Journal of Informatics and Web Engineering
Vol. 5 No. 1 (February 2026) eISSN: 2821-370X

An Empirical Evaluation of Machine Learning Methods and Text Classifiers for Sentiment Analysis of Online Consumer Reviews

Pei Qin Lo[1], Sew Lai Ng[2*], Li-xian Jiao[3]**
[1,2]Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Malaysia
[3] Computer Science Department, Jining Normal University, Ulaan Chab, Inner Mongolia Autonomous Region, 012000, China.
*Corresponding author: (slng@mmu.edu.my; ORCiD: 0000-0002-8610-581X)
**Corresponding author: (624292300@qq.com; ORCiD: 0009-0006-1735-910X)

Abstract - This study aims to identify the best predictive model for analysing online product reviews (OPRs) in the electronics industry, with a secondary focus on leveraging unstructured customer feedback to support product improvement. Using a dataset of 9,675 Oppo mobile phone reviews, this study employs three classification models—Random Forest, Support Vector Machine (SVM) and Logistic Regression–paired with Term Frequency-Inverse Document Frequency (TF-IDF) or bidirectional encoder representation transformer (BERT) as the embedding models to analyse customer sentiment and derive actionable insights. The methodology features a comprehensive analysis pipeline that includes text preprocessing with the Natural Language Toolkit (NLTK), feature extraction using) vectorization and BERT embeddings, and sentiment prediction through various classifiers. The results indicated that BERT was the most effective, achieving the highest accuracy, precision, recall, and F1-score. This superior performance stems from the Random's ability to handle high-dimensional, sparse data and effectively utilize the weighted word importance provided by TF-IDF, which makes it particularly well suited for sentiment classification tasks involving structured text representations. This study contributes to this field by providing an effective framework for analysing online reviews. This can help businesses understand customer needs for refining product offerings and laying the groundwork for future applications across different product categories.

Keywords—Sentiment Analysis, Online Product Review, Predictive Modelling, Text Classification, Natural Language Processing

Received: 09 April 2025; Accepted: 22 July 2025; Published: 16 February 2026

1. INTRODUCTION

In today's digital landscape, Online Product Reviews (OPRs) have served as a crucial resource for consumers and businesses, offering insights into customer preferences, satisfaction, and potential product improvements [1]. In the electronics industry, consumers frequently evaluate products based on key features such as battery life, application performance, and quality. Understanding this feedback can enable businesses to refine their offerings and remain competitive.

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

The extraction of useful information from large amounts of unstructured OPRs is extremely difficult [2], [3]. Customers are unable to assess product features when information is fragmented. Companies may lose the opportunity to understand where they are in the market. By implementing Natural Language Processing (NLP) techniques such as sentiment analysis, feature extraction, and trend analysis for consumer sentiment understanding, this study addresses these issues.

In addition, this study used different methods of key feature extraction from customer reviews. To conduct effective extraction, different types of greedy algorithms were selected, and their classification performance was evaluated on the OPRs. The effectiveness of each algorithm was analysed based on measurements, such as information comparability, representativeness, and diversity. This process enhances structured customer preference extraction, thereby allowing consumers to make effective decisions and strategies for their companies.

The objectives of this study were to investigate different machine learning techniques through comparative analysis and evaluate their methods for identifying patterns in OPRs. Existing analytical tools often rely on generic sentiment analysis or keyword frequencies, which may overlook comparative [4], [5]. By developing an approach to extract key customer preferences and facilitate feature-based comparisons, this study aims to transform unstructured data into actionable decision-making insights.

It is extremely difficult to extract useful information from a growing number of online reviews [1], [5]. Many online review analysis tools currently rely on sentiment analysis or measures that cannot discover certain customer preferences [6]. Moreover, some processes often do not possess the complexity required for detailed feature comparison. For instance, these tools can only offer a general overview and lack the depth required to help consumers make well-informed decisions when comparing two smartphones based on factors such as screen quality and battery efficiency [7]. Some machine learning models might not do well in identifying keywords, and this would reduce the precision of the analysis [6]. The high computational complexity and resource requirements of the current models limit their application. Therefore, it is less effective for real-time applications and large datasets [8].

To eliminate these restrictions, this study focuses on applying sentiment analysis to unstructured review data to extract key customer preferences [9]. Although it does not fully resolve all the limitations of existing models, it takes a step towards more structured preference extraction. In this study, Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) were applied to select the most relevant review sentences. The TF-IDF-based method prioritizes sentences rich in keywords to maximize feature coverage, whereas the BERT-based method leverages contextual embeddings to select diverse and semantically meaningful sentences. With such concerns being addressed, the study will provide customers with more informative data and the tools that they need to benefit businesses to have greater knowledge and address the needs of customers [2].

2. LITERATURE REVIEW

OPRs have become an important part of the online shopping experience. Reviews can provide useful information for both customers and businesses. These reviews can be found on e-commerce platforms, social media, and websites. Review platforms can provide valuable insights into customer experience with products and services. By referring to product ratings, it offered a detailed narrative about product strengths, weaknesses, and performance in real-world scenarios [9]. For customers, online reviews act as a source of trust and credibility, as they are perceived as authentic opinions from peers rather than polished marketing materials. This transparency helps customers evaluate whether a product aligns with their expectations before purchasing it. Beyond aiding customers, insights derived from online reviews are equally valuable to various stakeholders in the business ecosystem. Companies can study review trends to benchmark their products against competitors and identify areas that require improvement. By analysing sentiment patterns and recurring themes in customer feedback, businesses can gain a deeper understanding of how their products are perceived in the market, enabling them to make informed decisions for product enhancement and strategic development [10].

OPRs are typically presented in an unstructured format, making their analysis challenging [11]. Although structured data provide quantifiable insights, unstructured data are vast in volume and complexity, posing significant difficulties for computational interpretation [12]. Human language, with its nuances and implicit meanings, further complicates sentiment analysis [13]. Many existing sentiment analysis techniques focus on word-level sentiment identification but often struggle to extract meaningful insights from entire sentences or paragraphs [14]. NLP plays a crucial role in extracting sentiment from OPRs, allowing businesses to understand customer opinions regarding specific product features. Sentiment analysis typically categorizes textual data into positive, negative, or neutral classes, helping designers identify product aspects that require refinement [9]. Some of the words such as "like" and "hate" clearly indicate sentiment polarity [15]; however, challenges arise with some linguistic structures like negation can invert the intended sentiment and complicate machine interpretation. Various NLP algorithms are explored [16] to extract sentiments from product reviews and emphasized the value of tracking sentiment evolution over time to evaluate the impact of product changes. Similarly, "feature-based opinion summarization" method was introduced [17] to identify key product features by analysing noun usage, adjective proximity, and word frequency trends. Despite these advancements, implicit sentiment remains difficult to detect because machines struggle with statements where sentiment is not directly expressed such as "the camera will not fit easily in a pocket" highlighting the ongoing challenges in computational sentiment analysis.

Developments in NLP also present the possibility of a more profound understanding of customer preferences and product feedback beyond simple emotional classification. Although traditional sentiment analysis models are very good at detecting emotional polarities, they frequently struggle to comprehend the underlying causes, which restricts their applicability to decision makers and product designers [6]. Designers require analytical tools that not only reveal customers' emotions but also provide an explanation for those emotions to derive more actionable insights. To bridge this gap, a framework that distinguishes four essential components from product reviews: feature aspects (A), sentiment polarity (S), product characteristics (F), and detailed reasons (R) was developed [18]. Although extracting subtle characteristic aspects and causes remains a complex challenge, this framework provides a more structured understanding of consumer sentiments. In this case, Term Frequency Inverse Document Frequency (TF-IDF) is a widely used technique for handling massive unstructured textual data. TF-IDF was utilized [19] to detect repeated phrases and keywords from comments for the model to serve as a useful tool for discovering customer comment patterns. TF-IDF is very effective in identifying frequently discussed product attributes by estimating their relative importance within a large corpus. However, while TF-IDF does remarkably well in identifying relevant words, it fails to detect descriptive detail and finer linguistic cues that are crucial in capturing the emotional context of a customer in depth [9]. To address the limitations of quantitative keyword extraction over qualitative understanding, this limitation emphasizes the significance of adding more sophisticated models, such as those based on deep learning or context embedding, to TF-IDF. Ultimately, refining the use of the TF-IDF within a broader NLP framework will enhance its applicability in product analysis by helping businesses extract targeted and meaningful insights from many OPRs.

Various machine-learning models, such as Support Vector Machine (SVM), random forests, and logistic regression, have been widely used in sentiment analysis and aspect-based mining. These models can classify reviews by sentiment and pinpoint certain aspects of products so that businesses can position their products based on consumer choice [9]. In addition, transformer-based deep learning algorithms, such as BERT, have been proven to achieve improved performance in learning complex linguistic patterns and contextual features to support more delicate sentiment analysis and aspect extraction capable of handling implicit messages and semantic variation in customers' opinions [20].

SVM is widely used algorithms for text classification and sentiment analysis. It identifies an optimal hyperplane that separates sentiment classes while maximizing the margin [6]. SVM is robust and scalable, making it suitable for analysing large volumes of online reviews. However, it tends to reduce complex textual data to quantifiable metrics, potentially excluding key contextual details [6].

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their outputs to improve the classification accuracy. Random Forest is highly resistant to overfitting and excels in handling structured and unstructured data [20]. Random Forest model was successfully applied to e-commerce reviews [21], achieving high classification accuracy through feature selection and hyperparameter tuning. However, the Random's computational demands increase with the dataset size, making it less efficient for real-time applications.

Logistic Regression is a simple yet effective algorithm for binary and multiclass sentiment classification. It can model the probability of a review belonging to the sentiment class based on textual features [22]. Logistic Regression is computationally efficient and interpretable. These advantages render it suitable for small-to medium-sized datasets. However, the assumption of linear relationships between features and sentiment categories limits the ability to capture complex linguistic structures.

BERT is a powerful deep-learning sentiment classification model that learns contextual relationships in text via a self-attention mechanism [23]. BERT achieves state-of-the-art performance in a range of sentiment analysis tasks by considering the overall context from both directions and encoding complex word interactions. Although BERT requires only a small amount of labelled data, it is pre-trained on a large amount of text, enabling efficient transfer learning. Fine-tuning requires careful optimization to avoid overfitting, and its large number of parameters and computational requirements make it difficult to deploy it with limited resources.

3. DATA AND METHOD

In this study, the methodology established clear objectives and presented a comparative framework for evaluating the different preprocessing methods and classification techniques used to analyse unstructured online reviews. The primary goal was to determine the best model to effectively classify these reviews. The framework was designed to assess the performance of various approaches based on key performance indicators, including accuracy, precision, and recall. By systematically comparing these methods, this study identifies the most effective techniques for extracting meaningful insights from high-volume customer feedback, thereby enhancing the understanding of consumer sentiment and trends in OPRs.

Data.mendelay.com provides different types of datasets from e-commerce websites. From this website, a mobile phone with an Oppo branding dataset scraped from the Amazon API was selected for this project. The dataset has a total of 9675 reviews and attributes, such as reviewer name, review ID, Oppo phone types, phone ID, rating stars, review text, and date published. This dataset structure is particularly suited for the analysis objectives because it provides both structured and unstructured data. The combination of numerical ratings and detailed text reviews enables a comprehensive analysis of customer sentiment; product information allows for model-specific insights and comparisons. The size of the dataset provides a substantial sample for training reliable machine-learning models and extracting meaningful patterns in customer feedback.

Figure 1 shows the research flow of this study in analysing OPRs using NLP and machine-learning techniques. The process began with data collection, where review data were sourced from Mendeley.com, specifically extracting oppo phone reviews from Amazon. Data. Mendeley. com is a platform that offers diverse datasets from various e-commerce websites. In this project, a dataset of Oppo phone reviews crawled by the Amazon API was selected. The dataset contained 9675 reviews with attributes such as reviewer name, review ID, Oppo phone type, phone ID, star rating, review text, and date published. The dataset structure is suitable for analysis because it includes both structured and unstructured data, which allow comprehensive sentiment analysis. Additionally, product information allows for specific insights and comparisons. The size of the dataset provides a robust sample for training machine-learning models and identifying significant patterns in customer feedback.

Data preprocessing is a crucial step in processing raw text data for analysis, ensuring standardization of tokenization and feature extraction [9]. In this study, NLTK, a powerful Python library for text processing, was used to perform various preprocessing tasks. The process begins with text normalization. All review texts are converted into lowercase to maintain uniformity and avoid duplication due to case differences (e.g., “Battery life is EXCELLENT!” becomes “battery life is excellent.’). Next, tokenization splits text into smaller units, such as sentences and words, making it structured for machine learning models (e.g., “Battery lasts long and works perfectly” is tokenized into [“battery,” “lasts,” “long,”, “and,” works,” “perfectly”]). However, challenges such as merged tokens (e.g., “performanceoppo”) may need to be overcome because of inconsistencies in user-generated content. Lemmatization follows, reducing words to their base forms while preserving meaning (e.g., “running” → “run,” “batteries” → “battery”). Finally, the elimination of stop words that are plain words such as “and” “the,” “is” that give no meaningful information will further reduce the noise on computation making it more efficient (for example, “the battery lasts for a long time” becomes “battery lasts long time”).

Figure 2 shows the raw, unprocessed review data, which contain inconsistencies such as uppercase and lowercase variations, unnecessary punctuation, and unstructured text that can affect analysis accuracy. These raw data often include redundant words, typos, and formatting irregularities that can cause challenges for machine learning models in extracting meaningful insights.

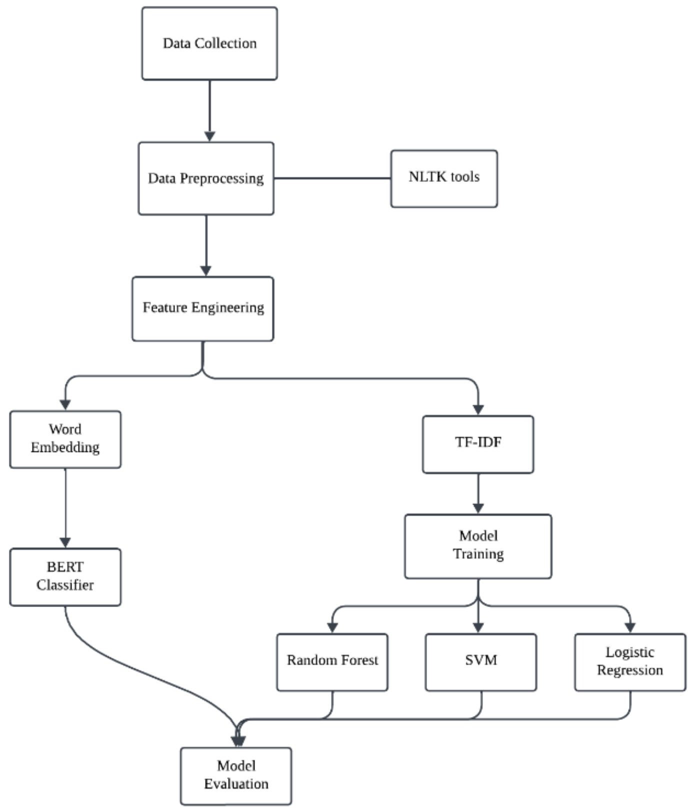

Figure 1. Research Flow

	Review_product	Rating_Star	Review
0	OPPO F3 Plus (Gold, 64 GB)	5	I bought OPPO F3 plus on 15.05.2017\n \n Thi...
1	OPPO F3 Plus (Gold, 64 GB)	5	Got it today. Build quality very good front ca...
2	OPPO F3 Plus (Gold, 64 GB)	5	Good phone I like it and I will give it 5 star...
3	OPPO F3 Plus (Gold, 64 GB)	5	Have used only for 3 weeks & thief snatched aw...
4	OPPO F3 Plus (Gold, 64 GB)	5	Very good quality\n \n \n\n READ MORE\n \n

Figure 2. Screenshot of Review Before Preprocessing

Figure 3 presents the same review data after the preprocessing steps, including text normalization, tokenization, lemmatization, and stop word removal. The processed text is now standardized, with all words converted to lowercase, irrelevant stop words removed, and words reduced to their root forms while maintaining their contextual meaning.

These refinements enhance the dataset quality, which helps to make the dataset more structured and suitable for sentiment analysis and classification tasks.

For feature extraction, BERT and TF-IDF were implemented concurrently to compare their effectiveness in representing sentiment classification text data. In the BERT token method, comments are first tokenized using BERT tokens, which tokenize text, add special tokens, and convert them into numerical input tensors. These include tokenization identifiers, attention masks, and segmentation identifiers. Tokenization was then input into a pre-trained text-box model, where the final state was embedded. These embeddings capture deeper semantics of the text and provide contextual numerical representations for each comment. Token embedding was used as a feature vector for classification in the input model.

	Review_product	Rating_Star	Review	word_tokens
0	OPPO F3 Plus (Gold, 64 GB)	5	bought oppo f plus best android mobile used pa...	[bought, oppo, f, plus, best, android, mobile,...
1	OPPO F3 Plus (Gold, 64 GB)	5	got today build quality good front camera capt...	[got, today, build, quality, good, front, came...
2	OPPO F3 Plus (Gold, 64 GB)	5	good phone like give star mak thxxx flipkart g...	[good, phone, like, give, star, mak, thxxx, fl...
3	OPPO F3 Plus (Gold, 64 GB)	5	used week thief snatched away phone train real...	[used, week, thief, snatched, away, phone, tra...
4	OPPO F3 Plus (Gold, 64 GB)	5	good quality read	[good, quality, read]

Figure 3. Review After Preprocessing

A traditional numerical text representation method was proposed using the TF-IDF method [6]. Each word is assigned a weight based on its importance. More weight is assigned to words that occur frequently in sentences, but rarely in other sentences. Frequently used words were assigned a lower weight. Each comment is transformed into a numerical vector depending on the significance of its terms to form a sparse matrix representation. Contextual embeddings were also performed to obtain the embedded word for the BERT classifier. These contextual embeddings are generated through the transformer of BERT, which processes each token in relation to all other tokens in the sequence. This allows the model to capture the meanings of words that vary based on the surrounding context and create word vectors.

Sentiment is measured at the sentence and word levels using different machine learning algorithms. To identify the most appropriate framework for analysing customer decision-making, in this study, different classification algorithms were implemented on TF-IDF features and BERT-based embeddings. Random Forest, SVM, Logistic Regression, and BERT classifier were the classification models chosen for classification. These algorithms learn from labelled and pre-processed data, and review ratings are aligned with the sentiment labels for each review. The purpose of this study is to discover the most accurate and trustworthy sentiment classification method by comparing the performance of each classifier with various feature engineering techniques.

To evaluate the performance of the classification models, a confusion matrix was used to compute key metrics, such as accuracy, precision, recall, and F1-score. After training all the classification algorithms on both BERT-based embeddings and TF-IDF features, each model was tested on the sample review, and the corresponding evaluation metrics were calculated. A confusion matrix was visualized using a heatmap, where darker colours represent higher prediction counts. By labelling on the y-axis and predicted labels on the x-axis, the matrix provides insights into correctly classified instances and misclassifications that help assess model performance and identify areas for improvement.

4. RESULTS AND DISCUSSIONS

4.1 Result of TF-IDF Pairs with Random Forest

Figure 4 shows the confusion matrix of the Random Forest classifier paired with TF-IDF, where the model correctly classified 477 negative, 84 neutral, and 2,075 positive reviews, indicating strong performance, particularly in recognizing positive sentiments. However, 105 negative reviews were misclassified as positive, while six were mislabelled as neutral, suggesting some difficulty in distinguishing negative from positive sentiments. Similarly, 112 neutral reviews were misclassified as positive and 27 were incorrectly labelled as negative, highlighting a challenge in accurately detecting neutral sentiments. Additionally, 16 positive reviews were misclassified as negative and one as neutral, showing minimal confusion for positive classifications. Overall, the low misclassification rates suggest that the model effectively captures sentiment distinctions, with positive reviews being the most accurately classified category. This strong performance can be attributed to the Random Forest ensemble learning approach, which aggregates predictions from multiple decision trees, reduces overfitting, and improves the classification accuracy. TF-IDF, as a feature extraction method, effectively captures keyword importance and frequency, enabling the model to identify dominant features of sentiment, particularly in positive reviews. However, the misclassification between negative and neutral sentiments could be attributed to the challenge of distinguishing these two similar sentiment categories, especially when they share overlapping keywords. However, the model's ability to accurately classify positive reviews suggests its robustness in handling more distinct sentiment expressions.

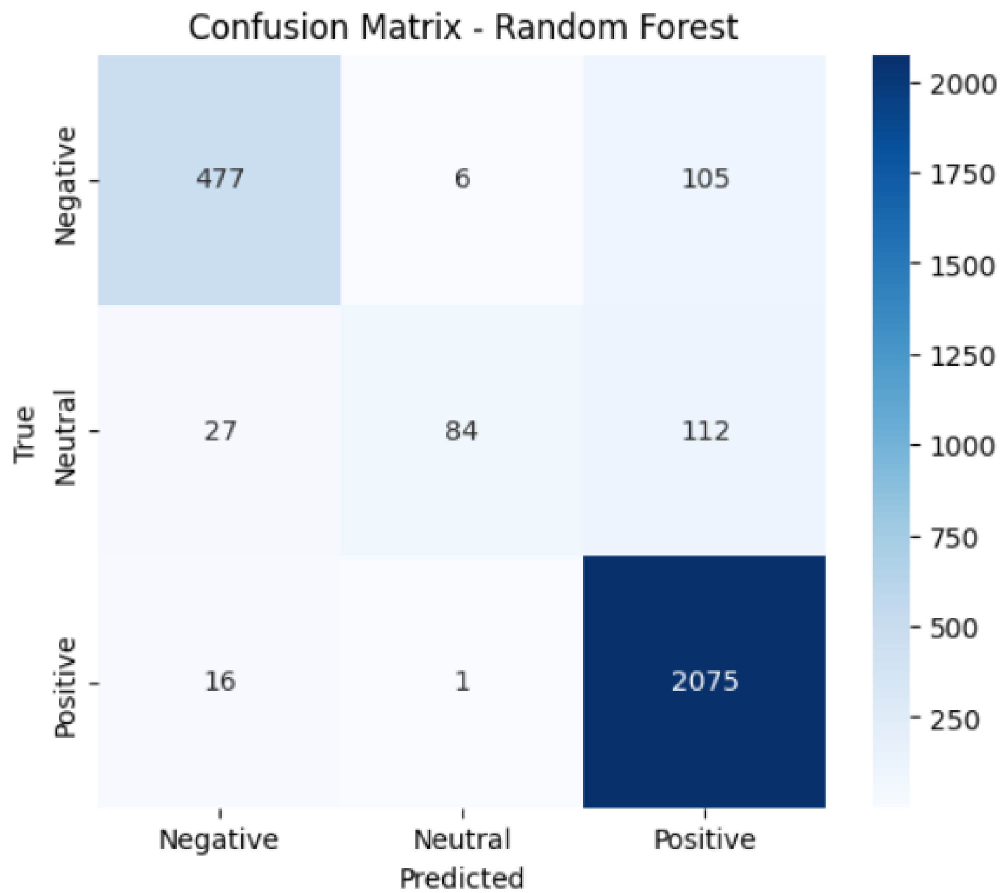

Figure 4. Confusion Matrix for Random Forest Paired with TF-IDF

4.2 Result of TF-IDF Pairs with Logistic Regression

Figure 5 presents the confusion matrix of the Logistic Regression model paired with TF-IDF, which illustrates its performance in classifying sentiment into negative, neutral, and positive categories. The model correctly classified 427 negative, nine neutral, and 2,056 positive reviews, indicating strong performance in detecting positive sentiment. However, 154 negative reviews were misclassified as positive and seven were mislabelled as neutral, suggesting that the model struggles to distinguish between negative and positive sentiments. Additionally, 161 neutral reviews were incorrectly identified as positive and 53 were misclassified as negative, indicating challenges in accurately detecting neutral sentiments. The misclassification of 30 positive reviews as negative and six as neutral suggests minor errors in identifying positive sentiments. Overall, while the model performs well in classifying positive reviews, it faces challenges in distinguishing between negative and neutral sentiments, leading to frequent misclassification. This result may stem from the linear nature of Logistic Regression, which simplifies decision boundaries and may not fully capture the complex, non-linear relationships between sentiment categories, particularly when distinguishing between closely related sentiments, such as negative and neutral. Although TF-IDF helps highlight significant keywords, it may not capture the subtleties of sentiment shifts in reviews, making it harder for the model to differentiate between these more nuanced sentiments. The dominance of positive reviews in the dataset may also lead the model to overfit towards positive sentiments, resulting in an increased tendency to misclassify negative and neutral sentiments as positive.

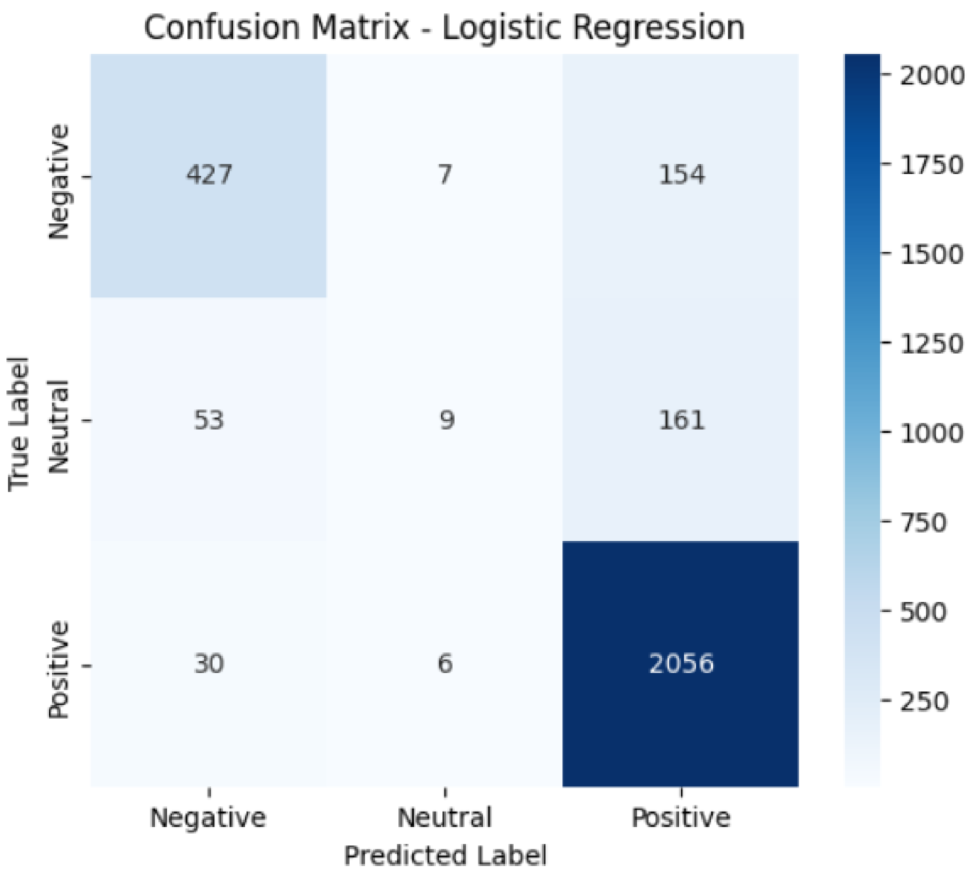

Figure 5. Confusion Matrix for Logistic Regression Paired with TF-IDF

4.3 Result of TF-IDF Pairs with SVM

Figure 6 presents the confusion matrix of the classifier paired with TF-IDF, which correctly classified 465 negative, 17 neutral, and 2,052 positive reviews, demonstrating strong accuracy in identifying positive sentiment. However, 116 negative reviews were misclassified as positive and seven as neutral, indicating that the model struggles to differentiate negative from positive sentiment. Similarly, 153 neutral reviews were mistakenly labelled as positive and 53 as negative, highlighting the challenges in recognizing neutral sentiments accurately. Additionally, 36 positive reviews were misclassified as negative and four as neutral, showing minor errors in distinguishing positive sentiments. Overall, the SVM model performs well in identifying positive reviews, but faces challenges in correctly classifying negative and neutral sentiments, often confusing neutral reviews with positive ones. This result can be attributed to SVM's tendency to maximize the margin between classes, which works well when there are clear distinctions between sentiment categories, such as positive sentiment, but struggles when the sentiment classes are more subtle, such as negative and neutral. Although TF-IDF effectively captures important terms, it may not account for the context in which words appear, leading to confusion between sentiments that share similar keywords or expressions. Additionally, SVM’s decision boundaries, especially when using a linear kernel, may not be flexible enough to distinguish fine-grained differences between negative and neutral sentiments, especially in cases where sentiment nuances are difficult to detect.

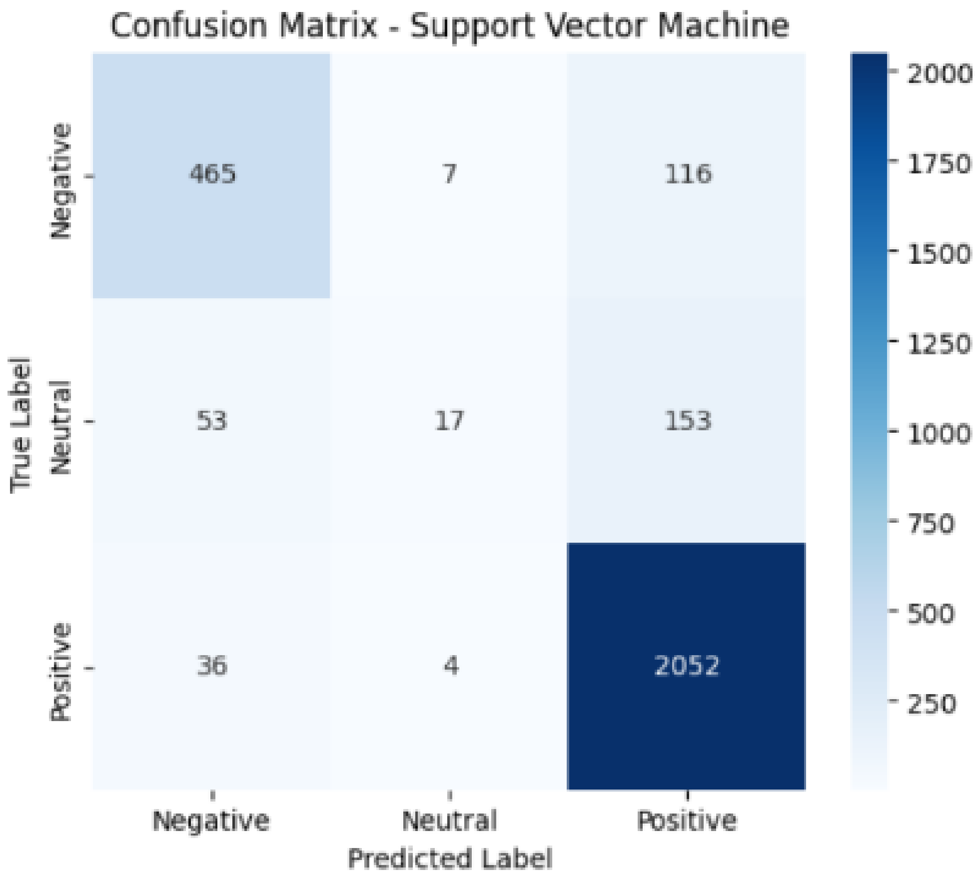

Figure 6. Confusion Matrix for SVM Paired with TF-IDF

4.4 Result of BERT Word Embedded and Classifier

Figure 7 presents the confusion matrix of the BERT model, which correctly classifies 502 negative, 99 neutral, and 2,043 positive reviews, demonstrating strong performance across all sentiment categories, particularly in identifying positive sentiment. However, 85 negative reviews were misclassified as positive (68) and neutral (17), indicating some challenges in distinguishing negative sentiments from others. Similarly, 114 neutral reviews were incorrectly labelled 88 as positive and 26 as negative, highlighting the difficulties in recognizing neutral sentiment, particularly when differentiating it from positive sentiments. Additionally, 60 positive reviews were misclassified 35 as negative and 25 as neutral, indicating minimal errors in positive sentiment recognition. Overall, the BERT model performed well across all sentiment categories, with its highest accuracy in positive sentiment classification and notable improvements in negative sentiment detection compared to traditional methods. This superior performance can be attributed to BERT's contextual embeddings, which capture semantic relationships and linguistic nuances beyond what static representations such as TF-IDF can achieve.

4.7 Discussion

Table 1 shows the performance of the TF-IDF paired with different classification models and the performance of the BERT model. The results show that TF-IDF performed well for all types of model sentiment classification, while the BERT-based model consistently outperformed the other models. Among the TF-IDF implementations, Random Forest achieved the best accuracy, precision, recall, and F1-score. Its high-quality performance when paired with TF-IDF explains the strength of its decision tree ensemble design in leveraging weighted word features from the TF-IDF vectorization technique [24]. The ability of Random Forest to handle sparse high-dimensional matrices without overfitting using bootstrap aggregation has shown a significant way towards such performance. However, the BERT-based models performed better than any TF-IDF configuration, which proves that contextual word embeddings that preserve linguistic subtleties and semantic relationships can provide a significant advantage over statistical frequency-of-words approaches for sentiment classification tasks.

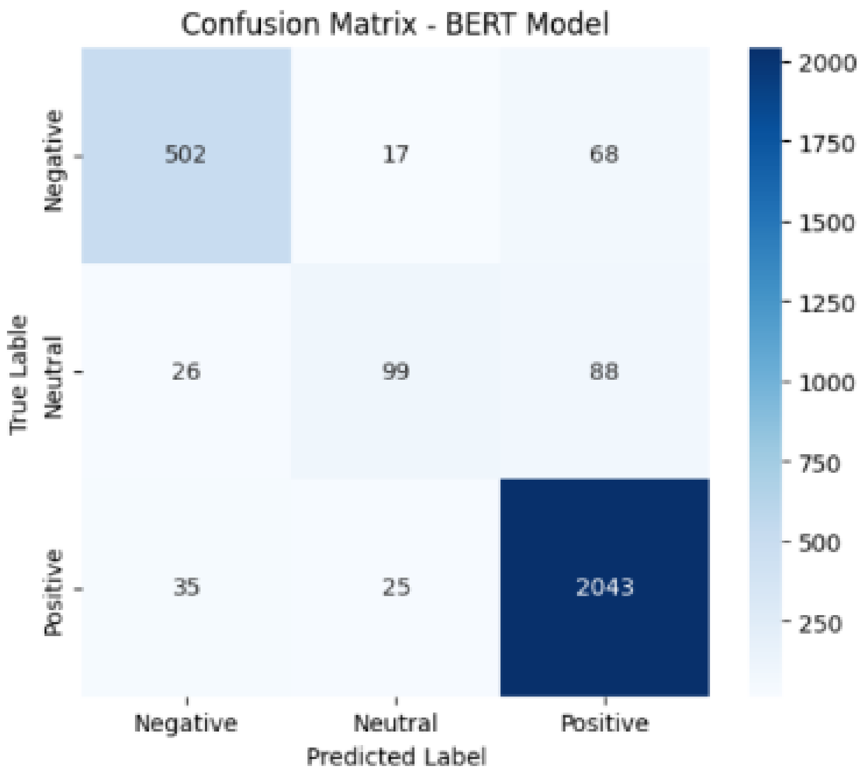

Figure 7. Confusion Matrix for BERT Model

Both Logistic Regression and SVM also performed well with TF-IDF. Both can be attributed to their optimization for linear separability and weighted feature importance, which aligns with TF-IDF's statistical word frequency approach [25]. This reciprocal complementariness stems from how TF-IDF transforms text into numerical vectors in which important, distinctive words receive higher weights. These algorithms work extremely well at identifying the decision boundaries in such high-dimensional spaces of features; thus, classifying sentiments by determining certain words is significant.

TF-IDF offers a well-organized, interpretable, and computationally effective text representation, which is suitable for traditional machine learning algorithms such as Random Forest, SVM, and Logistic Regression [9]. Among these traditional methods, Random Forest performed best with TF-IDF. Random Forest can leverage its capacity to handle feature importance, prevent overfitting, and detect nonlinear interactions. However, the BERT-deep learning model outperformed all TF-IDF implementations in all sentiment categories. BERT's contextual embeddings capture semantic depth and linguistic relationships that statistical models of word frequency, such as TF-IDF, cannot [26]. While Random Forest using TF-IDF provided the maximum performance of all traditional machine learning models, the BERT implementation captured greater accuracy, precision, recall, and F1-scores for every class of sentiment. The outcomes show that even if TF-IDF using conventional classifiers is computationally less demanding, deep learning technologies using contextual embedding, such as BERT, acquire significant performance benefits while performing extensive sentiment analysis tasks.

Table 1. Classification Result for TF-IDF-Based Models and BERT Models

Feature Engineering	Classification Model	Accuracy	Precision	Recall	F1-Score
TF-IDF	Random Forest	0.91	0.91	0.91	0.90
	Logistic Regression	0.86	0.83	0.86	0.83
	SVM	0.87	0.85	0.85	0.87
Word Embedding	BERT	0.91	0.91	0.91	0.91

5. CONCLUSION AND FUTURE WORK

OPRs are significant for consumer decision-making and business product innovation. However, the sheer volume and unstructured nature of comments hinder useful information extraction. This research resolves these issues by determining customer priorities using NLP methods such as sentiment analysis, feature extraction, and trend identification. Furthermore, this study delves into greedy algorithms based on structured feature comparisons to help consumers make decisions and improve their business strategies.

This study compared the effectiveness of BERT embeddings and TF-IDF in feature extraction and evaluated their performance across different classification approaches. Traditional machine learning models (Random Forest, SVM, and Logistic Regression) and TF-IDF vectorization were utilized, whereas BERT embeddings were implemented within a deep learning framework. The results show that both BERT and Random Forest with TF-IDF achieved excellent overall performance metrics, with BERT demonstrating superior capability in distinguishing between neutral and negative sentiments. The context sensitivity of BERT allows it to capture subtle linguistic features that are particularly significant in the effective classification of such challenging sentiment categories.

Overall, this study helps bridge unstructured and actionable insights. By improving the feature extraction and classification techniques, this study can extract and analyse customer preferences more effectively. This can help consumers make better decisions and companies optimize market strategies.

Future work should focus on methods that are currently difficult to classify. Researchers can attempt to develop more effective pre-processing methods. Advanced labelling strategies or background awareness normalization techniques can better capture subtle expressions of neutral sentiments. Advanced tokenization strategies or context-aware normalization techniques may better capture subtle expressions of neutral sentiment. In addition, exploring data augmentation strategies, such as generating virtual reviews or adopting synthetic data techniques, can help to balance the dataset and improve model robustness, especially for underrepresented sentiment categories. Future research can also combine deep learning-based BERT model fine-tuning or a hybrid model that combines the advantages of TF-IDF and contextual embedding to further improve the accuracy of sentiment classification and provide more detailed insights into customer preferences.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Lo Pei Qin: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Ng Sew Lai: Supervision, Writing – Review & Editing;
Li-xian Jiao: Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interest was disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. The source code is available at https://github.com/peiqinlo/Project

REFERENCES

[1] J. Nie, Z. Zhao, L. Huang, W. Nie, and Z. Wei, "Cross-domain recommendation via user-clustering and multidimensional information fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 868–880, 2023, doi: 10.1109/TMM.2021.3134161.

[2] S. Elzeheiry, W. A. Gab-Allah, N. Mekky, and M. Elmogy, "Sentiment analysis for e-commerce product reviews: Current trends and future directions," *Preprints*, 2023051649, 2023, doi: 10.20944/preprints202305.1649.v1.

[3] G. Zhao, Y. Luo, Q. Chen, and X. Qian, "Aspect-based sentiment analysis via multitask learning for online reviews," *Knowl.-Based Syst.*, vol. 264, p. 110326, 2023, doi: 10.1016/j.knosys.2023.110326.

[4] H. J. Alantari, I. S. Currim, Y. Deng, and S. Singh, "An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews," *Int. J. Res. Marketing*, vol. 39, no. 1, pp. 1–19, Mar. 2022, doi: 10.1016/j.ijresmar.2021.10.011.

[5] Z. Wen, Y. Chen, H. Liu, and Z. Liang, "Text mining based approach for customer sentiment and product competitiveness using composite online review data," *J. Theor. Appl. Electron. Commer. Res.*, vol. 19, no. 3, pp. 1776–1792, 2024, doi: 10.3390/jtaer19030087.

[6] R. Ireland, and A. Liu, "Application of data analytics for product design: Sentiment analysis of online product reviews," *CIRP J. Manuf. Sci. Technol.*, vol. 23, pp. 128–144, 2018, doi: 10.1016/j.cirpj.2018.06.003.

[7] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment analysis in e-commerce platforms: A review of current techniques and future directions," *IEEE Access*, vol. 11, pp. 90367–90382, 2023, doi: 10.1109/ACCESS.2023.3307308.

[8] G. K. Patra *et al.*, "A sentiment analysis of customer product review based on machine learning techniques in e-commerce," *J. Artif. Intell. Cloud Comput.*, vol. 2, no. 4, pp. 1–4, 2023, doi: 10.2139/ssrn.5153849.

[9] T. Shaik *et al.*, "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," *IEEE Access*, vol. 10, pp. 56720–56739, 2022.

[10] H. Yakubu, and C. K. Kwong, "Forecasting the importance of product attributes using online customer reviews and Google Trends," *Technol. Forecast. Soc. Change*, vol. 171, p. 120983, 2021.

[11] A. McAfee and E. Brynjolfsson, "Big data: The management revolution," *Harvard Bus. Rev.*, Oct. 2012. [Online]. Available: https://hbr.org/2012/10/big-data-the-management-revolution

[12] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Marketing Sci.*, vol. 31, no. 3, pp. 521–543, 2012.

[13] N. Archak, A. Ghose, and P. G. Ipeirotis, “Deriving the pricing power of product features by mining consumer reviews,” *Manage. Sci.*, vol. 57, no. 8, pp. 1485–1509, 2011.

[14] C. Manning, “Computational linguistics and deep learning,” *Comput. Linguist.*, vol. 41, no. 4, pp. 701–707, 2016.

[15] X. Fang, and J. Zhan, “Sentiment analysis using product review data,” *J. Big Data*, vol. 2, no. 5, pp. 1–14, 2015.

[16] J. Jin, Y. Liu, P. Ji, and H. Liu, “Understanding big consumer opinion data for market-driven product design,” *Int. J. Prod. Res.*, vol. 54, no. 10, pp. 3019–3041, 2016.

[17] M. Hu and B. Liu, “Mining opinion features in customer reviews,” in *Proc. AAAI*, vol. 4, no. 4, pp. 755–760, 2004.

[18] J. Jian, P. Ji, and C. Kwong, “What makes consumers unsatisfied with your products: Review analysis at a fine-grained level,” *Eng. Appl. Artif. Intell.*, vol. 47, pp. 38–48, 2016.

[19] W. Chung, and T. Tseng, “Discovering business intelligence from online product reviews: A rule-induction framework,” *Expert Syst. Appl.*, vol. 39, no. 15, pp. 11870–11879, 2012.

[20] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[21] N. Singh, and U. Jaiswal, “Sentiment analysis using machine learning: A comparative study,” *ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J.*, vol. 12, pp. 26785, 2023, doi: 10.14201/adcaij.26785.

[22] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proc. EMNLP*, pp. 79–86, 2002.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[24] L. Xiang, “Application of an improved TF-IDF method in literary text classification,” *Adv. Multimedia*, vol. 2022, no. 1, p. 9285324, 2021, doi: 10.1155/2022/9285324.

[25] F. Alzami, E. Udayanti, D. Prabowo, and R. Megantara, “Document preprocessing with TF-IDF to improve the polarity classification performance of unstructured sentiment analysis,” *Kinetik: Game Technol., Inf. Syst., Comput. Netw., Comput., Electron., Control*, vol. 5, no. 3, 2020, doi: 10.22219/kinetik.v5i3.1066.

[26] E. Alzahrani, and L. Jololian, “How different text-preprocessing techniques using the BERT model affect the gender profiling of authors,” *arXiv preprint arXiv:2109.13890*, 2021.

BIOGRAPHIES OF AUTHORS

	**Lo Pei Qin** is a final year student in Bachelor of Science Computer (Hons.) in Data Science in Multimedia University, Cyberjaya. He is passionate about studying data and analytics. He can be contacted at email: 1211101157@student.mmu.edu.my.
	**Ng Sew Lai** is an Assistant Professor in the Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia. Her research interests lie in the development and application of predictive modelling, statistical and econometric techniques, time series analysis, and risk management. She focuses on leveraging advanced modelling approaches to understand complex systems and inform decision-making in various domains. She can be contacted at email: slng@mmu.edu.my.
	**Li-xian Jiao** is a lecturer at Computer Science, Jining Normal University, China. She received her bachelor's degree in engineering from Jining Normal University, China in 2016. Later, she pursued and received her science master's degree from Liaoning Normal University, China in 2019. Her research interests are Information Security, Digital Image Forensics, Deep Learning etc. She can be contacted at email: 624292300@qq.com.