RV College of Engineering®

Go, change the world

ARTIFICIAL NEURAL NETWORKS AND DEEP LEARNING (AI253IA)

Intelligent Customer Support System with Sentiment Analysis

Presented by

Roshan Ninan John - 1RV22AI046
Safiya Farheen - 1RV22AI048
Dharshini M A - 1RV22AI069

Faculty Mentors: Dr. S Anupama Kumar & Dr. Somesh Nandi

Agenda

1. Introduction
2. Literature Survey
3. Summary of Literature Survey
4. Objectives
5. Requirement analysis – hardware and software specification
6. System architecture
7. Methodology
8. Module specification –
    a. Module 1 : Data Acquisition and Database Setup
        i. Input ii. Process iii output
    b. Module 2 : Model Development
    c. Module 3 : Integration and Testing

Introduction

Problem Definition:

- **Impact on Customer Service**: Modern businesses face increasing challenges in managing customer interactions, including high query volumes, maintaining consistency, and delivering personalized responses.
- **Challenges in Current Systems**: Traditional systems rely heavily on manual processes or rule-based chatbots that lack dynamic interaction capabilities. They struggle to process unstructured data like chat logs and fail to predict customer needs, leading to delayed responses and dissatisfaction.
- **Consequences**: Inefficiencies in handling queries result in customer frustration, operational delays, and loss of brand loyalty, especially as interaction volumes grow exponentially.

Statistical Reasoning:

- **Customer Expectations**: 75% of customers expect a reply within 5 minutes, with 77% considering consistent support crucial for brand loyalty.This statistic is mentioned in the Zendesk Customer Experience Trends Report 2020
- **Economic Impact**: Businesses lose millions annually due to inefficient support systems, with high operational costs from redundant human interventions.A report from A&OA reveals that 40% of businesses have lost sales due to inadequate IT support
- **Usage Trends**: Over 60% of consumers are comfortable with AI for handling routine queries, demonstrating the need for advanced AI-powered systems (PwC Voice of the Consumer Survey 2024)

Introduction

Solution Proposed:

An **AI-driven Intelligent Customer Support System** that integrates a **Sentiment Analysis Model** with an **ANN-based chatbot**.

- **Key Features**: Uses Natural Language Processing (NLP) to analyze customer queries, retrieves relevant data from structured and unstructured sources, and provides personalized responses.
- **Frontend Integration**: Implements Streamlit to seamlessly connect datasets ensuring scalable and real-time query handling.

Potential Impact:

- **Efficiency and Personalization**: Reduces response times by 50% while using predictive analytics to deliver tailored, accurate responses, enhancing customer satisfaction and operational performance.
- **Scalability and Cost-Effectiveness**: Automates routine inquiries, handles high interaction volumes without performance degradation, and reduces dependency on human support staff, offering a scalable and cost-efficient solution.

Stakeholders:

- **Businesses**: Benefit from operational efficiency and improved customer loyalty.
- **Customers**: Gain access to faster, more personalized support experiences.
- **Technology Integrators**: Leverage the system to deploy scalable AI-driven solutions across various industries.

Objectives

1. **Enhance Customer Query Response Efficiency**
   Enable the system to quickly and accurately process customer queries by integrating AI-driven chatbots, reducing response times and improving customer satisfaction.
2. **Asses Customer Complaints according to Sentiment**
   Implement machine learning models to personalize responses based on customers tones on reviews, including their previous interactions and preferences, enhancing the overall user experience.
3. **Automate Routine Customer Support Tasks**
   Automate common customer inquiries using AI-powered chatbots, reducing the need for human agents and allowing them to focus on more complex issues.
4. **Achieve High Accuracy in Customer Support**
   Develop a robust machine learning model to accurately understand and respond to a wide range of customer queries across various industries, ensuring effective problem resolution.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
1	Adaptive and Scalable Database Management with Machine Learning Integration: A PostgreSQL Case Study *(Maryam Abbasi 1 , Marco V. Bernardo 2,3 , Paulo Váz 3,4 , José Silva 3,4 and Pedro Martins 3,4,)*	Published in the journal Information, Volume 15, Issue 9, 2024, Article 574. Publisher: MDPI. DOI: 10.3390/info15090574	This paper presents a machine learning (ML) framework integrated with PostgreSQL to optimize query performance and manage workloads dynamically. It reduces query times by 42% and improves throughput by 74%, automating database tuning and adapting to workload changes. This approach minimizes manual administration while addressing tuning conflicts and ML performance challenges, providing a scalable solution for large-scale data management.
2.	Context-Aware NLP Models for Improved Dialogue Management *(Charlotte Dupont, Vinh Hoang)*	X. Gao, W. Zhu, J. Gao, and C. Yin, "F-PABEE: flexible-patience-based early exiting for single-label and multi-label text classification tasks," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023: IEEE, pp. 1-5.	Context-aware NLP models enhance dialogue systems by leveraging past interactions, user history, and contextual cues to provide coherent and personalized responses. Recent advancements include memory networks, transformer-based architectures, and multi-turn conversation tracking. These models are crucial for applications like customer service and virtual assistants but face challenges.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
3	The Effect of Using Chatbots at e-Commerce Services of Customer Satisfaction, Trust, and Loyalty *(Surjandy,Cadelina Cassandra)*	2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, Indonesia, 2023, pp. 268-273, doi: 10.1109/IC3INA60834.2023.10285799.	It discusses the influence of external factors, such as the COVID-19 pandemic, on the adoption and effectiveness of chatbots in e-commerce settings.The paper introduces the Q-factor model to assess the quality of chatbot interactions and their influence on customer perceptions. This model offers a unique perspective on evaluating chatbot effectiveness, focusing on reliability and system quality.
4.	Noised Consistency Training for Text Summarization *(Jeff Johnson, Matthijs Douze, Hervé Jégou)*	International Journal of Research (2022)Publication and Reviews. 1655-1659. 10.55248/gengpi.2022.3.4.12.	It proposes a method to improve text summarization models by training them to generate consistent outputs despite noised or perturbed inputs. The approach aims to enhance the robustness and generalization of summarization models, addressing issues like hallucination and sensitivity to slight input variations.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
5	Research and Development of an E-commerce with Sales Chatbot. *(Hossain, Mostaqim & Habib, Mubassir & Hassan, Mainuddin & Soroni, Faria & Khan, Mohammad. (2022)).*	Research Gate *557-564.* *10.1109/AIIoT54504.2022.9817272.*	The research discusses the chatbot's ability to integrate across multiple communication channels (social media, websites, and mobile apps)to enhance accessibility for users. It emphasizes the implementation of real-time analytics to monitor user interactions and gather feedback, enabling continuous improvement of performance and user satisfaction.The authors address the importance of incorporating security and privacy measures in the design, ensuring that user data is protected and compliance with data protection regulations
6.	E-Commerce Assistance with a Smart Chatbot using Artificial Intelligence *(Manikanta, M. & Rushi, J. & Lalitha, A. & Goud, B. & Suresh, V. & Daniya, Tyas)*	International Journal of Research (2022)Publication and Reviews. 1655-1659. 10.55248/gengpi.2022.3.4.12.	The research presents a web-based architecture for the e-commerce system, detailing how the integration of the chatbot within this architecture enhances accessibility and user experience across various devices and platforms, which is a distinct focus compared to previous studies. The paper emphasizes a user-centric design approach, showcasing how user feedback was incorporated into the development process to refine the chatbot's functionalities and interface Talks about load handling and efficient database management to handle high traffic.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
7	AI and Deep Learning-driven Chatbots: A Comprehensive Analysis and Application Trends *(Santosh K. Maher,Suvarnsing G. Bhable,Ashish R. Lahase,Sunil S. Nimbhore)*	Published in: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) Date of Conference: 25-27 May 2022 Conference Location:Madurai DOI:10.1109/ICICCS53718.2022.9788276	Chatbots, powered by AI and deep learning, are transforming industries like banking, healthcare, and retail by providing interactive, efficient user support. This paper analyzes chatbot applications, comparing their technologies, languages, and features. Chatbots are becoming essential tools for customer service, improving machine interaction and query resolution across various sectors.
8.	E-Commerce Assistance with a Smart Chatbot using Artificial Intelligence *(M. Rakhra et al.)*	2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2021, pp. 144-148, doi: 10.1109/ICIEM51511.2021.9445316.	The paper emphasizes the use of advanced NLP techniques to enable the chatbot to understand and respond to customer inquiries effectively, enhancing user interaction and satisfaction. Integrating chatbot with various e-commerce platforms, facilitating seamless transactions and customer support. It also briefs about user interface design that is essential for increasing usability.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
9	Development of Artificial Intelligence Based Chatbot Using Deep Neural Network. (Srinivasa Rao, Dr Dammavalam & Srikanth, K. & Pratyusha, J. & Sucharitha, M. & Tejaswini, M. & Ashwini, T)	Research Gate. 2021. 10.52458/978-93-91842-08-6-12.	The paper details the specific architecture of the deep neural network (DNN) used for the chatbot, including the number of layers, types of activation functions, and optimization techniques that enhance the model's performance in understanding and generating human-like responses.Highlights the techniques for data augmentation and the importance of a diverse dataset to improve the chatbot's ability to handle various user queries effectively.
10.	Chatbots in E-Commerce *(Ravi Kumar Chauhan, Manjot Arora, Yash Khadakban, Pranav Padmawar)*	IJARIIE-ISSN(O)-2395-439 6 Vol-7, Issue 1, 2021.	The research introduces specific metrics for measuring user engagement with chatbots in e-commerce settings. These metrics provide insights into user behavior and preferences, allowing businesses to optimize their chatbot interactions and improve customer retention rates The metrics are session duration,message count, user retention rate, user satisfaction score, drop off rate, intent recognition accuracy, first response time, feedback and improvement suggestions.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
11	Development of An e-commerce Sales Chatbot *(M.M. Khan)*	2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), Charlotte, NC, USA, 2020, pp. 173-176, doi: 10.1109/HONET50430.2020.9322667.	The research emphasizes the use of machine learning algorithms, specifically Support Vector Machines (SVM), for text categorization within the chatbot. This approach allows the chatbot to effectively understand and classify user queries, enhancing its ability to provide relevant responses and improve customer interactions, which is a distinct technical focus not previously discussed.It r briefs the implementation of Natural Language Understanding (NLU) techniques to enhance the chatbot's comprehension of user intents and context.
12.	*C*olBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT (*Omar Khattab, Matei Zaharia*)	43rd International ACM SIGIR conference on research and development in Information Retrieval, pp. 39-48. 2020.	ColBERT, is a novel ranking model that employs contextualized late interaction over deep LLMs (in particular, BERT) for effcient retrieval. By independently encoding queries and documents into one-grained representations that interact via cheap and pruning-friendly computations, ColBERT can leverage the expressiveness of deep LMs while greatly speeding up query processing. In addition, doing so allows using ColBERT for end-to-end neural retrieval directly from a large document collection. Our results show that ColBERT is more than 170× faster and requires 14,000× fewer FLOPs/query than existing BERT-based models, all while only minimally impact.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
13	A Bidirectional LSTM Model for Classifying Chatbot Messages (N. Lhasiw, N. Sanglerdsinlapachai and T. Tanantong)	Ayutthaya, Thailand, 2021, pp. 1-6, doi: 10.1109/iSAI-NLP54397.2021.9678173	explores the use of a Bidirectional Long Short-Term Memory (BiLSTM) model for classifying chatbot messages into five intention classes, particularly in the context of increased online communication during the COVID-19 pandemic. Implemented at Thammasat University, the model achieved an accuracy of 80% on the validation dataset, demonstrating its effectiveness in understanding user intent. This research contributes to the field of natural language processing by highlighting the potential of deep learning techniques to enhance chatbot functionality in educational settings.
14	Artificial Intelligence Approaches in Database Management Systems *(Krassimira Shvertner)*	Published in the Yearbook of the Faculty of Economics and Business Administration, Sofia University, Volume 18, Issue 1, 2020, Pages 303-326 DOI: 10.1109/ACCESS.2023.3263042	The paper explores the integration of AI with DBMS to address challenges like large data volumes and fast processing needs. This integration enables more efficient data management, intelligent processing, and automation of database maintenance tasks. Technologies like Exadata and SAP HANA leverage in-memory processing and advanced algorithms to improve performance and support autonomous databases

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
15	Chatbots: History, technology, and applications *(Eleni Adamopoulou ∗, Lefteris Moussiades)*	Published in Machine Learning with Applications, Volume 2, December 2020, Article 100006. DOI:10.1016/j.mlwa.2020.100006. Publisher: Elsevier.	This literature review explores the evolution, technologies, and applications of chatbots, highlighting their development stages, key approaches (pattern matching and machine learning), and design considerations. It discusses industrial use cases, risks, and mitigation strategies while suggesting ways to enhance chatbot intelligence.
16	Dense Passage Retrieval for Open-Domain Question Answering *(Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih)*	arXiv preprint arXiv:2004.04906 (2020).	Demonstrates that dense retrieval can outperform and potentially replace the traditional sparse retrieval component in open-domain question answering. Our empirical analysis and ablation studies indicate that more complex model frameworks or similarity functions do not necessarily provide additional values. As a result of improved retrieval performance, we obtained new state-of-the-art results on multiple open-domain question answering benchmarks.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
17	Billion-scale similarity search with GPUs *(Jeff Johnson, Matthijs Douze, Hervé Jégou*)	Published in: IEEE Transactions on Big Data (Volume: 7, Issue: 3, 01 July 2021), DOI:10.1109/TBDATA.2019.2921572,Date of Publication: 10 June 2019	This work enables applications that needed complex approximate algorithms before. For example, the approaches presented here make it possible to do exact k-means clustering or to compute the k-NN graph with simple brute-force approaches in less time than a CPU (or a cluster of them) would take to do this approximately.
18	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (*Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanov*a)	NAACL-HLT 2019, pages 4171–4186 Minneapolis, Minnesota, June 2 - June 7, 2019	Recent empirical improvements due to transfer learning with language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems. In particular, these results enable even low-resource tasks to benefit from deep unidirectional architectures. Our major contribution is further generalizing these findings to deep bidirectional architectures, allowing the same pre-trained model to successfully tackle a broad set of NLP tasks.

Literature Survey

Sl No	Author and Paper title	Details of Publication	Summary of the Paper
19	Attention Is All You Need (Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin)	31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. "Attention Is All You Need.(Nips), 2017." arXiv preprint arXiv:1706.03762 10 (2017): S0140525X16001837.	This research presented the Transformer, the first sequence transduction model based entirely on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention. For translation tasks, the Transformer can be trained significantly faster than architectures based on recurrent or convolutional layers. We are excited about the future of attention-based models and plan to apply them to other tasks.
20	Chatbots and conversational agents: A bibliometric analysis *(H. N. Io,C. B. Lee)*	Published in: 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) Date of Conference: 10-13 December 2017 Conference Location: Singapore DOI: 10.1109/IEEM.2017.8289883	Chatbots have evolved from rule-based systems to AI models that learn from experience. This research uses bibliometric analysis to explore past chatbot studies and identifies future research gaps, particularly driven by deep learning technology. The analysis suggests that deep learning will shape the future of chatbot development and provides recommendations for upcoming research.

Summary of Lit. Survey

1. Advancements in NLP Architectures

Transformers as a Foundation:

"Attention Is All You Need" introduced the Transformer, which replaced RNNs with a self-attention mechanism, leading to models like BERT and GPT. These models excel in contextual understanding and representation, making them pivotal for tasks like SQL generation and product recommendations.

Bidirectional Context Understanding:

BERT's bidirectional encoding enhances understanding of linguistic nuances, while GPT-3 demonstrates few-shot capabilities, highlighting its utility in domains with sparse labeled data.

2. Retrieval-Augmented Generation (RAG)

- RAG systems combine retrieval from structured and unstructured data sources with generative models to improve response accuracy.
- Dense Passage Retrieval and ColBERT emphasize efficiency in embedding-based retrieval, crucial for scalability in real-time applications like chatbots.

Summary of Lit. Survey

3. Vector Search in Information Retrieval

- Efficient similarity search methods like ScaNN and FAISS optimize large-scale vector operations. These tools enable near real-time retrieval, making them ideal for FAQs and product embeddings.
- The concept of Approximate Nearest Neighbor (ANN) is a recurring theme, reducing computational overhead without significant accuracy loss.

4. Domain-Specific Applications

- Research on Natural Language to SQL (NL2SQL) (e.g., Hofstätter et al.) addresses translating user queries into SQL, a key component for database interactions.
- Product recommendation systems leverage embeddings (BERT, Word2Vec) and cosine similarity to match user preferences with items. Hybrid models combining collaborative filtering and content-based filtering are highlighted.

Summary of Lit. Survey

General Observations

1. Effectiveness of Pre-trained Models

Pre-trained language models (e.g., BERT, GPT) outperform traditional models due to their contextual understanding. Fine-tuning improves domain-specific performance.

2. Integration of Retrieval and Generation

Combining retrieval (from structured databases or FAQs) with generation enhances chatbot capabilities, particularly for knowledge-intensive tasks.

3. Scalability and Efficiency

Optimized retrieval systems (e.g., ScaNN, FAISS) are essential for handling high-dimensional embeddings, especially in applications involving FAQs or recommendations.

Summary of Lit. Survey

Gaps Identified and Solutions

1. Scalability in Vector Databases

➔ **Gap**: Current vector search methods struggle with massive datasets in real-time settings.
➔ **Solution**: Use hybrid retrieval (dense and sparse representations) or memory-efficient indexing techniques like HNSW (Hierarchical Navigable Small World).

2. Accuracy in Natural Language to SQL

➔ **Gap**: NL2SQL systems face challenges with ambiguous or grammatically incorrect inputs.
➔ **Solution**: Incorporate fine-tuned models with error-handling mechanisms and leverage user feedback for iterative improvements.

3. Limited Context in Recommendations

➔ **Gap**: Traditional recommendation systems often overlook dynamic contexts (e.g., user mood, recent behavior).
➔ **Solution**: Use session-based recommendation techniques with recurrent architectures or fine-tuned BERT embeddings to capture evolving user preferences.

Summary of Lit. Survey

4. Efficiency in RAG Pipelines

- ➔ **Gap**: High computational costs due to dense retrieval and generative tasks.
- ➔ **Solution**: Employ lightweight retrieval models (e.g., ColBERT) and approximate techniques like ANN for faster responses.

5. Lack of Explainability

- ➔ **Gap**: Black-box nature of embeddings and generative models reduces trust in outputs.
- ➔ **Solution**: Incorporate interpretable embeddings (e.g., sparse or hybrid) and surface metadata from retrieved items.

Requirement Analysis

Hardware Requirements:

Processor: Intel i5 or equivalent for basic system operations and lightweight NLP processing.Recommended: Intel i7 or AMD Ryzen 7 for faster query handling and ANN model performance.

GPU: Minimum: NVIDIA GTX 1050 Ti for efficient model inference.Recommended: NVIDIA RTX 2060 or higher for faster training and real-time performance.

RAM: Minimum: 8 GB for development and testing purposes. Recommended: 16 GB for handling concurrent queries and managing large datasets.

Storage: Minimum: 256 GB SSD for application deployment. Recommended: 512 GB SSD or higher for storing model data, logs, and database files.

Networking: High-speed internet connection for real-time query handling and API communication.

Requirement Analysis

Software Requirements:

Programming Language: Python (Version 3.8 or above).

Libraries & Frameworks: NLP: NLTK (3.0), spaCy (3.0), or Hugging Face Transformers for natural language processing tasks. Deep Learning: TensorFlow (2.0) or PyTorch (1.10 or above).

Database Integration: PostgreSQL (Version 13.0 or above), MongoDB (Version 4.0 or above) for handling unstructured data.

Frontend: Streamlit (1.20 or above)

ORM: PostgreSQL for database interactions.

Integrated Development Environment (IDE): Jupyter Notebook 6.4.12 and above,PyCharm 2023.2 or above or VS Code 1.83.1 or above

Operating System: Windows 10/11, Ubuntu 20.04.1 and above (Linux), or macOS Monterey (12.0) or above for cross-platform compatibility.

Database Management Tools: pgAdmin (for PostgreSQL) and MongoDB Compass (for MongoDB)

Version Control: Git for version control and collaboration.

Additional Tools: Docker, Postman for API testing.

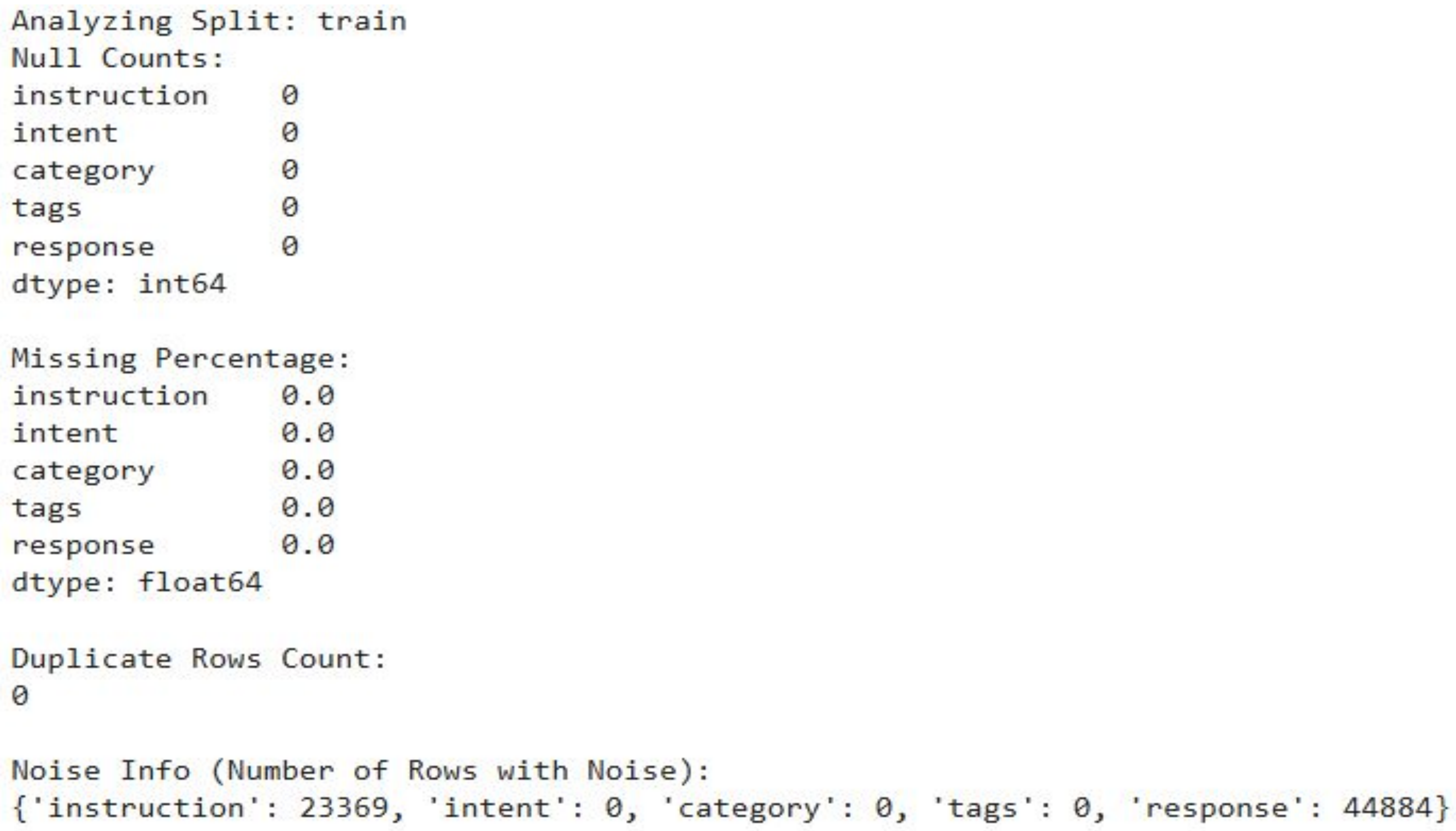
Dataset

Dataset Description for Sentiment Analysis

- ➢ Use Case: Sentiment Analysis
- ➢ Vertical: Retail (eCommerce)
- ➢ 50000 customer reviews assigned to 2 categories
- ➢ Categories of Reviews are labelled as positive ornegative
- ➢ 25000 reviews for each category

Requirement Analysis

Dataset Description



```
Analyzing Split: train
Null Counts:
instruction    0
intent         0
category       0
tags           0
response       0
dtype: int64

Missing Percentage:
instruction    0.0
intent         0.0
category       0.0
tags           0.0
response       0.0
dtype: float64

Duplicate Rows Count:
0

Noise Info (Number of Rows with Noise):
{'instruction': 23369, 'intent': 0, 'category': 0, 'tags': 0, 'response': 44884}
```

System Architecture

1. Input Layer:

User Input Channels:

➢ **Web Application**: Customers can input there queries and the chatbot will assess their intent or if its a complaint the sentiment of the complaint, the response is generated on the basis of the predicted intent or sentiment respectively.

Preprocessing:

Captures and processes text queries.

Performs normalization, tokenization, and language-specific handling.

System Architecture

2. Hidden Layers:

AI Processing Layer:

- ➢ **NLP Engine**: Performs intent recognition, Named Entity Recognition (NER), and sentiment analysis for semantic understanding.
- ➢ **RAG**: Contextual data ready for the Recommendation Engine.

Frontend Layer: Streamlit orchestrates real-time communication between dataset and the chatbot, ensuring efficient query handling and quick retrieval of data.

System Architecture

3. Output Layer:

- **Response Generation**: Combines the information retrieved from the knowledge database with the chatbots response to produce accurate, cohesive responses.
- **Intent detection and sentiment analysis**: Detect the intent of a query or the sentiment in a complaint.

Future Enhancement

Analytics and Monitoring: Tracks performance metrics such as query resolution times and customer satisfaction, logging errors to improve the system continuously.

System Architecture

Sentiment Analysis Model

The architecture of a Bidirectional LSTM (BiLSTM) model for sentiment analysis typically involves several key layers:

- **Input Layer:** Converts user text into sequences of numbers using tokenization, ensuring uniformity.
- **Embedding Layer:** Transforms these sequences into dense vectors, helping the model understand word meanings and relationships.
- **Bidirectional LSTM:** Reads text both forward and backward, capturing deeper context for better sentiment detection. It includes 64 LSTM units with a dropout rate of 0.3 to prevent overfitting.
- **Dense Layer:** Produces a sentiment score between 0 and 1, indicating whether the text is positive (closer to 1) or negative (closer to 0).

- **The model is trained on 50,000 IMDB movie reviews, using Binary Cross-Entropy Loss and the Adam optimizer. After training for 5 epochs, it effectively determines sentiment, helping the chatbot generate emotionally aware responses.**
  BiLSTM Model Accuracy: 0.8739799857139587

System Architecture

Transformer Architecture (Llama 3.2)

Bidirectional Processing:

- Unlike unidirectional models like GPT, BERT processes input sequences in both directions, capturing more context.

Layer Depth:

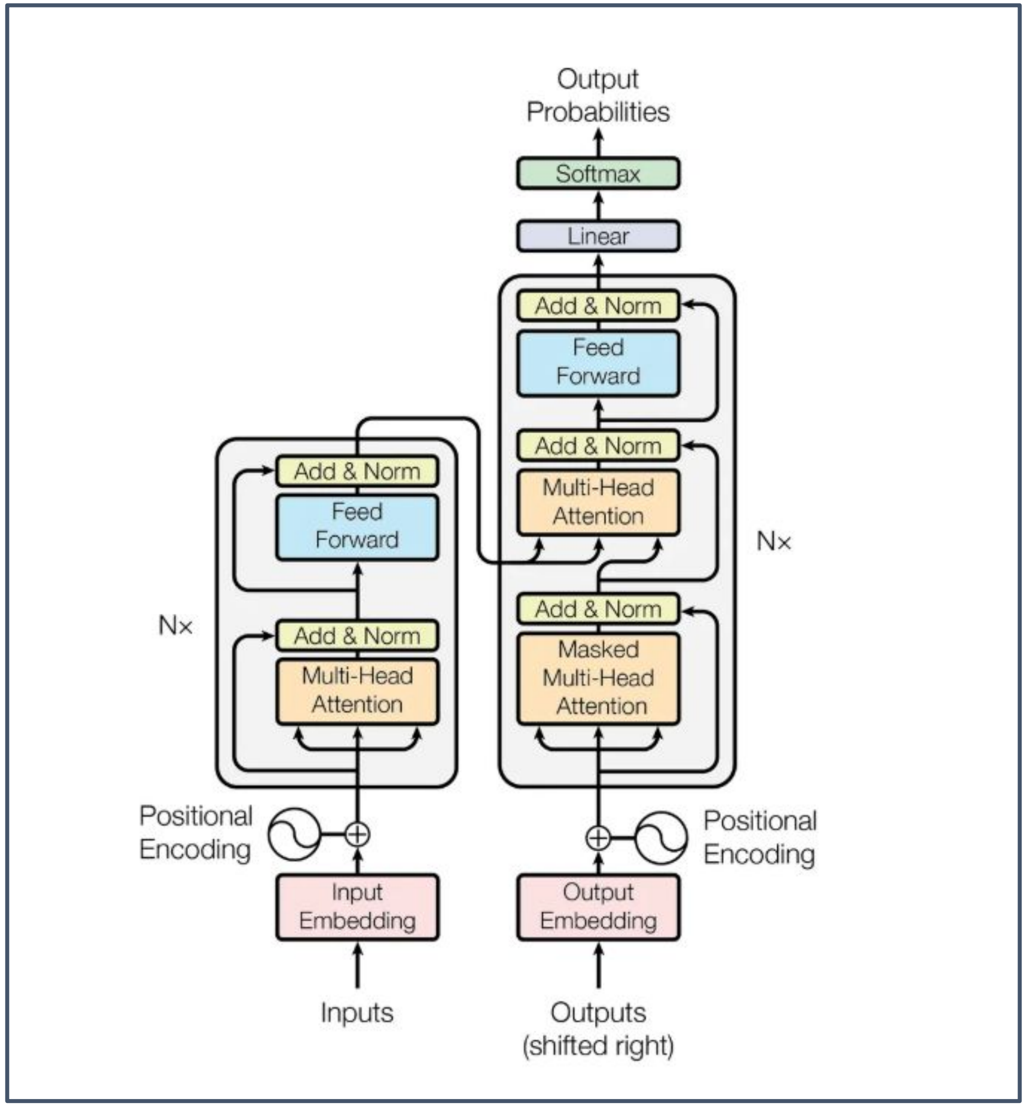
- The deep stack of encoder layers ensures complex patterns and relationships are learned.

Attention Mechanism:

- Self-attention ensures that even distant words in the sequence influence the contextual embedding of each token.

Flexibility:

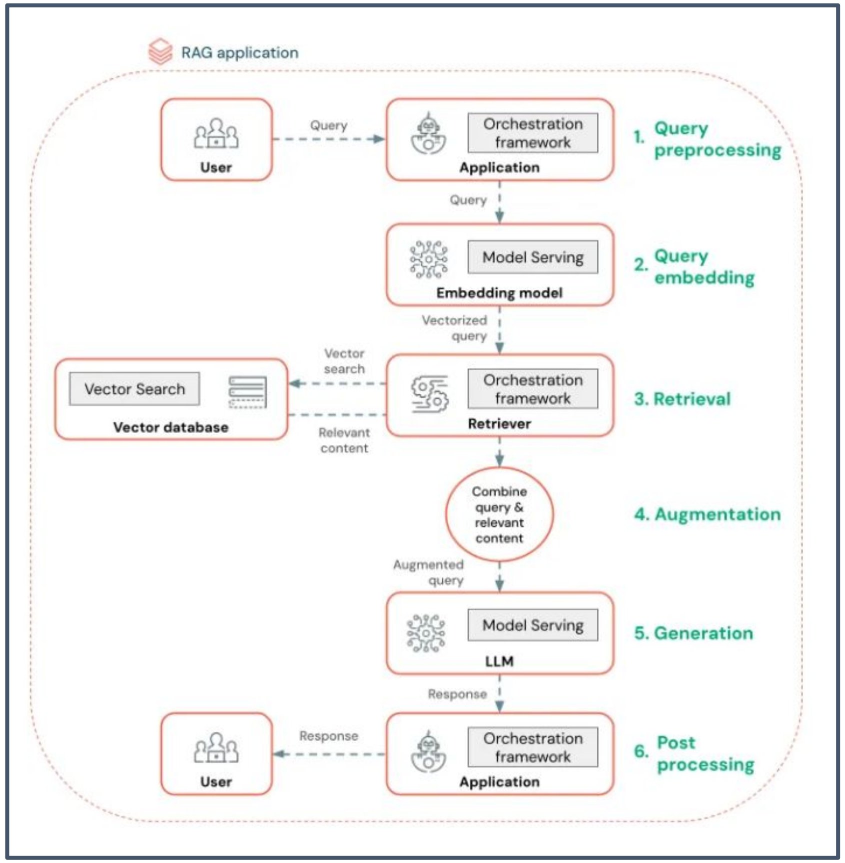
- BERT's embeddings can be fine-tuned for various NLP tasks, making it highly versatile.



System Architecture

RAG (Retrieval-Augmented Generation)

1. **Core Component:** Integrates structured and unstructured data with generative AI for a holistic response.
2. **Key Functions:**
   a. **Retrieval:**
      i. Fetches unstructured data from the NoSQL database (MongoDB), such as chat logs, support tickets, or user reviews.
      ii. Uses vector similarity search (Dense Passage Retrieval) to match semantically similar documents to the query.
   b. **Augmentation:**
      i. Combines the retrieved information with the structured data from PostgreSQL.
      ii. Ensures the generative model has contextually rich and accurate information for response generation(ColBERT).
   c. **Integration with Generative AI:**
      i. A generative model (e.g., GPT) uses the combined retrieved data to produce a comprehensive and context-aware response.
3. **Output:** Contextual data ready for the **Recommendation Engine**.



Methodology

Project Pipeline

User Query/Complaint Handling

- User submits a query through the chatbot.
- If it’s a complaint, it is passed to the Semantic Analysis Model.

Sentiment Analysis Model (Bi-LSTM)

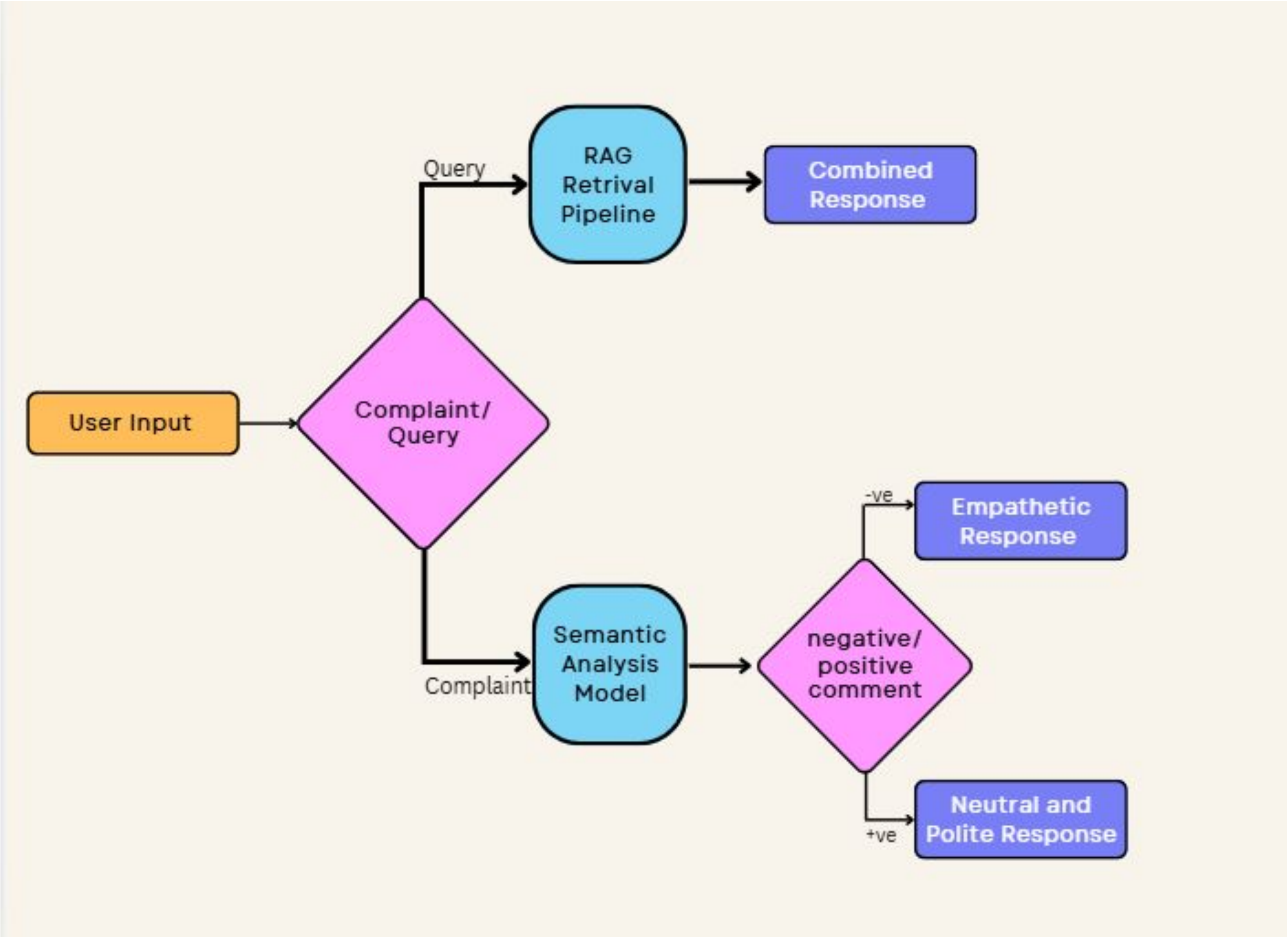
- Preprocess the query (tokenization, embedding, etc.).
- Predict the sentiment category.

RAG-Based Knowledge Retrieval

- If the intent requires knowledge-based retrieval, the query is passed to the RAG system.
- The system retrieves relevant context from the knowledge base.
- Generates a response using the retrieved information.

Response Generation

- If the query is simple (like "track my order"), the response is fetched from predefined intents.
- If complex, RAG refines the response based on retrieved knowledge.



Module Specification

Module 1 : Data Acquisition and Database Setup

Input:

- Collection and Generation of Customer Reviews, Complaints and Queries

Process:

- Data Cleaning and Validation
- Removal of stop words and special characters
- Making all words to lowercase and storing them as fixed length vector embeddings

Output:

Preprocessed and vectorised data ready for training and testing.

Module 2 : Model Development

- **NLP Model Development:** Focuses on understanding user inputs and generating appropriate responses.

Process:

1. **Sentiment Recognition:** Train model to identify the tone of a customers complaint(e.g., highly negative, neutral, positive).
2. **Embedding Creation for FAQs:** Generate vector embeddings for FAQs and tickets using a pre-trained model (e.g., Sentence Transformers).

- **RAG (Retrieval-Augmented Generation) Module:**Enhances chatbot responses by integrating database retrieval with LLM-generated outputs.

1. **Retrieve Context:**Query the database for relevant information (e.g., order details, FAQs).
2. **Combine with LLM:**Pass retrieved data as context to the language model to generate a more accurate and context-aware response.
3. **Generate Final Response:**Merge retrieved and generated content into a coherent response

Module 3 : Integration and Testing

1. Integration Steps

- **Establish Communication**:
  a. Input/output interface for each module.
- **Data Pipeline**:
  a. User Query → NLP → Sentiment Analysis or Intent Detection → Database → RAG → Chatbot Response

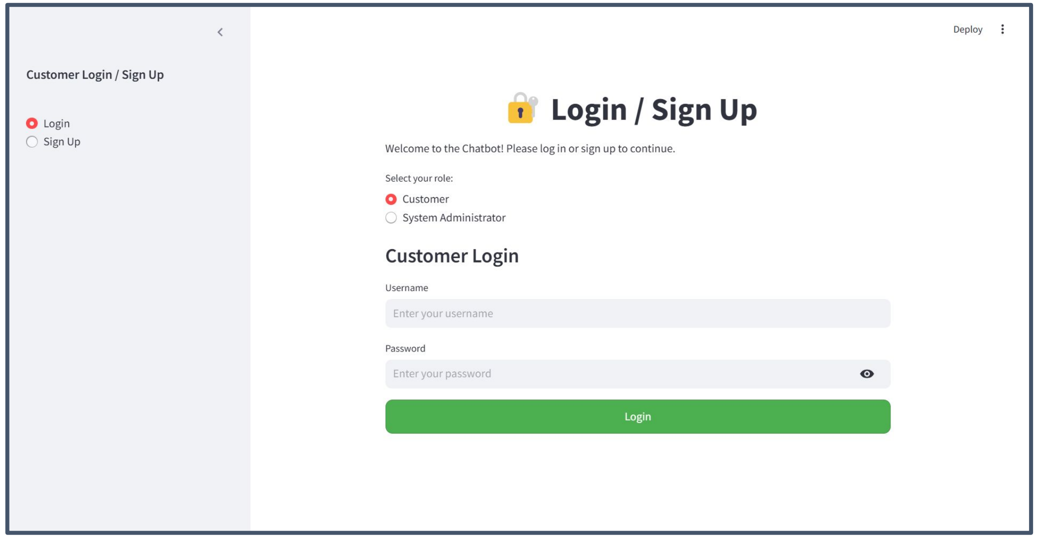
2. Module Testing

- **NLP Module**:
  a. **Unit Tests**: Test intents/entities extraction.
  b. **Edge Cases**: Handle ambiguous or typo-ridden queries.
- **Sentiment Analysis and Intent Detection**:
  a. **Unit Tests**: Verify generated output of each model.
  b. **Integration Tests**: Check responses for each category of query and complaint.
- **RAG**:
  a. **Testing Retrieval**: Validate context retrieval and similarity checks.
- **Response Generation**:
  a. **Unit Tests**: Test cosine similarity and output predictions.
  b. **Integration Tests**: Verify the response is generated taking all the outputs of the previous models into consideration.

Project Demo

1. Customer Login Interface

This figure displays the login page where users can authenticate as either a Customer or a System Administrator. Customers gain access to chatbot support services, while system administrators manage and analyze chatbot performance.



Customer Login / Sign Up

● Login
○ Sign Up

Deploy

Login / Sign Up

Welcome to the Chatbot! Please log in or sign up to continue.

Select your role:

● Customer
○ System Administrator

Customer Login

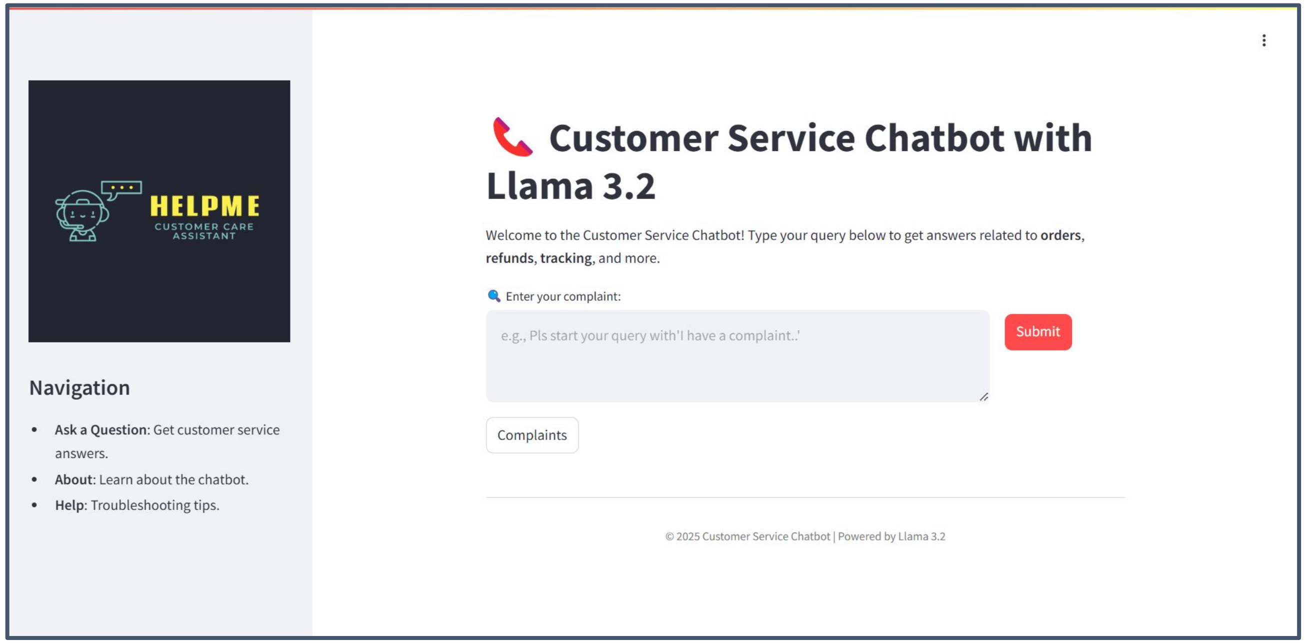
Username: Enter your username

Password: Enter your password

Login

2. Customer Chatbot Interface

After logging in as a customer, users can interact with the chatbot powered by Llama 3.2. The chatbot assists with common customer service tasks such as order tracking, complaints, and FAQs, leveraging RAG, FAISS, and BiLSTM for precise responses.



Navigation

- **Ask a Question**: Get customer service answers.
- **About**: Learn about the chatbot.
- **Help**: Troubleshooting tips.

☎ Customer Service Chatbot with Llama 3.2

Welcome to the Customer Service Chatbot! Type your query below to get answers related to **orders**, **refunds**, **tracking**, and more.

Enter your complaint:

e.g., Pls start your query with'I have a complaint.'

Submit

Complaints

© 2025 Customer Service Chatbot | Powered by Llama 3.2

THANK YOU !!