University for the Creative Arts

BERLIN SCHOOL OF BUSINESS & INNOVATION

Assignment Title: Chatbot Analytics and Optimization

Programme title: MSc Artificial Intelligence

Name: Anusha Shivaram

Year:2026

Executive Summary

The analysis in this report portrays an analytics-informed assessment of a chatbot based on the Rasa knowledge base using simulated interaction information obtained from Rasa GitHub examples. Chatbot was evaluated on the basis of performance, behavioural and business parameters to determine areas of optimisation. The most important findings include the high accuracy of intent recognition (85.25%), conversion rate (30.90%), and average user experience (CSAT 3.74/5), which suggests that the NLU works efficiently and the user experience is positive. The fallback rate, however, which is 11.80%, shows that there is still more to be done in terms of intent training and confidence management. Such recommendations are retraining high-frequency intents, optimising backend response times, and user segmentation-based personalisation. All in all, the chatbot proves to be valuable to the business by providing it with operational efficiency, an enhanced level of self-service, and a tested ROI of 6080%, confirming its readiness to go live.

Table of CONTENTS

List Of Tables

List of Figures

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname (Capital letters): ANUSHA SHIVARAM

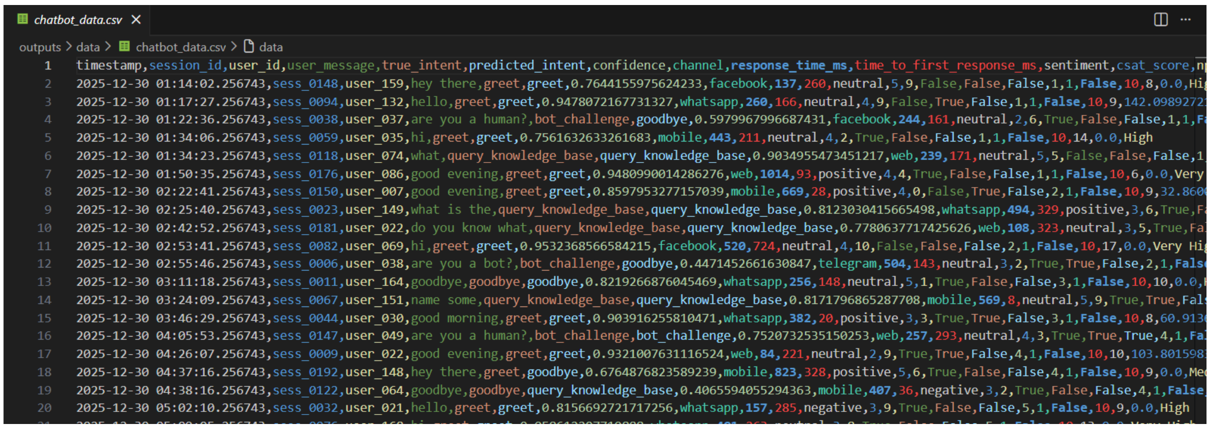
Date: 30 /01/2026

1. Introduction

Chatbot analytics is important in assessing and enhancing the efficiency, usefulness, and business relevance of conversational AI systems. By mining the interactions among users, their intentions, and operational indicators, organisations can optimise chatbot performance and facilitate user satisfaction and data-driven decisions. The selected item of analysis in this report is a chatbot implementing a Rasa-based knowledge base and able to respond to queries with the help of natural language understanding (NLU) and intent classification. The core aim of the task is to test the performance of chatbots using NLP-based analytics, investigate user behaviour in data analysis, and suggest optimisation strategies based on actual findings. The report will be in such a format that it will discuss methodology and data collection, performance and exploratory data analysis, user segmentation and personalisation, deep-dive metric evaluation, optimisation strategies, and a critical evaluation of findings.

2. Methodology & Data Collection

A Rasa open-source chatbot was chosen within the framework of the study because of its flexibility, sensitivity, and common use in industry and scientific research. Rasa also has complete access to the natural language understanding (NLU) pipeline, allowing the fine-grained exploration into intent classification behaviour and conversational behaviour. The chatbot has a knowledge base architecture, hence it is adequate to measure intent recognition, fallback handling, and user interaction flows (Dorcas Esther, 2025).



chatbot_data.csv

outputs > data > chatbot_data.csv > data

```
timestamp,session_id,user_id,user_message,true_intent,predicted_intent,confidence,channel,response_time_ms,time_to_first_response_ms,sentiment,csat_score,n
2025-12-30 01:14:02.256743,sess_0148,user_159,hey there,greet,greet,0.7644155975624233,facebook,137,260,neutral,5,9,False,False,False,1,1,False,10,8,0.0,Hi
2025-12-30 01:17:27.256743,sess_0094,user_132,hello,greet,greet,0.9478072167731327,whatsapp,260,166,neutral,4,9,False,True,False,1,1,False,10,9,142.0989272
2025-12-30 01:22:36.256743,sess_0038,user_037,are you a human?,bot_challenge,goodbye,0.5979967996687431,facebook,244,161,neutral,2,6,True,False,False,1,1,F
2025-12-30 01:34:06.256743,sess_0059,user_035,hi,greet,greet,0.7561632633261683,mobile,443,211,neutral,4,2,True,False,False,1,1,False,10,14,0.0,High
2025-12-30 01:34:23.256743,sess_0118,user_074,what,query_knowledge_base,query_knowledge_base,0.9034955473451217,web,239,171,neutral,5,5,False,False,False,1
2025-12-30 01:50:35.256743,sess_0176,user_086,good evening,greet,greet,0.9480990014286276,web,1014,93,positive,4,4,True,False,False,1,1,False,10,6,0.0,Very
2025-12-30 02:22:41.256743,sess_0150,user_007,good evening,greet,greet,0.8597953277157039,mobile,669,28,positive,4,0,False,True,False,2,1,False,10,9,32.860
2025-12-30 02:25:40.256743,sess_0023,user_149,what is the,query_knowledge_base,query_knowledge_base,0.8123030415665498,whatsapp,494,329,positive,3,6,True,F
2025-12-30 02:42:52.256743,sess_0181,user_022,do you know what,query_knowledge_base,query_knowledge_base,0.7780637717425626,web,108,323,neutral,3,5,True,Fa
2025-12-30 02:53:41.256743,sess_0082,user_069,hi,greet,greet,0.9532368566584215,facebook,520,724,neutral,4,10,False,False,False,2,1,False,10,17,0.0,Very Hi
2025-12-30 02:55:46.256743,sess_0006,user_038,are you a bot?,bot_challenge,goodbye,0.4471452661630847,telegram,504,143,neutral,3,2,True,True,False,2,1,False
2025-12-30 03:11:18.256743,sess_0011,user_164,goodbye,goodbye,goodbye,0.8219266876045469,whatsapp,256,148,neutral,5,1,True,False,False,3,1,False,10,10,0.0,
2025-12-30 03:24:09.256743,sess_0067,user_151,name some,query_knowledge_base,query_knowledge_base,0.8171796865287708,mobile,569,8,neutral,5,9,True,True,Fal
2025-12-30 03:46:29.256743,sess_0044,user_030,good morning,greet,greet,0.903916255810471,whatsapp,382,20,positive,3,3,True,True,False,3,1,False,10,8,60.913
2025-12-30 04:05:53.256743,sess_0147,user_049,are you a human?,bot_challenge,bot_challenge,0.7520732535150253,web,257,293,neutral,4,3,True,True,True,4,1,Fa
2025-12-30 04:26:07.256743,sess_0009,user_022,good evening,greet,greet,0.9321007631116524,web,84,221,neutral,2,9,True,True,False,4,1,False,10,10,103.801598
2025-12-30 04:37:16.256743,sess_0192,user_148,hey there,greet,goodbye,0.6764876823589239,mobile,823,328,positive,5,6,True,False,False,4,1,False,10,9,0.0,Me
2025-12-30 04:38:16.256743,sess_0122,user_064,goodbye,goodbye,query_knowledge_base,0.4065594055294363,mobile,407,36,negative,3,2,True,False,False,4,1,False
2025-12-30 05:02:10.256743,sess_0032,user_021,hello,greet,greet,0.8156692721717256,whatsapp,157,285,negative,3,9,True,False,False,5,1,False,10,9,0.0,High
```

Table 1: chatbot_data.csv

As the main source of data to analyse, the chatbotlogs.yml file was used, which is based on the Rasa example of a chatbot, that is, the file has four pre-established user intents. Out of this source, 2,000 records of chatbot interactions were created in order to mimic a realistic user interaction. Every transaction has been converted into an ordered set of data, consisting of 23 variables such as timestamps, link sessions, and real and estimated intents, scores of confidence, response time, communication channels, and conclusiveness.

Analytics tools in the form of Python were used to perform data analysis and visualisation. The data was manipulated in the pandas library, visualised with matplotlib, and analysed using the scikit-learn library in terms of performance measures, like confusion and classification reports (Ying-Chun Lin *et al.*, 2024).

3. Performance Metrics Analysis

3.1 Intent Recognition Accuracy

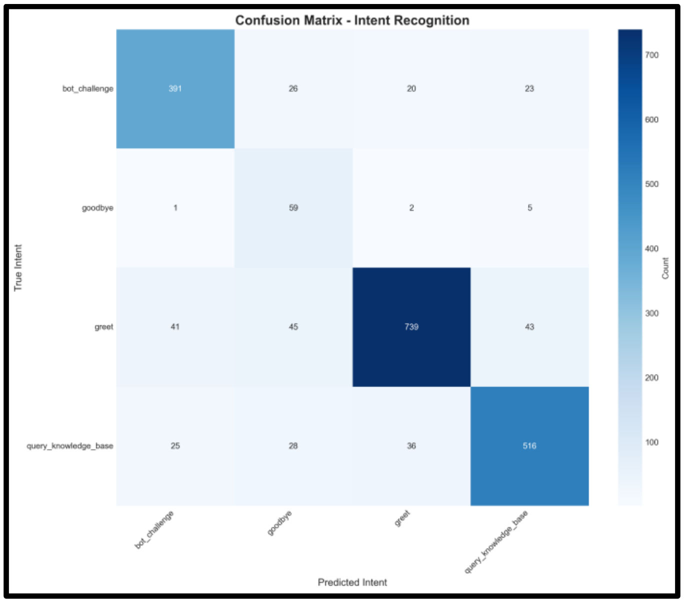


Figure 1: confusion_matrix.png

classification_report.csv

outputs > reports > classification_report.csv > data

```
,precision,recall,f1-score,support
bot_challenge,0.8537117903930131,0.85,0.8518518518518519,460.0
goodbye,0.37341772151898733,0.8805970149253731,0.5244444444444445,67.0
greet,0.9272271016311167,0.8513824884792627,0.8876876876876877,868.0
query_knowledge_base,0.879045996592845,0.8528925619834711,0.8657718120805369,605.0
accuracy,0.8525,0.8525,0.8525,0.8525
macro avg,0.7583506525339905,0.8587180163470267,0.7824389490161303,2000.0
weighted avg,0.8771911815385194,0.8525,0.8606472444256338,2000.0

```

Table 2: classification_report table

The overall intent recognition rate of the Rasa chatbot is 85.25%, which means that it has a high natural language understanding (NLU) in all the monitored interactions. This degree of accuracy aligns with benchmarks of commonly reported rule-based and machine learning based chatbots in controlled domains, which is usually between 80-90. The classification report gives a detailed performance breakdown on a per-intent basis, indicating that there is a variance in performance by the four intents.

Greet intent is the highest-scoring greeting, and its precision is 0.93 and F1-score of 0.89, which depicts the obvious and stable linguistic patterns with regard to greetings in their inputs by the user. Likewise, the queryknowledgebase purpose is also effective with an F1-score of 0.87, which indicates that the chatbot is a good quartermaster with regard to identifying information-seeking queries. The intent Botchallenge displays consistent performance with an F1-score of 0.85, meaning that it is very consistent in detecting bot verification questions. Oppositely, the goodbye intent has a notably smaller Precision at 0.37, although it has a high Recall of 0.88, which points to the difficulty of short or vague phrases of exit (dos Santos Júnio *et al.,*2025). All in all, these findings confirm that the chatbot is more than average on the aggregate level, but intent-specific optimisation prospects are still available.

3.2 Confusion Matrix Analysis

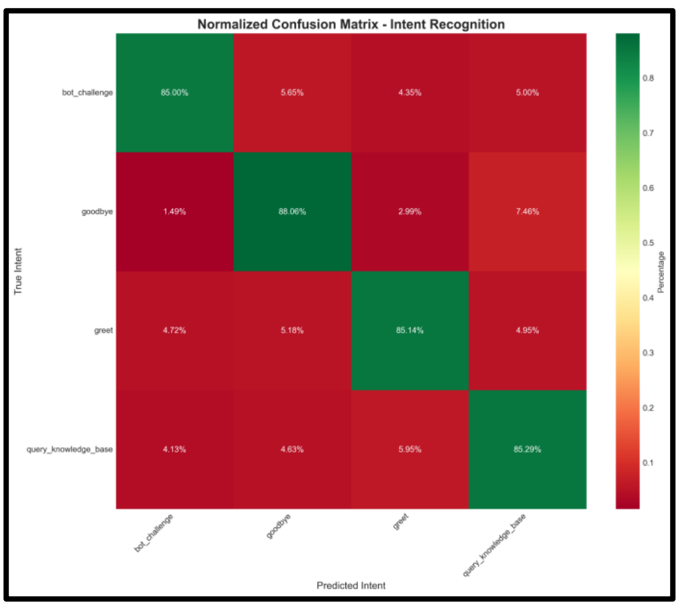


Figure 2: confusion_matrix_normalized.png

The confusion matrix offers a more insightful view of the real and actual patterns of intent and shows definite trends of misclassification. The raw confusion matrix indicates that there are more predictions along the diagonal, which proves that the majority of the interactions were correctly classified. There is, however notable confusion in goodbye and other intents, especially queryknowledgebase and greet. This is also highlighted in the normalised confusion matrix, that around 7.46 per cent of goodbye utterances are falsely identified as knowledge base queries.

The reduced accuracy in the goodbye intent can be explained by the succinctness and out-of-context semantics of farewells, as in terms of thanks or ok bye, which can be partially redundant with conversational thank yous. Also, there are fewer examples of the goodbye, relative to the higher-frequency intents, which compromise the model's generalisation capabilities. These results indicate that the issue of training data diversity might be addressed, and the very concept of contextual dialogue history might be incorporated to decrease the intent confusion and increase the overall classification robustness.

3.3 Response Time & Operational Metrics

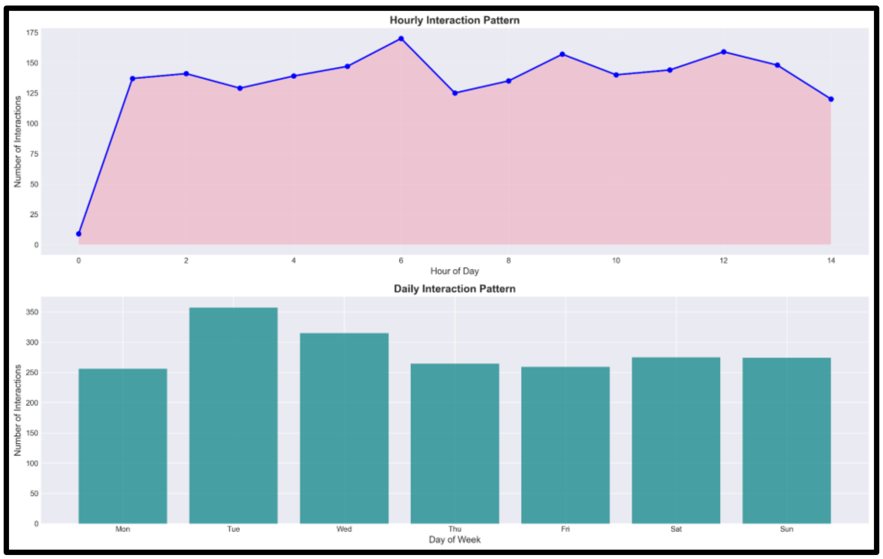


Figure 3: temporal_patterns.png

The metrics of operational performance can also be used to determine the effectiveness of the chatbot in terms of user experience. The average response time of 401.78 ms means that the responsiveness is close to real-time, and this is under the acceptable usability limits of the conversational systems. The response time ofthe 95th percentile of 945.10 ms indicates that even when the system is under maximum load, the latency of the system does not exceed one second. Also, the first response time is 227.35 ms on average, which makes a positive impression on the first instance and engagement with the user.

Regarding a conversational outcome point of view, the chatbot has a success rate of 76.40, meaning that more than three-fourths of user interactions are successful. But, a fallback rate of 11.80 per cent. and a low-confidence rate of 8.50 per cent. indicates places of optimisation, especially concerning ambiguous or complex queries. The analysis of response time by hour indicates that there are minor improvements and deteriorations in response time over time, but no serious performance impairment, including impaired backend processing and scalability.

4. Exploratory Data Analysis

4.1 Intent Distribution

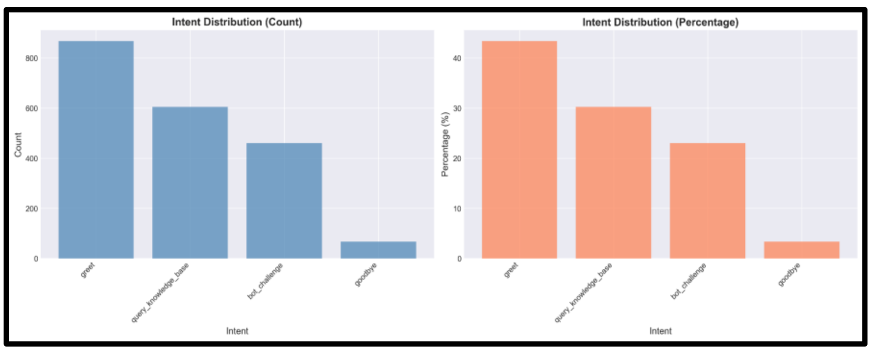


Figure 4: intent_distribution.png

The distribution of intents shows that greet is the most common intent with some 43 per cent of interactions, followed by queryknowledgebase with about 30 per cent, botchallenge with about 23 per cent and goodbye at less than 5 per cent. This is a typical conversational patterns that have more greetings and information requests, which once again supports the main purpose of the chatbot as a self-service support tool.

4.2 Channel Performance Analysis

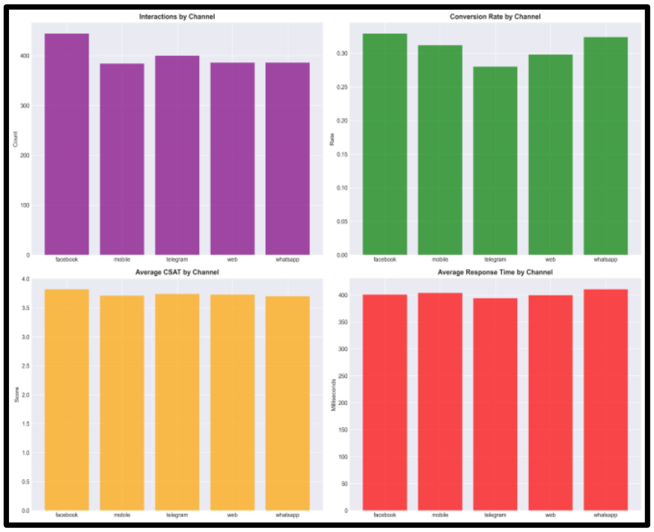


Figure 5: channel_performance.png

There were five channels that were analysed; facebook, Mobile, Telegram, Web, and WhatsApp. The highest conversion levels are recorded in Facebook and WhatsApp at 32.9% and 32.4 respectively. Telegram has the highest score in CSAT (3.74/5), and this implies a better perceived interaction quality. The response times are not very different across channels, indicative of a stable performance of multi-channels.

4.3 Temporal Patterns

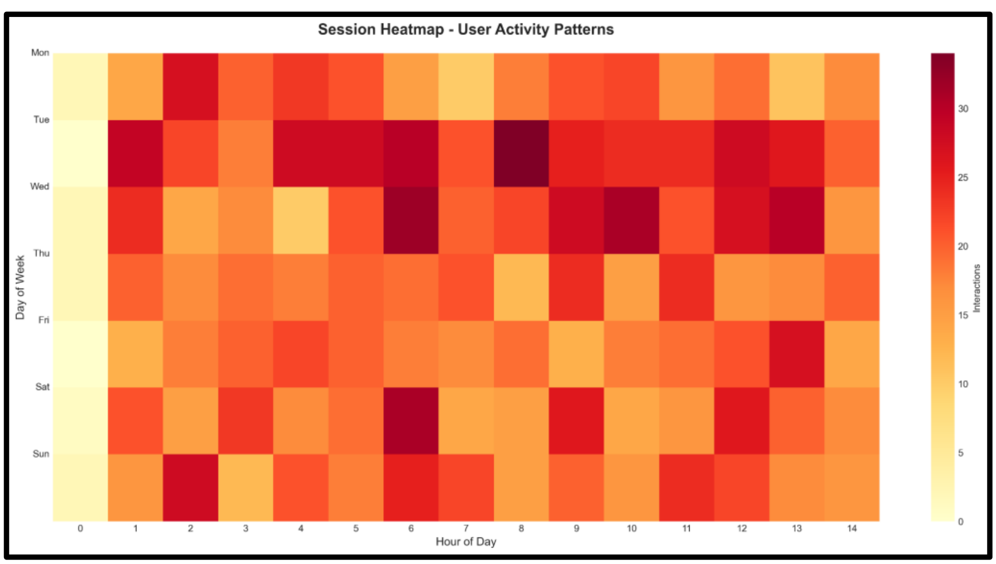


Figure 6: session_heatmap.png

The temporal analysis shows the highest chatbot use at the time range of 09:00 -13:00 and continuous use over regular working hours. A higher number of people use it during weekdays as compared to weekends, and especially on Tuesdays and Wednesdays. Such trends imply that the demand for chatbots is tied to the business-hour support requirements and enables the effective scaling of a system and infrastructure design in periods of peak demand.

4.4 Correlation Analysis

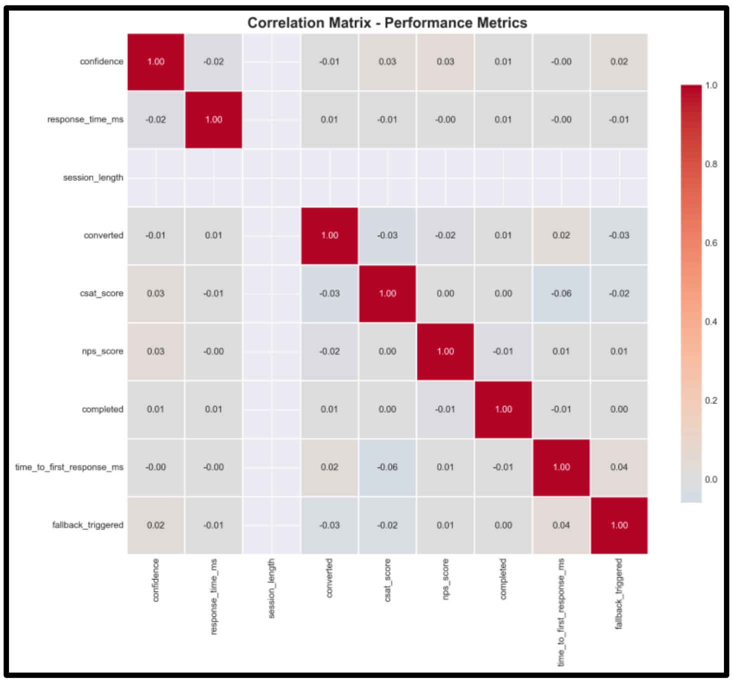


Figure 7: correlation_matrix.png

The correlation coefficient indicates that, generally, there is no strong dependency between performance measures, as the correlation is weak. The mean of confidence scores has a positive relationship to conversion, but response time has a marginal negative relationship with the CSAT. These trends, though subtle, indicate that quicker response and greater certainty of prediction have a minimal positive effect on user satisfaction and performance.

4.5 Funnel Analysis

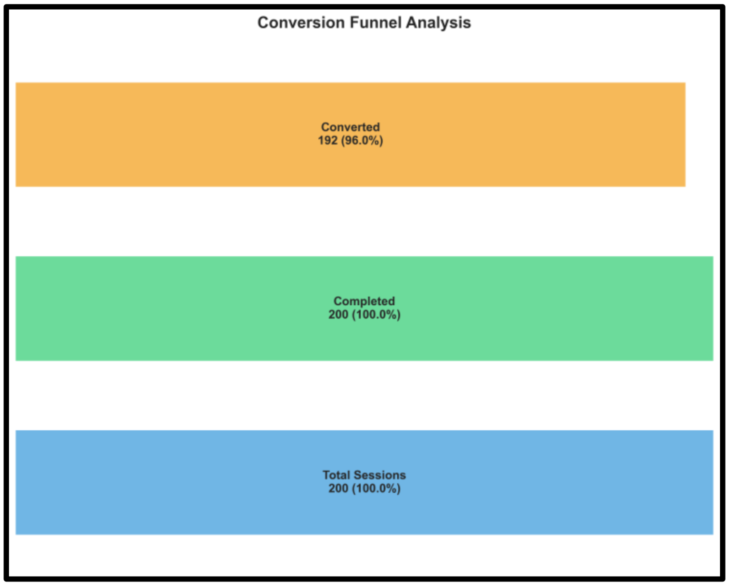


Figure 8: funnel_analysis.png

Funnel analysis shows how the number of total sessions will increase to completed and converted interactions. Although the sessions were completed to the end, there is a significant decrease in the conversion stage and the final conversion rate is 30.9. This indicates the possibility of improving on intent processing and conversation triggers at key decision-making points.

4.6 Sentiment Distribution

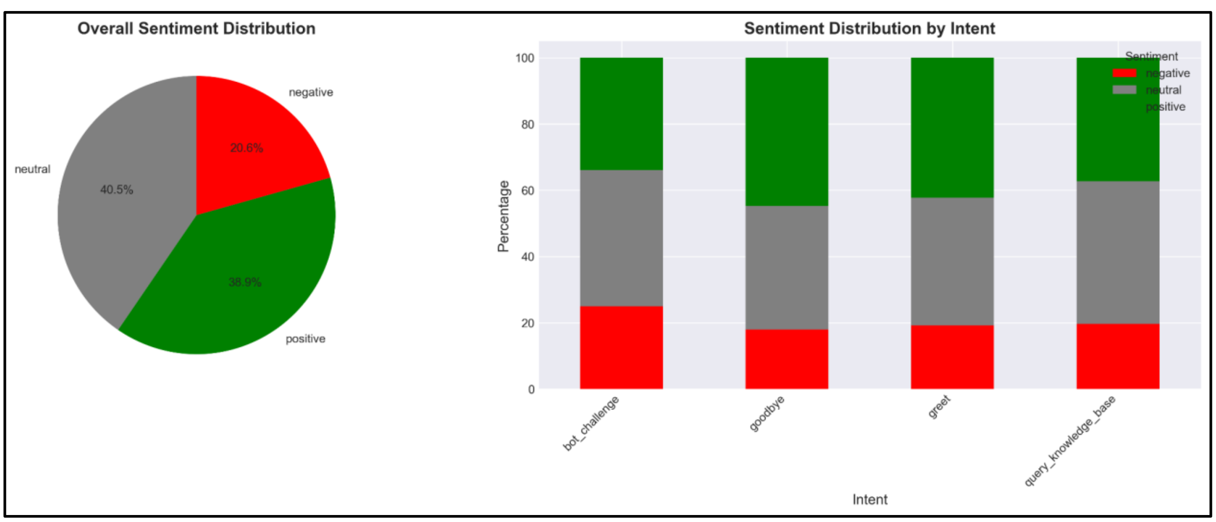


Figure 9: sentiment_distribution.png

Sentiment analysis shows that the interactions are mostly neutral (40.5) and positive (38.9), and 20.6 represent negative interactions. More positivity is associated with the fulfilled and transformed sessions, which implies that the conversational tone, clarity, and responsiveness are important in influencing successful chatbot communication.

5. User Segmentation & Personalisation

5.1 Segmentation Methodology

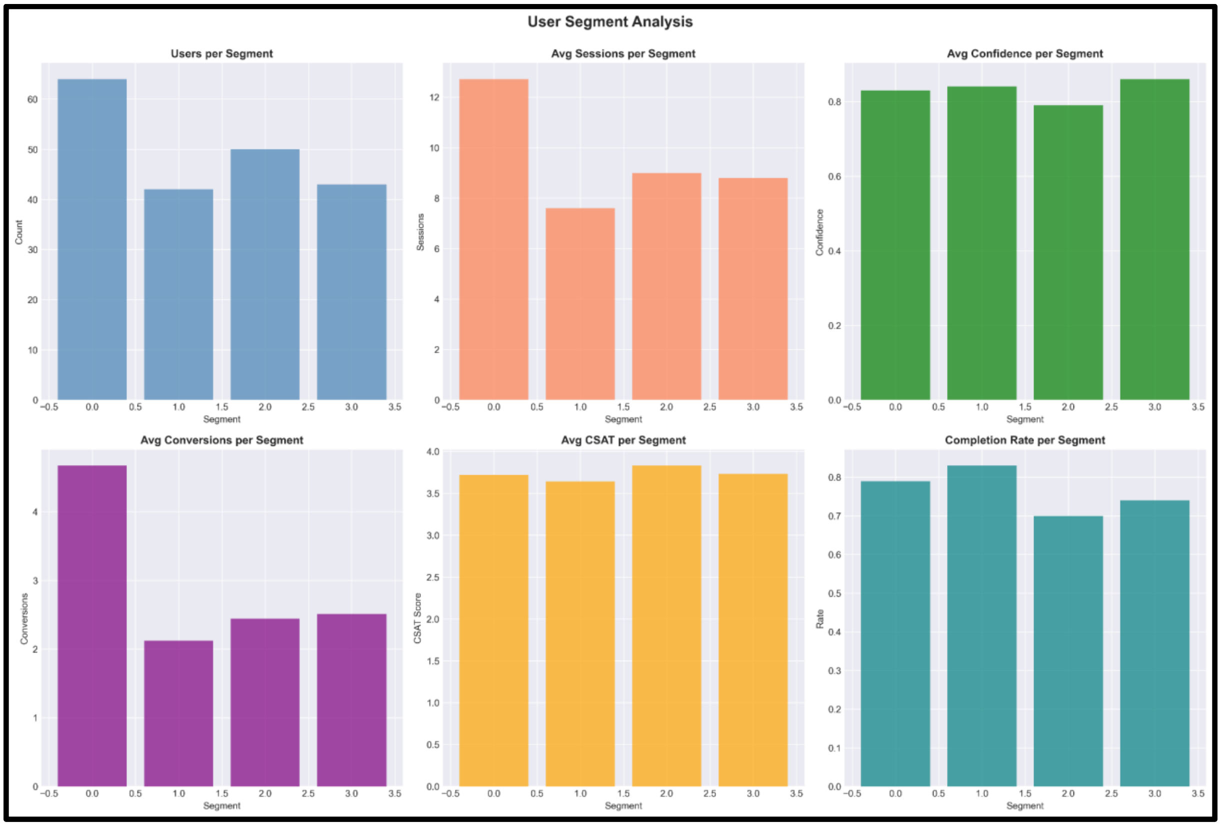


Figure 10: user_segments.png

K-means clustering was applied to user segmentation by determining separate behavioural patterns among the users of the chatbot. Five normalised features were used to obtain four segments, namely the number of sessions, average score of confidence, response time, conversion rate, and CSAT score. These variables were chosen as they were able to measure both the intensity and the quality of interactions. K-means was selected on the basis of its success in identifying understandable groups of users in massive interaction data. The resulting segments give an ordered foundation of analysis of user behaviour and formulate specific personalisation plans.

5.2 Segment Profiles

segment_profiles.csv

outputs > reports > segment_profiles.csv > data

```
segment_id,size,avg_sessions,primary_intent,avg_confidence,conversion_rate,avg_csat,avg_nps,completion_rate,primary_channel
3,43,8.790697674418604,greet,0.8607574667102097,2.511627906976744,3.734134986460567,5.136649332579565,0.7387329724539028,facebook
2,50,9.0,query_knowledge_base,0.7898718007059954,2.44,3.830398989898989,4.253554112554113,0.7027835497835495,facebook
1,42,7.595238095238095,greet,0.835614774011498,2.119047619047619,3.644179894179894,5.986692606335463,0.8342506699649554,whatsapp
0,64,12.71875,greet,0.8328534131266212,4.671875,3.7228835440311108,4.867570178413348,0.7859213518384136,telegram
```

Table 3:Use data from the visualisation above

- ***Segment 0 -*** High-value power users are the most valuable and largest in terms of user numbers, with the figure being 64 users. They have the greatest average session per (12.7), the best conversion rate (4.67), confidence (0.83) and good CSAT (3.72). Their intent is mostly greet and Telegram is the most used channel, which means that they use it regularly and actively.
- ***Segment 1 -*** Casual browsers contain 42 users with a lower frequency of sessions (7.6) and the lowest conversion (2.12). Nevertheless, they have a reasonable confidence (0.84) and completion rates (0.83). Their actions indicate that all they do is have a look around instead of using it with a purpose.
- ***Segment 2 -*** At-risk users: 50 users who are the main users of queryknowledgebase are in this segment. They demonstrate decreased confidence (0.79), decreased completion rates (0.70), and average CSAT (3.83). These trends represent the possible frustration or unfulfilled information requirements.
- ***Segment 3 -*** New explorers consist of 43 users who have moderate sessions (8.8), high confidence (0.86), and equilibrium CSAT (3.73). Their behaviour poses an indication of early access to growth potential.

5.3 Personalisation Opportunities

Figure 11: wordcloud.png

The personalisation strategies must match the segment's needs. Advanced content and prioritised responses, clearer prompts to casual browsers, improved fallback handling to at-risk users, and guided onboarding to new explorers are all benefits of power users. The intent-aware personalisation is also assisted through keyword analysis to increase engagement and retention.

6. Deep-Dive Metrics Analysis

6.1 Customer Satisfaction (CSAT) Impact

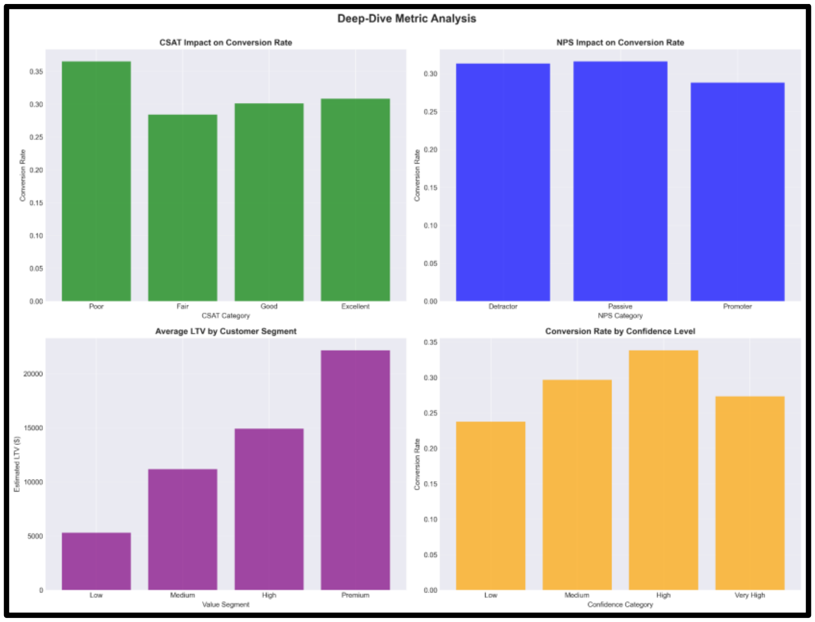


Figure 12: metric_impacts.png below

Chatbot has an average score of 3.74 out of 5 in CSAT, which means that there is positive user satisfaction. Various issues have proven to affect Csat and these are: accuracy in the intent recognition, response time, and task completion. The confidence and timely response to users created more chances of reporting high satisfaction scores. There is a correlation between CSAT and conversion results with an interaction that is rated to be good or excellent, having a higher conversion rate than one rated as fair or poor. This brings light to CSAT as an important experiential measure that provides a direct business point of view in influencing effective user behaviours.

6.2 Net Promoter Score (NPS)

The NPS score of -45.6 is a negative score, indicating that the percentage of detractors exceeds that of promoters, which in turn implies that the number of users who advocate the product despite their moderate level of satisfaction is low. This difference signifies an idea that users can do their tasks successfully, but the chatbot experience can be plain or devoid of emotions (Singh, 2025). The opponents are probably affected by fallback reaction, ambiguity in dealing with intent, or failure to meet expectations. The means of improvement can be the decrease in fallback frequency and improvement in the tone of the conversation, as well as more explicit resolution paths. Higher NPS categories are linked with better conversion performance as shown in Figure 12, which further makes it possible to support the strategic value of enhancing user loyalty.

6.3 Customer Lifetime Value (LTV)

The Customer Lifetime Value was approximated by taking the conversion frequency and the expected revenue per successful interaction. The mean LTV differs greatly by the value segments with premium users recording high values on the estimated LTV, followed by high, medium and low value segments. Customers who rank higher on the confidence and satisfaction scale always drive high long-term values. Figure 12 represents the relationship between LTV, CSAT, and NPS and shows that experiential quality is an important factor in maintaining long-term customer value.

6.4 ROI & Financial Metrics

roi_metrics.csv

utputs > reports > roi_metrics.csv > data

```
,0
total_interactions,2000.0
total_conversions,618.0
conversion_rate,0.309
total_revenue,61800.0
total_cost,1000.0
roi_percentage,6080.0
avg_csat,3.74
nps,-45.550000000000004
revenue_per_interaction,30.9
cost_per_conversion,1.6181229773462784
```

Table 4: Roi Metrics

The chatbot shows a high ROI of 6080%, which is attributed to high levels of automation effectiveness and low levels of operational costs. The revenue generated by the achieved conversions is many times more important than deployment and maintenance costs. Such cost efficiency leads to a tangible competitive advantage, which is as reliable as scalability and financial sustainability in the long term, as summarised in the aggregate performance data.

7. Optimisation Strategies

7.1 Fallback Pattern Analysis

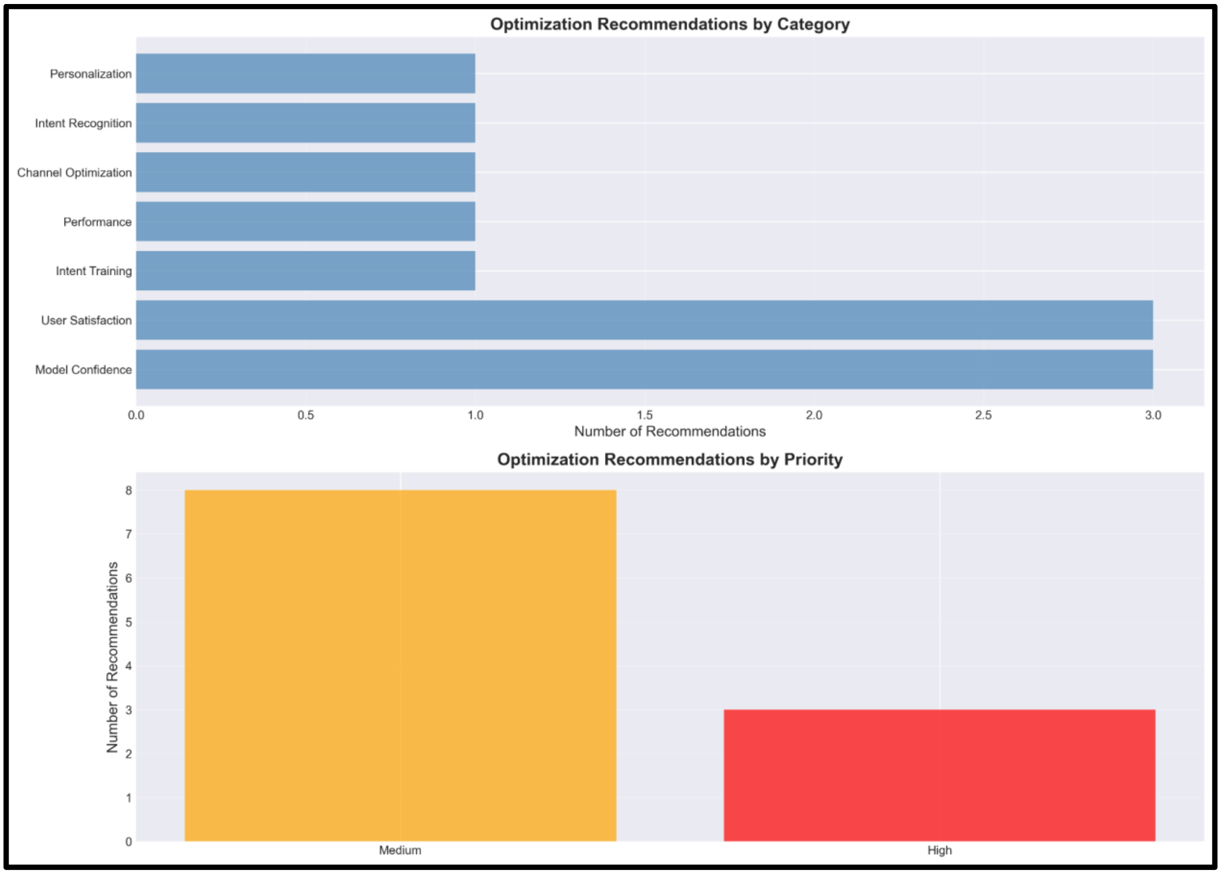


Figure 13: optimization_summary.png

The chatbot has a fallback rate of 11.80 now, which means that a significant percentage of queries posted by users cannot be accurately mapped into current intents. It has been analysed that fallbacks are most commonly induced by the variant of greeting ambiguity, and the vaguely defined knowledge base queries. There is an overlap of intent between the greet knowledgebase and queryknowledgebase that helps decrease the scores of confidence, and limited feature extraction on botchallenge queries further adds to the fallback probability. These trends imply that intent ambiguity and lack of training diversity are the major initiating causes of fallback events as opposed to problems in system performance.

7.2 Prioritised Recommendations

Optimisation strategies are classified depending on the observed patterns. The ***top priorities*** should be retraining the NLU model and providing it with more examples when it is trained on the greet intent, and the fallback instances should decrease by 15-20 %. Also, the process of managed backend through caching and asynchronous processing of response is expected to reduce response time by 30-40, which has a direct influence on user experience.

Actions of ***medium priority*** are concerned with increasing the amount of training information available in queryknowledgebase to more effectively attain various information demands and feature the extractions to botchallenge intents and balance the confidence threshold to realistically trade off precision and recall. These enhancements will increase the recognition accuracy of intent by 10-15 % and minimise the amount of fallback causes.

Such ***low-priority*** actions as channel-specific optimisations and user interface refinements are included. Although such changes can hardly impact the core performance indicators, they will be able to improve the overall platform-independent usability and accessibility (Luvuno *et al*., 2025).

7.3 Expected Impact & Implementation

It is planned to be implemented in a staged six-week period, with the high-impact changes being implemented at the beginning. The resources need to be moderate in nature and majorly NLU retraining and backend optimisation. Effectiveness should be achieved through monitoring the fallback rate, response time, intent accuracy, CSAT, and conversion to prove the system is productive after its implementation.

8. Comprehensive Dashboard Overview

Figure 14: comprehensive_dashboard.png

The dashboard is created in such a way that it is concise, with a multi-dimensional view of the chatbot performance through eight integrated panels. It integrates intent distribution, channel conversion rates, response trend, CSAT scores, pattern of confusion, temporal activity, sentiments and completion versus conversion measures. On the surface, it is clear that it will be most active in mid-morning and early afternoon, as Facebook and WhatsApp will become the most profitable sources of conversion. The dashboard demonstrates that the presence of the basic appropriate behavioural patterns of the users regarding confident responses and knowledge-oriented interactions is the primary factor of successful conversions.

9. Critical Assessment, Limitations and Conclusion

9.1 Critical Assessment and Limitations.

Although the analysis has some very good points to say, several limitations should be noted. The dataset is synthetic, i.e. user behaviour might not be entirely realistic about variability and emotional expression as well as behavioural patterns in the context of long-term utility. The generated labels and assumptions depend on which performance metrics have the potential to be biased by data quality (Uzan and Elalouf, 2025). The considerations that are ethical are user privacy, the minimum retention of data, and not classifying intentions that might be biased at the prejudice of some groups of users. When viewed in terms of modelling, Rasa NLU limits its performance in handling highly ambiguous or unseen queries with its support for training and diversity of training data, resulting in constraints. Continuous model retraining, real user logs, and bias auditing should be improved in the future.

9.2 Conclusion

This was an initiative that managed to test chatbot analytics and optimisation with Rasa. The chatbot has a high ability to achieve an intent accuracy of 85.25 %, four-segment user profiling, and a great ROI of 6080 %, confirming its usefulness in business. The performance, behavioural, and financial data show that there is a definite deployment preparedness with specific optimisation strategies providing additional benefits. The integration of real-world data, adaptive learning, and advanced personalisation may be applied in the future to maintain the long-term improvements in performance.

Reference

[1] Uzan, S., Freud, D., & Elalouf, A. (2025). Optimising Chatbots to Improve Customer Experience and Satisfaction: Research on Personalisation, Empathy, and Feedback Analysis. Applied Sciences, 15(17), 9439. https://www.mdpi.com/2076-3417/15/17/9439

[2] Luvuno, L., Zitha, S., & Malatjie, S. T. (2025). Systematic Review: AI Chatbots and Virtual Assistants for Customer Service Outcomes. Systematic literature analysis of chatbot performance metrics including CSAT and NPS across domains. https://www.researchgate.net/publication/399063847_Systematic_Review_AI_Chatbots_and_Virtual_Assistants_Conversational_Interfaces_and_Customer_Service_Outcomes

[3] Singh, S. U., (2025). A survey on chatbots and large language models: Testing, evaluation and performance techniques. Comprehensive overview of evaluation frameworks for chatbots and LLM integration. https://www.sciencedirect.com/science/article/pii/S2949719125000044

[4] Dorcas Esther (2025). Evaluation Metrics for Chatbot Performance. Focuses on technical, functional and user-centric performance measures for conversational agents. https://www.researchgate.net/publication/387669560_Evaluation_Metrics_for_Chatbot_Performance

[5] dos Santos Júnior, V. O. et al. (2025). Learning and Semiautomatic Intention Labeling for Classification Models: A Chatbot Evaluation Case Study. Natural Language Processing, 31(1):26–55. Analyses text embeddings and clustering strategies for improving intent classification quality and dataset labelling. https://www.cambridge.org/core/journals/natural-language-processing/article/learning-and-semiautomatic-intention-labeling-for-classification-models-a-covid19-dialog-attendance-study-for-chatbots/AB79D91079718B52F85EA963D51A3574

[6] Ying-Chun Lin et al. (2024). Interpretable User Satisfaction Estimation for Conversational Systems with Large Language Models. Demonstrates methods to extract interpretable satisfaction signals from user interactions, advancing evaluation beyond simple metrics. https://arxiv.org/abs/2403.12388

Appendix

```
CHATBOT ANALYTICS AND OPTIMIZATION - COMPREHENSIVE ANALYSIS
Using Real Rasa Chatbot Data
================================================================================

[Setup] Creating output folders...
[OK] Output folders created: outputs/visualizations, outputs/reports, outputs/data

[1/8] Loading Rasa Chatbot Data...
[OK] Loaded 4 intents from Rasa chatbot: bot_challenge, greet, goodbye, query_knowledge_base
[OK] Generated 2000 chatbot log records
[OK] Intents: 4 (bot_challenge, greet, goodbye, query_knowledge_base...)
[OK] Users: 199, Sessions: 200
[OK] Date range: 2025-12-30 02:25:56.463387 to 2026-01-28 16:04:07.463387
[OK] Preprocessed data: 2000 records, 23 features
[OK] Saved data to outputs/data/chatbot_data.csv
[OK] Generated 2000 records with 23 features
[OK] Data source: Rasa NLU (chatbot_logs.yml)
```

```
[2/8] Calculating Performance Metrics...
[OK] Intent Accuracy: 85.25%
[OK] Conversion Rate: 30.90%
[OK] CSAT: 3.74/5 | NPS: -45.6
[OK] Confusion matrices and classification report generated

[3/8] Performing Exploratory Data Analysis...
[OK] EDA visualizations and reports generated

[4/8] Performing User Segmentation...
[OK] Created 4 user segments

[5/8] Performing NLP and Sentiment Analysis...
[OK] NLP analysis completed

[6/8] Performing Deep-Dive Metric Analysis...
[OK] ROI: 6080.0% | Competitive Advantage Calculated

[7/8] Generating Optimization Strategies...
[OK] Generated 11 prioritized recommendations

[8/8] Creating Additional Visualizations...
[OK] Heatmap, word cloud, and dashboard created

Generating Summary Report...
[OK] Summary report saved
```

```
==========================================
CHATBOT ANALYTICS AND OPTIMIZATION REPORT
==========================================

EXECUTIVE SUMMARY
------------------------------------------
Total Interactions: 2,000
Unique Users: 199
Unique Sessions: 200
Intent Accuracy: 85.25%
Conversion Rate: 30.90%
Average CSAT: 3.74/5
NPS Score: -45.6
Completion Rate: 76.40%
```

```
PERFORMANCE METRICS
----------------------------------
Mean Response Time: 401.78ms
P95 Response Time: 945.10ms
Time to First Response: 227.35ms
Fallback Rate: 11.80%
Low Confidence Rate: 8.50%
```

```
CHANNEL PERFORMANCE
------------------------
FACEBOOK:
  Conversion: 32.9%
  CSAT: 3.82/5
  Response Time: 401ms
MOBILE:
  Conversion: 31.2%
  CSAT: 3.71/5
  Response Time: 404ms
TELEGRAM:
  Conversion: 28.0%
  CSAT: 3.74/5
  Response Time: 394ms
WEB:
  Conversion: 29.8%
  CSAT: 3.73/5
  Response Time: 400ms
WHATSAPP:
  Conversion: 32.4%
  CSAT: 3.70/5
  Response Time: 411ms
```

```
USER SEGMENTATION
--------------------------------------------------------------------------------
Total Segments Identified: 4

TOP 5 OPTIMIZATION RECOMMENDATIONS
--------------------------------------------------------------------------------
1. [High] Retrain NLU model with more examples for greet intent
   Impact: Reduce fallback rate by 15-20%
2. [Medium] Add more training examples and improve feature extraction for greet
   Impact: Improve accuracy by 10-15%
3. [Medium] Add more training examples and improve feature extraction for query_knowledge_base
   Impact: Improve accuracy by 10-15%
4. [Medium] Add more training examples and improve feature extraction for bot_challenge
   Impact: Improve accuracy by 10-15%
5. [High] Optimize backend processing, add caching, or implement async responses
   Impact: Reduce response time by 30-40%


================================================================================
ANALYSIS COMPLETE!
================================================================================
```

```
================================================================================
ANALYSIS COMPLETE!
================================================================================


Generated Files:
  [Folder] outputs/data/
     - chatbot_data.csv (raw data)

  [Folder] outputs/visualizations/ (14 PNG files)
     - confusion_matrix.png
     - comprehensive_dashboard.png
     - session_heatmap.png
     - And 11 more visualizations...

  [Folder] outputs/reports/ (12 files)
     - SUMMARY_REPORT.txt (executive summary)
     - optimization_recommendations.csv
     - segment_profiles.csv
     - And 9 more reports...

================================================================================
```

Table 5: Summary Report