Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

Analyzing Chatbot Interaction Patterns and Customer Satisfaction Classification Using Text Data

Dhafin Rizky Trisandy[1*], Rizka Dhini Kurnia[2], Purwita Sari[3], Ari Wedhasmara[4], Apriansyah Putra[5]
[1,2,3,4,5] Information System, Faculty Of Computer Science, Universitas Sriwijaya, Indonesia
[1*]09031282328076@student.unsri.ac.id, [2]Rizkadhini@gmail.com, [3]wita@ilkom.unsri.ac.id,
[4]a.wedhasmara@gmail.com, [5]Apriansyah@unsri.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 16-10-2026
Accepted: 25-10-2026
Published: 01-11-2026

Keywords:
Chatbot; Customer Satisfaction; Machine Learning; Support Vector Machine; Text Classification; TF-IDF.

Brilliance: Research of Artificial Intelligence is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

This study aims to identify recurring customer interaction patterns in the chatbot service of CV Mawar and independently evaluate text-based customer satisfaction classification using the Flipkart customer support dataset.

The primary dataset consisted of 137 chatbot interaction sessions from CV Mawar. After removing 10 sessions with empty transcripts, 127 sessions were manually coded and analyzed descriptively. Because reliable customer satisfaction labels were limited, a separate Flipkart dataset containing customer remarks and Customer Satisfaction (CSAT) scores was used for machine learning. Text classification was performed using Term Frequency–Inverse Document Frequency (TF-IDF) and a linear Support Vector Machine (SVM). The model was evaluated using accuracy, balanced accuracy, precision, recall, F1-score, and a confusion matrix.

Delivery estimates and product information or usage were the most frequent categories in CV Mawar, each representing 19.69% of the coded sessions. Only three usable sessions contained genuine service ratings, all rated “Good,” while no usable written feedback was available. The Flipkart classification model achieved 79.62% accuracy, 81.92% balanced accuracy, and a macro F1-score of 78.31%, outperforming the majority-class baseline accuracy of 68.74%.

The findings highlight common customer information needs in chatbot interactions and demonstrate the potential of text classification for analyzing customer feedback. However, the limited satisfaction ratings at CV Mawar and the use of an independent Flipkart dataset restrict conclusions about satisfaction among CV Mawar customers.

INTRODUCTION

The rapid growth of digital commerce has increased the demand for customer service that can respond quickly and consistently to customer needs. Artificial intelligence (AI) chatbots have become an important technology in this area because they can automate customer interactions and provide assistance without relying entirely on human agents. However, implementing a chatbot does not automatically result in a satisfactory customer experience. The quality of communication, responsiveness, information provided, and handling of service problems can influence how customers perceive the service. (Hsu & Lin, 2023) highlighted the importance of conversational quality and service recovery, while (Jiang et al., 2022b) examined the relationships between chatbot communication, customer satisfaction, and engagement.

Customer satisfaction in chatbot services is closely connected to the interactions customers experience. (Xu et al., 2023) showed that temporal and conversational cues can influence perceived warmth, competence, and customer satisfaction. Similarly, (Sari & Adinda, 2023) examined customer experiences with a chatbot used by a telecommunications company in Indonesia, highlighting the different needs and problems that can arise during chatbot use. These findings suggest that understanding customer service requires examining not only the technology itself but also the actual requests and issues reflected in customer interactions.

Many previous studies have investigated chatbot satisfaction through surveys, interviews, or controlled experiments. Although these methods provide valuable information about customer perceptions, actual chatbot transcripts can offer a complementary perspective by revealing recurring requests and interaction patterns. (Følstad & Taylor, 2021) proposed a framework for analyzing customer service chatbot dialogues to investigate user experience. Examining real-world transcripts is therefore useful for identifying the types of information customers seek and the interaction categories that occur most frequently. Nevertheless, transcript patterns alone cannot establish customer satisfaction without sufficient ratings or other reliable evaluation data.

Text classification provides another approach to analyzing customer feedback. (Hadwan et al., 2022) applied machine learning to classify satisfaction-related feedback, while (Sumantiawan et al., 2023) and (Nugroho et al., 2023) examined Support Vector Machine (SVM)-based approaches for classifying customer-generated text. More broadly,

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

(Aldunate et al., 2022) investigated customer satisfaction using natural language processing and deep learning. These studies demonstrate the potential of textual data and machine learning for customer feedback analysis. However, chatbot transcript analysis and machine-learning-based satisfaction classification are often examined separately.

This study addresses this gap through two complementary but independent analyses. The primary dataset consists of 137 chatbot interaction sessions collected from CV Mawar, of which 127 contained usable transcripts after data cleaning. These transcripts are manually coded to identify recurring customer interaction patterns. Available service ratings are examined as supplementary information, although their limited availability restricts conclusions about satisfaction. The secondary dataset is obtained from Flipkart and contains customer remarks and Customer Satisfaction (CSAT) scores. These remarks are processed using Term Frequency–Inverse Document Frequency (TF-IDF) and classified using a linear Support Vector Machine (SVM).

The two datasets are analyzed independently and are not combined for prediction. Accordingly, the objectives of this study are to identify recurring customer interaction patterns in the CV Mawar chatbot service and to evaluate the performance of text-based customer satisfaction classification using the Flipkart dataset. By providing descriptive evidence from actual chatbot conversations alongside an independent classification experiment, this study seeks to offer practical insights into customer service improvement while recognizing the limitations of the available satisfaction data.

LITERATURE REVIEW

AI chatbots are increasingly used in customer service, but their use does not automatically ensure customer satisfaction. Service quality, information quality, system quality, conversational quality, and responsiveness have been identified as important factors in chatbot-based customer experiences (Hsu & Lin, 2023; Jenneboer et al., 2022; Jiang et al., 2022a) . Customer satisfaction may also vary according to interaction conditions, including response style and how service failures are handled (Ruan & Mezei, 2022; Xu et al., 2023; Zhu et al., 2023).

Most previous studies examine chatbot satisfaction using surveys, experiments, or interviews. In contrast, (Følstad & Taylor, 2021) demonstrated that actual chatbot transcripts can be analyzed to identify interaction patterns and user experience issues. This approach provides a useful basis for examining real customer–chatbot conversations.

Text classification has also been applied to customer satisfaction analysis. (Hadwan et al., 2022) demonstrated the use of machine learning for satisfaction classification from textual feedback, while (Sumantiawan et al., 2023) and (Nugroho et al., 2023) showed the effectiveness of TF-IDF and SVM for classifying customer-generated text.

Despite these developments, chatbot transcript analysis and text-based satisfaction classification are often studied separately. This study addresses the gap by analyzing 137 real chatbot sessions from CV Mawar to identify recurring interaction patterns and using a separate Flipkart customer support dataset to classify satisfaction through TF-IDF and SVM. The two datasets are analyzed independently and interpreted together to provide complementary insights into customer satisfaction.

Other studies have explored customer satisfaction through textual data using different machine learning approaches. (Aldunate et al., 2022) used natural language processing and deep learning to identify customer satisfaction drivers from open-ended survey responses. (Wulandari et al., 2025) applied TF-IDF and Support Vector Machine-based classification to analyze aspects of service quality and sentiment in Shopee application reviews. Their study provides a relevant methodological comparison for the present research, which uses TF-IDF and a linear SVM to classify customer satisfaction labels from customer support remarks.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

METHOD

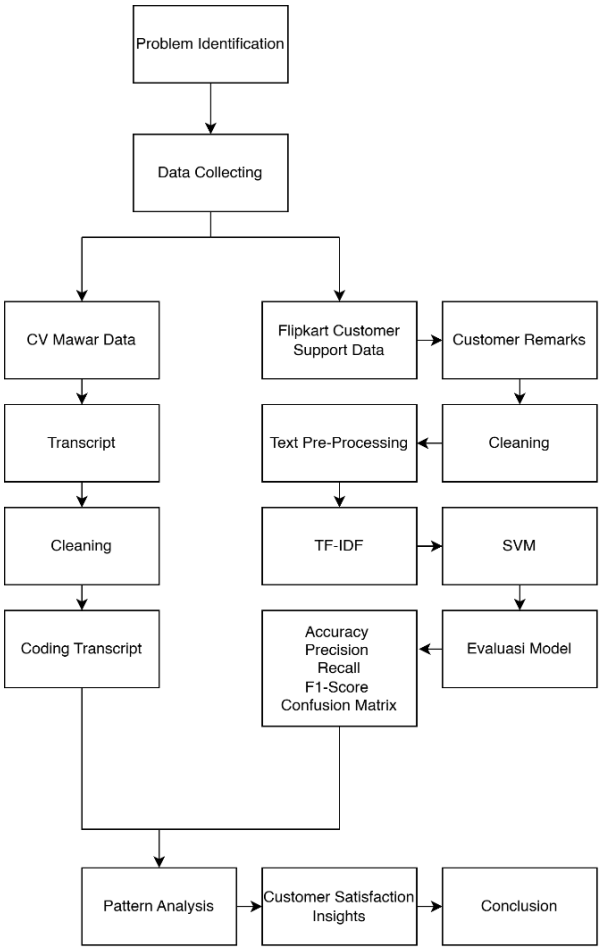


Figure 1. Research Workflow for the Two Independent Analytical Streams

Figure 1 summarizes the workflow: CV Mawar transcripts are cleaned and manually coded for descriptive pattern analysis, while the independent Flipkart stream preprocesses customer remarks with TF-IDF, applies a linear SVM, and evaluates classification performance. The datasets are analyzed separately and are not combined to predict satisfaction for CV Mawar.

This study uses two complementary but independent analytical approaches. The primary data consist of 137 chatbot interaction sessions obtained from CV Mawar. The secondary data are obtained from the publicly available Flipkart Customer Support dataset, which contains customer remarks and Customer Satisfaction (CSAT) scores. The transcript analysis follows (Følstad & Taylor, 2021), while the machine learning approach is supported by previous customer satisfaction classification studies (Hadwan et al., 2022; Sumantiawan et al., 2023).

The procedure begins with data cleaning. For CV Mawar, empty and duplicate sessions are removed, and each session is manually coded to identify recurring topics and interaction patterns. The results are summarized using frequencies and percentages. Available service ratings and feedback are examined as supporting evidence, although the limited number of rated sessions restricts further statistical conclusions.

For the Flipkart dataset, records with missing customer remarks or CSAT scores are removed. The text is preprocessed through lowercasing and the removal of URLs, numbers, special characters, and unnecessary spaces. The CSAT scores are grouped into Not Satisfied (1–3) and Satisfied (4–5), based on the original five-point scale. After preprocessing, the final modeling dataset contains 28,153 records. To prevent identical cleaned remarks from appearing in both the training and testing sets, a group-aware stratified splitting procedure is applied using StratifiedGroupKFold with five folds, shuffling, and a random state of 42. The first fold is used as the testing set, resulting in 25,018 training records (88.86%) and 3,135 testing records (11.14%). The resulting split ratio differs from a conventional 80:20 split because identical cleaned remarks are kept within the same group. Model hyperparameters are selected through three-fold grouped cross-validation on the training data, using macro F1-score as the optimization metric.

(Madyatmadja et al., 2023) reported an accuracy of 89.2% using a linear Support Vector Machine (SVM) and discussed 85% as a practical minimum accuracy benchmark for alternative classifiers in their specific dataset. In this

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

study, 85% is considered a reference benchmark rather than a universal acceptance criterion. Model performance is also evaluated using precision, recall, F1-score, and a confusion matrix to provide a more comprehensive assessment of classification performance (Hand & Anagnostopoulos, 2023).

RESULT

This section presents the findings of the two analytical approaches described in the Method. The results of the CV Mawar transcript analysis are presented first, followed by the results of the Flipkart text classification.

CV Mawar Data Preparation and Service Evaluation

The CV Mawar workbook contained 137 chatbot sessions. Ten sessions had empty transcripts and were removed, and no duplicate session identifiers were found among the remaining records. As a result, 127 sessions were retained and coded. The visitor name column was removed from the analysis dataset because it was not required for the research. Table 1 summarizes the data preparation and the available service evaluation information.

Table1. Data Preparation and Service Evaluation Information of the CV Mawar Dataset

Item	Count
Source sessions	137
Sessions with empty transcripts (removed)	10
Usable and coded sessions	127
Sessions with a genuine service rating in the source	4
Usable sessions with a genuine service rating	3
Sessions with written service feedback	0
Sessions transferred to the seller	30
Sessions not transferred to the seller	107

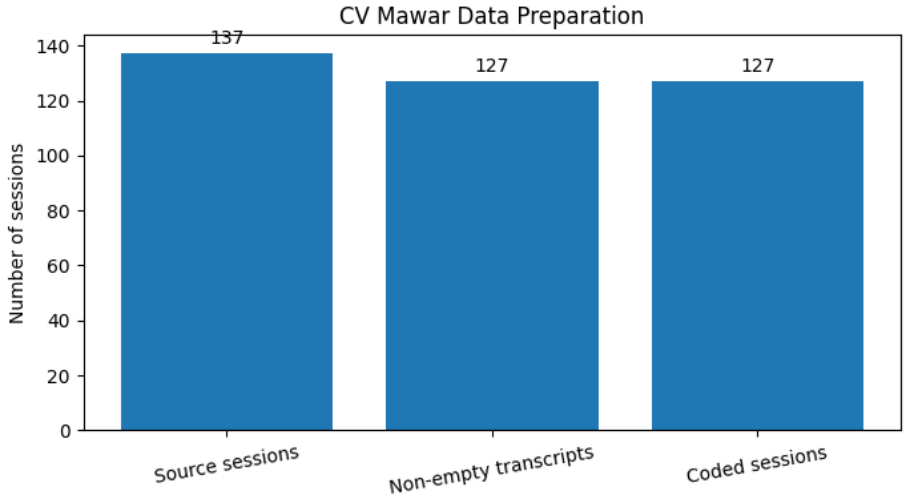


Figure 2. CV Mawar Session Counts Before and After Transcript Cleaning

Table 1 summarizes both the cleaning process and the availability of service-evaluation evidence. Removing 10 sessions with empty transcripts reduced the CV Mawar dataset from 137 source sessions to 127 usable sessions, and no duplicate session identifiers were found. Figure 2 visualizes these counts across the source sessions, non-empty transcripts, and coded sessions. Only four source sessions contained a genuine service rating, of which three had usable transcripts; all three were rated “Baik” (“Good”), and no usable written feedback was available. In addition, 30 sessions (21.9%) were transferred to the seller, and each had a recorded transfer reason. These figures describe dataset completeness and workflow outcomes; the limited ratings and absence of usable written feedback do not support conclusions about overall customer satisfaction.

Recurring Interaction Patterns in CV Mawar

The manual coding assigned each of the 127 sessions to one primary transcript category, producing ten categories. Table 2 presents the frequency distribution across these categories.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

Table 2. Distribution of Primary Transcript Categories in CV Mawar Sessions

Transcript Category	Sessions	Percentage (%)
Order Shipping/Delivery Estimate	25	19.69
Product Information/Usage	25	19.69
No Clear Request/Automated Message	20	15.75
Product Availability/Stock	19	14.96
Order Processing/Status Request	15	11.81
Store Location/Contacting the Seller	9	7.09
Order Complaint/Problem	8	6.30
Payment/Transaction Method	3	2.36
Cancellation/Return/Exchange	2	1.57
Message Outside the Service Context	1	0.79
Total	127	100.00

Table 2 shows that delivery-estimate inquiries and product information/usage were tied as the most frequent topics, with 25 sessions each (19.69%). Product availability/stock was next (19 sessions; 14.96%), while 20 sessions (15.75%) contained no clear request or only automated messages. Together, the leading categories indicate that customers often sought practical information about products, stock, and delivery; the no-clear-request group also shows that a notable share of records did not reveal an identifiable customer need.

Beyond the leading categories in Table 2, order processing/status requests accounted for 15 sessions (11.81%). The remaining categories were each below 10%: store location or contacting the seller (7.09%), order complaints (6.30%), payment methods (2.36%), cancellation, return, or exchange (1.57%), and one message outside the service context (0.79%). Taken together, product information and availability inquiries accounted for 44 sessions (34.65%), while order-related categories (shipping, processing, complaints, and cancellation) accounted for 50 sessions (39.37%).

Because only three coded sessions had a genuine rating, the cross-tabulation of categories and ratings was descriptive only. Two rated sessions belonged to the product information/usage category and one to the product availability/stock category, all rated “Baik”. These counts are too small to support any statistical comparison between categories and satisfaction.

Flipkart Data Preparation and Label Distribution

The Flipkart Customer Support dataset contained 85,907 records. The customer remarks field was empty in 57,165 records (66.54%), while the CSAT score was complete for all records. After removing records with missing remarks, 28,742 records remained, and 589 more were removed because their remarks became empty after text preprocessing. The final modeling dataset therefore contained 28,153 records. The original CSAT scores ranged from 1 to 5, and the grouped labels are shown in Table 3.

Table 3. Label Distribution and Data Split of the Flipkart Dataset

Item	Not Satisfied (CSAT 1–3)	Satisfied (CSAT 4–5)	Total
Records after removing missing remarks, before text preprocessing	6,472 (22.52%)	22,270 (77.48%)	28,742
Final modeling dataset after preprocessing	6,371 (22.63%)	21,782 (77.37%)	28,153
Training data	5,391 (21.55%)	19,627 (78.45%)	25,018
Testing data	980 (31,26%)	2.155 (68,74%)	3.135

The dataset remained imbalanced toward the Satisfied class. After preprocessing, the final modeling dataset contained 28,153 records. The group-aware split produced 25,018 training records (88.86%) and 3,135 testing records (11.14%). No identical cleaned remarks were shared between the two sets. However, the class distribution differed between training and testing. The Not Satisfied class accounted for 21.55% of the training data and 31.26% of the testing data, while the Satisfied class accounted for 78.45% and 68.74%, respectively. This difference should be considered when interpreting the model's classification performance.

SVM Classification Performance

The grid search selected C = 0.1 as the best parameter, with a macro F1-score of 0.7714 in grouped cross-validation on the training data. On the testing data, the model reached an accuracy of 79.62%, compared with 68.74% for the majority-class baseline and the 85% reference benchmark. Table 4 presents the overall evaluation metrics, and Table 5 presents the per-class results.

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

Table 4. Overall Performance of the TF-IDF and SVM Model on the Testing Data

Metric	Value
Majority baseline accuracy	0.6874
SVM accuracy	0.7962
Balanced accuracy	0.8192
Macro F1-score	0.7831
Weighted precision	0.8362
Weighted recall	0.7962
Weighted F1-score	0.8031
Reference accuracy benchmark	0.8500

Table 5. Classification Report per Class

Class	Precision	Recall	F1-score	Support
Not Satisfied	0.62	0.88	0.73	980
Satisfied	0.93	0.76	0.84	2,155
Macro average	0.78	0.82	0.78	3,135
Weighted average	0.84	0.80	0.80	3,135

The model detected most Not Satisfied records (recall 0.88) but with lower precision (0.62). For the Satisfied class, precision was high (0.93) and recall was lower (0.76). Figure 3 shows the confusion matrix.

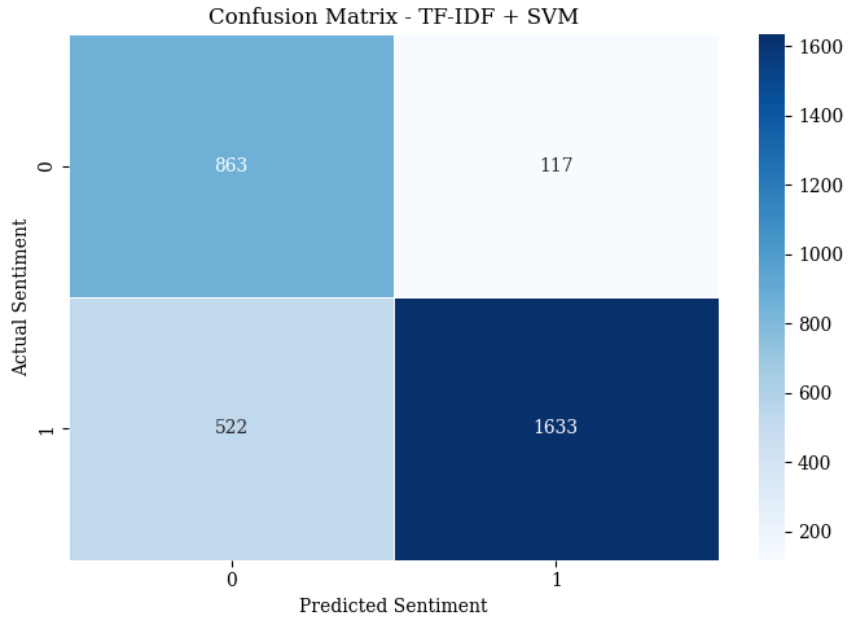


Figure 3. Confusion Matrix of the TF-IDF and SVM Model on the Testing Data (n = 3,135)

Of the 980 Not Satisfied records, 863 were classified correctly and 117 were classified as Satisfied. Of the 2,155 Satisfied records, 1,633 were classified correctly and 522 were classified as Not Satisfied. In total, 639 testing records were misclassified, and most errors (522) were Satisfied records predicted as Not Satisfied.

Terms Associated with Each Class

The coefficients of the linear SVM were examined to identify the terms with the strongest weights. Table 6 shows the terms with the highest weights for each class.

Table 6. Terms with the Strongest Weights for Each Class

Class	Terms with the strongest weights (coefficient)
Satisfied	great (1.218), not bad (1.152), best (1.143), thanks for (1.136), good service (1.113), hope (1.090), thankyou (0.965), thank (0.953)
Not Satisfied	not (−2.254), worst (−1.690), bad (−1.414), didn (−1.256), poor (−1.250), pathetic (−1.223), bed (−1.073), doesn (−1.003)

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

Table 6 shows that the SVM associated the Satisfied class with positive evaluations and expressions of gratitude, including “great,” “best,” and “good service.” For the Not Satisfied class, negative evaluations such as “worst,” “bad,” and “poor” received strong weights; “not” had the largest absolute coefficient (−2.254), highlighting the importance of negation in the learned patterns. The fragments “didn” and “doesn” remained after apostrophes were removed during preprocessing, indicating that tokenization could be improved to preserve contractions and their intended meaning. These coefficients represent statistical associations learned by the model, not causal factors influencing customer satisfaction.

.DISCUSSION

The findings provide two complementary perspectives on chatbot-based customer service, customer interaction patterns in CV Mawar and the application of text classification to customer satisfaction analysis using the Flipkart dataset. Although the datasets were analyzed independently, their findings help explain the value and limitations of transcript analysis and machine learning in customer service evaluation.

The CV Mawar analysis shows that customers primarily used the chatbot to inquire about delivery estimates, product information, and product availability. Delivery estimates and product information or usage were the most frequent categories, each accounting for 19.69% of the 127 coded sessions. Product availability represented 14.96%, while order processing or status requests accounted for 11.81%. These findings suggest that customers frequently rely on chatbot services to obtain practical information during the purchasing process. Therefore, improving the clarity of product descriptions, stock availability, delivery estimates, and order-status information may be beneficial for CV Mawar's customer service. This is consistent with previous research emphasizing the importance of conversational quality, responsiveness, and service recovery in chatbot-based customer experiences (Hsu & Lin, 2023; Jiang et al., 2022b).

Another finding is that 20 sessions (15.75%) contained no clear request or consisted of automated messages. This indicates that some recorded interactions provided insufficient information to identify a specific customer need. Examining such conversations may help improve the quality of chatbot records and the identification of customer requests. Consistent with (Følstad & Taylor, 2021), analyzing actual chatbot transcripts can provide insights into interaction patterns that may not be captured through surveys alone. However, the frequency of a particular topic does not necessarily indicate dissatisfaction or poor service quality; additional evidence about response quality and interaction outcomes would be required.

The relationship between interaction patterns and customer satisfaction remains uncertain in the CV Mawar dataset. Only three usable sessions contained genuine service ratings, and all were rated “Baik.” Furthermore, no usable written service feedback was available. Consequently, the available data were insufficient to determine whether specific interaction categories were associated with higher or lower satisfaction. Although 30 sessions were transferred to the seller, this information alone does not demonstrate chatbot failure or customer dissatisfaction, because the final outcomes of those interactions were not established. These findings should therefore be interpreted as descriptive evidence of customer needs rather than direct measurements of satisfaction.

The testing set represented 11.14% of the final modeling dataset because the group-aware split kept identical cleaned remarks together. Furthermore, the proportion of Not Satisfied records was higher in the testing set than in the training set. These characteristics should be considered when interpreting the evaluation metrics and comparing the model's performance with other studies.

The separate Flipkart classification experiment demonstrates the potential of TF-IDF and a linear Support Vector Machine (SVM) for classifying customer remarks. The model achieved 79.62% accuracy, exceeding the majority-class baseline of 68.74%. It also obtained a balanced accuracy of 81.92% and a macro F1-score of 78.31%. However, its accuracy remained below the study's 85% reference benchmark. This comparison should be interpreted cautiously because model performance depends on the dataset, class definitions, preprocessing procedures, and evaluation design. The benchmark is therefore a reference point rather than a universal standard for classification quality.

The class-level results reveal different performance characteristics. For the Not Satisfied class, recall was 0.88, while precision was 0.62. This means the model identified a large proportion of not-satisfied records but also incorrectly assigned this label to some satisfied records. For the Satisfied class, precision reached 0.93, while recall was 0.76. The confusion matrix showed that 863 of 980 Not Satisfied records and 1,633 of 2,155 Satisfied records were correctly classified. These results demonstrate why accuracy alone is insufficient to evaluate the model. In customer service applications, identifying dissatisfied customers is important, but excessive false positives may also lead to unnecessary follow-up. The model may therefore support the initial screening of customer remarks, although its predictions require careful interpretation.

The terms with the strongest SVM weights also illustrate the textual patterns learned by the model. Positive expressions such as “great,” “best,” and “good service” were associated with the Satisfied class, while terms such as “worst,” “bad,” and “poor” were associated with the Not Satisfied class. These associations are consistent with the evaluative language commonly found in customer feedback. Nevertheless, model weights indicate statistical associations rather than causal factors influencing satisfaction. The remaining fragments “didn” and “doesn” also

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

suggest that text preprocessing could be improved to preserve meaningful expressions and negation.

The main contribution of this study is the combination of real-world chatbot transcript analysis and a separate machine-learning classification experiment. The two approaches provide complementary insights without assuming that a model developed using Flipkart data can directly predict satisfaction among CV Mawar customers. Several limitations should be acknowledged. The CV Mawar analysis relies on manual coding and contains very few genuine satisfaction ratings, while the Flipkart model uses a different dataset with different characteristics. Future research should improve the availability of customer ratings and written feedback, validate transcript coding through multiple annotators, and refine text preprocessing. A model intended for CV Mawar should be developed and evaluated using sufficient, reliable satisfaction-labeled data from that service.

CONCLUSION

This study examined customer interaction patterns in the CV Mawar chatbot service and evaluated TF-IDF with a linear Support Vector Machine for customer satisfaction classification using a separate Flipkart customer support dataset. Of the 137 CV Mawar source sessions, 127 contained usable transcripts and were manually coded. Delivery estimates and product information or usage were the most frequent interaction categories, each accounting for 19.69% of the coded sessions. These findings suggest that product information, stock availability, delivery estimates, and order-status communication are important areas to consider when improving chatbot-supported customer service.

The Flipkart classification model achieved 79.62% accuracy, 81.92% balanced accuracy, and a macro F1-score of 78.31%, outperforming the majority-class baseline. The model achieved relatively high recall for the Not Satisfied class but lower precision, indicating that some satisfied records were incorrectly classified as not satisfied. The findings demonstrate the potential of text classification to support customer feedback analysis, while also showing the importance of evaluating performance across multiple metrics.

The study contributes a complementary approach to understanding chatbot interactions and customer satisfaction classification. However, the limited satisfaction ratings in the CV Mawar dataset prevent reliable conclusions about the relationship between interaction patterns and customer satisfaction. In addition, the Flipkart model cannot be assumed to generalize to CV Mawar because the datasets were analyzed independently. Future research should collect more complete customer satisfaction ratings and written feedback, improve the reliability of transcript coding, and evaluate classification models using data specific to CV Mawar. These improvements would provide a stronger basis for assessing chatbot performance and supporting data-driven customer service improvements.

REFERENCES

Aldunate, Á., Maldonado, S., Vairetti, C., & Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *Expert Systems with Applications*, *209*, 118309. https://doi.org/10.1016/J.ESWA.2022.118309

Følstad, A., & Taylor, C. (2021). Investigating the user experience of customer service chatbot interaction: a framework for qualitative analysis of chatbot dialogues. *Quality and User Experience*, *6*(1). https://doi.org/10.1007/s41233-021-00046-5

Hadwan, M., Al-Sarem, M., Saeed, F., & Al-Hagery, M. A. (2022). An Improved Sentiment Classification Approach for Measuring User Satisfaction toward Governmental Services' Mobile Apps Using Machine Learning Methods with Feature Engineering and SMOTE Technique. *Applied Sciences (Switzerland)*, *12*(11). https://doi.org/10.3390/app12115547

Hand, D. J., & Anagnostopoulos, C. (2023). Notes on the H-measure of classifier performance. *Advances in Data Analysis and Classification*, *17*(1), 109–124. https://doi.org/10.1007/s11634-021-00490-3

Hsu, C. L., & Lin, J. C. C. (2023). Understanding the user satisfaction and loyalty of customer service chatbots. *Journal of Retailing and Consumer Services*, *71*, 103211. https://doi.org/10.1016/J.JRETCONSER.2022.103211

Jenneboer, L., Herrando, C., & Constantinides, E. (2022). The Impact of Chatbots on Customer Loyalty: A Systematic Literature Review. In *Journal of Theoretical and Applied Electronic Commerce Research* (Vol. 17, Number 1, pp. 212–229). MDPI. https://doi.org/10.3390/jtaer17010011

Jiang, H., Cheng, Y., Yang, J., & Gao, S. (2022a). AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior*, *134*, 107329. https://doi.org/10.1016/J.CHB.2022.107329

Madyatmadja, E. D., Sianipar, C. P. M., Wijaya, C., & Sembiring, D. J. M. (2023). Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost. *Informatics*, *10*(4). https://doi.org/10.3390/informatics10040084

Nugroho, R. A., Warsito, B., & Hakim, A. R. (2023). Performance Comparison of Support Vector Machine and Random Forest Using Delta TF-IDF on Sentiment Classification of Tiket.com Reviews. *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 632–637. https://doi.org/10.1109/ISITIA59021.2023.10221142

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 6, **Number** 4, November 2026
https://doi.org/10.47709/brilliance.v6i4.9810

Ruan, Y., & Mezei, J. (2022). When do AI chatbots lead to higher customer satisfaction than human frontline employees in online shopping assistance? Considering product attribute type. *Journal of Retailing and Consumer Services*, *68*, 103059. https://doi.org/10.1016/J.JRETCONSER.2022.103059

Sari, H., & Adinda, R. (2023). Examining Customer Experience in Using a Chatbot. *International Journal of Asian Business and Information Management*, *14*(1). https://doi.org/10.4018/IJABIM.322438

Sumantiawan, D. I., Suseno, J. E., & Syafei, W. A. (2023). Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction. *J. Sistem Info. Bisnis*, *13*(1), 1–9. https://doi.org/10.21456/vol13iss1pp1-9

Wulandari, D. A., Bachtiar, F. A., & Indriati, I. (2025). Aspect Based Sentiment Analysis on Shopee Application Reviews Using Support Vector Machine. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *15*(02), 99. https://doi.org/10.24843/lkjiti.2024.v15.i02.p03

Xu, Y., Niu, N., & Zhao, Z. (2023). Dissecting the mixed effects of human-customer service chatbot interaction on customer satisfaction: An explanation from temporal and conversational cues. *Journal of Retailing and Consumer Services*, *74*, 103417. https://doi.org/10.1016/J.JRETCONSER.2023.103417

Zhu, Y., Zhang, J., & Liang, J. (2023). Concrete or abstract: How chatbot response styles influence customer satisfaction. *Electronic Commerce Research and Applications*, *62*, 101317. https://doi.org/10.1016/J.ELERAP.2023.101317

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.