INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

E-commerce Product Review Classification using Neural Network-Based Approach

Fahrendra Khoirul Ihtada [a,1], Zainal Abidin [a], Cahyo Crysdian [a.2]

[a] *Department of Informatics, Maulana Malik Ibrahim State Islamic University, Lowokwaru, Malang, Indonesia*
Corresponding author: [1]fahrendra.khoirul@gmail.com; [2]cahyo@ti.uin-malang.ac.id

***Abstract*—E-commerce has become an integral part of how people shop, with the rise of customer reviews on various platforms. These reviews provide important insights into product, customer service, and delivery. The growing volume of e-commerce reviews makes manual sorting time-consuming and error-prone for business owners. This study aims to classify e-commerce reviews into three categories: product, customer service, and delivery. The data was collected from e-commerce customer reviews on Tokopedia and labeled using crowdsourcing for ground truth. To classify the reviews, a Neural Network is performed with various numbers of node and learning rate. TF-IDF is also used for feature extraction to capture important features from the review data. From nine test scenarios, model B3 with 50 nodes in the first hidden layer and a learning rate of 0.1 provided the best performance with an accuracy of 65.85%, precision of 62.27%, recall of 58.61%, and f1-score of 59.71%. Validation using K-Fold Cross Validation shows an average accuracy of 64.17% at k=10. Word analysis with TF-IDF identified dominant words in each category. The B3 model is not yet able to classify reviews perfectly, due to the large and unbalanced dataset, less complex model architecture, and less effective TF-IDF preprocessing. However, this study shows potential for better classification in the future. With optimization, this model can be very useful for e-commerce business owners to gain insight from customer reviews and can help them to identify aspects that will lead to customer satisfaction and trust.**

***Keywords*—Classification; product reviews; neural network.**

Manuscript received 27 Jun. 2024; revised 17 Nov. 2024; accepted 9 Feb. 2025. Date of publication 30 Nov. 2025.

I. Introduction

Online shopping, or e-commerce, has rapidly grown in recent years, becoming one of the most popular internet activities. According to Mordor Intelligence, Indonesia's e-commerce market size is USD 90.35 billion in 2025 and is forecast to nearly double to USD 185.71 billion by 2030 [1].

It is also the increase of online customer users. It's important to understand the impact that online reviews have on this growing number of e-commerce users. According to [2], the number of e-commerce users in Indonesia is predicted to continue to increase significantly. The e-commerce user base in Indonesia is expected to grow by 33.5 million between 2024 and 2029, reaching 99.1 million users by 2029 [2].

This growing number of users also indicates an increase in online reviews on e-commerce. The large number of product reviews has a great influence on customer buying decisions in the future [3]. The product reviews will greatly help customers who are unsure whether to buy the product or not [4]. The product reviews become the primary and trustworthy information for customers. In contrast to product descriptions that can be manipulated, product reviews are obtained from the experiences of previous customers. It shows that the product reviews alone are real and describe the actual information about the product and the online store itself [5].

Before buying a product, consumers often look for and gather information first, one of which is through product reviews [6]. A survey found that 93.5% of consumers rely on product reviews, which reflect other customers' experiences, before making a purchase [7], [8]. Therefore, customer reviews are very important in increasing the trust and reputation of business owners in e-commerce.

With the rise in e-commerce users, businesses must actively manage reviews to understand customer expectations and improve service quality [9]. One way is to classify reviews into categories using text classification techniques.

Researchers have performed many studies on text classification with different approaches. Aslam *et al.* [10] used XGBoost and Random Forest to classify app reviews into categories like bug reports and user experience, achieving optimal prediction accuracy. Chen *et al.* [11] conducted sentiment classification using a Gated Recurrent Neural Network with some modifications (GRNN-SR). Their results show that the proposed GRNN-SR model manages to effectively capture sentimental relationships and outperforms the traditional GRNN base model in sentiment classification tasks. Afzaal *et al.* [12] performed aspect classification on tourist destination reviews. They used several machine learning methods and also a combination of two feature extraction, which are post tagging and N-grams. The best accuracy was obtained with the Multinomial Naïve Bayes (MNB) method. Zhang *et al.* [13] performed multiclass sentiment classification on e-commerce reviews. They compared the BERT model with the directed-weighted model they proposed. Overall, the BERT model results are better than the Directed Weight. Noori performed sentiment classification on customer reviews into 2 classes, which are positive and negative. TF-IDF and PCA were used for feature extraction. The results show that NB has the best results compared to other machine learning methods such as SVM, KNN, and Neural Networks. Nasiri & Budi [14] conducted research focusing on aspect category detection in Indonesian e-commerce app reviews to support better sentiment analysis. It compares traditional machine learning models like SVM with deep learning approaches such as GRU+CNN. The GRU+CNN model achieved the highest accuracy, showing strong performance in identifying multiple aspects like app quality, promos, and payment options. Alamoudi *et al.* [36]conducted sentiment classification research on restaurant review data on the Yelp website. They compared 3 models, including, machine learning, deep learning, and transfer learning. Feature extraction in their research is Bag of Word, TF-IDF, and Gloce word embedding. The best results were obtained in the ALBERT model. Yu *et al.* [15] classified social media aspects related to hurricane natural disasters. They classify into 5 classes. CNN has the best performance compared to SVM and Logistic Regression.

Research on Indonesian customer reviews classification has also been conducted over the past few years. But the existing research for this topic mostly focused on sentiment analysis, such as research conducted by [16] and [17] that classify the reviews into binary sentiment (positive and negative) or a study by [18] that classifies the reviews into tertiary sentiment (positive, negative, neutral). While sentiment analysis provides an overall picture of customer satisfaction, indicating whether they like or dislike something, this information does not delve deeper into the specific aspects customers pay attention to.

This study seeks to bridge a gap by categorizing reviews into three primary aspects: product, customer service, and shipping/delivery—areas directly aligning with core e-commerce components yet underexplored in existing studies on e-commerce feedback. The product category describes customer satisfaction with the quality, performance, and conformity of the product to the description given. The customer service category covers the interaction between the customer and the service provided, such as friendliness, responsiveness, and other things related to customer service [19]. Meanwhile, the shipping/delivery category involves aspects such as speed of delivery, condition of goods when received, and timeliness of delivery.

For this purpose, a deep learning method known as neural networks is utilized. The application of neural networks enables more effective handling of the complexity of textual data compared to traditional methods [20]. Neural networks have the capability to capture intricate patterns and relationships within the data, facilitating a more accurate and comprehensive analysis of customer reviews [21]. By leveraging the power of neural networks, this study extends beyond simple sentiment classification, offering detailed insights into specific aspects of product reviews, customer service interactions, and delivery experiences.

TF-IDF is also employed as the feature extraction technique for the text reviews. TF-IDF is selected for its simplicity, interpretability, and computational efficiency, making it a practical choice for baseline modelling [22]. Unlike more advanced methods such as word embeddings (Word2Vec, GloVe) or transformer-based models (BERT), TF-IDF does not require extensive computational resources or large amounts of training data [23]. It operates by weighting words based on their frequency in a document (Term Frequency) and their rarity across all documents (Inverse Document Frequency), effectively emphasizing words that are particularly relevant to specific reviews while downweighting common words that appear frequently across all reviews [24]. While advanced methods like word embeddings and transformers excel at capturing semantic relationships and contextual nuances [25], [26], TF-IDF offers a straightforward and effective way to represent text data, particularly in scenarios where computational resources or labeled data are limited.

The developed model is then integrated into a website platform that can be applied in real-world scenarios, so that business owners can systematically classify customer reviews based on their specific aspects. Using this intelligent system, businesses can gain more comprehensive insights into areas that require improvement, as it focuses on analysing customer feedback efficiently. In addition, the automation provided by this system significantly reduces the time and resources needed for manual classification, thus allowing businesses to streamline their review management processes. This research is expected to make a meaningful contribution in improving customer review management on e-commerce platforms, which can ultimately improve decision-making and customer satisfaction.

II. Materials and Method

In this section, an outline of the implementation of the machine learning framework for e-commerce product review classification will be presented. Starting from data, text preprocessing, feature extraction, modeling, and its variations. The subchapter will explain in more detail.

A. Data Collection

This study uses Tokopedia e-commerce review data. Review data consists of 1024 data from several product categories such as fashion, electronics, sports, software, mobile phones, and others. The review data in this study is only limited to the Indonesian language. Initially, the data was obtained without having ground truth aspects. Therefore, in this study, labeling is performed first. Labeling is done by crowdsourcing to public respondents in each review data to categorized into one of three classes, which are product, customer service, and shipping/delivery. Crowdsourcing was selected due to scalability and cost-effectiveness. To ensure reliability, responses were validated by taking the most commonly selected label for each review. This approach allowed for accurate label assignment while capturing varied perspectives from public respondents.

The product class describes customer satisfaction with the quality, performance, and suitability of the product to the description given. The customer service class describes interaction between customers with the services provided, such as friendliness, responsiveness, and other things related to customer service. Meanwhile, the shipping/delivery class describes the aspect of speed delivery, condition of goods when received, and timeliness of delivery.

From the results of Crowdsource, the category with the dominating number of votes for each review is taken. Thus, the dataset has ground truth that will be used to train and test the classification model in this study. Table I shows an example of an e-commerce product review dataset.

TABLE I
DATA PRODUCT REVIEW E-COMMERCE

No	Product Review	Label
1	"Kirain bisa make stik ini buat gta v" *("I thought I could make this stick for GTA V")*	Product
2	"Barangnya berfungsi dengan baik mantap apalagi harganya untuk kurir gosend terlalu lama sampai 2 hari bintang 3 untuk kurir" *("The item works great especially the price for GoSend courier is too long up to 2 days, 3 stars for courier")*	Product
3	"Barang sudah sampai tujuan dalam kondisi baik proses Pengiriman sangat cepat terima kasih" *("The goods have reached their destination in good condition, the shipping process is very fast, thank you.")*	Shipping/Delivery
4	"Baik recomendasi order disini aja bca bagus" *("Good recommendation, just order here bca is good")*	Customer Service
5	"Packingnya keren rapi dan aman pakai bubble wrap. semoga awet 😉" *("The packing is cool neat and safe with bubble wrap. hopefully it will be last long 😉")*	Customer Service

B. Text Preprocessing

Text preprocessing is a very important stage in getting valuable information from text data [9], [27], [28], [29]. This study involves case folding, punctuation removal, tokenization, stopword, removal, and stemming.

Case folding converts all the letters in the data into lowercase letters to make them more uniform since uppercase and lowercase letters do not have such a high contrast in meaning [14], [30]. Punctuation removal removes and emoticons removal removes all punctuation marks and emoticon in the data [12], [31]. Punctuation removal is done to clean up unnecessary elements in the sentence so that it can be more focused on the content of the sentence content [31]. Emoticons appear in reviews introduce noise in data, which their presence often convey sentiment or mood. Emoticons removal allowed the model to focus on the textual content alone, aiming for clearer, more contextually focused predictions. Tokenization is a process of separating the word in a sentence[9]. Commonly, tokenization uses the space character in the sentence as a separator [30]. Stopword removal is a stage in text preprocessing by removing words that are common and have less meaning [9], [14], [31]. These words are stopwords. In the Indonesian language, examples of stopwords are "dan", "di", "dari", "yang". Stopwords usually appear in a very large number without depending on a particular language. This stopword stage has the aim of increasing focus on important words so as to reduce the process at the next stage [31]. Stemming is a process in text processing where words are converted into their basic form by removing the suffixes and prefixes on the word [9], [14], [32]. With stemming, the variation of words with similar meanings can be reduced by making them uniform and considered identical. By then, it will reduce the dimensionality of the text data and make the processing more effective [32].

TABLE II
PREPROCESSING

No	Before Text Preprocessing	After Text Preprocessing
1	"barang sudah saya terima thanks" *("I have received the item thanks")*	"barang terima thanks" *("item received thanks")*
2	"Fast ResponseFriendly Recommended Seller"	"fast response friendly recommended seller"
3.	"Seller fast respon Produk sudah diterima buat stok. Thanks…" *("Seller fast response Product has been received for stock. Thanks...")*	"seller fast respon produk terima stok thanks" *("seller fast response product receive stock thank you")*
4	"Sangat memuaskann dan barang sesuai deskripsi sudah 3 kali beli disini dan selalu puas recommended sellerer!" *("Very satisfying and the goods according to the description have bought 3 times here and always satisfied recommended seller!")*	"memuaskann barang sesuai deskripsi 3 kali beli puas recommended sellerer" *("satisfying goods according to description 3 times buy satisfied recommended seller")*
5	"Packingnya keren rapi dan aman pakai bubble wrap. semoga awet 😉" *("The packing is cool neat and safe with bubble wrap. hopefully it will last 😉")*	"Packing keren rapi aman pakai buble wrap semoga awet" *("Packing cool neat safe using buble wrap hopefully durable")*

C. Feature Extraction

Feature extraction was performed to capture important features from the data to build the final feature set, which will be used as input to the model [33].

In this study, the feature extraction used is the Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF is one of the algorithms that is often used in text mining research. TF-IDF transforms text into a numerical vector that represents the importance of words or terms in the dataset. TF-DF measures how often a word appears in the document [9], [14], [33], [34], [35], [36]. In this case, each document is a product review. TF-IDF has the concept that if a word appears in one or a few documents, then the word is important and it should have a high TF-IDF value. Otherwise, If the word appears often in all or most documents, then the word is considered unimportant and has a low TF-IDF value [14], [35].

TF-IDF consists of two main components, which are Term Frequency (TF) and Inverse Document Frequency (IDF). the formulas TF-IDF can be seen in (1), (2), and (3).

$$TF_{(t,r)} = \frac{N_{t,r}}{\sum_{i=0}^{n} N_{i,r}} \quad (1)$$

Where:
$N_{t,r}$: number of occurrences of term *t* in review *r*
$\sum_{i=0}^{n} N_{i,r}$: length of review *r*

$$IDF_{(t)} = log \frac{|R|}{\sum r_t \in R} \quad (2)$$

Where:
$|R|$: total of all reviews
$\sum r_t \in R$: number of occurrences of reviews *r* containing term t

$$TFIDF_{(t,r)} = TF_{(t,r)} \times IDF_{(t)} \quad (3)$$

The output of the TF-IDF vector consists of numbers that represent the weight of each word in the review based on its frequency of occurrence and importance in the dataset [37]. In this study, the vector length of this TFIDF vector is 1085 for each review. This TFIDF vector will be input into the neural network input layer. The number of neural network input layer nodes will be set according to the length of the TFIDF vector.

D. Modelling

In this study, the dataset is divided into an 80:20 split ratio, which of 1024 data is divided into 819 training data and 205 testing data. This 80:20 split is widely used in machine learning as it balances enough data for training and testing, especially for limited size of datasets like ours. In order to prevent bias in the model, an equal distribution was applied for each class in the training data and testing data[33]. The training data is then fed into the Neural Network model for the training phase and then the model is evaluated for performance using the testing data.

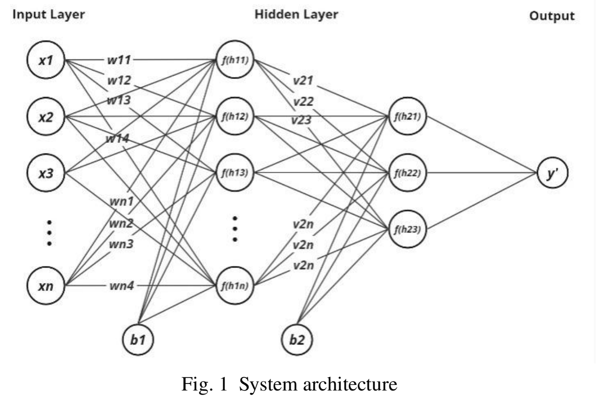


Fig. 1 System architecture

The neural network model in this study visualized in Fig. 1, uses a total of 1085 input nodes, which have been adjusted to the vector length from the previous TF-IDF process. For nodes in the output layer use as many as 3 nodes representing the output class, which are products, customer service, and shipping/delivery.

To find the optimal model parameters, this study conducted several scenarios on two parameters in the Neural Network model, which are the number of nodes in the hidden layer of the neural network, and the learning rate. The number of nodes in the hidden layer determines the complexity of the model and its ability to learn patterns in the data. The learning rate controls the speed at which the model learns and updates its weights and biases in the training process. The proposed model is shown in Table III.

TABLE III
PROPOSED MODELS

Model	Feature Extraction	Number of Hidden Layer Nodes	Learning Rate
A1	TF-IDF	25	0.01
A2	TF-IDF	25	0.05
A3	TF-IDF	25	0.1
B1	TF-IDF	50	0.01
B2 (Proposed)	TF-IDF	50	0.05
B3	TF-IDF	50	0.1
C1	TF-IDF	100	0.01
C2	TF-IDF	100	0.05
C3	TF-IDF	100	0.1
IB_A	IndoBERT	25	0.00003
IB_B	IndoBERT	50	0.00003
IB_C	IndoBERT	100	0.00003

Variations in the number of nodes of the hidden layer are shown in three options, namely 25, 50, and 100. The selection of the number of nodes is determined based on calculations using formula 4 and formula 5.

$$Nh = \sqrt{nm} \quad (4)$$

Where :
Nh : the number of neurons in the hidden layer
n : number of input features
m : number of output class

Variations in the number of nodes of the hidden layer are shown in three options, namely 25, 50, and 100. Based on formula (4) [38], the number of hidden layer nodes with 1085 inputs and 3 class outputs results in 57 hidden layer nodes.

$$Nh = \frac{\sqrt{1+8N_i}}{2} - 1 \quad (5)$$

Where :
Nh : the number of neurons in the hidden layer
N_i : number of input features

Meanwhile, based on the formula in (5)[39], with an input of 1085, the number of hidden layer nodes is 46. In this study, a value between the two results was taken, which is 50 nodes. Variations of 25 and 100 were chosen to see how the model behaves with different numbers of first hidden layer nodes. In the learning rate, three variations are to be used, which are 0.01, 0.05, and 0.1. The study conducted by [40] state that a high learning rate may cause the model to diverge, while a very low rate can slow or prevent convergence. The optimal rate depends on the problem and network architecture. This research will use a learning rate of 0.01 as the low learning rate value, 0.1 as the high learning rate value, and 0.05 as an additional comparison to see how the model behaves in the training process.

III. Result and Discussion

In this study, our proposed model will be evaluated using a confusion matrix. The proposed confusion matrix is 3x3 adjusted to the number of output classes. From the confusion matrix, accuracy, precision, recall, and 1-score can be obtained. From these variables, it can be seen how well the performance of each model is. The best model will be selected and validated using k-fold cross-validation with several variations of the k value. This is to determine how the model behaves when looking at new data.

A. *Model Performance Evaluation*

In this section, the training and testing phases of the model are performed. The results of the training process provide the performance values of the model, including cost, accuracy, epoch, and the duration of the training process. The cost value represents how well the model can adapt to the training data. Accuracy represents the model's ability to predict the class correctly. Information on the number of epochs and training time is important to evaluate the efficiency and effectiveness of the model training process.

TABLE IV
Training result

Model	Training			
	Cost	**Epoch**	**Time in seconds**	**Time in minutes**
A1	0.166438	9963	388.33	6.47
A2	0.098499	4987	283.68	4.73
A3	**0.083149**	**3618**	**214.31**	**3.57**
B1	0.165462	10009	817.38	13.62
B2	0.102014	4775	428.86	7.15
B3	0.088147	3379	283.35	4.72
C1	0.165794	9922	1575.18	26.25
C2	0.103425	4723	772.01	12.87
C3	0.090012	3332	1167.37	19.46

The models are distinguished by the number of nodes in the hidden layer and the learning rate. The letter code in the model name indicates the number of nodes in the hidden layer, A for 25, B for 50, and C for 100. The number in the model name indicates the learning rate of the model. Number 1 is for 0.01, number 2 is for 0.05, and number 3 is for 0.001.

There is a significant time difference between each model in Table IV. Models with low learning rates (models coded as 1) tend to take longer than models with high learning rates (models coded as 3) because the learning rate affects the speed of the training process on the model.

A significant difference in time is also seen in the number of hidden layer nodes. Models with a higher number of hidden layer nodes tend to take longer than models with a lower number of hidden layer nodes. The number of neural network layer nodes indicates the complexity of the model. From this analysis, the more complex the model, the longer the time required for the training process.

However, model complexity does not guarantee better performance. The best model in the training process is model B3, which has the smallest number of hidden layer nodes, 25, and has the largest learning rate of 0.1. Model A3 is the most efficient model compared to the other models. Model A3 gets the smallest cost value, which is 0.083149 for cost. Model A3 also takes the shortest time compared to the other models, which is only about 214.13 seconds or 3.57 minutes.

After the training phase, the test scenario models are evaluated on the testing data. At this stage, the performance of the model is measured using testing data that has never been seen before by the model during the training phase. The evaluation results of each model are shown in Table V.

TABLE V
Evaluation metrics across models

Model	Precision	Recall	F1-Score	Accuracy
A1	0.56109	0.51508	0.51968	0.62927
A2	0.58353	0.55458	0.56080	0.63902
A3	0.60796	0.57261	0.58148	0.65366
B1	0.55798	0.50340	0.50635	0.62439
B2	0.59819	0.561209	0.57036	0.63902
B3	**0.62272**	**0.58616**	**0.59719**	**0.65854**
C1	0.59620	0.53758	0.54452	0.65366
C2	0.579486	0.54477	0.55215	0.63415
C3	0.59794	0.56626	0.57432	0.64390

The testing phase is performed on testing data. It can be seen that all models have the results of each metric that are not significantly different from each other. The closer to the value of 1.0 in the precision, recall, F1-score, and accuracy metrics, the better the performance. The results show that the precision metric reaches a value in the range of 0.55 - 0.62. In the recall metric, the value obtained is relatively low in the range of 0.51 - 0.58. Also in the F1-score metric, the value is obtained in the range of 0.51 - 0.59. Meanwhile, the accuracy metric obtained a slightly larger value than the other metrics, which is in the range of 0.62 - 0.65.

The accuracy of model B3 is 65.85% suggest that correctly classifies most data but still has room for improvement, as a significant portion of instances remain misclassified. This value is still relatively low. Model B3 also has precision, recall, and F1-score values that are still relatively low, indicates that Model B3 struggles to capture complex patterns in product reviews text, limiting its predictive performance.

One key factor contributing to this lower performance is the limited of dataset size, which made it difficult for model to see diverse pattern across different aspects (product, customer service, and delivery). In addition, the imbalance in class representation may led model to be more likely to predict certain classes, which impacted precision, recall, and F1-scores, especially for minority classes. These limitations reduced the generalizability of the model, explaining the lower scores and limited ability to capture the complex linguistic patterns that often characterize product reviews.

Additionally, TF-IDF as feature extraction, focuses on word frequency and term relationships of documents, but cannot capture the rich contextual information embedded in product reviews. This limitation restricts the model’s capacity to fully understand specific aspects nuances, which are crucial for effectively classifying product reviews.

B. *Cross Validation*

To validate the performance of the model on all data distributions, k-fold cross-validation is performed. This study uses 3 k-fold cross-validation scenarios with each k-fold value, which are 4, 5, and 10. Based on research conducted by [41] The k-fold value of 10 provides a good balance between bias and variance, making it a choice that is often used in model evaluation practice. The value of k = 10 produces prediction errors that are almost unbiased compared to smaller or larger k values. Values of k=4 and k=5 were also used to see how the model behaves with different k values. The model that will be used in k-fold cross-validation is model B3.

TABLE VI
K-FOLD CROSS VALIDATION RESULTS IN B3 MODEL

K	Average Accuracy	Standard Deviation
4	62.59%.	0.0449
5	63.29%	0.0532
10	64.17%	0.0580

From the k-fold cross-validation results, the best average accuracy was obtained at $k = 10$ with an average accuracy of 64.17%. Although this indicates that the model performs slightly better with higher k values, the overall accuracy is still relatively low. Although there is an increase in accuracy with increasing k values, the value obtained is still relatively low. The low performance of the model indicates that the model is less practical and less efficient for real-world applications due to its large probability of error. This underscores the need for further model tuning or alternative approaches to improve performance and reliability in practical scenarios.

The standard testing accuracy is higher (65.854%) because the neural network is only tested on one data set, allowing the model to focus and recognize patterns more easily, although this increases the risk of overfitting. In k-fold, the model is tested on multiple pieces of data, making it more generalized, but also decreasing accuracy. At (K = 4) and (K = 5), the accuracy is lower (62.59% and 62.29%) with a small standard deviation, indicating stable but suboptimal performance due to less data trained. At (K = 10), the accuracy increases (64.17%) but the standard deviation is higher (0.0580), which means that the model is more capable of generalization but the results are less consistent across folds. This higher standard deviation indicates that the convergence of the model is not stable across all folds. The accuracy of standard testing is better because the model does not need to adapt to variations in the data, making it easier to learn specific patterns.

C. Word Analysis

The feature extraction process in this research uses TF-IDF. The value of TF-IDF can be used to analyze words. The TF-IDF value can be used to identify the most dominating or representative words in a review class. This word analysis is applied to understand customer preferences and ratings on products, customer service, and shipping/delivery classes. The most dominating words in the product class are shown in Table VII and visualized in Fig. 2.

TABLE VII
TF-IDF RANKING IN PRODUCT ASPECT

No	Word	TF-IDF Score
1	“Bagus” *(good)*	0.08822
2	“Sesuai” *(suitable)*	0.08082
3	“Mantap” *(great)*	0.05956
4	“Barang” *(goods)*	0.05645
5	“Produk” *(product)*	0.05169
6	“Baik” *(good)*	0.05042
7	“Deskripsi” *(description)*	0.04703
8	“Terima” *(receive)*	0.04359
9	“Fungsi” *(function)*	0.04112
10	“Harga” *(price)*	0.03931

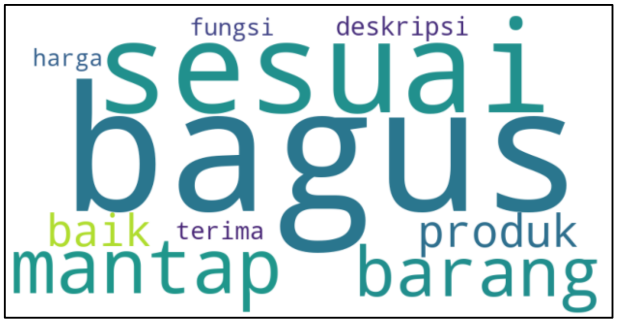


Fig. 2 Wordcloud product aspect

Words such as “bagus” (good), “sesuai”(suitable), “mantap” (great) are the 3 words that dominate or are most relevant to the product class. These words have a high TF-IDF value for the product class. This shows that aspects such as the quality of the item, the suitability of the description, and also customer satisfaction have a considerable influence on the assessment of the product by customers.

Then, the most dominating words in the customer service class are shown in Table VIII and visualized in Fig. 3.

TABLE VIII
TF-IDF RANKING IN PRODUCT ASPECT

No	Word	TF-IDF Score
1	“Respon” *(response)*	0.08489
2	“Cepat” *(fast)*	0.08005
3	“Barang” *(goods)*	0.07574
4	“Seller” *(seller)*	0.07318
5	“Fast” *(fast)*	0.06868
6	“Rapi” *(neat)*	0.06190
7	“Gan” *(dude)*	0.05862
8	“Recommended” *(recommended)*	0.05629
9	“Kirim” *(send)*	0.05115
10	“Mantap” *(great)*	0.04891

recommended
respon
barang
seller
rapi
cepat
kirim
fast
mantap
gan

Fig. 3 Wordcloud customer service aspect

Words such as “cepat” (fast), “respon”(response), “barang”(goods) are the 3 words that dominate or are most relevant to the customer service class. These words have a fairly high TF-IDF value in the customer service class. This shows that aspects such as how the seller responds, the speed of the seller in serving customers, and also the quality of goods received by customers have a considerable influence on the assessment of customer service.

In addition, the most dominating words in the delivery class are shown in Table IX and visualized in Fig. 4.

TABLE IX
TF-IDF RANKING IN SHIPPING ASPECT

No	Word	TF-IDF Score
1	“Cepat“ *(fast)*	0.15582
2	“Kirim” *(sende)*	0.11844
3	“Terima” *(receive)*	0.08296
4	“Selamat” *(safe)*	0.07486
5	“Barang” *(goods)*	0.07431
6	“Kasih” *(give)*	0.07010
7	”Aman” *(safe)*	0.05379
8	“Coba” *(try)*	0.05348
9	“Sampai” *(arrive)*	0.04849
10	“Fungsi” *(function)*	0.04672

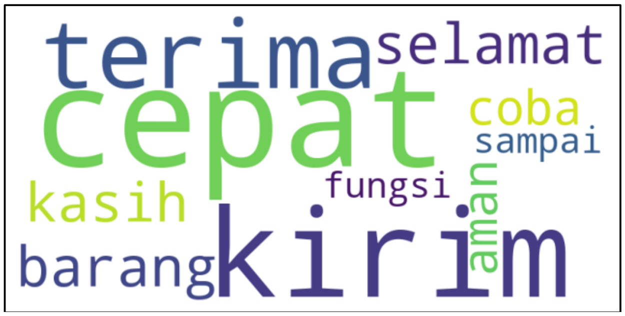
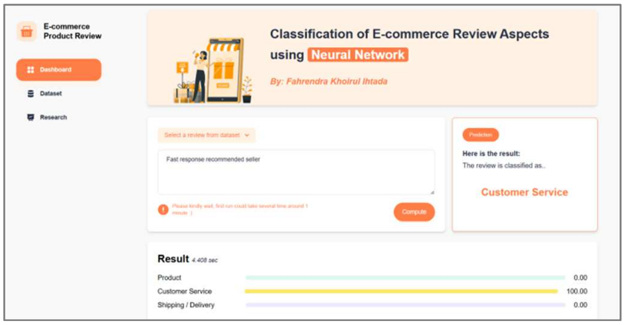


Fig. 4 Wordcloud shipping aspect

Words such as “cepat” (fast), “kirim”(send), “terima"(receive) are the 3 words that dominate or are most relevant to the delivery class. These words have a fairly high TF-IDF value in the delivery class. This shows that aspects such as speed and accuracy of delivery are highly considered by customers in assessing the delivery aspect. The presence of the words "safe" and "secure" also shows that the safety and security of the delivery of goods is an important concern of customers.

The results of word analysis using TF-IDF values provide a deeper understanding of the aspects that are often considered by customers. With this analysis, sellers or e-commerce businesses can utilize it to identify what aspects need to be improved or evaluated in their online stores according to customer feedback through the reviews given.

The best model, referred to as Model B3, has been integrated into the website. As shown in Figures 5, 6, and 7, the website interface allows users to input product reviews directly.



Classification of E-commerce Review Aspects using Neural Network

By: Fahrendra Khoirul Ihtada

Fast response recommended seller

Please kindly wait, first run could take several time around 1 minute :)

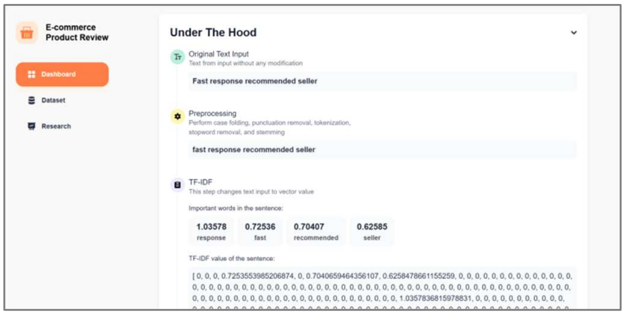
Here is the result:
The review is classified as..

Customer Service

Result 4.408 sec

Product	0.00
Customer Service	100.00
Shipping / Delivery	0.00

Fig. 5 Website interface 1



Under The Hood

Original Text Input
Text from input without any modification

Fast response recommended seller

Preprocessing
Perform case folding, punctuation removal, tokenization, stopword removal, and stemming

fast response recommended seller

TF-IDF
This step changes text input to vector value

Important words in the sentence:

1.03578	0.72536	0.70407	0.62585
response	fast	recommended	seller

TF-IDF value of the sentence:

```
[ 0, 0, 0, 0.7253553985206874, 0, 0.7040659464356107, 0.6258478661155259, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1.0357836815978831, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

Fig. 6 Website interface 2

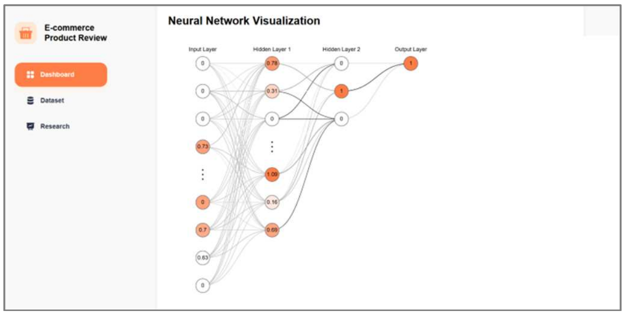


Fig. 7 Website interface 3

The website dashboard contains a header banner, text fields for inputting reviews and also an output section as shown in Fig 5. Once inputted, the website will run Model B3 to process the input and display the classification output. The website provides a detailed step-by-step visualization of the processing workflow shown in Fig. 6 from raw data input to various stages of data transformation and model inference, resulting in the final classification. In addition, the website also features a visualization of the neural network architecture used by Model B3 shown in Fig. 7, allowing users to intuitively understand the underlying process.

IV. CONCLUSIONS

This study aims to build a product review classification into three aspects including product, customer service, and shipping/delivery. A Neural Network with TF-IDF feature extraction served as the classification model.

Model B3 with the hidden layer node count of 50 nodes and a learning rate of 0.1, provides the best performance compared to other models. Model B3 achieved 62.27% precision, 58.61% recall, 59.71% F1-score, and 65.85% accuracy. Validation of model B3 using k-fold cross-validation obtained average accuracy results that were not much different. The best results are obtained at k = 10 with an average accuracy value of 64.17%.

Based on word analysis using TF-IDF, it is found that in the product class, three words are dominant, namely "good", "suitable", and "great". In the customer service class, the three dominant words are "response", "fast", and "goods". And finally, in the delivery class, the three dominant words are "fast", "send", and "receive". These words are often used by customers in giving reviews on product ratings, customer service, and delivery of online stores on e-commerce.

Based on the results of the confusion matrix and k-fold cross-validation on model B3, it shows that model B3 still cannot classify e-commerce product review data perfectly.

This model accuracy did not meet expectation, primarily due to limitations in dataset, which was small and imbalanced, and the proposed model architecture struggled to capture complex pattern with the data. Additionally, the TF-IDF feature extraction was less effective for capturing rich contextual information embedded in product reviews. This contextual information is often crucial for identifying specific aspects discussed in the reviews, which TF-IDF alone cannot address.

Future research should explore the use of contextual feature extraction techniques, such as word embedding transformer-based encoding like the BERT or word embeddings from Word2Vec or GloVe. These methods capturing semantic relationships and contextual meaning, helping the model understand sentiments and specific aspects more effectively.

Lastly, real-world e-commerce product reviews often cover multiple aspects, thus the need for a multi-label classification approach. Traditional single-label classification methods, such as the one used in this study, have limitations in their ability to handle complex reviews. Multi-label classification, on the other hand, allows the model to assign multiple labels or aspects to each review.

Acknowledgment

This research would not have been possible without the support of many parties. We are particularly grateful to Maulana Malik Ibrahim State Islamic University for providing the necessary resources that enabled us to conduct this research.

References

[1] Mordor Intelligence, "Indonesia e-commerce market size, growth & industry analysis, 2030," Mordor Intelligence, 2025. [Online]. Available: https://www.mordorintelligence.com/industry-reports/indonesia-ecommerce-market.

[2] Statista Research Department, "Number of users of e-commerce in Indonesia from 2020 to 2029," Statista, 2023. [Online]. Available: https://www.statista.com/forecasts/251635/e-commerce-users-in-indonesia.

[3] T. M. A. Rizki, "The influence of product reviews, trust, and marketing content on TikTok on Jiniso's product purchase decisions," *Formosa J. Sustain. Res.*, vol. 2, no. 4, pp. 899-918, Apr. 2023, doi:10.55927/fjsr.v2i4.3613.

[4] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Shitharth, "Towards improving e-commerce customer review analysis for sentiment detection," *Sci. Rep.*, vol. 12, no. 1, p. 21437, Dec. 2022, doi:10.1038/s41598-022-26432-3.

[5] D. R. Patil and N. L. Rane, "Customer experience and satisfaction: Importance of customer reviews and customer value on buying preference," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 5, no. 5, pp. 1-12, May 2023, doi: 10.56726/irjmets36460.

[6] R. Alshweesh and S. Bandi, "The impact of e-commerce on consumer purchasing behavior: The mediating role of financial technology," *Int. J. Res. Rev.*, vol. 9, no. 2, pp. 479-499, Feb. 2022, doi:10.52403/ijrr.20220261.

[7] S. J. Payne and A. Howes, "E-commerce," in *Adaptive Interaction*. Springer, 2013, pp. 123-145, doi: 10.1007/978-3-031-02199-2_7.

[8] Power Reviews, "Survey: The ever-growing power of reviews (2023 edition)," Power Reviews, 2023. [Online]. Available: https://www.powerreviews.com/power-of-reviews-2023/.

[9] B. Noori, "Classification of customer reviews using machine learning algorithms," *Appl. Artif. Intell.*, vol. 35, no. 8, pp. 567-588, 2021, doi:10.1080/08839514.2021.1922843.

[10] N. Aslam, W. Y. Ramay, K. Xia, and N. Sarwar, "Convolutional neural network based classification of app reviews," *IEEE Access*, vol. 8, pp. 185 949-185 961, 2020, doi: 10.1109/access.2020.3029634.

[11] C. Chen, R. Zhuo, and J. Ren, "Gated recurrent neural network with sentimental relations for sentiment classification," *Inf. Sci.*, vol. 502, pp. 268-278, Oct. 2019, doi: 10.1016/j.ins.2019.06.050.

[12] M. Afzaal, M. Usman, and A. Fong, "Tourism mobile app with aspect-based sentiment classification framework for tourist reviews," *IEEE Trans. Consum. Electron.*, vol. 65, no. 2, pp. 233-242, May 2019, doi:10.1109/TCE.2019.2908944.

[13] S. Zhang, D. Zhang, H. Zhong, and G. Wang, "A multiclassification model of sentiment for e-commerce reviews," *IEEE Access*, vol. 8, pp. 189 513-189 526, 2020, doi: 10.1109/ACCESS.2020.3031588.

[14] D. F. Nasiri and I. Budi, "Aspect category detection on Indonesian e-commerce mobile application review," in *Proc. Int. Conf. Data Softw. Eng. (ICoDSE)*, 2019, pp. 1-6, doi:10.1109/ICoDSE48700.2019.9092619.

[15] M. Yu, Q. Huang, H. Qin, C. Scheele, and C. Yang, "Deep learning for real-time social media text classification for situation awareness-using hurricanes Sandy, Harvey, and Irma as case studies," *Int. J. Digit. Earth*, vol. 12, no. 11, pp. 1230-1247, Nov. 2019, doi:10.1080/17538947.2019.1574316.

[16] C. N. V. Jyothirmai and Y. M. Madhavi, "Sentiment analysis of customer product reviews using machine learning," *Int. J. Res. Anal. Rev.*, vol. 11, no. 2, pp. 846-855, 2024.

[17] A. Iqbal *et al.*, "Sentiment analysis of consumer reviews using deep learning," *Sustainability*, vol. 14, no. 17, p. 10844, Sep. 2022, doi:10.3390/su141710844.

[18] A. Patel, P. Oza, and S. Agrawal, "Sentiment analysis of customer feedback and reviews for airline services using language representation model," *Procedia Comput. Sci.*, vol. 218, pp. 2459-2467, 2023, doi: 10.1016/j.procs.2023.01.221.

[19] M. Zakirin, "Influence of product quality, service quality and completeness on customer satisfaction on Mie Soponyono producer," *Int. J. Rev. Manage. Bus. Entrepreneurship*, vol. 1, no. 2, pp. 281-294, Dec. 2021, doi: 10.37715/rmbe.v1i2.2434.

[20] P. L. Prasanna and D. R. Rao, "Text classification using artificial neural networks," *Int. J. Eng. Technol.*, vol. 7, no. 3, pp. 1-6, 2018.

[21] O. Bellar, A. Baina, and M. Ballafkih, "Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers," *Mathematics*, vol. 12, no. 15, p. 2403, Aug. 2024, doi: 10.3390/math12152403.

[22] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 30, Dec. 2019, doi: 10.1186/s13673-019-0192-7.

[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513-523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[24] A. Gothankar, L. Gupta, N. Bisht, S. Nehe, and M. Bansode, "Extractive text and video summarization using TF-IDF algorithm," *Int. J. Res. Anal. Rev.*, vol. 9, no. 1, pp. 434-435, 2022, doi:10.22214/ijraset.2022.40775.

[25] L. K. Senel, I. Utlu, V. Yucesoy, A. Koc, and T. Cukur, "Semantic structure and interpretability of word embeddings," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1769-1779, Oct. 2018, doi: 10.1109/TASLP.2018.2837384.

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

[27] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manage.*, vol. 50, no. 1, pp. 104-112, 2014, doi: 10.1016/j.ipm.2013.08.006.

[28] S. Chatterjee, D. Goyal, A. Prakash, and J. Sharma, "Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application," *J. Bus. Res.*, vol. 131, pp. 815-825, Jul. 2021, doi: 10.1016/j.jbusres.2020.10.043.

[29] J. F. Kusuma and A. Chowanda, "Indonesian hate speech detection using IndoBERTweet and BiLSTM on Twitter," *JOIV: Int. J. Inform. Visualization*, vol. 7, no. 3, pp. 773-780, 2023, doi:10.30630/joiv.7.3.1035.

[30] J. H. Jaman and R. Abdulrohman, "Sentiment analysis of customers on utilizing online motorcycle taxi service at Twitter with the support vector machine," in *Proc. 3rd Int. Conf. Elect. Eng. Comput. Sci. (ICECOS)*, 2019, pp. 231-234, doi:10.1109/ICECOS47637.2019.8984483.

[31] G. C. Banks, H. M. Woznyj, R. S. Wesslen, and R. L. Ross, "A review of best practice recommendations for text analysis in R (and a user-friendly app)," *J. Bus. Psychol.*, vol. 33, no. 4, pp. 445-459, 2018, doi:10.1007/s10869-017-9528-3.

[32] N. M. G. D. Purnamasari, M. A. Fauzi, Indriati, and L. S. Dewi, "Cyberbullying identification in Twitter using support vector machine and information gain based feature selection," *Indones. J. Elect. Eng. Comput. Sci.*, vol. 18, no. 3, pp. 1494-1500, 2020, doi:10.11591/ijeecs.v18.i3.pp1494-1500.

[33] Y. Q. Lim and Y. L. Loo, "Characteristics of multiclass suicide risks tweets through feature extraction and machine learning techniques," *JOIV: Int. J. Inform. Visualization*, vol. 7, no. 4, pp. 2297-2305, Dec. 2023, doi: 10.62527/joiv.7.4.2284.

[34] M. Liang and T. Niu, "Research on text classification techniques based on improved TF-IDF algorithm and LSTM inputs," *Procedia Comput. Sci.*, vol. 208, pp. 460-470, 2022, doi: 10.1016/j.procs.2022.10.064.

[35] G. A. Dalaorao, A. M. Sison, E. Aguinaldo, and R. P. Medina, "Integrating collocation as TF-IDF enhancement to improve classification accuracy," in *Proc. IEEE 4th Int. Conf. Trends Electron. Informat. (ICOEI)*, 2019, pp. 1-6, doi:10.1109/TSSA48701.2019.8985458.

[36] E. S. Alamoudi and N. S. Alghamdi, "Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings," *J. Decis. Syst.*, vol. 30, no. 2-3, pp. 259-281, 2021, doi: 10.1080/12460125.2020.1864106.

[37] B. P. Zen, I. Susanto, and D. Finaliamartha, "TF-IDF method and vector space model regarding the COVID-19 vaccine on online news," *SinkrOn*, vol. 6, no. 1, pp. 69-79, Oct. 2021, doi:10.33395/sinkron.v6i1.11179.

[38] M. I. C. Rachmatullah, J. Santoso, and K. Surendro, "A novel approach in determining neural networks architecture to classify data with large number of attributes," *IEEE Access*, vol. 8, pp. 204 728-204 743, 2020, doi: 10.1109/access.2020.3036853.

[39] K. Yotov, E. Hadzhikolev, S. Hadzhikoleva, and S. Cheresharov, "Finding the optimal topology of an approximating neural network," *Mathematics*, vol. 11, no. 1, p. 217, Jan. 2023, doi:10.3390/math11010217.

[40] V. Dharanalakota, A. A. Raikar, and P. K. Ghosh, "Improving neural network training using dynamic learning rate schedule for PINNs and image classification," *Mach. Learn. Appl.*, vol. 21, p. 100697, Sep. 2025, doi: 10.1016/j.mlwa.2025.100697.

[41] D. Berrar, "Cross-validation," in *Encyclopedia of Bioinformatics and Computational Biology*, vol. 1, Elsevier, 2019, pp. 542-545, doi:10.1016/B978-0-12-809633-8.20349-X.