as NB, SVM, and RF perform significantly worse. Our results reveal that such models perform worse than deep learning approaches. NB has the lowest accuracy (76.32%) because it is based on word frequency and assumes feature independence, which hinders its ability to recognize contextual relationships. Since SVM works well in high-dimensional spaces, it achieves a better accuracy of 80.19% but it doesn't have any semantic understanding of the text. RF provides an accuracy of 78.41% because we use ensemble learning, but it fails to capture the dependency of words in the first sentence. On the other hand, BERT understands the context of words in both directions (i.e., left, and right), which is a significant improvement over these models. The CNN-LSTM hybrid model we developed also outperforms all non-transformer-based models across the board, with an accuracy of 85.67%. Although this model works well, BERT's self-attention mechanism allows for more detailed extraction of both local and global dependencies,

which explains its better performance. It is noteworthy that even the non-fine-tuned BERT model achieves 83.45% accuracy, underscoring the need for fine-tuning for domain adaptation. Fine-tuning allows the model to absorb the linguistic structure of the dataset, how it differs between categories, etc. Moreover, although we are dealing with fine-tuning of BERT for sentence embedding, we can observe an improvement of the absolute F1-score from 0.835 (non-fine-tuned BERT) to 0.915 (fine-tuned BERT), indicating a better precision-recall balance.

The fine-tuned BERT model outperforms other approaches because it effectively captures contextual meaning through its self-attention mechanism. Unlike traditional models such as NB or SVM, which rely on statistical features, BERT processes words in relation to their surrounding context. Additionally, fine-tuning allows the model to adapt to the specific linguistic patterns of the Reuters dataset, improving classification accuracy and F1-score.

B. Error Analysis

Even though the fine-tuned BERT model achieves a high classification accuracy, it still doesn't provide a perfect classification. An in-depth analysis of misclassified instances suggests three main reasons for errors:

- The first concern is ambiguous articles, since some news stories cover multiple categories such as an article that covers finance and politics. In such situations, BERT can only assign one class of the multiple classes to which the article belongs. Multi-label classification instead of a single-label classification could alleviate this problem.
- The second challenge is infrequent categories. The Reuters dataset contains some classes that are underrepresented, meaning that the classes corresponding to some topics have a limited number of training samples. This results in the model not being able to learn meaningful patterns for these categories. In future work, we could try data augmentation or few-shot learning techniques to alleviate this problem.
- The third limitation has to do with long articles, where BERT consumes input with a maximum sequence length of 300 tokens. Some items from news articles exceed this limit, leading to truncation, where important information may be cut off. For longer texts, hierarchical transformer models or document-level BERT adaptations may help improve classification accuracy.

A closer analysis of the misclassified examples reveals that about 15% of the misclassifications belong to several overlapping categories, such as finance and politics. Additionally, underrepresented classes, such as "strategic-metal" and "housing-starts" had error rates above 20% due to the limited training samples. A confusion matrix analysis showed that most misclassifications occurred between semantically similar categories. Addressing this issue could involve strategies like multi-label classification or hierarchical categorization.

C. Computational Cost and Trade-offs

Fine-tuned BERT provides the highest classification accuracy but has a significant computational cost in terms of training time, memory overhead, and hardware requirements. BERT consumes far more resources than machine learning classifiers or even some deep learning models, which is an important deployment factor. To optimize fine-tuning for real-world deployment, techniques such as model pruning, quantization, and knowledge distillation can be applied. Model pruning reduces unnecessary parameters, quantization lowers precision to speed up inference, and knowledge distillation allows training a smaller student model from a large BERT model. Recent advancements, such as Low-Rank Adaptation (LoRA), also allow efficient fine-tuning without updating the entire model, making deployment in resource-constrained environments more feasible.

An important drawback of transformer-based models such as BERT is the long training time, particularly when fine-tuning on larger datasets. While techniques such as NB, SVM, or RF can be trained on a regular CPU in minutes to hours, fine-tuning BERT on a GPU can take hours or even days, depending on the size of the dataset and the network architecture, as shown in Table II.

TABLE II. COMPUTATIONAL COST OF DIFFERENT MODELS

Model	Training time (GPU)
Fine-tuned BERT	~3 hours
Non-fine-tuned BERT	~1 hour
NB	~10 minutes
SVM	~45 minutes
RF	~30 minutes
CNN-LSTM hybrid	~2 hours

As can be seen from the table, BERT-based models need to be trained on the data for a much longer time than traditional methods. The non-fine-tuned BERT does not require intensive training time (~1 hour), but it does not surpass the fine-tuned BERT results. In contrast, standard models such as NB require minutes of training, but with much lower performance. Future work could explore lightweight alternatives such as DistilBERT or ALBERT to achieve a trade-off between accuracy and computational efficiency.

IV. CONCLUSION

In this study we investigated the ability of a fine-tuned Bidirectional Encoder Representations from Transformers (BERT)-based model to perform news classification on the Reuters-21578 dataset. We demonstrate that the fine-tuned BERT outperforms not only classical machine learning algorithms, but also the non-fine-tuned BERT, achieving a classification accuracy of 91.77%. These improvements are due to BERT's ability to handle bidirectional context and effectively understand complex semantic relationships that standard models do not exploit. While BERT achieved state-of-the-art results in both pre-training and fine-tuning tasks, its accuracy improvements came at the cost of an extra layer of computational complexity with increased training time and memory requirements for the larger networks, making BERT

infeasible for low-resource contexts. Nevertheless, breakthroughs in areas such as model pruning, quantization, and knowledge distillation suggest a pathway for improving efficiency without very much loss of accuracy.

Further studies could be conducted to adapt lightweight transformer models for real-time systems and to explore multi-label classification strategies to improve accuracy for news items belonging to multiple categories. Based on these factors, this study illustrates the revolutionary effect that transfer learning has had on the Natural Language Processing (NLP) landscape and reaffirms BERT's capacity as an effective approach for automatic news classification. Future work could focus on optimizing BERT for real-time news classification using lightweight transformer models such as DistilBERT. In addition, exploring hybrid approaches that combine BERT with Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) layers could further improve classification accuracy. Multi-label classification techniques could also be considered to address ambiguity in news categorization.

REFERENCES

[1] A. K. Mehta and S. Kumar, "Comparative Analysis and Optimization of Spam Filtration Techniques Using Natural Language Processing," in *2024 International Conference on Communication, Computer Sciences and Engineering*, Gautam Buddha Nagar, India, 2024, pp. 1005–1010, https://doi.org/10.1109/IC3SE62002.2024.10593598.

[2] H. Singh, S. Sood, H. Maity, and Y. Kumar, "Exploring Pre-processing Strategies and Feature Extraction in practical aspect for Effective Spam Detection," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*, Gwalior, India, 2024, pp. 1–6, https://doi.org/10.1109/IATMSI60426.2024.10502863.

[3] T. Vo, "A topic-driven graph-of-words convolutional network for improving text classification: Array," *Journal of Science and Technology on Information and Communications*, vol. 1, no. 1, pp. 10–19, Mar. 2022.

[4] O. M. Ahmed, L. M. Haji, A. M. Ahmed, and N. M. Salih, "Bitcoin Price Prediction using the Hybrid Convolutional Recurrent Model Architecture," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11735–11738, Oct. 2023, https://doi.org/10.48084/etasr.6223.

[5] L. M. Haji, O. M. Mustafa, S. A. Abdullah, and O. M. Ahmed, "Enhanced Convolutional Neural Network for Fashion Classification," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16534–16538, Oct. 2024, https://doi.org/10.48084/etasr.8147.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.

[7] H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT-Like Decoders for Generation," *iRADIOLOGY*, Jan. 2025, https://doi.org/10.1002/ird3.115.

[8] Z. Lin, J. Xie, and Q. Li, "Multi-modal news event detection with external knowledge," *Information Processing & Management*, vol. 61, no. 3, May 2024, Art. no. 103697, https://doi.org/10.1016/j.ipm.2024.103697.

[9] B. A. Baltes, Y. Cardinale, and B. Arroquia-Cuadros, "Automated Fact-checking based on Large Language Models: An application for the press," in *Proceedings of the 1st Workshop on COuntering Disinformation with Artificial Intelligence*, Santiago de Compostela, Spain, 2019, pp. 40–53.

[10] B. Florea and A. Iftene, "The News in Brief - Leveraging Machine Learning and Artificial Intelligence in News Clustering, Summarization and Evaluation," in *18th International Conference on Linguistic Resources and Tools for Processing the Natural Language*, Brasov, Romania, 2023.

[11] S. Mishra, P. Shukla, and R. Agarwal, "Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified Datasets," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, Mar. 2022, Art. no. 1575365, https://doi.org/10.1155/2022/1575365.

[12] H.-G. Yoon, H.-J. Song, S.-B. Park, and K. Y. Kim, "A Personalized News Recommendation using User Location and News Contents," *Applied Mathematics & Information Sciences*, vol. 9, no. 2L, pp. 439–449, Apr. 2015, https://doi.org/10.12785/amis/092L19.

[13] S. A. Abdullah, M. I. Salih, and O. M. Ahmed, "Improving Sentiment Classification using Ensemble Learning," *International Journal of Informatics, Information System and Computer Engineering*, vol. 6, no. 2, pp. 200–211, Dec. 2025, https://doi.org/10.34010/injiiscom.v6i2.13921.

[14] H. Alamoudi *et al.*, "Arabic Sentiment Analysis for Student Evaluation using Machine Learning and the AraBERT Transformer," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11945–11952, Oct. 2023, https://doi.org/10.48084/etasr.6347.

[15] H. Dabjan and M.-B. Kurdy, "Efficient Streamlined Online Arabic Web Page Classification Using Artificial Bee Colony Optimization," *Iraqi Journal of Science*, vol. 65, pp. 7220–7236, Dec. 2024, https://doi.org/10.24996/ijs.2024.65.12.35.

[16] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, "Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms," *PeerJ Computer Science*, vol. 8, Mar. 2022, Art. no. e914, https://doi.org/10.7717/peerj-cs.914.

[17] B. Yosra and M. Hakim, "Enhancing Twitter Sentiment Analysis Using Hybrid Transformer and Sequence Models," *Japan Journal of Research*, vol. 6, no. 1, Nov. 2024, Art. no. 089, https://doi.org/10.33425/2690-8077.1163.

[18] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, Sep. 2023, Art. no. 100026, https://doi.org/10.1016/j.nlp.2023.100026.

[19] D. Uribe, E. Cuan, and E. Urquizo, "Fine-Tuning of BERT models for Sequence Classification," in *2022 International Conference on Mechatronics, Electronics and Automotive Engineering*, Cuernavaca, Mexico, 2022, pp. 140–144, https://doi.org/10.1109/ICMEAE58636.2022.00031.

[20] G. W. Tatchum, A. J. N. Nzeko'o, F. S. Makembe, and X. Y. Djam, "Class-Oriented Text Vectorization for Text Classification: Case Study of Job Offer Classification," *Journal of Computer Science and Engineering*, vol. 5, no. 2, pp. 116–136, Aug. 2024.

[21] M. M. Alnaddaf and M. S. Başarslan, "Sentiment Analysis Using Various Machine Learning Techniques on Depression Review Data," in *2024 8th International Artificial Intelligence and Data Processing Symposium*, Malatya, Turkiye, 2024, pp. 1–5, https://doi.org/10.1109/IDAP64064.2024.10710889.

[22] X. Jiang, X. Ding, C. Liu, X. Li, Y. Zhang, and S. Wang, "Research on the application of computer-aided deep learning model in natural language processing," *Journal of Electronics and Information Science*, vol. 9, no. 3, pp. 153–159, Nov. 2024, https://doi.org/10.23977/jeis.2024.090320.

[23] S. Mikaeelzadeh and A. Mirzaei, "A Review of Natural Language Processing Models for Classifying Non-Functional Software Requirements." Researchgate, Dec. 03, 2024. [Online]. Available: https://www.researchgate.net/publication/384441061_A_Review_of_Natural_Language_Processing_Models_for_Classifying_Non-Functional_Software_Requirements.

[24] M. A. Wani, A. A. A. El-Latif, M. ELAffendi, and A. Hussain, "AI-based Framework for Discriminating Human-authored and AI-generated Text," *IEEE Transactions on Artificial Intelligence*, vol. 1, no. 01, pp. 1–15, Dec. 2024, https://doi.org/10.1109/TAI.2024.3516713.

[25] L. Lilli *et al.*, "Lupus Alberto: A Transformer-Based Approach for SLE Information Extraction from Italian Clinical Reports," in *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, 2024, pp. 510–516.

[26] Z. Z. Chen *et al.*, "A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law." arXiv, Nov. 21, 2024, https://doi.org/10.48550/arXiv.2405.01769.

[27] O. Jokinen, "Document-level embeddings and graph-augmented learning for Finnish articles," M.S. thesis, Faculty of Science, University of Helsinki, Finland, 2024.

[28] R. Abilio, G. P. Coelho, and A. E. A. da Silva, "Evaluating Named Entity Recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset," *Applied Soft Computing*, vol. 166, Nov. 2024, Art. no. 112158, https://doi.org/10.1016/j.asoc.2024.112158.

[29] D. D. Lewis, "Reuters-21578 Text Categorization Test Collection." [Online]. Available: https://www.daviddlewis.com/resources/testcollections/reuters21578/.