Journal Computer and Technology
Vol. 3, No. 1, July 2025, pp. 20-25
ISSN : 3048-1880 (Online)
DOI : https://doi.org/10.69916/comtechno.v3i1.355

FAKE REVIEW DETECTION ON DIGITAL PLATFORMS USING THE ROBERTA MODEL: A DEEP LEARNING AND NLP APPROACH

Zulpan Hadi[1], Lalu Moh. Nurkholis[*2], Bahtiar Imran[3], Selamet Riadi[4], Emi Suryadi[5]

[1,3,4,5]Rekayasa Sistem Komputer, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram
[2]Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Universitas Teknologi Mataram
Email: [1]zlpnhadi@gmail.com, [2]lalunurkholis1967@gmail.com, [3]bahtiarimranlombok@gmail.com, [4]didiriadijumantoro@gmail.com, [5]emisuryadi@gmail.com

ARTICLE HISTORY
Received: 21.06.2025
Revised: 07.07.2025
Published: 21.07.2025

ABSTRACT

Fake reviews have emerged as a serious threat to the integrity of digital platforms, particularly in e-commerce and online review sites. This study explores the application of RoBERTa (Robustly Optimized BERT Approach), a transformer-based architecture optimized for natural language processing (NLP), in automatically detecting fake reviews. The methodology includes data collection from online platforms, contextual feature extraction using RoBERTa embeddings, model training through supervised learning, and evaluation using classification metrics such as accuracy, precision, recall, and F1-score. The training results indicate a significant convergence trend in the training loss, while the validation loss remains relatively unstable, reflecting challenges in model generalization. Nevertheless, experimental results demonstrate that RoBERTa outperforms other approaches such as Logistic Regression PU, K-NN with EM, and LDA-BPTextCNN, achieving an accuracy of 86.25%. These findings highlight RoBERTa's strong potential in detecting manipulative content and underscore its value as an essential tool in building a transparent and trustworthy digital ecosystem.

Keywords: fake reviews, roBERTa, NLP, deep learning, classification text

1. INTRODUCTION

Fake reviews have become an increasingly pressing issue in today's digital era, especially within the realms of e-commerce and online review platforms. These deceptive statements, published online to artificially boost the reputation of a product, service, or company, often fail to reflect genuine customer experiences [1], [2], [3]. In recent years, the growing prevalence of fake reviews has not only misled consumers' purchasing decisions but also eroded public trust in digital platforms [3]. With advancements in technology and machine learning—particularly in natural language processing (NLP)—fake review detection has emerged as a highly active research area. A cutting-edge approach in this domain involves the use of RoBERTa (Robustly Optimized BERT Approach), a transformer-based model proven effective in analyzing and identifying linguistic patterns indicative of deception in reviews [4], [5].

Recent studies reveal that RoBERTa can achieve accuracy rates of up to 97% in detecting fake reviews, underscoring its capacity to capture deeper linguistic nuances and detect subtle anomalies in review content [6]. Sentiment analysis data often indicate that fake reviews tend to be "too good to be true" or inconsistent with the experiences shared by other users. This creates a disconnect between the textual content and the emotional tone expressed in such reviews [7]. In this context, deep learning techniques like RoBERTa offer significant advantages in processing unstructured data and capturing emotional context, which is often missed by traditional methods [8], [9]. To address this challenge, many review platforms have begun implementing AI-based detection systems to identify and remove fake content. Research by Sajid et al. has shown that hybrid approaches and transformer-based models such as RoBERTa offer promising results in the battle against misleading information [3], [9].

Several comparative studies have evaluated RoBERTa against other transformer models like BERT and XLNet in the domain of fake review detection. For instance, Kanmani and Surendiran demonstrated that transformer models—including RoBERTa—significantly outperform traditional machine learning techniques [8]. Additionally, research by Alsaad highlighted the high effectiveness of RoBERTa in filtering fake reviews, noting that although the model requires longer training times, the resulting accuracy justifies the computational cost [6].

In a systematic review on fake review detection, Mohawesh et al. observed that while various methods have been proposed, transformer-based models are gaining attention for their ability to tackle linguistic complexity and contextual ambiguity in review classification tasks [2].

Nonetheless, challenges persist. Hussain et al. found that many traditional approaches continue to fall short when confronted with increasingly sophisticated deception techniques [10]. This has led current research efforts to focus on developing hybrid strategies and integrating psychological factors, as noted by Hu et al., who emphasized that reviewer credibility and sentiment are crucial in determining review authenticity[11].

From a broader perspective, studies show that not only has NLP technology evolved, but so too have the manipulation techniques used in fake reviews to influence consumer behavior. Wu et al. emphasized the importance of understanding the complex features of fake reviews to improve algorithmic detection methods [12]. Clearly, models like RoBERTa open new avenues in text recognition technologies, with the potential to combat consumer manipulation on online platforms [13].

In conclusion, fake reviews represent an evolving digital threat that demands serious attention from researchers, marketers, and platform administrators. By leveraging advancements in detection technologies and sentiment analysis, we can foster a more transparent and trustworthy digital ecosystem—ultimately reducing the harmful impact of fake reviews in online marketplaces.

2. RESEARCH METHODS

Detecting fake reviews has become increasingly essential to safeguard the integrity of online marketplaces and protect consumers from deceptive content. In recent years, deep learning methods—particularly those grounded in natural language processing (NLP)—have shown substantial progress in addressing this issue. A leading model in this area is RoBERTa (Robustly Optimized BERT Approach), which has outperformed earlier transformer architectures such as BERT across various NLP tasks, including sentiment analysis and text classification [16].

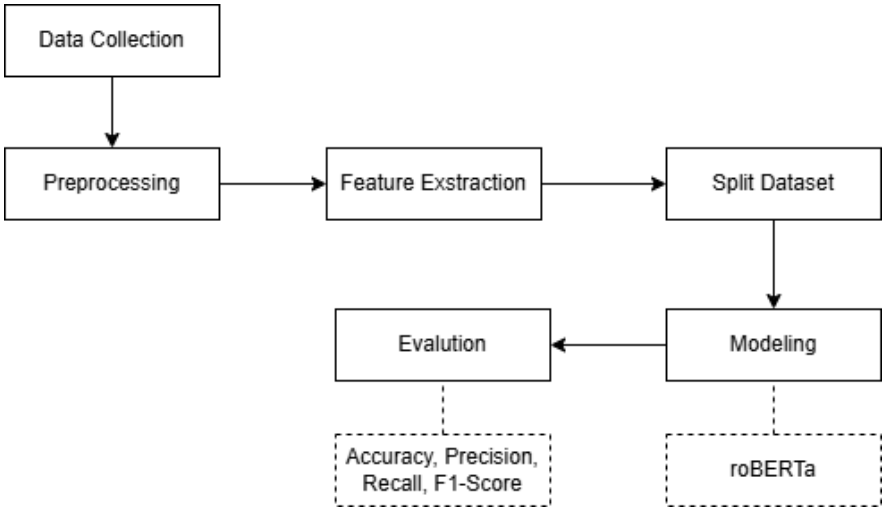

Figure 1. Research Flow

2.1. Data Collection

In this study, the authors utilized a hotel review dataset originally developed by Ott et al. [14], which has also been employed in prior research by Rout et al. [15]. Initially, Ott et al. constructed a dataset consisting solely of positively framed fake reviews. These deceptive reviews were gathered via Amazon Mechanical Turk (AMT), where crowd workers were instructed to write favorable reviews as if they were hotel marketing managers. The dataset targeted the 20 most popular hotels in the Chicago area, as listed on TripAdvisor. Through the AMT platform, a total of 400 positive fake reviews were collected. Additionally, 400 genuine reviews were obtained from sources such as TripAdvisor and Yelp. Although early classification results were promising, the dataset suffered from class imbalance, as it lacked negatively framed fake reviews. To address this issue, Ott et al. later enhanced the dataset by incorporating 400 negative fake reviews and an additional 400 authentic reviews, ultimately producing a balanced dataset containing both genuine and fake reviews across positive and negative sentiments.

2.2. Preprocessing

RoBERTa does not require aggressive text cleaning such as stemming or stopword removal because its tokenizer is designed to handle various forms of natural language contextually. However, some light preprocessing can still be applied to improve data consistency, such as removing unnecessary extra spaces or eliminating irrelevant non-alphabetic characters using regex techniques. Lowercase normalization is sometimes applied in other models, but for RoBERTa, this step is unnecessary since its tokenizer is case-sensitive by default and can handle both uppercase and lowercase letters according to their contextual usage.

2.3. Feature Extraction

RoBERTa operates by tokenizing textual input and generating contextual embeddings for each word using its transformer-based architecture. These tokenized text inputs are fed into the RoBERTa model, which outputs high-dimensional embeddings that capture the semantic meaning of each review [16]. These embeddings are critical, as they encode rich linguistic features that help distinguish between genuine and deceptive reviews.

2.4. Model Training

After the data is preprocessed and the features are extracted, the next step is to train the RoBERTa model on the dataset. This involves fine-tuning the pre-trained RoBERTa model on the specific task of fake review detection, applying a supervised learning approach where the model learns to classify reviews as genuine or fake based on labeled training data [19]. This is an iterative process that requires hyperparameter optimization to achieve better accuracy, often involving techniques such as cross-validation [10]. To support this training process, the next step is to configure the training parameters using TrainingArguments. In this stage, several important settings are defined, such as the output directory to save the training results (output_dir), the evaluation strategy to be performed at each epoch (eval_strategy="epoch"), and the learning rate set to 2e-5. In addition, the batch size for training and evaluation processes is set to 16 samples per device (per_device_train_batch_size=16, per_device_eval_batch_size=16), the total number of training epochs is set to 10 (num_train_epochs=10), and a weight decay of 0.01 is applied to reduce the risk of overfitting. Logging is directed to a specific folder (logging_dir="./logs"), and reporting to external platforms such as Weights & Biases (W&B) is disabled (report_to="none").

2.5. Evaluation Metrics

Evaluating the performance of the RoBERTa-based model is a crucial step. Researchers commonly use metrics such as accuracy, precision, recall, and F1-score to assess the model's effectiveness in detecting fake reviews [17]. In addition, the confusion matrix serves as a valuable visualization tool, offering insights into the number of false positives and false negatives, and further clarifying the model's classification behavior across different review categories.

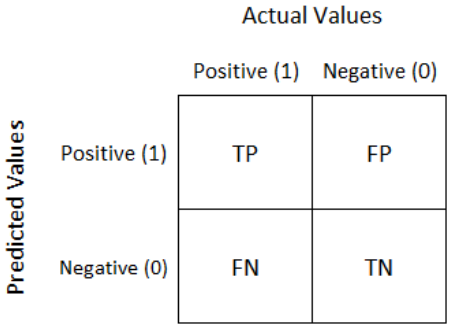

Figure 2. Confusion Matrix

To better illustrate how classification performance is measured, consider the analogy of a pregnancy test:

a. True Positive (TP): The test result is positive, and the person is indeed pregnant.
b. False Positive (FP): The test result is positive, but the person is not pregnant.
c. False Negative (FN): The test result is negative, but the person is actually pregnant.
d. True Negative (TN): The test result is negative, and the person is indeed not pregnant.

A classification model's effectiveness can be evaluated using a confusion matrix, which captures these four possible outcomes. Several key metrics are then derived from this matrix to assess model performance:

1. Accuracy

   Accuracy measures the proportion of correct predictions (both true positives and true negatives) out of the total number of predictions. A high accuracy indicates that the model performs well in general.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots (11)$$

2. Recall

Recall, also known as sensitivity or true positive rate, evaluates the model's ability to correctly identify all relevant instances (i.e., actual positives). It is especially important when missing a positive case is costly:

$$Recall = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (12)$$

3. Precision

Precision indicates how many of the predicted positive instances are truly positive. It reflects the model's ability to avoid false positives:

$$Precision = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (13)$$

4. F1-Score

The F1-Score is the harmonic mean of precision and recall, providing a balanced metric that is particularly useful when dealing with imbalanced datasets. It gives a single score that considers both false positives and false negatives.

$$F1 - Score = 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \dots\dots\dots (14)$$

3. RESULTS AND DISCUSSION

In evaluating a machine learning model, a comprehensive analysis of performance metrics and experimental outcomes is essential to determine the method's overall effectiveness and efficiency. Accordingly, monitoring the progression of loss values during training, along with a quantitative comparison of competing methods using evaluation metrics such as accuracy, recall, precision, and F1-score, serves as a fundamental basis for assessing the model's capabilities.

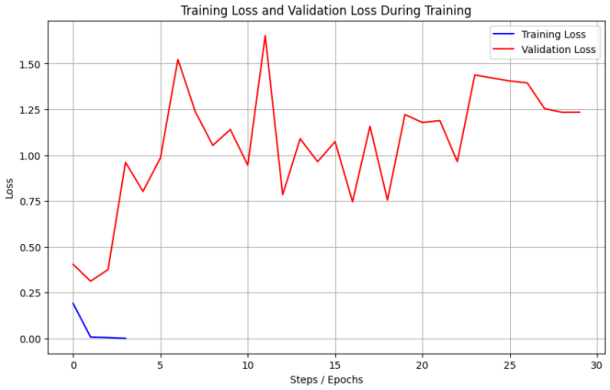

Figure 3. Training Loss and Validation Loss During Traning

In the context of machine learning model development, analyzing Training Loss and Validation Loss metrics plays a fundamental role in assessing both model performance and generalization capability. These metrics, typically visualized by the blue and orange curves respectively in a loss graph, illustrate the learning dynamics of the model over thirty training iterations—commonly referred to as epochs.

Our initial attention focuses on the blue curve, which represents the Training Loss. This curve reflects how effectively the model internalizes patterns and structures inherent in the training dataset through repeated iterations. A decreasing loss value indicates reduced prediction error on the training data. Notably, during the first six epochs, the training loss is marked as "No log," which may be attributed to the model's initialization phase or specific logging configurations. After this initial phase, the blue curve shows a substantial and continuous decline, converging toward zero. This convergence strongly suggests that the model is achieving high accuracy and has effectively learned from the training data.

Conversely, the orange curve depicts the Validation Loss, which is critical for evaluating the model's generalization ability—its capacity to make accurate predictions on previously unseen data. Unlike the relatively smooth downward trajectory of the training loss, the validation loss exhibits notable volatility. Although there are moments of decline (e.g., around epoch 2, between epochs 9 and 11, and near epoch 17), the overall curve is characterized by substantial fluctuations and several noticeable spikes, such as those at epochs 4, 6, 7, and 12. Moreover, a pronounced divergence is observed between the training loss—which approaches zero—and the validation loss, which remains consistently higher. This gap highlights potential overfitting, where the model performs exceptionally well on training data but struggles to generalize to new, unseen data.

Following this observation of training and validation loss trends—illustrating convergence and generalization behavior—the next stage involves a comparative evaluation of model performance based on quantitative results from multiple tested approaches. Table 1 presents experimental findings that demonstrate the superior performance of the proposed RoBERTa model, clearly outperforming alternative methods previously reported in the literature.

Table 1. Expreriment Result

Ref	Methods	Accuracy	Recall	Precision	F1-Score
[15]	Logistic Regresion PU)	0.83	-	-	-
	K-NN with (EM)	0.83	-	-	-
[18]	LDA-BPTextCNN	0.859	0.86	0.86	0.859
Proposed Method	roBERTa	0.8625	0.87	0.88	0.86

The comparison table clearly highlights the superiority of the proposed method, which employs the RoBERTa model, particularly in terms of classification accuracy. According to the data presented, RoBERTa achieves the highest accuracy score of 0.8625, outperforming all other baseline methods listed in the table.

For instance, techniques reported in reference [16], such as Logistic Regression (PU) and K-NN with EM, yield lower accuracies of 0.83. Similarly, the LDA-BPTextCNN approach from reference [17] records an accuracy of 0.859, which is slightly below that of RoBERTa.

This higher accuracy indicates that the RoBERTa model has a greater overall rate of correct predictions compared to the other competing methods. In practical terms, this means RoBERTa is more reliable in handling classification tasks, consistently producing accurate results across a wide range of cases—making it the most effective model based on this comparative evaluation. It is important to note that all the methods compared were tested using the same dataset, ensuring that the accuracy comparison is fair and consistent, without any bias arising from differences in data sources or dataset characteristics

4. CONCLUSION

This study demonstrates that the RoBERTa model delivers superior performance in the task of fake review detection, excelling in both accuracy and classification precision. By leveraging contextual embeddings and its transformer-based architecture, RoBERTa effectively captures subtle linguistic cues and patterns often missed by traditional approaches. Although the Validation Loss curve exhibits noticeable fluctuations, suggesting challenges in generalization, the model's overall performance remains consistently higher than that of the baseline methods. With an achieved accuracy of 0.8625, RoBERTa emerges as a promising solution for empowering digital platforms in their efforts to combat fake reviews and maintain consumer trust. Furthermore, this research opens avenues for future exploration, particularly in developing hybrid approaches and incorporating emotional or psychological features to enhance the robustness and reliability of fake review detection systems.

BIBLIOGRAPHY

[1] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating Sentiment-Preserving Fake Online Reviews Using Neural Language Models and Their Human- and Machine-based Detection," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.09177

[2] R. Mohawesh, S. Xu, M. Springer, M. Al-Hawawreh, and S. Maqsood, "Fake or Genuine? Contextualised Text Representation for Fake Review Detection," Academy and Industry Research Collaboration Center (AIRCC), Dec. 2021, pp. 137–148. doi: 10.5121/csit.2021.112311.

[3] A. Melleng, A. J. Loughrey, and P. Deepak, "Sentiment and emotion based text representation for fake reviews detection," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2019, pp. 750–757. doi: 10.26615/978-954-452-056-4_087.

[4] A. P. Rifai *et al.*, "DETECTION MODEL FOR FAKE NEWS ON COVID-19 IN INDONESIA," *ASEAN Engineering Journal*, vol. 13, no. 4, pp. 119–126, 2023, doi: 10.11113/aej.V13.19648.

[5] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Front Hum Neurosci*, vol. 17, Dec. 2023, doi: 10.3389/fnhum.2023.1292010.

[6] Maysara Mazin Badr Alsaad, "Transformer-Based Language Deep Learning Detection of Fake Reviews on Online Products," *Journal of Electrical Systems*, vol. 20, no. 3, pp. 2368–2378, May 2024, doi: 10.52783/jes.4083.

[7] M. Puttarattanamanee, L. Boongasame, and K. Thammarak, "A Comparative Study of Sentiment Analysis Methods for Detecting Fake Reviews in E-Commerce," *HighTech and Innovation Journal*, vol. 4, no. 2, pp. 349–363, Jun. 2023, doi: 10.28991/HIJ-2023-04-02-08.

[8] S. Kanmani and S. Balasubramanian, "Leveraging Readability and Sentiment in Spam Review Filtering Using Transformer Models," *Computer Systems Science and Engineering*, vol. 45, no. 2, pp. 1439–1454, 2023, doi: 10.32604/csse.2023.029953.

[9] T. Sajid *et al.*, "Analysis and Challenges in Detecting the Fake Reviews of Products using Naïve Bayes and Random Forest Techniques," Jun. 23, 2023. doi: 10.21203/rs.3.rs-2302761/v1.

[10] M. A. Al-Garadi *et al.*, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges," *IEEE Access*, vol. 7, pp. 70701–70718, 2019, doi: 10.1109/ACCESS.2019.2918354.

[11] S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth, and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," *IEEE Access*, vol. 8, pp. 26172–26189, 2020, doi: 10.1109/ACCESS.2020.2971087.

[12] Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decis Support Syst*, vol. 132, May 2020, doi: 10.1016/j.dss.2020.113280.

[13] P. Gupta, "Leveraging Transfer learning techniques-BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection MSc Research Project Data Analytics."

[14] M. Ott, C. Cardie, and J. T. Hancock, "Negative Deceptive Opinion Spam," Association for Computational Linguistics, 2013. [Online]. Available: http://plagiarisma.net

[15] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting Semi-Supervised Learning for Online Deceptive Review Detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017, doi: 10.1109/ACCESS.2017.2655032.

[16] N. Mkwananzi and H. Smuts, "Guidelines for Detecting Cyberbullying in Social Media Data Through Text Analysis," *International Journal of Social Media and Online Communities*, vol. 15, no. 1, pp. 1–13, Sep. 2023, doi: 10.4018/IJSMOC.330533.

[17] C. P. Barlett, C. Bennardi, S. Williams, and T. Zlupko, "Theoretically Predicting Cyberbullying Perpetration in Youth With the BGCM: Unique Challenges and Promising Research Opportunities," *Front Psychol*, vol. 12, Sep. 2021, doi: 10.3389/fpsyg.2021.708277.

[18] N. Cao, S. Ji, D. K. W. Chiu, M. He, and X. Sun, "A deceptive review detection framework: Combination of coarse and fine-grained features," *Expert Syst Appl*, vol. 156, Oct. 2020, doi: 10.1016/j.eswa.2020.113465.