**Fake Review Detection in E-Commerce**

Muhammad Daniyal1, a), Prabha Kumaresan1, 2, b), Mohammad Shadab Khan1, 3, c)

*1 Faculty of Computing & Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

*2 Centre of Natural Language Processing, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

*3 Centre of Advanced Analytics, CoE for Artificial Intelligence, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia.*

*c) corresponding author: shadab.khan@mmu.edu.my*

*a) 1221304153@student.mmu.edu.my*

*b) prabha.kumaresan@mmu.edu.my*

**Abstract.** The rapid expansion of e-commerce has led to a huge number of online reviews, which have a significant impact on consumers' purchasing decisions. The increasing number of fake reviews, which are meant to deceive potential customers, substantially harms the credibility of these platforms. This study investigates the effectiveness of transformer-based models, BERT for text classification in order to identify fake reviews. Accurately distinguishing between computer-generated (CG) and original review (OR) reviews is the primary goal. This was accomplished by implementing a structured preprocessing pipeline that included tokenization, Stopword removal, emoji removal, and text cleaning. BERT has exhibited well performance in terms of accuracy (92%) and F1-score (92%). From these, it can be concluded that the transformer model works well, and further research will focus on other transformer models, on deployment in real time, and on hyperparameter adjustment.

## **INTRODUCTION**

With the rapid rise of e-commerce, online reviews now have a greater influence on a consumer's purchase decision. However, with fake reviews on the platform, users could get confused and dilute the credibility of the platform. Such acts could manipulate product ratings, damage the reputation of companies, especially small businesses that depend heavily on genuine feedback, and diminish customer trust.

Such algorithms are still limited in scalability and accuracy due to lack of labelled data and prevailing complex language patterns of fraudulent content [1],[2]. Review manipulation takes a new method from time to time, which goes undetected in current methods, thus allowing highly sophisticated fraud to go through the crevices of detection. Hence, this problem demands the need for more refined automated systems that can carefully differentiate fake reviews from the real ones. Without resolution to these problems, there might indeed be a breach of equal opportunity, meaningful user experiences, and, eventually, the very integrity of e-commerce ecosystems.

The goal of the present research is to utilize natural language processing (NLP) and self-supervised machine learning for building a reliable fake review detection system. The key objectives involve reducing the need for labelled data, evaluating the generalizability of the model to various datasets, and tuning preprocessing pipelines (tokenization, text cleaning, and feature extraction) to better analyze textual patterns. Increasing classification accuracy without compromising the efficiency of computations is the primary concern of the project to enable scalability in real-world usage. Moreover, to enhance customer trust in internet reviews, enable well-informed buying decisions, and assist platform with efficient tools for dealing with fraud activity. In addressing the linguistic complexity as well as the data imbalance present in current e-commerce environments, the system design places a strong emphasis on adaptation.

## **LITERATURE REVIEW**

The foundation for detection of fake reviews began with research in 2008 [3], in which the existence of spam content on internet platforms was highlighted. The research served as a motivation to other researcher to develop automated systems using machine learning. Using supervised classification algorithms such as SVM and Naive Bayes, they achieve up to 97% accuracy [1]. However, problems involving dataset consistency and scalability exist. The results point up the use of hybrid approaches and strong pre-processing to address the noisy data.

Later studies went into deeper architectures. Contextual patterns were captured more accurately by recurrent neural networks (RNNs) and transformer-based models on benchmark datasets [4]. Likewise, the adjustment of pre-trained language models to detect fake reviews was documented with a 91% accuracy rate in another study [5]. Their studies concentrated on the use of contextual embeddings and hyperparameter optimization to improve classification.

Efficiency and generalizability are the main goals of recent research. During their evaluation of lightweight transformer architectures, it was found that performance declines with small size dataset, creating trade-offs across accuracy and model scale [6]. An ensemble technique of several transformer models was suggested by Rami Mohawesh et al., which performed much better than earlier approaches and achieved over 94% accuracy [7]. All of these findings suggest that while deep learning is still imperfect with respect to data imbalance and processing costs, it is the better approach to locating nuanced levels of language detail. Another study [8] uses the SELC model which is a corpus-based and lexicon-based combination for sentiment classification. Phase 1 balances positive and negative reviews by iterative lexicon growth with ratio control, while Phase 2 refines ambiguous reviews using SVM trained on Phase 1 outputs. When evaluated on Chinese product reviews, SELC minimised domain variance and obtained an 89.35% F1-score, which is 6.63% better than previous studies. Tests conducted on datasets that were unbalanced (up to an 8:2 positive/negative ratio) demonstrated strong performance (F1: 86.68–89.70%), demonstrating its domain versatility and capacity to mitigate bias through negation expansion and ratio control.

Building on these advancements, the suggested model stands out for achieving the ideal balance between performance and computational efficiency . Although RoBERTa-LSTM achieves great contextual accuracy (93.13%) and SELC excels in domain adaptation, these methods are limited by scalability or complexity. Thus, this study prioritises deployability without imposing significant computing overhead by addressing data imbalance through streamlined preprocessing and deliberate oversampling, resulting in 92% accuracy using the standard BERT architecture. This establishes BERT as a practically optimal approach that manages the trade-offs between resource efficiency, generalisation, and accuracy that have been highlighted throughout the research.

Firstly, this research is explained in Figure 1, which was conducted within the Anaconda Jupyter Notebook with Python’s libraries. Table 1 depicts the device specification that were used in the research.

## **METHODOLOGY**

**FIGURE 1**. The proposed method's comprehensive framework

**TABLE 1.** Device’s Specification

|  |  |
| --- | --- |
| **Processor** | 11th Gen Intel(R) Core(TM) i7- 11800H @ 2.30GHz 2.30 GHz |
| **RAM** | 8 GB |
| **GPU** | NVIDIA GeForce RTX3050 Ti |

This study used a systematic framework that included data preprocessing, model training, and evaluation. First, a dataset of 40,433 labelled e-commerce reviews from Kaggle was analysed to identify linguistic patterns. Preprocessing included removing duplicates, lowercasing, stripping HTML tags, URLs, punctuation, and emojis. Stopwords were eliminated to focus on meaningful content, and message lengths were analysed to standardise input sequences. There were 37,511 reviews in the dataset after preprocessing, that includes 19255 Original Reviews (OR) and 18256 Computer-Generated (CG). To solve class imbalance, oversampling was used to provide a fair split of genuine and fake reviews.

Text data was tokenised using BERT specific techniques which are Next Sentence Prediction (NSP) and Masked Language Modelling (MLM), resulting in subword units that are compatible with transformer configurations. Sequences were trimmed to 128 tokens for consistency and because of computational power. The dataset was divided into three sets: training (70%), validation (15%), and test (15%), with a similar label distribution. Stratified sampling ensured unbiased splits, which were necessary for model evaluation.

Pre-trained BERT was optimised for binary classification. The model used a custom dense layer with softmax activation to produce probabilities. Model training used the features of the Adam optimiser with a learning rate of 1e-5 and categorical cross-entropy loss. Model were trained for three epochs with a batch size of eight to achieve a balance between computational efficiency and learning stability. Attention masks helped model focus on relevant text parts during training.

Accuracy, precision, recall, and the F1 score were used to evaluate performance. The loss and accuracy curves graphs represented training dynamics. Validation metrics were monitored to prevent overfitting. Ideally, the best result for a model is indicated from its balance of accuracy, computational efficiency, and generalisability. Table 2 displays a classification report of the proposed research model.

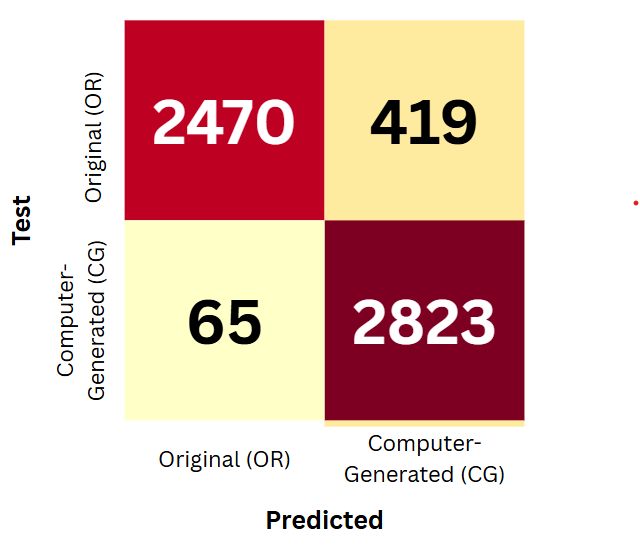
**RESULTS**

**TABLE 2.** Classification report from the model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0 (OR)** | 0.97 | 0.85 | 0.91 | 2889 |
| **1 (CG)** | 0.87 | 0.98 | 0.92 | 2888 |

The evaluation found differences in performance between the class as shown in Table 2 on original review and fake review. With a 92% accuracy rate, the model demonstrates its strong language understanding capacity. With a precision of 0.97 for OR and 0.87 for CG, it showed a high level of confidence in accurately classifying genuine reviews while maintaining an acceptable rate of fake review identification. It was successful in capturing the majority of the CG reviews, as shown by its recall of 0.85 for OR and 0.98 for CG. The precision and recall performances were balanced, as indicated by the F1-scores of 0.91 for OR and 0.92 for CG.

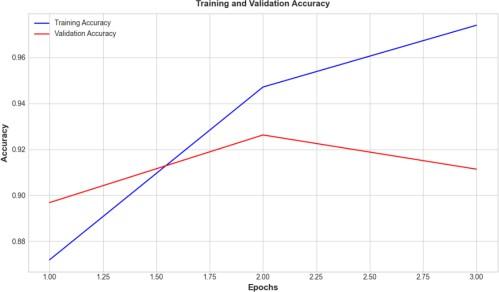
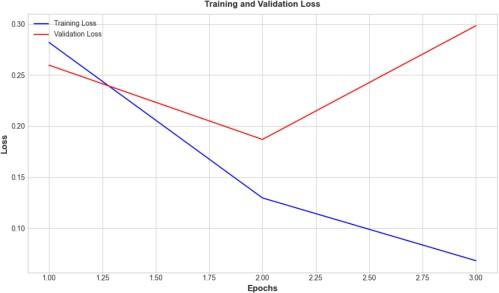
As shown in Figure 2, BERT identified 2470 OR and 2823 CG correctly, but it misclassified 65 CG reviews as OR and 419 OR reviews as CG. This shows a slightly higher rate of false positives for OR reviews but a strong performance in identifying CG reviews with a high true positive rate.



**FIGURE 2**. Confusion matrices result from the model

Effective learning was demonstrated by the model's training loss, which showed a steady decline over three epochs from about 0.30 to just under 0.10. But as epoch 2 took closer, the validation loss at first declined before slowly increasing in the final stages, indicating a possible overfitting issue.

As shown in Figure 3, BERT showed strong learning on the training set as its training accuracy increased from roughly 0.88 to over 0.96. Although it also improved, the validation accuracy did not keep up with the training accuracy's rate of improvement. It peaked at 0.92 and then began to decline slightly after epoch 2, which proves the loss graph's overfitting finding.



**FIGURE 3**. Model training and validation loss graph and accuracy graph

Based on benchmark studies as shown in Table 3, the proposed BERT model outperformed known techniques, achieving comparable results with 92% accuracy on the Kaggle dataset. Although it falls behind the RoBERTa-LSTM ensemble (93.15% on Deception), this result slightly surpasses the refined BERT (91% on Hotels+) and SELC model (89.35% on Chinese reviews) by 1.15% to 2.65%. The model's computational efficiency and balanced precision-recall trade-off (F1-scores: 0.91 OR, 0.92 CG) demonstrate its suitability for scaled implementation in e-commerce systems.

**TABLE 3.** Comparison performance between established models

|  |  |  |  |
| --- | --- | --- | --- |
| **Study** | **Model** | **Accuracy** | **Dataset** |
| [8] | SELC | 89.35% | Chinese product |
| [5] | Fine-Tuned BERT | 91% | Hotels+ |
| [3] | RoBERTa + LSTM | 93.13 | Deception |
| *This Study* | *BERT Base* | *92%* | *Kaggle* |

# **CONCLUSION AND FUTURE WORK**

Transformer-based models definitely have a high potential to improve the integrity of online review platforms, as demonstrated by the implementation and results of training models for fake review identification. BERT demonstrated well performance with high accuracy, achieving a balanced performance in precision and recall despite having minor overfitting issues [9]. This suggests that the model has a comprehension of the linguistic nuances that distinguish Computer-Generated Reviews from Original Review.

In future research, the present model would be enhanced, especially with hyperparameter tuning for generalization and less overfitting. Other machine learning [10] and transformer-based models, such as ROBERTA and ALBERT, will also be explored to circumvent this issue. Model compression techniques would also be considered to improve the efficiency. More specifically, factorised embedding parameterisation, which reduces dimensionality by breaking down the embedding matrix into smaller matrices, and cross-layer parameter sharing, which minimises different parameters between layers, will be studied. These techniques aim to reduce memory usage and speed up training, allowing deployment in production settings with limited resources while preserving detection performance.The challenges of practical implementation will also be analyzed, including how to incorporate these models into applications for real-time review analysis. This will entail adapting the model to work within the constraints of a production environment, taking into account factors such as real-time processing capability, user interface design, and API integration. To improve and widen the usage of these models in real-world scenarios, the current study will also look into the ethical implications of such systems in terms of consumer trust.

# **REFERENCES**

1. C. Silpa, P. Prasanth, S. Sowmya, Y. Bhumika, C.H.S. Pavan, and M. Naveed, “Detection of Fake Online Reviews by using Machine Learning,” in 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), (IEEE, Uttarakhand, India, 2023), pp. 71–77.
2. Wang, S., Yu, S., & Liu, P. “Combining BERT and graph convolutional networks for enhanced fake review detection. In Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence” (2025, February) pp. 24-28.
3. R. Mohawesh, H.B. Salameh, Y. Jararweh, M. Alkhalaileh, and S. Maqsood, “Fake review detection using transformer-based enhanced LSTM and RoBERTa,” International Journal of Cognitive Computing in Engineering 5, 2024, pp. 250–258.
4. N. Jindal, and B. Liu, “Opinion spam and analysis,” in Proceedings of the International Conference on Web Search and Web Data Mining - WSDM ’08, (ACM Press, Palo Alto, California, USA, 2008), p. 219.
5. D. Refaeli, and P. Hajek, “Detecting Fake Online Reviews using Fine-tuned BERT,” in 2021 5th International Conference on E-Business and Internet, (ACM, Singapore Singapore, 2021), pp. 76–80.
6. P. Gupta, S. Gandhi, and B.R. Chakravarthi, “Leveraging Transfer learning techniques- BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection,” in Forum for Information Retrieval Evaluation, (ACM, Virtual Event India, 2021), pp. 75–82.
7. R. Mohawesh, S. Xu, M. Springer, M. Al-Hawawreh, and S. Maqsood, “Fake or Genuine? Contextualised Text Representation for Fake Review Detection,” in Natural Language Processing, (Academy and Industry Research Collaboration Center (AIRCC), 2021), pp. 137–148.
8. L. Qiu, W. Zhang, C. Hu, and K. Zhao, “SELC: a self-supervised model for sentiment classification,” in Proceedings of the 18th ACM Conference on Information and Knowledge Management, (ACM, Hong Kong China, 2009), pp. 929–936.
9. J. Jayaram, Y. Kulkarni, L.V. Ganesh, P. Naveen, and E.A. Anaam, “Treatment Recommendation using BERT Personalization,” Journal of Informatics and Web Engineering **3**(3), 41–62 (2024).
10. K. Venkatesh, S. U. Jain, S. L. A. Ancy, N. M. Prabhu, D. Reena, and A. S. Kumar, “Detecting fake reviews for e-commerce platform using machine learning,” in Proc. 2025 Int. Conf. Comput. Robot., Test. Eng. Eval. (IEEE, ICCRTEE, May 2025), pp. 1–6.