Advancing Sentiment Analysis: Evaluating RoBERTa against Traditional and Deep Learning Models

Pongsathon Pookduang
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand
pongsathon.po@kkumail.com

Rapeepat Klangbunrueang
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand
rapeepat.klangbunrueang@kkumail.com

Wirapong Chansanam
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Thailand
wirach@kku.ac.th (corresponding author)

Tassanee Lunrasri
Department of Information Systems, Faculty of Business Administration and Information Technology, Rajamangala University of Technology, Khon Kaen, Thailand
tassanee.so@rmuti.ac.th

Received: 23 November 2024 | Revised: 15 December 2024 | Accepted: 1 January 2025

 DOI: https://doi.org/10.48084/etasr.9703

ABSTRACT

This research evaluates the performance of various sentiment analysis models, including traditional machine learning approaches (Naive Bayes, KNN, CART), a deep learning model (LSTM), and the transformer-based model RoBERTa using an Amazon book reviews dataset. ROBERTa outperformed all other models, achieving an accuracy of 96.30% and an F1-score of 98.11%, underscoring its superior ability to process complex and semantically diverse textual data. Traditional models, while computationally efficient, demonstrated limitations in capturing nuanced textual relationships, and the LSTM model, although competitive, faced scalability challenges and overfitting issues. These results demonstrate how transformer-based architectures such as RoBERTa offer advantages in real-world applications, particularly in e-commerce and social media sentiment analysis. This study underscores the superior capabilities of RoBERTa for sentiment analysis, particularly in processing semantically diverse and context-rich textual data that traditional models struggle to capture. Future work will explore optimizing RoBERTa's computational efficiency and expanding its applications to multilingual and cross-domain sentiment analysis tasks.

Keywords-sentiment analysis; RoBERTa; Amazon book reviews; deep learning; machine learning models

I. INTRODUCTION

Researchers and practitioners use sentiment analysis as a vital tool to investigate opinions and emotions across diverse digital platforms, encompassing applications in business, politics, and public health [1, 2]. With the exponential growth of online data, sentiment analysis enables researchers and practitioners to extract valuable insights and leverage this information for informed strategic decision-making [3]. Despite its potential, one of the main challenges in sentiment analysis is addressing linguistic diversity and handling imbalanced datasets [4, 5]. Imbalanced data, where certain sentiment classes (e.g., positive or negative) significantly outweigh others, remains a pervasive issue. To mitigate this, data augmentation techniques, including the use of embedding such as GloVe, have been employed to generate linguistically

diverse samples, enhancing the representativeness of datasets [6, 7]. These methods enrich the dataset and bolster the models' capacity to handle imbalances effectively [7].

Sequential models, including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have proven to be instrumental in capturing long-term dependencies within textual data. However, their reliance on sequential data processing often results in higher computational overhead than transformer-based models, which can process data in parallel [5]. Transformer architectures, exemplified by the RoBERTa model, excel at capturing nuanced word meanings and contextual relationships with remarkable precision [8]. Recent studies have proposed hybrid architectures, such as RoBERTa-GRU and RoBERTa-LSTM, which synergize the strengths of sequential and transformer models. These hybrids enhance the ability to capture long-term dependencies while improving computational efficiency [5, 8].

The integration of Transformer-based architectures, particularly RoBERTa, has significantly advanced the field of Natural Language Processing (NLP). In [9] the hybrid RoBERTa-BiLSTM was proposed, which leverages RoBERTa's ability to generate meaningful word embeddings and BiLSTM's strength to capture long-term dependencies. The experiments demonstrated superior performance over traditional models, achieving 80.74%, 92.36%, and 82.25% accuracy on the Twitter US Airline, IMDb, and Sentiment140 datasets, respectively. In [10], stress tests were carried out on transformer-based models, including RoBERTa, to assess their robustness in NLI and QA tasks. The findings showed that although these models exhibit enhanced robustness compared to recurrent neural networks, they remain susceptible to certain adversarial inputs, highlighting the need for further refinement. Collectively, these studies underscore the efficacy of RoBERTa in various NLP tasks while also pointing out areas where traditional and deep learning models can be further improved to achieve greater robustness and accuracy.

The applicability of sentiment analysis extends across multiple domains. For instance, it has been utilized to evaluate user opinions on e-commerce platforms [4] and social media platforms such as Twitter and YouTube, which are critical for gauging public sentiment on a variety of issues [11, 12]. In public health, sentiment analysis has facilitated the evaluation of public attitudes toward epidemics, such as COVID-19 and Monkeypox, through hybrid models such as CNN-LSTM [12, 13]. Tourism also benefits from sentiment analysis, which has been applied to assess tourist reviews of destinations. Such insights enable service providers and policymakers to enhance offerings and promote tourism more effectively [14, 15]. Advances in sentiment analysis have also involved the development of innovative techniques and models. Feature selection approaches such as Term Frequency-Inverse Document Frequency (TF-IDF), and information gain have demonstrated efficacy in improving model performance [6]. Deep neural network architectures, including Convolutional Neural Networks (CNNs) and LSTMs, have further contributed to enhancing both accuracy and processing speed when applied to complex datasets [1, 8]. In parallel, user-centric tools, such as web-based applications to predict sentiment scores on e-commerce platforms and identify discrepancies between user reviews and ratings, have facilitated greater accessibility and engagement [4].

In summary, sentiment analysis continues to evolve by integrating cutting-edge techniques, robust tools, and domain-specific applications. These advances enable researchers and practitioners to capture textual meaning and context with increasing accuracy while optimizing data processing and decision-making across diverse fields [8]. Therefore, sentiment analysis has demonstrated its versatility and impact across diverse fields, from business and politics to public health and tourism. Integration of advanced techniques, such as hybrid models that combine sequential and transformer-based architectures, has significantly enhanced the accuracy and efficiency of sentiment analysis. Innovations in data augmentation, feature selection, and deep learning architectures have further addressed linguistic diversity and data imbalance, enabling models to process complex datasets with precision. Furthermore, the development of user-friendly tools has expanded the accessibility and applicability of sentiment analysis, fostering greater engagement and practical utility. As the field continues to evolve, these advances improve the ability to extract meaningful insights from textual data and empower strategic decision-making, driving innovation across a wide array of domains.

II. MATERIALS AND METHODOLOGY

This study utilizes a combination of traditional machine learning, deep learning, and transformer-based models for sentiment analysis of Amazon book reviews. Figure 1 shows a research method diagram illustrating a systematic four-phase approach for a sentiment analysis or text classification project:

1) Dataset Collection:
    a) Data is loaded from Kaggle
    b) Involves metrics and PSD revision data
2) Preparation:
    a) Feature extraction using TF-IDF
    b) Word embedding implementation
    c) Sentiment labeling using a 1-5 scale rating system
    d) Split data into training (70%) and testing (30%) sets
3) Building and testing multiple models:
    a) Naive Bayes (NB)
    b) KNN
    c) CART
    d) LSTM
    e) RoBERTa
4) Performance comparison and summary based on:
    a) Accuracy
    b) Precision

c) Recall

d) F1-Score

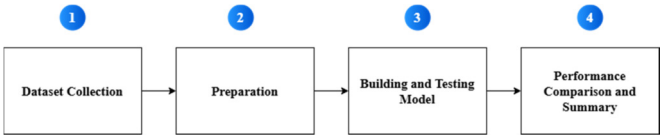

Fig. 1. Research method framework.

The research follows a comprehensive sequential workflow that comprises four main phases: dataset collection, preparation, model development, and performance evaluation. Starting with data collection from Kaggle [16], the process goes through careful preparation steps including feature extraction using TF-IDF, word embedding implementation, and sentiment labeling on a 6-point scale, followed by a 70-30 train-test split. The method then progresses to model building and testing, implementing five different algorithms and comparing them. The final phase involves a thorough performance evaluation using standard metrics, culminating in a comprehensive summary of results that enables an objective comparison of model effectiveness in the given task.

Although this study used TF-IDF for feature extraction, future implementations could benefit from combining TF-IDF with n-grams or leveraging word embeddings to improve the traditional models' capacity to handle nuanced text data.

A. Dataset Collection and Preparation

This study uses a dataset of Amazon book reviews [16] associated with information such as reviewer names, locations, review dates, titles, and descriptions of the evaluated products. Each review also included a numerical rating ranging from 0 to 5, where 0 indicated that the book was not recommended, and 5 indicated a high recommendation. Reviews with fewer than three stars were classified as negative, while those with three or more stars were categorized as positive. Ambiguous reviews with unclear polarity were removed from the dataset.

B. Word Embedding and Feature Extraction

For feature extraction, this study used two methods:

- TF-IDF was applied to convert the cleaned text into a numerical format suitable for traditional machine-learning models. A feature set of 5,000 words was created from the training data, representing the importance of each word within the overall document context.
- Word Embedding for deep learning: The text was tokenized using Keras's tokenizer, and the tokenized sequences were padded to a uniform length of 100 words. This process allowed deep learning models, such as LSTM, to efficiently process the text.

C. Sentiment Labeling

Sentiment labels were generated based on the review ratings. Ratings of 3 and above were labeled as positive (1), while ratings below 3 were labeled as negative (0). Additionally, the NLTK's Sentiment Intensity Analyzer was used to analyze text polarity and verify the sentiment labels.

D. Machine Learning Models

Multinomial Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, particularly effective for text classification and document categorization tasks where features represent word counts or frequencies [17, 18]. The algorithm assumes conditional independence between features and uses the multinomial distribution to model the probability of observing specific word counts, making it computationally efficient and surprisingly accurate despite its simplicity in NLP applications [19].

Decision Tree (DT) algorithms operate through recursive binary splitting of the feature space to create a tree-like structure of decision rules that minimize impurity measures such as the Gini index or entropy, making them effective for both classification and regression tasks [20, 21]. The mathematical formulation for the DT optimization problem aims to find the optimal split at each node to maximize information gain, expressed as:

$$InformationGain = I(parent) - \sum_{j=1}^{k} (n_j/n) * I(j)$$

where $I()$ represents the impurity measure (Gini or entropy), parent is the current node being split, k is the number of child nodes after splitting, n_j is the number of instances in child node j, and n is the total number of instances in the parent node. For entropy:

$$I(node) = -\sum_{i=1}^{c} p_i * log2(p_i)$$

For Gini:

$$I(node) = 1 - \sum_{i=1}^{c} p_i^2$$

where c is the number of classes and p_i is the proportion of instances belonging to class i in the node.

KNN is a non-parametric, instance-based learning algorithm that classifies or predicts values based on the majority vote or weighted average of the k closest training examples in the feature space [22, 23]. The algorithm functions by calculating distances between data points and making predictions based on the characteristics of neighboring points, with its effectiveness heavily dependent on the choice of distance metric and the value of k, making it particularly useful in pattern recognition and data mining applications [24]. The distance calculation and prediction for classification is given by:

$$\hat{y} = mode(\{y_i : i \in Nk(x)\})$$

and for regression:

$$\hat{y} = (1/k) * \Sigma\big(i \in Nk(x)\big)y_i$$

The distance metrics are as follows:

- Euclidean: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- Manhattan: $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$
- Minkowski: $d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$

where $\hat{y}$ is the predicted value, $Nk(x)$ represents the k nearest neighbors of point x, y_i is the target value of neighbor i, x_i and y_i are feature values of points x and y, n is the number of features, and $\cdot p$ is the power parameter for the Minkowski distance. For each model, performance metrics such as accuracy, precision, recall, and F1-score were calculated.

E. Deep Learning Model

LSTM networks are specialized Recurrent Neural Networks (RNNs) designed to overcome the vanishing gradient problem in traditional RNNs by incorporating memory cells with controlled information flow through input, forget, and output gates [25, 26]. This architecture enables LSTMs to learn long-term dependencies in sequential data, making them particularly effective for tasks such as NLP, time series prediction, and speech recognition, where they can selectively remember or forget information over extended sequences [27]. The gate and cell equations are:

- Forget gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- Input gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- Candidate cell state: $\tilde{c}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$
- Cell state update: $c_t = f_t * c_{t-1} + i_t * \tilde{c}_t$
- Output gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- Hidden state: $h_t = o_t * tanh(c_t)$

where σ represents the sigmoid function, $tanh$ is the hyperbolic tangent function, denotes matrix multiplication, $*$ represents element-wise multiplication, $[h_{t-1}, x_t]$ indicates the concatenation of h_{t-1} and x_t, W and b are weight matrices and bias vectors, respectively, t denotes the current time step, $t-1$ represents the previous time step, f_t, i_t, o_t are the forget, input, and output gates, c_t is the cell state, h_t is the hidden state, x_t is the input at time t, W_f, W_i, W_c, and W_o are the respective weight matrices, and b_f, b_i, b_c, and b_o are the respective bias vectors.

F. Transformer-based Model: RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is an enhanced variant of BERT that improves upon the original architecture through dynamic masking, larger batch sizes, and longer training on more diverse data, while removing the next sentence prediction objective [28, 29]. The model leverages the transformer architecture's self-attention mechanism and employs masked language modeling for pre-training, achieving state-of-the-art performance across various NLP tasks through its robust optimization strategies and enhanced training method [30]. The proposed architecture implements a sophisticated attention and output calculation framework, fundamentally based on the multi-head attention mechanism. The calculation of attention and output in the proposed architecture adheres to the following framework. The multi-head attention mechanism is expressed as:

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h)\mathrm{W}^{\mathrm{O}}$$

where each individual attention head is defined as:

$$head_i = Attention(QW_i, KW_i, VW_i)$$

The attention function itself operates as:

$$Attention(Q,K,V) = softmax(QKT/\sqrt{d_k})V$$

where Q, K, and V represent the query, key, and value matrices, respectively. The learnable parameter matrices QW, KW, VW, and WO facilitate model optimization. The term d_k corresponds to the dimension of the key vectors. The final output of the attention mechanism is computed using a residual connection followed by layer normalization:

$$FinalOutput = LayerNorm(MSA + FFN)$$

where Multi-head Self-Attention (MSA) and the Feed Forward Network (FFN) outputs are combined. For the training process, the loss function is derived from the masked language model objective, defined as:

$$MaskedLanguageModelLoss = -\sum_{i \in M} log\ P(xi|\tilde{x})$$

where M represents the set of masked token positions, x_i is the original token, and $\tilde{x}$ denotes the corrupted input sequence. This framework ensures effective learning of contextualized representations while preserving robust training dynamics.

G. Evaluation Metrics

The performance of each model was rigorously evaluated using a suite of evaluation metrics. These metrics serve as critical mathematical tools for quantifying the effectiveness of machine learning models, offering a comprehensive analysis of their performance across multiple dimensions, including accuracy, precision, and discriminative capacity [31, 32]. These metrics form the cornerstone of model evaluation by providing quantitative insights, facilitating meaningful comparisons, and guiding optimization efforts. Each metric contributes a distinct perspective, ranging from fundamental measures of classification accuracy to more sophisticated evaluations that address challenges such as class imbalance and ranking performance [33]. Each model's performance was evaluated using the following metrics.

Accuracy evaluates the proportion of correct predictions and is calculated as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

Precision measures the ratio of correctly predicted positive observations to the total predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Recall, also known as sensitivity or true positive rate, quantifies the ratio of correctly predicted positive observations to all actual positives:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

F1-score is a harmonic mean of Precision and Recall that provides a balanced measure for datasets with imbalanced class distributions:

$$\text{F1} - \text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision+Recall}}$$

where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, FN denotes False Negatives, TPR is the True Positive Rate, and FPR is the False Positive Rate.

III. RESULTS AND DISCUSSION

A. Comparison of Model Performance

Table I shows the comparative analysis of model performance metrics. RoBERTa emerged as the best performer, demonstrating exceptional capabilities across all evaluation criteria with remarkable scores of 96.20% accuracy, 99.77% precision, 96.40% recall, and 98.06% F1-score. LSTM achieved 88.37% accuracy, 99.63% precision, 88.65% recall, and 93.82% F1-score, while IG-KNN maintained similar metrics with 87.17% accuracy, 99.63% precision, 87.45% recall, and 93.14% F1-score. Despite achieving high precision at 99.74%, the IG-NB model performed lower on other metrics with 82.28% accuracy, 82.42% recall, and 90.26% F1-score. IG-CART recorded the lowest overall performance with 80.87% accuracy, 99.47% precision, 81.22% recall, and 89.42% F1-score. These results show a clear hierarchy in model performance, with RoBERTa significantly outperforming other models, particularly in maintaining high scores across all metrics. In contrast, the remaining models showed varying effectiveness with consistently high precision but lower accuracy and recall metrics.

TABLE I. COMPARISON OF MODEL PERFORMANCE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	91.20	99.64	91.48	95.39
RoBERTa	96.30	99.77	96.51	98.11
IG-NB	90.33	99.52	90.72	94.92
IG-CART	90.98	99.64	91.27	95.27
IG-KNN	89.24	99.52	89.63	94.31

The RoBERTa model outperformed the others with the highest accuracy and F1-score. LSTM also showed strong results, achieving notable accuracy and F1-score, while traditional models such as NB and DT had lower accuracy and recall due to the complexity of the text data. Figure 2 illustrates the comparative performance analysis between the LSTM and RoBERTa models through their respective confusion matrices. The LSTM model demonstrated moderate performance with 812 correct positive predictions but showed significant limitations with 104 false negatives and only one correct negative class prediction, alongside three false positives. In contrast, the RoBERTa model exhibited superior performance across all metrics, accurately predicting 883 positive cases with only 33 false negatives while achieving two correct negative class predictions with only two false positives. This comprehensive comparison reveals RoBERTa's notable advantages over LSTM, particularly its ability to maintain balanced prediction accuracy across both positive and negative classes, making it a more reliable choice for complex classification tasks such as sentiment analysis. The stark difference in false negative rates (104 for LSTM versus 33 for RoBERTa) highlights RoBERTa's enhanced ability to handle intricate data patterns and its overall superior performance in real-world applications where accurate classification is crucial.

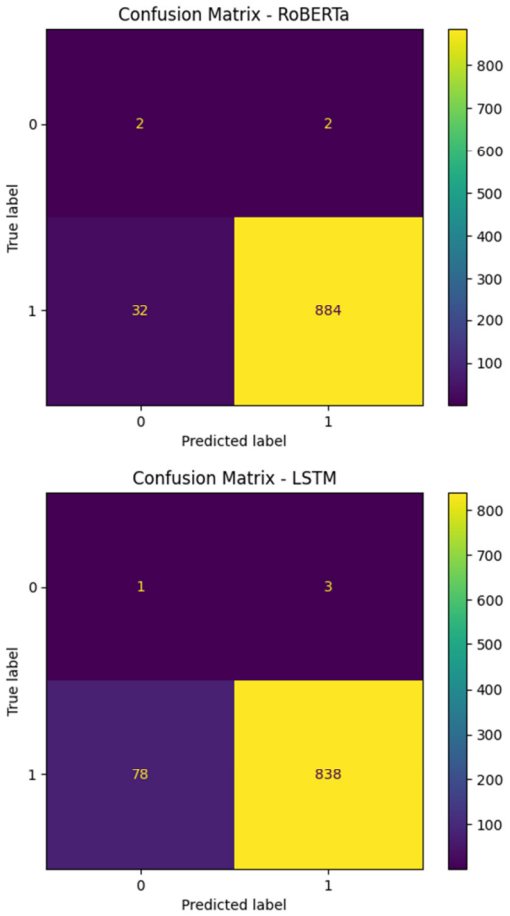

Fig. 2. Confusion matrices for LTSM and RoBERTa.

B. Performance Comparison of RoBERTa and LSTM Models

Figures 3 to 6 show a comparative analysis of RoBERTa and LSTM, revealing distinct performance patterns and learning dynamics across their training epochs. The RoBERTa model demonstrated superior performance with more stable convergence, showing oscillatory validation loss behavior that reached its optimal point (~0.105) at epoch five while achieving peak accuracy (~0.965) at epoch 4. The model's validation accuracy maintained an upward trend despite a brief decline at epoch 1, indicating effective generalization capabilities. In contrast, the LSTM model exhibited more complex behavioral patterns, with training loss consistently decreasing from ~0.8 to ~0.5 while maintaining stable validation loss between 0.3 and 0.4. Although LSTM's training accuracy steadily improved from ~0.55 to ~0.7, with validation accuracy fluctuating between 0.9-1.0, the notable gap between training and validation metrics suggested potential overfitting issues. Each model exhibited distinct strengths: RoBERTa maintained stable convergence, while LSTM effectively captured long-term dependencies. RoBERTa showed a more balanced performance between training and validation phases with smaller loss variations, while LSTM, despite achieving high validation accuracy, indicated a need for additional regularization strategies to optimize its performance. These findings suggest that RoBERTa's architecture is better suited

for achieving stable and generalizable results, while LSTM might benefit from architectural modifications to address its overfitting tendencies.

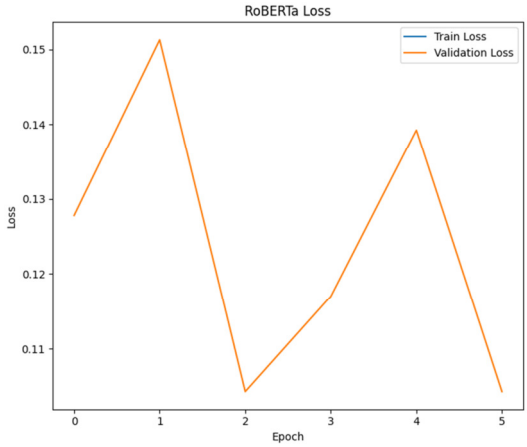

Fig. 3. Training dynamics (loss) for RoBERTa models over epochs.

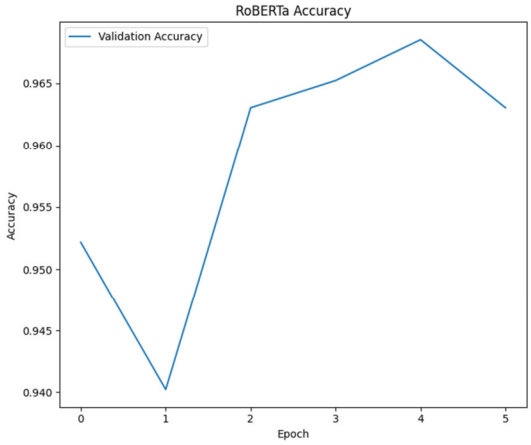

Fig. 4. Training dynamics (accuracy) for RoBERTa models over epochs.

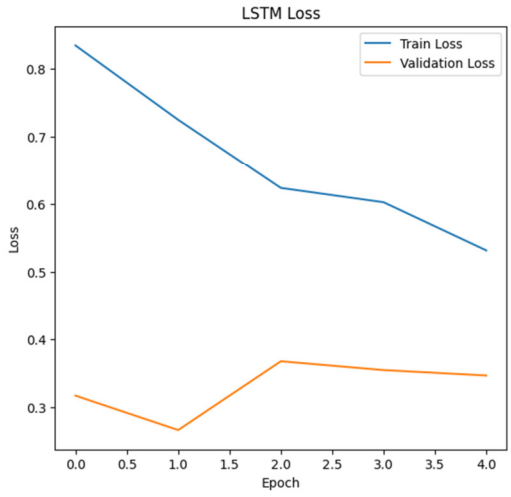

Fig. 5. Training dynamics (loss) for LSTM model over epochs.

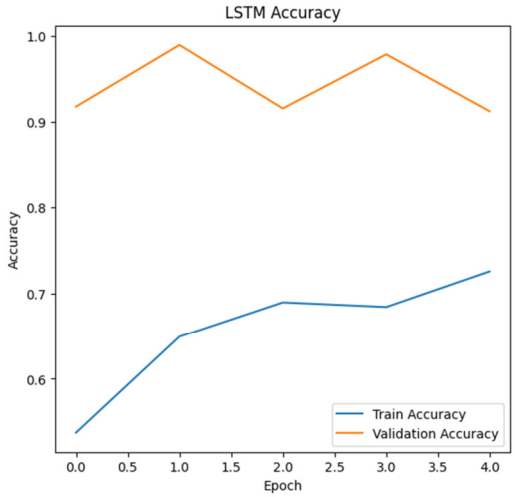

Fig. 6. Training dynamics (accuracy) for LSTM models over epochs.

C. Parameter Optimization for RoBERTa and LSTM

The parameter optimization process for RoBERTa and LSTM was carefully conducted to ensure optimal performance and mitigate overfitting. For RoBERTa, the learning rate was set to $2x10^5$ with a warm-up period covering 10% of the training steps, followed by a decay scheduler to stabilize convergence. A batch size of 32 and a maximum sequence length of 128 tokens were found to balance efficiency and contextual coverage. The AdamW optimizer, with a weight decay of 0.01 and a dropout rate of 0.1, was employed to further enhance generalization. The model was fine-tuned for 5 epochs, with early stopping applied if validation loss plateaued. In the case of LSTM, a learning rate of $1x10^{-3}$ with adaptive reduction was optimal, combined with a batch size of 64 and a sequence length of 100 tokens. The LSTM architecture incorporated 128 hidden units and utilized a dropout rate of 0.3 along with a recurrent dropout rate of 0.2 to control overfitting. The Adam optimizer facilitated efficient learning, and early stopping was employed after 20 epochs if no improvement was observed in the validation loss. These optimization strategies ensured robust model performance and convergence stability for sentiment analysis tasks.

IV. DISCUSSION

This study demonstrates the superior performance of the RoBERTa model compared to traditional machine learning models and LSTM-based deep learning approaches in sentiment analysis of Amazon book reviews. RoBERTa achieved an accuracy of 96.30% and an F1-score of 98.11%, confirming its effectiveness in capturing complex semantic relationships. These results align with prior research emphasizing the strengths of transformer-based models in processing nuanced textual data [5, 28, 34].

In contrast, traditional models such as NB, CART, and KNN, which rely on feature extraction methods such as TF-IDF, showed lower accuracy scores ranging from 89.24% to 90.98%. Although these methods are suitable for basic classification tasks, they struggle in capturing contextual meaning [6, 35]. For instance, the NB model, despite its high

precision (99.52%), exhibited lower recall (90.72%) and an F1-score of 94.92%, indicating its limited generalization across diverse sentiment classes [7, 18, 36]. The LSTM model performed better than traditional approaches, achieving an accuracy of 91.20% and an F1-score of 95.39%. Its strength lies in capturing long-term dependencies in sequential data [25]. However, its sequential nature limits scalability and efficiency, making it less effective than transformer models such as RoBERTa, which process data in parallel. Additionally, the LSTM model showed overfitting tendencies, a limitation noted in previous studies [26, 27]. Comparative studies [6] reinforce RoBERTa's superiority, showing its robustness across all evaluation metrics, with an accuracy of 96.20%, precision of 99.77%, recall of 96.40%, and an F1-score of 98.06%. Although LSTM models and IG-based models demonstrated competitive precision, they fell short in accuracy and recall. For instance, IG-KNN achieved accuracies between 84.98% and 87.17%, and IG-NB displayed precision above 99% but inconsistent accuracy (82.28% to 87.23%). These results suggest that although IG-based models are computationally efficient, they are best suited for simpler tasks.

RoBERTa's success can be attributed to its multihead self-attention mechanism and dynamic masking, allowing it to capture semantic complexity without extensive feature engineering [28]. This makes it ideal for applications that require deep contextual understanding, such as e-commerce sentiment analysis, where accurate interpretation of customer opinions is critical [37]. However, RoBERTa's high computational demands pose challenges for deployment in resource-limited environments. Future research should explore optimization techniques such as knowledge distillation, model quantization, and pruning to reduce resource requirements. Furthermore, hybrid models that combine RoBERTa with LSTM or GRU have shown promise [5, 8], leveraging RoBERTa's contextual strengths with the ability of recurrent models to capture temporal dependencies. Expanding this research to cross-domain and multilingual datasets could further assess RoBERTa's adaptability and scalability. Incorporating advanced feature extraction techniques such as n-grams or word embeddings into traditional models may also improve their performance in handling more complex text data. In summary, RoBERTa's ability to balance accuracy and contextual understanding makes it a powerful tool for sentiment analysis tasks. However, trade-offs between performance and computational efficiency must be considered based on application needs. ROBERTa is ideal for high-accuracy tasks, while traditional models remain valuable for low-resource and real-time scenarios.

V. CONCLUSION

This study highlights the superior performance of the RoBERTa model in the sentiment analysis of Amazon book reviews, achieving an accuracy of 96.30% and an F1-score of 98.11%. RoBERTa outperformed traditional machine learning models (NB, KNN, CART) and deep learning approaches (LSTM), demonstrating its exceptional ability to process complex and semantically diverse textual data. These findings confirm RoBERTa as an ideal tool for practical applications, such as e-commerce sentiment analysis and public opinion monitoring, where precision in understanding user sentiment is critical. The results underscore the transformative potential of transformer-based architectures in addressing linguistic complexity and imbalanced datasets. However, the study also highlights the high computational costs associated with RoBERTa, contrasting with the efficiency of traditional models and the balanced performance of LSTM models. Future research should focus on optimizing RoBERTa's computational efficiency using techniques such as model distillation, quantization, or hybrid approaches to enhance accessibility for broader applications.

Although this study focused on Amazon book reviews in English, the method could be extended to other domains, including social media, healthcare, and tourism. Further research should evaluate the models on multilingual and cross-domain datasets to assess their robustness and adaptability to diverse linguistic and contextual nuances. Additionally, exploring hybrid models that integrate transformer-based architectures with sequential models may offer further performance improvements. In conclusion, RoBERTa's accuracy and contextual processing capabilities establish it as a leading model for sentiment analysis. By addressing its computational demands and expanding its application to different languages and domains, RoBERTa can be a versatile and powerful tool for real-world sentiment analysis tasks.

ACKNOWLEDGEMENT

The author wishes to express profound gratitude to a remarkable intellectual contributor, whose unwavering support and insights have been invaluable throughout this endeavor. Special recognition is due to Tassanee Lunrasri, whose guidance, knowledge, and dedication have greatly enriched the depth and quality of this work.

REFERENCES

[1] A. Alsaeedi and M. Zubair, "A Study on Sentiment Analysis Techniques of Twitter Data," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, 2019, https://doi.org/10.14569/IJACSA.2019.0100248.

[2] W. Chansanam and K. Tuamsuk, "Thai Twitter Sentiment Analysis: Performance Monitoring of Politics in Thailand using Text Mining Techniques," *International Journal of Innovation*, vol. 11, no. 12, 2020.

[3] S. Sweta, "Application of Sentiment Analysis in Diverse Domains," in *Sentiment Analysis and its Application in Educational Data Mining*, S. Sweta, Ed. Singapore: Springer Nature, 2024, pp. 19–46.

[4] N. Shrestha and F. Nasoz, "Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings," *International Journal on Soft Computing, Artificial Intelligence and Applications*, vol. 8, no. 1, pp. 01–15, Feb. 2019, https://doi.org/10.5121/ijscai.2019.8101.

[5] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, https://doi.org/10.1109/ACCESS.2022.3152828.

[6] A. S. Lv, D. Babu.M, A.Manonmani, Y. M. Reeja, M. S. S, and A. R. Kumar, "An Efficient Approach in Selection of Information-Gaining Features Using Sentiment Analysis," *Journal of Computational Analysis and Applications (JoCAAA)*, vol. 33, no. 05, pp. 719–725, Sep. 2024.

[7] J. M. T. Habib and A. A. Poguda, "Comparison of Deep Learning Sentiment Analysis Methods, Including LSTM and Machine Learning," *Open Education*, vol. 27, no. 4, pp. 60–71, Aug. 2023, https://doi.org/10.21686/1818-4243-2023-4-60-71.

[8] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences*, vol. 13, no. 6, Jan. 2023, Art. no. 3915, https://doi.org/10.3390/app13063915.

[9] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis." arXiv, Jun. 01, 2024, https://doi.org/10.48550/arXiv.2406.00367.

[10] C. Aspillaga, A. Carvallo, and V. Araujo, "Stress Test Evaluation of Transformer-based Models in Natural Language Understanding Tasks." arXiv, Mar. 27, 2020, https://doi.org/10.48550/arXiv.2002.06261.

[11] A. Rawat, H. Maheshwari, M. Khanduja, R. Kumar, M. Memoria, and S. Kumar, "Sentiment Analysis of Covid19 Vaccines Tweets Using NLP and Machine Learning Classifiers," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, May 2022, pp. 225–230, https://doi.org/10.1109/COM-IT-CON54601.2022.9850629.

[12] O. Iparraguirre-Villanueva *et al.*, "The Public Health Contribution of Sentiment Analysis of Monkeypox Tweets to Detect Polarities Using the CNN-LSTM Model," *Vaccines*, vol. 11, no. 2, Feb. 2023, Art. no. 312, https://doi.org/10.3390/vaccines11020312.

[13] K. K. Mohbey, G. Meena, S. Kumar, and K. Lokesh, "A CNN-LSTM-Based Hybrid Deep Learning Approach for Sentiment Analysis on Monkeypox Tweets," *New Generation Computing*, vol. 42, no. 1, pp. 89–107, Mar. 2024, https://doi.org/10.1007/s00354-023-00227-0.

[14] A. K. Laturiuw and Y. A. Singgalen, "Sentiment Analysis of Raja Ampat Tourism Destination Using CRISP-DM: SVM, NBC, DT, and k-NN Algorithm," *Journal of Information Systems and Informatics*, vol. 5, no. 2, pp. 518–535, May 2023, https://doi.org/10.51519/journalisi.v5i2.490.

[15] D. Suryadi and J. T. Sabarman, "Analyzing Restaurants in Tourism Destinations Through Online Reviews Using Topic Modeling and Sentiment Analysis," in *2023 IEEE 9th Information Technology International Seminar (ITIS)*, Batu Malang, Indonesia, Oct. 2023, pp. 1–6, https://doi.org/10.1109/ITIS59651.2023.10419923.

[16] "Amazon Books Reviews." Kaggle, Accessed: Jan. 04, 2025. [Online]. Available: https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews.

[17] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, 1998, vol. 752, no. 1, pp. 41–48.

[18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, 1st ed. Cambridge University Press, 2008.

[19] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Machine Learning: ECML-98*, 1998, pp. 4–15, https://doi.org/10.1007/BFb0026666.

[20] S. L. Salzberg, "C4.5: Programs for Machine Learning," *Machine Learning*, vol. 16, no. 3, pp. 235–240, Sep. 1994, https://doi.org/10.1007/BF00993309.

[21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Routledge, 2017.

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, https://doi.org/10.1109/TIT.1967.1053964.

[23] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992, https://doi.org/10.1080/00031305.1992.10475879.

[24] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, Art. no. 1883, https://doi.org/10.4249/scholarpedia.1883.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[26] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, Oct. 2000, https://doi.org/10.1162/089976600300015015.

[27] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM networks," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Montreal, Canada, 2005, vol. 4, pp. 2047–2052, https://doi.org/10.1109/IJCNN.2005.1556215.

[28] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, Jul. 26, 2019, https://doi.org/10.48550/arXiv.1907.11692.

[29] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019, https://doi.org/10.48550/arXiv.1810.04805.

[30] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020, pp. 2158–2170, https://doi.org/10.18653/v1/2020.acl-main.195.

[31] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." arXiv, Oct. 11, 2020, https://doi.org/10.48550/arXiv.2010.16061.

[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009, https://doi.org/10.1016/j.ipm.2009.03.002.

[33] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, https://doi.org/10.5121/ijdkp.2015.5201.

[34] T. Ngootip, "Enhancing Network Intrusion Detection in Cloud Computing Using a Deep Boltzmann Machine and LightGBM Ensemble Model: A Performance Evaluation on the NSL-KDD Dataset," *Sociolytics Journal*, vol. 1, no. 1, pp. 1–7, Sep. 2024.

[35] S. Pansayta and W. Chansanam, "Thai COVID-19 patient clustering for monitoring and prevention: data mining techniques," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, Mar. 2024, Art. no. 256, https://doi.org/10.11591/ijai.v13.i1.pp256-265.

[36] P. Manorom, U. Detthamrong, and W. Chansanam, "Comparative Assessment of Fraudulent Financial Transactions using the Machine Learning Algorithms Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Random Forest," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15676–15680, Aug. 2024, https://doi.org/10.48084/etasr.7774.

[37] M. K. Myee, R. D. C. Rebekah, T. Deepa, G. D. Zion, and K. Lokesh, "Detection of Depression in Social Media Posts using Emotional Intensity Analysis," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16207–16211, Oct. 2024, https://doi.org/10.48084/etasr.7461.