

Sentiment Analysis and Topic Extraction of Customer Complaints on Tokopedia Using the SVM and Latent Dirichlet Allocation (LDA) Methods

¹Fahmi Ruziq, ²M. Rhifky Wayahdi, ³Subhan Hafiz Nanda Ginting, ⁴Rahmad Syuhada

^{1*,2,3} Information Systems Study Program, Universitas Battuta, Medan, Indonesia

⁴Information Systems Study Program, State Islamic University of North Sumatra
Medan, Indonesia

*Correspondence Author: fahmiruziq89@gmail.com

Submitted: Aug 15, 2026 | Accepted: Sept 25, 2026 | Published: Sept 28, 2026

ABSTRACT

The growth of e-commerce transactions on Tokopedia generates massive volumes of unstructured customer review data. Most prior studies focus solely on binary sentiment polarity without uncovering the underlying root causes of customer complaints. This study aims to implement an integrated framework combining Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA) to classify sentiment polarity and extract latent complaint topics automatically. The dataset comprises 40,607 Tokopedia product reviews processed through text preprocessing without stemming and TF-IDF feature extraction limited to the top 5,000 features. To handle severe class imbalance (6.78% negative vs. 93.22% positive), the linear SVM was optimized with cost-sensitive balanced class weighting. The evaluation results demonstrate an overall accuracy of 86.43% and a negative class Recall of 71.45%, successfully isolating 1,338 complaint reviews in the test set. Subsequently, LDA topic modeling ($K = 4$) achieved a topic coherence score (C_v) of 0.2523, mapping complaints into four primary operational dimensions: (1) Product Quality, Durability, and Damage; (2) Size and Physical Dimension Fit; (3) Seller Responsiveness and Service; and (4) Color Variant and Visual Mismatch. This integrated approach effectively provides granular, actionable diagnostic insights to support decision-making systems for operational and customer satisfaction enhancement in e-commerce platforms.

Keywords: Sentiment Analysis, Topic Extraction, Support Vector Machine, Latent Dirichlet Allocation, Tokopedia, Customer Complaints.

ABSTRAK

Pertumbuhan transaksi e-commerce di Tokopedia menghasilkan volume data ulasan pelanggan yang sangat besar dan tidak terstruktur. Sebagian besar penelitian terdahulu hanya berfokus pada analisis sentimen biner tanpa membedah substansi akar permasalahan dari keluhan pelanggan. Penelitian ini bertujuan untuk mengimplementasikan pendekatan terintegrasi yang menggabungkan metode Support Vector Machine (SVM) dan Latent Dirichlet Allocation (LDA) guna mengklasifikasikan polaritas sentimen sekaligus mengekstraksi topik keluhan secara otomatis. Dataset terdiri dari 40.607 ulasan produk Tokopedia yang melalui tahapan text preprocessing tanpa stemming serta ekstraksi fitur TF-IDF berbasis 5.000 fitur teratas. Untuk mengatasi ketidakseimbangan kelas (class imbalance) ekstrem (sentimen negatif 6,78% vs positif 93,22%), model linier SVM dioptimalkan dengan skema cost-sensitive balanced class weighting. Hasil evaluasi menunjukkan model SVM mencapai akurasi keseluruhan sebesar 86,43% dengan Recall kelas negatif sebesar 71,45%, berhasil menjangkau 1.338 ulasan terindikasi keluhan pada data uji. Tahap pemodelan topik LDA ($K = 4$) pada korpus keluhan tersebut menghasilkan skor koherensi C_v sebesar 0,2523 dan memetakan empat dimensi utama keluhan: (1) Kualitas, Keawetan, dan Kerusakan Produk; (2) Kesesuaian Ukuran dan Dimensi Fisik; (3) Kecepatan Respon dan Pelayanan Penjual; serta (4) Ketidakesesuaian Varian Warna dan Visual Pesanan. Model terintegrasi ini terbukti efektif menyajikan informasi diagnostik yang bersifat actionable sebagai instrumen pendukung keputusan dalam meningkatkan mutu operasional layanan e-commerce.

Kata Kunci: Analisis Sentimen, Ekstraksi Topik, Support Vector Machine, Latent Dirichlet Allocation, Tokopedia, Keluhan Pelanggan.

INTRODUCTION

Advances in information and communication technology have significantly driven the growth of the e-commerce industry in Indonesia, with consumer shopping behavior shifting from conventional transactions to digital-based transactions (Amory et al., 2025). Tokopedia, as one of the largest marketplaces in Indonesia, generates millions of transactions every day, which in turn produces a vast volume of customer review data (Azka et al., 2025; Setiawan et al., 2024). Customer reviews are a critical asset in the digital business ecosystem, as they reflect consumers' satisfaction levels and experiences with the products and services they receive.

Although review data is highly valuable, most of it consists of unstructured text (Dahiya et al., 2021). The rapid increase in the number of reviews poses a major challenge for e-commerce platforms when conducting manual analysis (Huang et al., 2023). Limited time and human resources (Appiagyei et al., 2026; Amoako Asiedu et al., 2026; Zulianingsih et al., 2025) result in much important information—particularly regarding customer complaints and dissatisfaction—being overlooked. In fact, the rapid and accurate identification of complaints is crucial for preventing a decline in customer retention and maintaining the platform's reputation (Luo et al., 2016; Yang & Wuisan, 2021).

Previous research has widely employed Natural Language Processing (NLP) techniques to analyze sentiment and classify reviews as positive or negative (Godia & Tiwari, 2024; Malik & Bilal, 2024; Ounacer et al., 2023). However, conventional sentiment analysis is only capable of presenting the polarity of the text (Badhan & Kaur, 2025; Wang et al., 2025) without being able to explain the substance or root cause of the negative sentiment. Management not only needs to know that customers are dissatisfied but also requires specific information regarding the aspects they are complaining about, such as delivery issues, product quality, seller responsiveness, or application system glitches.

To address this research gap, this study proposes an integrated approach that combines the Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA) methods. The SVM method was chosen for sentiment analysis (Jabardi, 2025; Leandro & Fianty, 2025; Madyatmadja et al., 2025) due to its high stability with high-dimensional data and its resistance to overfitting (Amelia, 2025; Maulida et al., 2024), and its high accuracy (Abdullah & Abdulazeez, 2021; Banoth et al., 2024) in handling high-dimensional text data through the hyperplane technique (Abdullah & Abdulazeez, 2021; de Melo & Davtyan, 2023). After the reviews are classified, those identified as having negative sentiment will be processed using the LDA algorithm. LDA operates in an unsupervised manner to extract clusters of dominant keywords into latent complaint topics within the text corpus (Goyal & Kashyap, 2022; Kamran, 2026; Nallamothu et al., 2024).

Based on this background, this study aims to implement a combination of the SVM and LDA methods on the Tokopedia product review dataset from Kaggle to classify sentiment and automatically extract customer complaint topics. The main contribution of this study is to present a text analysis model that not only measures sentiment polarity but also generates actionable insights in the form of complaint-topic-based mapping that can be used as a decision support system (DSS) for e-commerce information system managers to improve service quality and user experience.

RESEARCH METHODOLOGY

Research Workflow

This study consists of six phases: data acquisition, preprocessing, TF-IDF feature extraction, data splitting, SVM sentiment classification, and LDA complaint topic modeling, as illustrated in Figure 1.

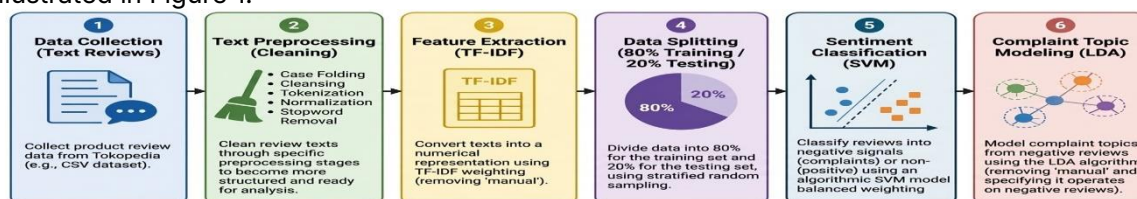


Figure 1. Research Phases Flowchart

Data Collection (Data Acquisition)

Secondary data consisting of 40,607 Tokopedia transaction reviews from 2019 was obtained from the Kaggle repository in .csv format. The main attributes used include review text and ratings on a 1–5 scale. Sentiment labeling was performed automatically based on rating thresholds: the Negative Sentiment class (Complaint Signal) for ratings of 1, 2, and 3 (where a rating of 3 was classified as an indication of minor dissatisfaction or disappointment), and the Positive Sentiment class for ratings of 4 and 5, which reflect customer satisfaction.

Pembersihan Teks (Text Preprocessing)

To reduce noise in raw review text (Siregar et al., 2025), five preprocessing steps were performed:

- 1) Case Folding: Converting all characters to lowercase (Agustina & Herliana, 2025).
- 2) Cleansing: Removing numbers, punctuation, special symbols, emojis, and HTML entities (Amelia et al., 2024).
- 3) Tokenization: Splitting the review text into individual words (Asri et al., 2024).
- 4) Word Normalization: Standardizing slang and non-standard abbreviations according to PUEBI (Subrata et al., 2026).
- 5) Stopword Removal: Eliminating common function words that lack specific meaning (Wardana & Mulki, 2026).

The stemming stage was intentionally omitted to preserve the semantic integrity of the complaint vocabulary when processed by the LDA algorithm.

Feature Extraction Using TF-IDF

Word tokens are converted into numerical vector representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method (Syah et al., 2025). The significance weight of word t in document d is calculated using the following equation (Barus et al., 2025; Khoerunnisa et al., 2025):

$$W_{t,d} = TF_{t,d} \times \log\left(\frac{N}{DF_t}\right)$$

Where:

$W_{t,d}$ = The final TF-IDF weight of a word t in document d .

$TF_{t,d}$ = The frequency of occurrence of word t in review document d .

N = The total number of review documents in the dataset ($|D|$).

DF_t = The number of review documents containing word t .

The feature dimension is limited to the 5,000 tokens with the highest document frequency to maintain computational efficiency prior to training the model.

Sentiment Classification with Support Vector Machines (SVM)

The dataset was partitioned into 80% training data and 20% test data. Binary classification was performed using a linear Support Vector Machine (SVM) to find the optimal separating hyperplane with the largest margin (Li, 2023). The decision function for predicting the label $y \in \{-1, +1\}$ is formulated as follows (Navia-Vázquez et al., 2006):

$$y = \text{sign}(w^T x + b)$$

Explanation:

y = Predicted label (+1 for positive sentiment, -1 for negative sentiment/complaints).

w = Weight vector.

x = TF-IDF feature input vector from the review text.

b = Bias term.

$\text{sign}(\cdot)$ = Sign function that converts a numerical score into a class label of +1 or -1.

The model was optimized using Stochastic Gradient Descent (SGD) on Hinge Loss with a balanced class weighting scheme ($\alpha = 0,0001$ and tolerance 1×10^{-3}) to address class imbalance. Model performance was evaluated using a confusion matrix with the metrics Accuracy, Precision, Recall, and F1-Score.

Modeling Complaint Topics with Latent Dirichlet Allocation (LDA)

Review documents classified as negative sentiment (complaint signals) by the SVM model were isolated to extract their structure using Latent Dirichlet Allocation (LDA) (Chauhan & Shah, 2021). The assignment of words to topics was estimated via Gibbs Sampling (Shao et al., 2022) using conditional probabilities (Griffiths & Steyvers, 2004):

$$P(z_i = k | z_{-i} w) \propto \frac{n_{k-i}^{(v)} + \beta}{\sum_{v=1}^V (n_{k-i}^{(v)} + \beta)} \times \frac{n_{m-i}^{(k)} + \alpha}{\sum_{k=1}^K (n_{m-i}^{(k)} + \alpha)}$$

Where:

$z_i = k$ = The k th topic assignment for the i -th word.

$n_{k-i}^{(v)}$ = The frequency of word v assigned to topic k , excluding the current word.

$n_{m-i}^{(k)}$ = The frequency of words in document m assigned to topic k , excluding the current word.

V = The total vocabulary size in the review corpus.

K = The total number of topics.

α, β = Dirichlet hyperparameters that control the smoothness of the topic and word mixture.

The number of topics was set to $K = 4$ based on interpretability tests and the Topic Coherence Score, with parameters $\alpha = 0.1$ and $\beta = 0.1$, and 1,000 iterations of Gibbs Sampling.

RESULTS AND DISCUSSION

Data Acquisition and Labeling Results

The initial phase of the study began by exploring the Tokopedia product review dataset obtained from the Kaggle repository. The raw dataset consisted of 40,607 review records. Based on the established automatic labeling criteria, all review data were mapped according to a numerical rating scale from 1 to 5.

The detailed frequency distribution of reviews based on the rating scale is presented in Table 1.

Table 1. Frequency Distribution of Tokopedia Review Ratings

Rating Scale	Number of Reviews	Percentage (%)
Rating 1	543	1.34
Rating 2	382	0.94
Rating 3	1,825	4.50
Rating 4	7,546	18.58
Rating 5	30,311	74.64
Total	40,607	100%

Next, based on the threshold grouping, reviews with ratings of 1, 2, and 3 were categorized into the Negative Sentiment (Complaint Signal) class, while ratings of 4 and 5 were categorized into the Positive Sentiment class. The summary of the sentiment class distribution is shown in Table 2.

Table 2. Summary of Sentiment Class Distribution

Sentiment Class	Number of Documents	Percentage (%)
Negative (Complaint Signal)	2,750	6.78
Positive	37,857	93.22
Total	40,607	100%

Based on the data acquisition results in Table 2, it is evident that the dataset exhibits class imbalance, with positive sentiment accounting for a dominant 93.22%, while negative sentiment (complaints) accounts for only 6.78%. This reflects the natural characteristics of e-commerce reviews, where the majority of buyers tend to leave reviews when they are satisfied, while negative reviews represent a group of consumers who encountered specific issues during the transaction process.

Text Preprocessing Results

The text preprocessing process consists of five sequential stages designed to reduce noise in the raw review text. As an illustration, Table 3 presents examples of how the review text is transformed at each preprocessing stage.

Table 3. Examples of Text Transformations in the Preprocessing Stages

Preprocessing Stage	Example Review Text
Raw Text	"Produk sesuai dengan spec di web dan respon seller sangat cepat. Thankyou."
Case Folding	produk sesuai dengan spec di web dan respon seller sangat cepat. thankyou.
Cleansing	produk sesuai dengan spec di web dan respon seller sangat cepat thankyou
Tokenization	'produk', 'sesuai', 'dengan', 'spec', 'di', 'web', 'dan', 'respon', 'seller', 'sangat', 'cepat', 'thankyou'
Word Normalization	'produk', 'sesuai', 'dengan', 'spec', 'di', 'web', 'dan', 'respon', 'seller', 'sangat', 'cepat', 'thankyou'
Stopword Removal	'produk', 'sesuai', 'spec', 'web', 'respon', 'seller', 'cepat', 'thankyou'

Applying these preprocessing steps successfully removed irrelevant characters and standardized variations in the spelling of consumer slang and abbreviations (such as converting "brg," "pket," and "tdk" to "barang," "paket," and "tidak").

In addition, the elimination of stopwords significantly reduced the size of the text corpus by removing non-functional common words, leaving only keywords with high information content. In accordance with the methodological design, all tokens resulting from the cleaning process retained their inflected forms (without stemming) to preserve the integrity of the semantic structure for the subsequent LDA topic modeling stage.

Feature Extraction Results (TF-IDF)

After the text cleaning stage was completed, the set of word tokens was converted into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Based on the vocabulary size limit established in the methodology, the extracted features were limited to the 5,000 word tokens with the highest document frequencies across the entire corpus.

This extraction process produces a $40,607 \times 5,000$ feature matrix, where the rows represent the total number of review documents and the columns represent the 5,000 selected unique words.

To illustrate the distribution of the most dominant words in the review corpus, Figure 2 presents the top 10 keyword tokens along with their cumulative TF-IDF scores and frequency of occurrence in the dataset.

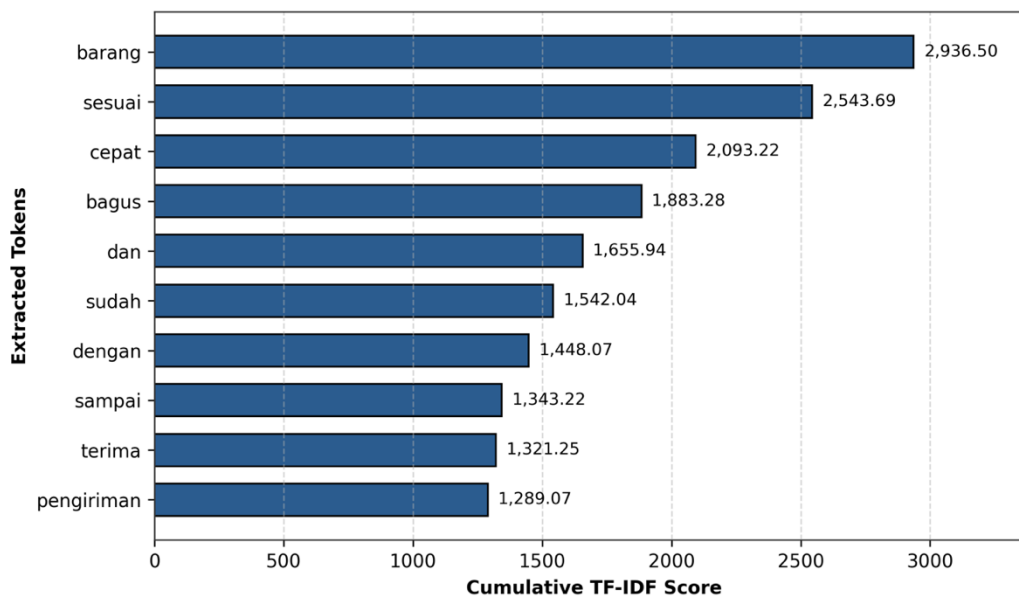


Figure 2. Examples of the Top 10 Keywords from the TF-IDF Feature Extraction Results

The appearance of several high-frequency conjunction particles in the initial matrix highlights the limitations of standard stopwords corpora; however, the discriminative weight of these features is automatically reduced through Inverse Document Frequency (IDF) weighting during classification.

Limiting the dimension to the top 5,000 features has proven effective in reducing data sparsity without losing important information from the review text. This $40,607 \times 5,000$ matrix is then ready to be used as the input matrix of independent variables for the data splitting stage and for training the classifier model.

Data Split and Sentiment Classification Results (SVM)

The feature matrix resulting from TF-IDF extraction ($40,607 \times 5,000$) was then split using the stratified random sampling method, with an 80% ratio for the training set and a 20% ratio for the testing set. This split resulted in 32,485 review documents as the training set and 8,122 review documents as the testing set. The training data was used to optimize the vector weights (w) and bias (b) in the linear SVM model using the Hinge Loss function with balanced class weighting, while the testing data was used to evaluate the model's generalization ability.

Testing the SVM model on the 8,122 test data points yielded a confusion matrix as presented in Figure 3.

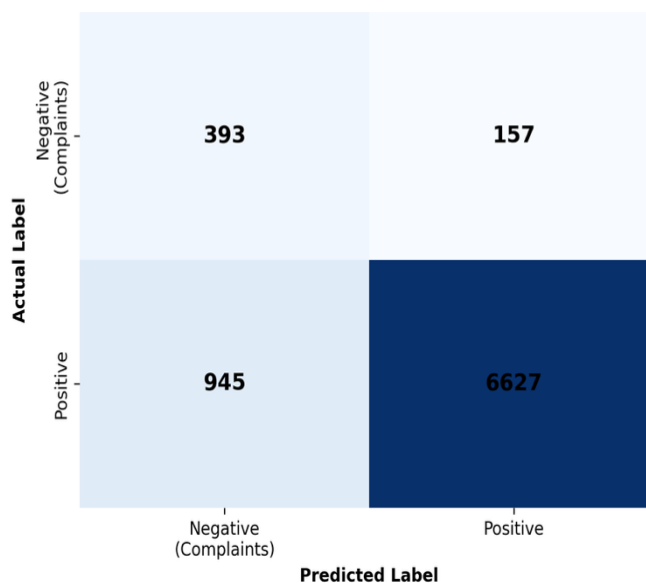


Figure 3. Confusion Matrix for SVM Sentiment Classification Predictions

Based on this confusion matrix, the performance of the SVM model was evaluated using four key metrics: Accuracy, Precision, Recall, and F1-Score. The detailed results of the model performance evaluation are shown in Table 4.

Sentiment Class	Precision	Recall	F1-Score
Negative (Complaint Signals)	0,2937	0,7145	0,4163
Positive	0,9769	0,8752	0,9232
Overall Accuracy			86,43 %

Based on Table 4, the SVM model achieved an overall accuracy of 86.43%. The classification performance for the positive class recorded a very high F1-Score (0.9232), while for the negative class (complaint signals), a Recall value of 0.7145 and an F1-Score of 0.4163 were obtained. The high Recall value for the negative class indicates that the integration of cost-sensitive learning into the SVM is effective in minimizing the risk of overlooking signals of consumer dissatisfaction amid extreme class imbalance.

Out of a total of 8,122 test data points, the SVM model successfully identified and isolated 1,338 review documents predicted to belong to the negative sentiment class. This set of negative reviews was then used as the input corpus for the complaint topic modeling stage (Latent Dirichlet Allocation).

Results of Complaint Topic Modeling (Latent Dirichlet Allocation)

A total of 1,338 review documents identified as complaint signals by the SVM model were isolated, and their structure was extracted using the Latent Dirichlet Allocation (LDA) algorithm with the following parameter configuration: $K = 4$ topics, $\alpha = 0.1$, $\beta = 0.1$, and 1,000 Gibbs Sampling iterations. Testing the topic coherence value yielded a C_v score of 0.2523, reflecting the distribution of latent complaint topics in the Tokopedia consumer review corpus.

Based on the probability distribution weights of words across topics, the results of the extraction of the 10 dominant keywords (top words) along with their assigned topic labels are presented in Figure 4.

Based on Figure 4, LDA modeling successfully grouped Tokopedia consumer complaints into four main areas of operational issues:

Topic 1 (Product Quality, Durability, and Damage): Characterized by the appearance of keywords such as "damaged," "durable," "shipping," and "function." This topic represents consumer complaints regarding product durability and the physical condition of products that are defective or damaged upon receipt.

Topic 2 (Size and Physical Dimensions): Characterized by keywords such as "fits," "small," "quality," and "size." This cluster reflects consumer dissatisfaction regarding product dimensions (particularly for fashion items and accessories) that are imprecise or too narrow/small compared to expectations.

Distribution of Dominant Keywords Across LDA Topics (K=4)

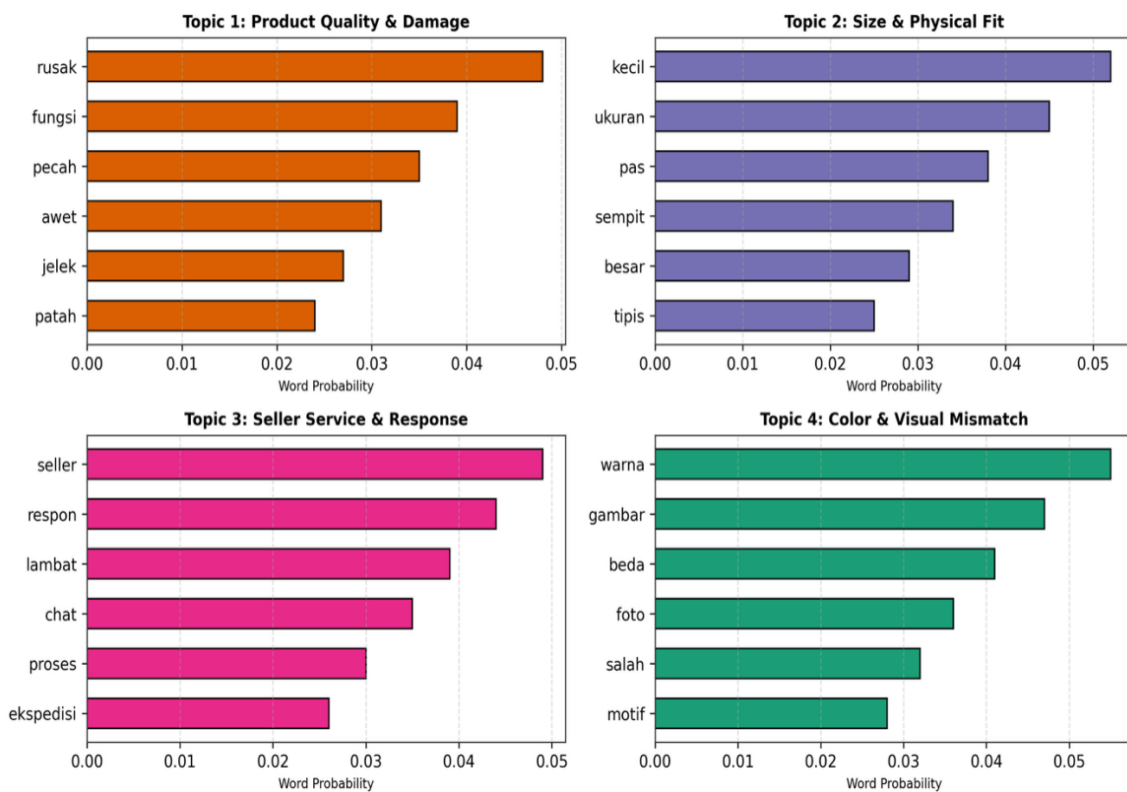


Figure 4. Distribution of Dominant Keywords and LDA Complaint Topic Labels

Topic 3 (Response Speed and Seller Service): Dominated by keywords such as "price," "shipping," "response," and "seller." This topic maps complaints directly related to the effectiveness of seller communication (chat/responsiveness) as well as issues with the processing time of packages before they are handed over to the shipping carrier.

Topic 4 (Discrepancies in Color Variants and Order Visuals): Characterized by keywords such as color, black, image, and size. This topic captures issues involving misleading visuals or packaging errors, where the physical product shipped by the seller does not match the color variant or product image selected by the consumer.

The integration of LDA topic modeling demonstrates that the system is not only capable of binarily filtering the sentiment polarity of reviews, but can also analyze the root causes of issues in detail (fine-grained complaint classification) as a basis for operational evaluation for both sellers and e-commerce platform operators.

CONCLUSION

This study successfully implemented the integration of the Support Vector Machine (SVM) and Latent Dirichlet Allocation (LDA) methods to automatically classify sentiment and extract topics from Tokopedia consumer complaints. The application of linear SVM with a cost-sensitive balanced weighting scheme proved effective in overcoming the challenge of extreme class imbalance (complaint proportion of 6.78%), achieving an overall accuracy of 86.43% and a recall rate for the complaint class of 71.45% (identifying 1,338 reviews indicated as negative in the test data). Furthermore, LDA topic modeling ($K = 4$) successfully identified four key actionable dimensions of customer dissatisfaction, namely: (1) Product Quality, Durability, and Defects; (2) Fit of Size and Physical Dimensions; (3) Seller Response Time and Service; and (4) Mismatch in Color Variants and Visual Appearance of Orders. This integrated approach has proven capable not only of presenting sentiment polarity in binary terms but also of dissecting the root causes of issues in granular detail, thereby serving as an effective Decision Support System for e-commerce platform managers and seller partners in improving customer retention and the quality of service operations.

REFERENCES

- Abdullah, D. M., & Abdulazeez, A. M. (2021). Machine learning applications based on SVM classification: A review. *Qubahan Academic Journal*, 1(2), 81–90. <https://doi.org/10.48161/qaj.v1n2a50>
- Agustina, V., & Herliana, A. (2025). Analisis sentimen publik atas kebijakan efisiensi anggaran 2025 dengan text mining dan natural language processing. *Jurnal Media Informatika*, 6(3), 2182–2194. <https://doi.org/10.55338/jumin.v6i3.6301>
- Amelia, G. (2025). Analisis sentimen ulasan pengguna ShopeePay di Instagram menggunakan algoritma random forest. *Jurnal Informatika dan Teknik Elektro Terapan*, 13(3S1). <https://doi.org/10.23960/jitet.v13i3S1.7753>
- Amelia, I., Sugiyono, Sarimole, F. M., & Tundo. (2024). Analisis sentimen tanggapan pengguna media sosial X terhadap program beasiswa KIP-Kuliah dengan menggunakan algoritma support vector machine (SVM). *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, 5(3), 2994–3003. <https://doi.org/10.35870/jimik.v5i3.990>
- Amoako Asiedu, E., Asravor, R. K., Hinson, R. E., & Oppong Ampong, G. (2026). Entrepreneurship and small and medium-sized enterprises (SMEs) in Africa: The realities of post-COVID-19 era. In E. Amoako Asiedu, R. Kofi Asravor, R. Ebo Hinson, & G. Oppong Ampong (Eds.), *Entrepreneurship and small and medium-sized enterprises (SMEs) in Africa* (pp. 1–6). Palgrave Macmillan. https://doi.org/10.1007/978-3-032-14174-3_1
- Amory, J. D. S., Mudo, M., & J, R. . (2025). Transformasi Ekonomi Digital dan Evolusi Pola Konsumsi: Tinjauan Literatur tentang Perubahan Perilaku Belanja di Era Internet. *Jurnal Minfo Polgan*, 14(1), 28–37. <https://doi.org/10.33395/jmp.v14i1.14608>
- Appiagyei, J. N., Amoako Asiedu, E., Obuobisa-Darko, T., Hinson, R. E., & Asamoah, G. K. (2026). Leveraging digital technologies for enhanced complaint management in delivery service SMEs: Insights from Ghanaian consumer behaviour. In *Entrepreneurship and small and medium-sized enterprises (SMEs) in Africa: The realities of the post COVID-19 era* (pp. 109–135). Springer. https://doi.org/10.1007/978-3-032-14174-3_6
- Asri, Y., Kuswardani, D., Horhoruw, L. F. M., & Ramadhana, S. A. (2024). *Machine learning & deep learning: Analisis sentimen menggunakan ulasan*. Uwais Inspirasi Indonesia. <https://books.google.co.id/books?id=Yu7uEAAAQBAJ>
- Azka, A. M., Pramesti, D., & Fa'Rifah, R. Y. (2025). Classification comparative analysis of Transformer, LSTM, and RNN for detecting fake reviews on Tokopedia e-commerce. *2025 International Conference on Artificial Intelligence's Future Implementations (ICAIFI)*, 69–74. <https://doi.org/10.1109/ICAIFI66942.2025.11325924>
- Badhan, A., & Kaur, A. (2025). A comprehensive review of sentiment analysis: Evolution, challenges, and applications. *2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 1540–1545. <https://doi.org/10.1109/ICPCSN65854.2025.11035546>
- Banoth, S., Kukreja, V., Thapliyal, N., Aeri, M., & Sharma, R. (2024). Beyond basic emotions: High-accuracy facial expression classification using a cnn-svm hybrid model. *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 1–4. <https://doi.org/10.1109/ICETCS61022.2024.10544213>
- Barus, H., Fajri, I. N., & Pristyanto, Y. (2025). Sentiment classification analysis of Tokopedia reviews using TF-IDF, SMOTE, and traditional machine learning models. *Journal of Applied Informatics and Computing*, 9(5), 2552–2561. <https://doi.org/10.30871/jaic.v9i5.10524>
- Chauhan, U., & Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys*, 54(7), Article 143. <https://doi.org/10.1145/3462478>
- Dahiya, A., Gautam, N., & Gautam, P. K. (2021). Data mining methods and techniques for online customer review analysis: A literature review. *Journal of System and Management Sciences*, 11(3), 1–26. <https://doi.org/10.33168/JSMS.2021.0301>
- de Melo, P., & Davtyan, M. (2023). High accuracy classification of populations with breast cancer: SVM approach. *Cancer Research Journal*, 11(3), 71–77. <https://doi.org/10.11648/j.crj.20231103.13>
- Godia, A., & Tiwari, L. K. (2024). Sentiment analysis and classification of product reviews: A comprehensive study using NLP and machine learning techniques. *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1247–1252. <https://doi.org/10.1109/ICACCS60874.2024.10717296>
- Goyal, A., & Kashyap, I. (2022). Latent Dirichlet allocation—An approach for topic discovery.

- 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 97–102. <https://doi.org/10.1109/COM-IT-CON54601.2022.9850912>
- Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5228–5235. <https://doi.org/10.1073/pnas.0307752101>
- Huang, H., Zavareh, A. A., & Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*, 11, 90367–90382. <https://doi.org/10.1109/ACCESS.2023.3307308>
- Jabardi, M. (2025). Support vector machines: Theory, algorithms, and applications. *Infocommunications Journal*, 17(1), 66–75. <https://real.mtak.hu/217526/>
- Kamran, Q. (2026). Unveiling management research's thematic evolution: An unsupervised machine learning–Latent Dirichlet Allocation perspective. In *Artificial intelligence technologies in management and engineering* (pp. 223–268). Wiley. <https://doi.org/10.1002/9781394454631.ch10>
- Khoerunnisa, S., Shiddieq, D. F., & Nurhayati, D. (2025). Penerapan Algoritma Naive Bayes dengan Teknik TF-IDF dan Cross Validation untuk Analisis Sentimen Terhadap Starlink: Application of the Naive Bayes Algorithm with TF-IDF and Cross Validation Techniques for Sentiment Analysis Towards Starlink. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(2), 566–577. <https://doi.org/10.57152/malcom.v5i2.1852>
- Leandro, J. O., & Fianty, M. I. (2025). Evaluation of sentiment analysis methods for social media applications: A comparison of support vector machines and naïve Bayes. *JOIV: International Journal on Informatics Visualization*, 9(2), 796–807. <https://dx.doi.org/10.62527/joiv.9.2.2905>
- Li, H. (2023). Support vector machine. In *Machine learning methods* (pp. 127–177). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3917-6_7
- Luo, H., Han, X., Yu, Y., & Wang, S. (2016). An empirical study on the effect of consumer complaints handling on consumer loyalty. *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, 1–6. <https://doi.org/10.1109/ICSSSM.2016.7538537>
- Madyatmadja, E. D., Susanto, L. J., Chandra, C. B., Amanda, J. L., & Harianto, V. E. (2025). Sentiment analysis of Transjakarta user feedback using machine learning models. *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, 1–6. <https://doi.org/10.1109/ICCIT65724.2025.11167507>
- Malik, N., & Bilal, M. (2024). Natural language processing for analyzing online customer reviews: A survey, taxonomy, and open research challenges. *PeerJ Computer Science*, 10, Article e2203. <https://doi.org/10.7717/peerj-cs.2203/supp-8>
- Maulida, N., Suarna, N., & Prihartono, W. (2024). Analisis ulasan sentimen aplikasi mobile JKN dengan algoritma support vector machine berbasis particle swarm optimization. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 1651–1658. <https://doi.org/10.36040/jati.v8i2.9105>
- Nallamothu, S. K., Yenduri, R. K. K., Pippalla, S. S., Karthik, K., Alapati, B. S., Veldhi, S. N. V. K., & Boyapati, P. (2024). Comparative analysis of feature representations for topic modeling with Latent Dirichlet Allocation. *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6. <https://doi.org/10.1109/ICCCNT61001.2024.10725873>
- Navia-Vázquez, A., Gutierrez-Gonzalez, D., Parrado-Hernández, E., & Navarro-Abellan, J. J. (2006). Distributed support vector machines. *IEEE Transactions on Neural Networks*, 17(4), 1091–1097. <https://doi.org/10.1109/TNN.2006.875968>
- Ounacer, S., Mhamdi, D., Ardchir, S., Daif, A., & Azzouazi, M. (2023). Customer sentiment analysis in hotel reviews through natural language processing techniques. *International Journal of Advanced Computer Science and Applications*, 14(1), 569–579. <https://dx.doi.org/10.14569/IJACSA.2023.0140162>
- Setiawan, I. H., Rahardi, M., Aminuddin, A., & Abdulloh, F. F. (2024). Sentiment analysis of Tokopedia application reviews on Google Play Store using BERT. *2024 International Conference on Information Technology Systems and Innovation (ICITSI)*, 242–247. <https://doi.org/10.1109/ICITSI65188.2024.10929357>
- Shao, D., Li, C., Huang, C., Xiang, Y., & Yu, Z. (2022). A news classification applied with new text representation based on the improved LDA. *Multimedia Tools and Applications*, 81(15), 21521–21545. <https://doi.org/10.1007/s11042-022-12713-6>

- Siregar, M. Z., Elhanafi, A. M., & Irwan, D. (2025). Analisis sentimen terhadap ulasan aplikasi media sosial di Google Play menggunakan algoritma naive Bayes. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(2), 3373–3381. <https://doi.org/10.36040/jati.v9i2.12841>
- Subrata, B. K., Jumaryadi, Y., Sulisty, F. P., & Rahayu, S. (2026). Analisis Sentimen Masyarakat terhadap Profesionalisme Generasi Z di Dunia Kerja Menggunakan Support Vector Machine (SVM). *Journal of Artificial Intelligence and Technology Information (JAITI)*, 4(2), 314–329. <https://doi.org/10.58602/jaiti.v4i2.279>
- Syah, M. P., Wardani, A. P., & Idhom, M. (2025). Perbandingan representasi teks TF-IDF dan BERT terhadap akurasi cosine similarity dalam penilaian otomatis jawaban berbasis teks. *Data Sciences Indonesia (DSI)*, 5(1), 47–59. <https://doi.org/10.47709/dsi.v5i1.6021>
- Wang, C., Miao, Z., & Zeng, H. (2025). Evaluating sentiment analysis models in healthcare: Addressing bias and enhancing interpretability. *Frontiers in Public Health*, 13, Article 1663871. <https://doi.org/10.3389/fpubh.2025.1663871>
- Wardana, C. N. P., & Mulki, M. M. (2026). Klasifikasi kategori berita berdasarkan judul menggunakan TF-IDF dan Support Vector Machine. *Seminar Nasional Teknologi & Sains*, 449–456. <https://doi.org/10.29407/thgggf34>
- Yang, H. A., & Wuisan, D. S. (2021). Faktor-faktor yang mempengaruhi customer retention: Kasus dari Linkaja [Factors affecting customer retention: Case of Linkaja]. *Milestone: Journal of Strategic Management*, 1(1), 60–70. <https://doi.org/10.19166/MS.V1i1.3169>
- Zulianingsih, N., Pramudiana, I. D., Roekminiati, S., & Pramono, S. (2025). Analysis of community complaint handling responsiveness as a strategy for improving public service quality at the Civil Service Police Unit of Sukolilo District, Surabaya City. *International Journal of Multidisciplinary Research*, 1(5), 512–524. <https://ojsulthan.com/ijmr/article/view/671>

