Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Comparison of TF-IDF and Word2Vec Feature Representations for Emotion Classification of Tokopedia E-Commerce Review Using LinearSVC

Fitriyani Azzahra[1]*, Bambang Irawan[2], Ahmad Faqih[3], Denni Pratama[4], Dian Ade Kurnia[5]

[1,2,3,4,5]*Department of Informatics Engineering, STMIK IKMI Cirebon*
Fitriyaniazzahra317@gmail.com[1*], Bambangirawan2000@yahoo.com[2], ahmadfaqih367@gmail.com[3], pratamadenni@gmail.com[4], dianade2012@gmail.com[5]

Abstract

This study aims to compare the performance of TF-IDF and Word2Vec feature representations for emotion classification of Tokopedia e-commerce reviews using the LinearSVC algorithm. The dataset used is PRDECT-ID, which consists of 5,400 Indonesian-language reviews labeled with positive and negative emotions. The preprocessing stages include case folding, non-alphabet character cleaning, slang normalization, stopword removal, Sastrawi stemming, and emoji handling. Feature extraction was performed using TF-IDF and Word2Vec, after which the models were trained using LinearSVC and evaluated through 5-Fold Cross Validation and holdout testing. The experimental results show that TF-IDF achieves better performance, with an accuracy of 0.65, a macro-F1 score of 0.645, and a Cohen's Kappa value of 0.294. Meanwhile, Word2Vec attains an accuracy of 0.58 and a macro-F1 score of 0.540. These findings indicate that TF-IDF is more effective for short and informal texts characteristic of Indonesian e-commerce reviews.

Keywords: *TF-IDF, Word2Vec, LinearSVC, Emotion Classification, Tokopedia*

1. Introduction

The rapid development of digital technology has significantly increased online shopping activities in Indonesia, particularly through e-commerce platforms such as Tokopedia. Each transaction conducted by users generates product reviews that contain opinions, experiences, and emotional expressions regarding the quality of products or services received. These reviews represent valuable information for sellers and platforms to understand customer satisfaction. However, the large volume of user-generated reviews makes manual analysis inefficient. Therefore, automated methods based on Natural Language Processing (NLP) are required to accurately classify emotions in textual reviews [1].

Previous studies on emotion and sentiment analysis in Indonesian-language reviews have applied various approaches. Hadju and Jayadi reported that the combination of TF-IDF and Support Vector Machine (SVM) achieved high accuracy in product review analysis [2]. Rahmawati et al. argued that Word2Vec is capable of capturing semantic context better than frequency-based methods, although it requires a large and clean corpus to achieve optimal performance [3]. Furthermore, Prasetyo and Nugroho highlighted that SVM models are effective for emotion classification, but their performance is highly dependent on the feature representation used [4]. Meanwhile, Saputra emphasized the importance of Indonesian-language preprocessing, as e-commerce reviews often contain informal words, abbreviations, and emojis that may degrade prediction quality if not handled properly [5].

Despite extensive research on sentiment and emotion classification, most existing studies focus on a single feature representation without explicitly comparing TF-IDF and Word2Vec on the same dataset. In addition, previous works have not thoroughly examined the effectiveness of these feature representations using the LinearSVC algorithm. This limitation reveals a research gap regarding the comparative performance of frequency-based and semantic-based representations for Indonesian e-commerce review emotion classification [3][4].

Therefore, this study addresses the research gap by explicitly comparing TF-IDF and Word2Vec feature representations for emotion classification of Indonesian e-commerce reviews using LinearSVC. The novelty of this research lies in the implementation of comprehensive preprocessing steps and the evaluation of both feature representations using accuracy, macro-F1, and Cohen's Kappa metrics under Cross Validation and Holdout Test schemes.

2. Methods

This research method is designed to address the problem of emotion classification in Tokopedia e-commerce reviews by comparing two feature representations, namely TF-IDF and Word2Vec, using the Linear Support Vector Classifier (LinearSVC) algorithm. The overall methodological workflow is visualized in Figure 1, which presents the research flowchart consisting of the following stages: problem identification, dataset collection, text preprocessing, TF-IDF and Word2Vec feature extraction, LinearSVC model training, model evaluation, and conclusion drawing. All procedures are structured to ensure reproducibility by other researchers, referring to commonly published methods in previous studies, with only relevant modifications applied based on the characteristics of the data [1][2].

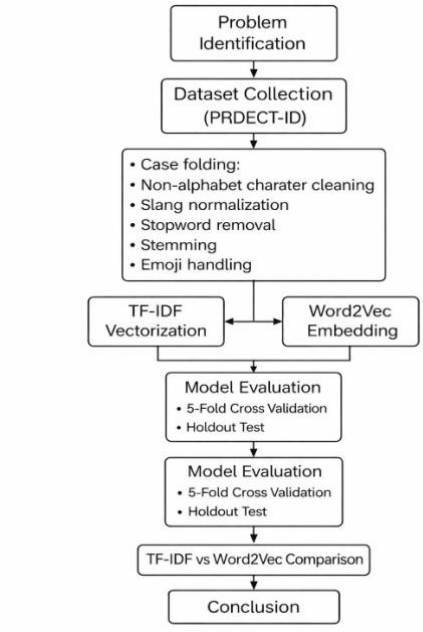

Figure 1. Research Flowchart

2.1. Dataset

The dataset used in this study is PRDECT-ID, consisting of 5,400 Indonesian-language Tokopedia reviews labeled as positive and negative emotions. The dataset is selected because it represents the informal language style commonly found in e-commerce product reviews. The data are imported using the pandas library and inspected to ensure there are no duplicate or missing values.

2.2. Text Preprocessing

The preprocessing stage was conducted to improve text quality before being used as model input. The preprocessing procedures followed standard approaches in natural language processing [3][4], with several modifications to accommodate the characteristics of informal Indonesian-language text. The preprocessing steps consisted of:

1. Case Folding
   All characters were converted to lowercase to eliminate semantic differences caused by capitalization variations.
2. Non-Alphabetic Character Removal
   Numbers, punctuation marks, symbols, URLs, and other irrelevant characters were removed, as they do not contribute to emotion analysis.
3. Slang Normalization
   Non-standard words were converted into their standard forms based on an Indonesian slang dictionary, for example, “bgt” to “banget” and “gk” to “tidak”.
4. Stopword Removal
   Common words that do not carry significant meaning for the classification task were removed using an Indonesian stopword list [5].
5. Sastrawi Stemming
   Words were reduced to their root forms using the Sastrawi library, following the Indonesian stemming methodology [6].
6. Emoji Handling
   Emojis were converted into textual representations using the *demojize* function so that they could be processed as emotional features.

2.3. Feature Extraction

2.3.1. TF-IDF

The Term Frequency–Inverse Document Frequency (TF-IDF) method measures the importance of a word based on its frequency within a document and across the entire corpus [1]. Feature extraction was performed using Tf-IdfVectorizer with the following parameters:

1. n-grams: unigram–bigram
2. minimum document frequency (min_df): 5
3. Indonesian stopwords: enabled
4. output: a sparse matrix containing more than 12,000 features

TF-IDF was selected because it is well suited for short and concise texts such as e-commerce reviews.

2.3.2. Word2Vec

Word2Vec was used to map words into 300-dimensional vectors using the skip-gram approach [2]. The model was trained using the gensim library with the following parameters:

1. vector_size: 300
2. window: 5
3. min_count: 2
4. sg: 1 (skip-gram)

Document representations were generated by calculating the average embedding of all words within each review.

2.4. Classification Model

The classification model employed in this study was the Linear Support Vector Classifier (LinearSVC), an algorithm that has been proven effective for high-dimensional text data [7]. The parameters used were:

1. C = 1.0
2. loss = hinge
3. max_iter = default

The model was trained twice, each time using TF-IDF and Word2Vec features, respectively, to enable an objective performance comparison.

2.5. Model Evaluation

Model evaluation was conducted using two approaches to ensure robust results:

1. Stratified 5-Fold Cross Validation
   The dataset was divided into five folds while maintaining balanced label proportions in each fold. This evaluation reduces bias risk and provides a stable estimation of model performance.
2. Holdout Test (80:20 Split)
3. The dataset was split into 80% training data and 20% testing data to measure the model's generalization performance.

2.6. Evaluation Metrics

The model performance was measured using the following metrics:

1. Accuracy, which measures the proportion of correct predictions
2. Macro-F1 Score, used because the dataset contains two classes with potential class imbalance [8]
3. Cohen's Kappa, which measures the level of agreement between predictions and true labels [9]

3. Results and Discussion

This section presents the results of emotion classification model testing using two feature representations, namely TF-IDF and Word2Vec, evaluated on the LinearSVC model with two evaluation schemes: 5-Fold Cross Validation and an 80:20 Holdout Test. The results are presented in the form of tables and figures to clarify the discussion. All experiments were conducted on the PRDECT-ID dataset, which consists of 5,400 Indonesian-language reviews.

3.1. Cross Validation Results (5-Fold)

The model was evaluated using 5-Fold Stratified Cross Validation. The complete results of the average macro-F1 performance are presented in Table 1.

Table 1: 5-Fold Cross Validation Results

Fold	SVM Linear	SVM RBF	Naïve Bayes	Logistic Regression	Random Forest
1	0.659	0.659	0.659	0.659	0.672
2	0.648	0.648	0.648	0.648	0.651
3	0.644	0.644	0.644	0.644	0.657
4	0.644	0.644	0.644	0.644	0.671

5	0.672	0.672	0.672	0.672	0.678
Average Macro-F1 Score	0.657	0.657	0.657	0.657	0.666
Standard Deviation (±)	0.011	0.011	0.011	0.011	0.011

The model was trained using Word2Vec with mean pooling. Multinomial Naive Bayes was not evaluated because Word2Vec produces negative values.

Table 2: Cross Validation Results (Word2Vec)

Model	Mean Macro-F1	Std Dev
Random Forest	0.6659	0.0119
Logistic Regression	0.5682	0.0063
Linear SVM	0.5680	0.0072

3.2. Holdout Test Results (80:20)

The holdout evaluation was conducted to assess the model's generalization ability when tested on previously unseen data. The results are presented in Table 3.

Table 3: Holdout Test Results (TF-IDF + LinearSVC)

Label	**Precision**	**Recall**	**F1-Score**	**Support**
Negative	0.65	0.73	0.69	564
Positive	0.66	0.56	0.61	516
Accuracy	-	-	0.65	1080
Macro Avg (F1)	-	-	0.645	-
Cohen's Kappa	-	-	0.294	-

In the holdout evaluation, the Linear SVM model based on Word2Vec achieved an accuracy of 54.35%, a macro-F1 score of 0.54, and a Cohen's Kappa value of 0.081. The low Kappa score indicates a weak level of agreement between the model's predictions and the actual labels. These results are presented in Table 4.

Table 4: Holdout Test Results (Word2Vec)

Label	Precision	Recall	F1-Score	Support
Negative	0.5579	0.6063	0.5811	564
Positive	0.5246	0.4748	0.4984	516
Accuracy	-	-	0.5435	1080
Macro Avg (F1)	-	-	0.540	-
Cohen's Kappa	-	-	0.0815	-

Based on these results, it can be concluded that Word2Vec is less optimal for e-commerce review texts, which tend to be short and lack strong sentence context. Since Word2Vec performs better on texts with longer contextual information and well-structured sentences, this approach is unable to effectively capture emotional nuances in short reviews.

3.3. Performance Comparison Results and Visualization

The results of the evaluation using 5-Fold Cross Validation are presented in Table 5. Based on these results, TF-IDF demonstrates superior performance compared to Word2Vec across all evaluation metrics.

Table 5: 5-Fold Cross Validation Results

Method	**Accuracy**	**Macro-F1**	**Kappa**
TF-IDF + LinearSVC	0.64	0.637	0.280
Word2Vec + LinearSVC	0.56	0.523	0.081

The observed performance differences indicate that TF-IDF is more stable in processing short and informal texts, as frequency-based features are better suited to the sentence structures commonly found in e-commerce reviews. In contrast, Word2Vec requires a large corpus to generate consistent and high-quality embeddings. Since the PRDECT-ID dataset is relatively small, the resulting Word2Vec representations are less optimal.

A comparative holdout evaluation was conducted to assess the model's generalization capability when tested on previously unseen data. The results are presented in Table 6.

Table 6: Comparative Holdout Test Results

Method	Accuracy	Macro-F1	Kappa
TF-IDF + LinearSVC	0.65	0.645	0.294
Word2Vec + LinearSVC	0.58	0.540	0.081

Based on these results, TF-IDF remains superior, achieving a macro-F1 score that is 10.5% higher than that of Word2Vec. The Cohen's Kappa value for TF-IDF also indicates a better level of agreement between the predicted and actual labels.

Figure 2 presents a performance comparison of the LinearSVC model using two different feature representations, namely TF-IDF and Word2Vec. According to the visualization, TF-IDF consistently produces higher accuracy and macro-F1 values than Word2Vec. TF-IDF achieves an accuracy of 0.65 and a macro-F1 score of 0.645, whereas Word2Vec attains an accuracy of 0.58 and a macro-F1 score of 0.540. This substantial performance gap indicates that TF-IDF is more effective in capturing the characteristics of short and informal texts such as Tokopedia reviews, enabling the model to classify emotions more reliably. Conversely, Word2Vec tends to be less effective due to its reliance on larger corpora to produce high-quality word embeddings. Overall, the visualization reinforces the conclusion that TF-IDF is the superior feature representation in this study.

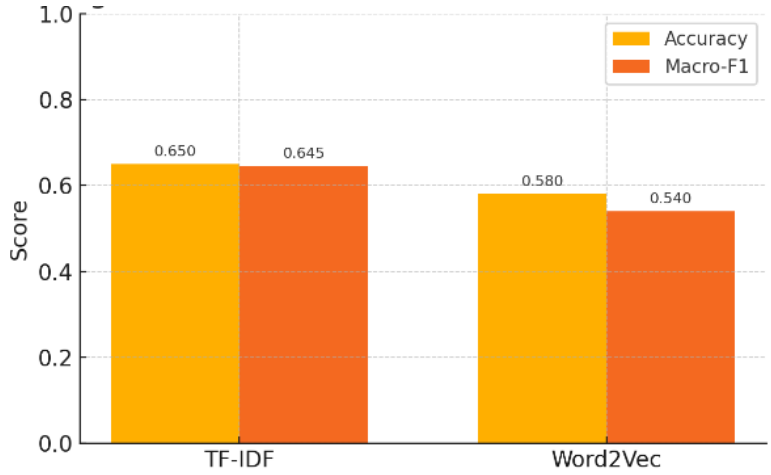

Figure 2: Performance Comparison Visualization

3.4. Discussion

The experimental results indicate that TF-IDF is the most effective feature representation for emotion classification in Tokopedia reviews. This superiority is primarily influenced by the characteristics of review texts, which are generally short, ranging from 5 to 20 words. Under such conditions, frequency-based approaches like TF-IDF are able to capture important information more efficiently. In addition, the language patterns found in e-commerce reviews—often informal, containing abbreviations, and exhibiting variations of non-standard words—do not significantly degrade TF-IDF performance, as this method does not rely on semantic contextual understanding between words. In contrast, Word2Vec is highly dependent on the quality and size of the corpus to produce stable and representative word embeddings. The PRDECT-ID dataset used in this study is relatively small and therefore insufficient to support optimal Word2Vec training. Furthermore, LinearSVC as a classification algorithm performs very well on sparse data such as TF-IDF vectors, making their combination particularly effective. The higher Cohen's Kappa values achieved by TF-IDF also indicate better prediction consistency compared to Word2Vec. Therefore, based on both quantitative analysis and data characteristics, it can be concluded that TF-IDF is superior and more suitable as a feature representation for emotion classification tasks on Indonesian e-commerce reviews.

4. Conclusion

Based on the results of this study, it can be concluded that the TF-IDF feature representation is the most effective method for emotion classification in Tokopedia e-commerce reviews compared to Word2Vec. Through a series of experiments using the LinearSVC model, TF-IDF consistently achieved higher accuracy, macro-F1, and Cohen's Kappa values under both the 5-Fold Cross Validation and Holdout Test schemes, thereby answering the research question regarding the most appropriate feature representation for emotion classification tasks. The effectiveness of TF-IDF is largely influenced by the short and informal nature of review texts, making frequency-based approaches more suitable than embedding-based methods that require large corpora to achieve stability. Thus, this study demonstrates that the combination of TF-IDF and LinearSVC provides a more optimal and reliable solution for analyzing emotions in user reviews on Indonesian e-commerce platforms.

References

[1] D. Ariani, S. Putri, and T. Ramadhani, "Adaptive preprocessing for Indonesian language in consumer text analysis," *Journal of Information Systems*, 2023.

[2] A. Assiroj, D. Rahayu, and R. Firmansyah, "Comparative study of SVM, logistic regression, and naïve Bayes for Indonesian text classification,"

Indonesian Journal of Computing, 2023.
[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
[4] S. Hadju and R. Jayadi, "Sentiment analysis of e-commerce reviews using TF-IDF and SVM," *Jurnal Teknologi Informasi Indonesia*, 2021.
[5] R. Hidayat and P. Sari, "Consumer emotion analysis on e-commerce platforms," *Jurnal Informatika Nusantara*, 2022.
[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson, 2023.
[7] R. Kurniawan and M. Aji, "Characteristics of Indonesian slang language in NLP," *Jurnal Linguistik Komputasional*, 2022.
[8] R. Kurniawan and A. Nugroho, "Consumer review analytics for digital marketing strategy," *Journal of Digital Business*, 2023.
[9] S. Kusumaningrum, D. Pratiwi, and T. Siregar, "Performance comparison of SVM, naïve Bayes, and CNN in Indonesian text classification," *Journal of Data Science*, 2022.
[10] R. Pane *et al*., "Ensemble methods for Indonesian text emotion classification," *Journal of Intelligent Systems*, 2023.
[11] Romadhony *et al*., "Sentiment analysis on a large Indonesian product review dataset," 2024.
[12] L. Koto *et al*., "Indonesian informal text normalization," in *Proceedings of the Asian Language Processing Conference*, 2020.
[13] Putra *et al*., "Machine learning metrics for text classification," 2021.
[14] R. Kurniawan and A. Nugroho, "PRDECT-ID dataset documentation," 2023.