Intelligent Customer Support System with Sentiment Analysis

Roshan Ninan John
Department of Artificial Intelligence and machine learning
RV College of Engineering (VTU affiliated)
Bangalore, India
roshanninanj.ai22@rvce.edu.in

Safiya Farheen
Department of Artificial Intelligence and machine learning
RV College of Engineering (VTU affiliated)
Bangalore, India
safiyafarheen.ai22@rvce.edu.in

Dharshini M A
Department of Artificial Intelligence and machine learning
RV College of Engineering (VTU affiliated)
Bangalore, India
dharshinima.ai22@rvce.edu.in

Abstract— This project presents an AI-powered Intelligent Customer Support Chatbot designed to deliver real-time, personalized, and context-aware assistance. Built using Llama 3.2, Retrieval Augmented Generation (RAG), and FAISS, the chatbot efficiently retrieves relevant information from a knowledge base to handle customer queries such as order tracking, refunds, and product inquiries. By incorporating Bidirectional Long Short-Term Memory (BiLSTM) and sentiment analysis, the system analyzes both user intent and emotional tone, ensuring empathetic and human-like interactions. The chatbot automates routine tasks, reducing the dependency on human agents while maintaining seamless customer engagement. A user-friendly Streamlit interface enables easy accessibility, making interactions smooth and efficient. During testing, the chatbot demonstrated high accuracy and reduced response times, making it a scalable and effective solution for modern customer service needs.

Keywords— AI Chatbot, Customer Support Automation, Natural Language Processing, Llama 3.2, Retrieval Augmented Generation, FAISS, BiLSTM, Sentiment Analysis, Streamlit.

I. INTRODUCTION

In today's digital era, customer support plays a vital role in maintaining customer satisfaction and brand loyalty. Traditional support models struggle to keep up with rising customer expectations, with 75% of customers expecting a response within five minutes and 77% considering consistent service crucial for loyalty. With customer interactions projected to increase by 30–50% annually, businesses need scalable solutions to manage high query volumes efficiently while maintaining service quality.

This project introduces an Intelligent Customer Support Chatbot leveraging Llama 3.2, Retrieval Augmented Generation (RAG), and FAISS for precise knowledge base retrieval and contextual understanding. The system incorporates Bidirectional Long Short-Term Memory (BiLSTM) and sentiment analysis to enhance user intent detection and emotional understanding, ensuring personalized responses that adapt to customer needs. Implemented through a Streamlit interface, the chatbot automates routine tasks like order tracking, refunds, and FAQs, providing an efficient solution for modern customer service challenges. The system is built for scalability, smoothly integrating across web, mobile, and social media environments to address a variety of user needs while maintaining consistent performance and response quality across all platforms.

II. LITERATURE SURVEY

Recent advancements in AI-powered customer support systems have shown significant progress in addressing the growing demands of digital customer service. Vaswani et al. [19] introduced the transformative Transformer architecture, enabling more efficient and context-aware customer interactions. Devlin et al. [18] presented BERT, demonstrating how deep bidirectional architectures can effectively tackle diverse NLP tasks, particularly beneficial for understanding customer queries.
In the domain of retrieval systems, Karpukhin et al. [16] introduced Dense Passage Retrieval, while Khattab and Zaharia [12] proposed ColBERT, offering efficient passage search capabilities crucial for knowledge retrieval. Johnson et al. [17]'s work on billion-scale similarity search using GPUs enabled rapid response times in large-scale applications.
For sentiment analysis and intent detection, Lhasiw et al. [13] demonstrated the effectiveness of Bidirectional LSTM models in classifying chatbot messages, achieving 80% accuracy in understanding user intent. Surjandy and Cassandra [3] analyzed the impact of chatbots on customer satisfaction and loyalty in e-commerce settings, while Hossain et al. [5] explored multi-channel integration capabilities.
The evolution of chatbot technologies has been well-documented by Adamopoulou and Moussiades [15], providing insights into their historical development and applications. Khan [11] specifically addressed e-commerce applications, emphasizing the importance of machine learning algorithms for text categorization and natural language understanding. Our project builds upon these foundations to provide a more robust and scalable solution for modern customer support needs, addressing challenges in real-time performance and service quality.

III. SYSTEM ARCHITECTURE

A. *Semantic Analysis Model*

This model helps the chatbot understand the sentiment behind user queries using a Bidirectional LSTM (BiLSTM) network. It processes text sequentially, capturing both past and future context to enhance accuracy.

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.
The accuracy of Bidirectional LSTM model is 0.8739.

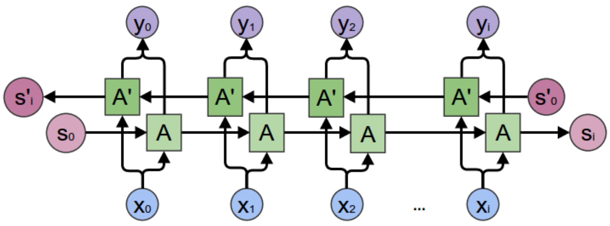

Fig 1: Bidirectional LSTM Architecture[21]

B. *Llama 3.2 Language Model*

Llama 3.2 is a Transformer-based Large Language Model (LLM) designed for advanced natural language understanding and generation. Built on the self-attention mechanism, it can analyze complex sentence structures and generate human-like responses with high accuracy.

Multi-Head Self-Attention: The model processes multiple parts of a sentence simultaneously, allowing it to grasp intricate contextual relationships and long-range dependencies.
Feed-Forward Network: Enhances understanding by refining patterns in text, ensuring coherent and well-structured responses.
Positional Encodings: Helps the model retain the sequence and structure of input text, making responses more contextually relevant.
Pre-Training & Fine-Tuning: Llama 3.2 is trained on a diverse dataset, learning grammar, meaning, and conversational flow. It is then fine-tuned on customer support queries, improving its ability to handle domain-specific questions.
Context-Aware Responses: By integrating retrieved information from the knowledge base, it generates highly relevant, factual, and personalized replies.
Scalability & Adaptability: The model is designed to handle large-scale interactions, adapting to different industries and user needs with minimal retraining.

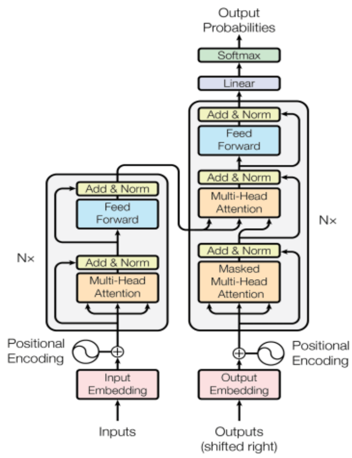

Fig 2: The Transformer Model Architecture[19]

C. *Retrieval-Augmented Generation (RAG) Pipeline*

RAG enhances chatbot accuracy by retrieving relevant knowledge before generating responses.

Retrieval Phase: Converts user queries into dense vector embeddings using Sentence Transformers (e.g., all-MiniLM-L6-v2). These embeddings are matched against a FAISS index, retrieving the most relevant documents.

Generation Phase: Combines these retrieved documents with the user's question and feeds them into Llama 3.2, which generates a well-informed and natural response.
This hybrid approach allows the chatbot to answer user queries with both precision and creativity, ensuring accurate, engaging, and real-time interactions.

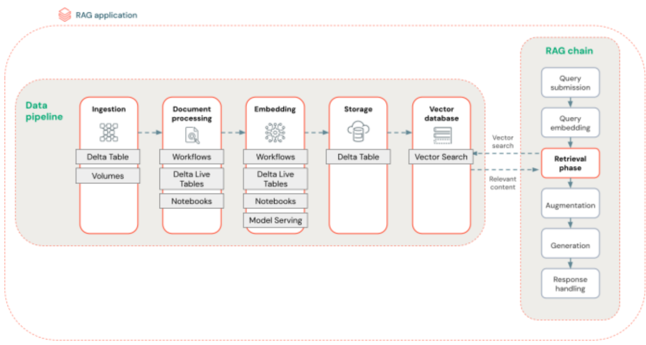

Fig 3: RAG Architecture[20]

IV. METHODOLOGY

The architecture of the e-commerce chatbot system is divided into four primary modules: Data Collection and

Preprocessing, Bi-directional Long Short Term Memory (BiLSTM) Model Implementation, Sentiment Analysis Integration, and Multi-channel Integration. Each module is vital to ensuring the system delivers accurate, personalized, and empathetic responses.

A. *Data Collection and Preprocessing*

The dataset used includes e-commerce queries, FAQs, and customer reviews. Data preprocessing involves tokenization, stop-word removal, and the use of pre-trained word embeddings like Word2Vec. These steps help transform raw data into meaningful input for machine learning models, improving their ability to understand and generalize queries.

B. *Bi-directional Long Short Term Memory (BiLSTM) Model Implementation*

The BiLSTM model is the core of the system. It processes text bidirectionally, considering both past and future words, which allows for a deeper understanding of user queries. The model focuses on intent recognition to identify the user's needs and emotion recognition to detect the emotional tone of the query, adapting the chatbot's responses accordingly.

C. *Sentiment Analysis Integration*

The Sentiment Analysis module identifies the emotional context of the user's message—whether they are happy, frustrated, or neutral. This analysis enables the chatbot to tailor its responses, offering empathetic replies for frustrated users and engaging ones for satisfied customers.

D. *Multi-channel Integration*

The chatbot is designed to provide a consistent user experience across various platforms, including websites, mobile applications, and social media. This multi-channel approach ensures that users can interact with the chatbot seamlessly, no matter where they are. It supports API-based integrations, enabling businesses to embed it into existing support systems effortlessly. Additionally, real-time synchronization across channels ensures that users receive uniform responses, enhancing overall engagement and satisfaction.

E. *Testing, Validation, and Deployment*

The system undergoes thorough testing and validation to ensure accuracy and relevance of responses. Once validated, the system is deployed, often using cloud services or containerized platforms like Docker. Continuous monitoring and version control ensure smooth operation and future improvements.
This modular design ensures the e-commerce chatbot is efficient, responsive, and capable of providing personalized support across various channels.

V. Results and Discussions

Customer Login Interface

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.

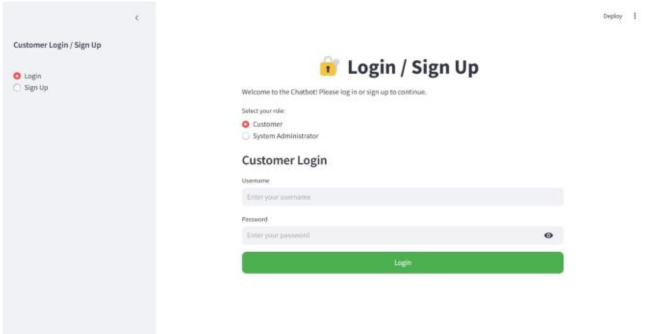

Fig 4: Customer Login page

System Administrator Login Interface

This screen allows administrators to log in. Admins can access analytics, monitor trends, and optimize chatbot performance based on user interactions.

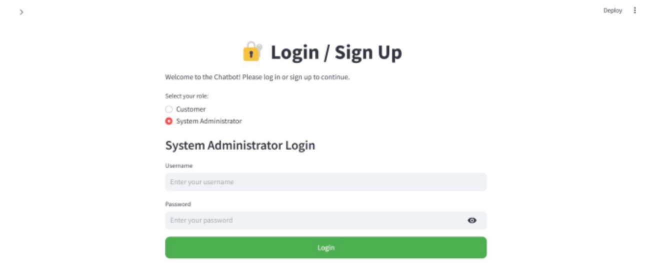

Fig 5: System Administrator Login page

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.

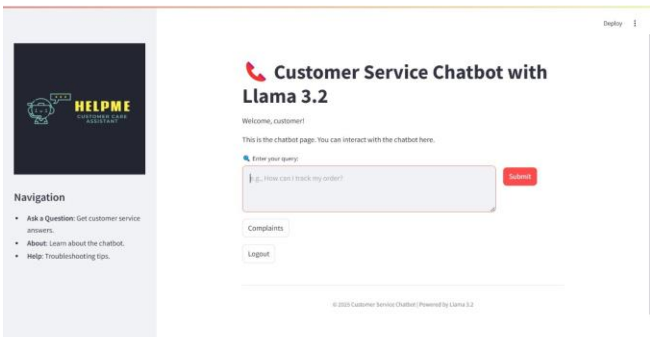

Fig 6: Chatbot

Admin Dashboard

The admin dashboard provides insights into chatbot performance, including user engagement trends, query resolution efficiency, and sentiment analysis metrics. This helps optimize responses and improve customer satisfaction.

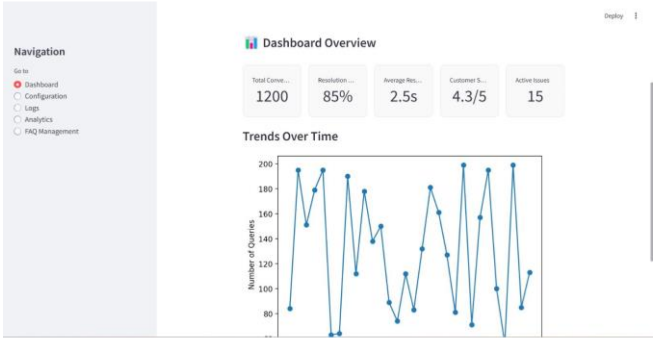

Fig 7: Admin Dashboard

VI. CONCLUSION

The Intelligent Customer Support Chatbot project showcases AI's transformative role in customer service. Utilizing Llama 3.2, Retrieval Augmented Generation (RAG) with FAISS, and a Streamlit interface, the system delivers rapid, context-aware responses while ensuring accessibility and ease of use. By automating repetitive tasks and handling high interaction volumes, it enhances efficiency, reduces operational costs, and minimizes human-agent reliance, allowing businesses to allocate resources more strategically.

The chatbot significantly improves customer satisfaction by providing fast query resolution for order tracking, refunds, complaints, and FAQs. Its scalability and continuous learning capabilities enable it to adapt to evolving user demands, ensuring sustained performance as customer expectations grow. Additionally, integration with CRM systems and multi-channel support (including websites, mobile apps, and social media) ensures a seamless user experience.

This project underscores AI's increasing role in modern customer service, paving the way for automated, scalable, and high-quality support systems that enhance both customer experience and business operations.

VII. FUTURE ENHANCEMENTS

To further enhance the Intelligent Customer Support Chatbot, several key improvements can be made:

1. Model Enhancement: Upgrading to advanced models like GPT-4 or T5 and fine-tuning on domain-specific datasets would improve the chatbot's ability to handle complex queries and provide more precise, contextually aware responses. Integrating transfer learning techniques could further refine its performance in specialized industries.
2. Dataset Expansion: Expanding the chatbot's knowledge base with a broader set of FAQs, industry-specific data, and real-time updates would enhance its ability to handle diverse queries. Incorporating multilingual support would enable businesses to serve a wider, global audience, ensuring seamless communication across different languages.
3. Performance Optimization: Deploying lightweight AI frameworks like TensorFlow Lite and implementing model pruning and quantization would reduce computational load, ensuring faster response times while optimizing memory usage. This would be particularly beneficial for mobile and edge device deployment, enhancing real-time performance.
4. User Experience & Deployment: Developing a dedicated mobile app, browser extension, and voice recognition features would make the chatbot more accessible across devices. Integration with CRM systems would improve customer interaction tracking, helping businesses provide personalized support based on previous interactions.
5. Cross-Platform Scalability: Deploying the chatbot on a cloud-based infrastructure would ensure scalability and high availability, allowing it to handle increasing user loads without performance degradation. Multilingual capabilities would further enhance its reach, making it a versatile and globally accessible support tool.

These enhancements will significantly boost the chatbot's accuracy, scalability, and usability, making it an even more powerful, efficient, and customer-centric AI assistant for businesses.

REFERENCES

[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, Information, Volume 15, Issue 9, 2024, Article 574.Publisher: MDPI. DOI: 10.3390/info15090574

[2] Context-Aware NLP Models for Improved Dialogue Management ,Charlotte Dupont, Vinh Hoang, Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (pp. 191-200),

[3] The Effect of Using Chatbots at e-Commerce Services of Customer Satisfaction, Trust, and Loyalty, Cassandra, Cadelina, *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. IEEE, 2023, doi: 10.1109/IC3INA60834.2023.10285799.

[4] Liu, Junnan, Qianren Mao, Bang Liu, Hao Peng, Hongdong Zhu, and Jianxin Li. "Noised consistency training for text summarization,doi*:2105.13635* (2021).

[5] Research and Development of an E-commerce with Sales Chatbot, Hossain, Mostaqim & Habib, Mubassir & Hassan, Mainuddin & Soroni, Faria & Khan, Mohammad, 2022 IEEE World AI IoT Congress (AIIoT) (pp. 557-564). IEEE

[6] E-commerce assistance with a smart chatbot using artificial intelligence, Rakhra, Manik, Gurram Gopinadh, Narasimha Sai Addepalli, Gurasis Singh, Shaik Aliraja, Vennapusa Siva Ganeshwar Reddy, and Muthumula Navaneeswar Reddy, 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 144-148). IEEE. DOI: 10.1109/IC3INA60834.2023.10285799

[7] AI and deep learning-driven chatbots: a comprehensive analysis and application trends, Maher, Santosh K., Suvarnsing G. Bhable, Ashish R. Lahase, and Sunil S. Nimbhore, 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 994-998). IEEE.

[8] E-commerce assistance with a smart chatbot using artificial intelligence, Rakhra, M., Gopinadh, G., Addepalli, N.S., Singh, G., Aliraja, S., Reddy,

V.S.G. and Reddy, M.N, 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM) (pp. 144-148). IEEE.
[9] Development of artificial intelligence based chatbot using deep neural network, Rao, Dammavalam Srinivasa, K. Lakshman Srikanth, J. Noshitha P. Pratyusha, M. Sucharitha, M. Tejaswini, and T. Ashwini, SCRS conference proceedings on intelligent system, DOI: 10.52458/978-93-91842-08-6-12.
[10] Development of an e-commerce sales Chatbot, Khan, Mohammad Monirujjaman, 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET) (pp. 173-176). IEEE, doi: 10.1109/MCE.2021.3083976
[11] Colbert: Efficient and effective passage search via contextualized late interaction over bert, Khattab, Omar, and Matei Zaharia, Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, DOI: 13.52458/978-93-91842-08-6-42.
[12] A bidirectional LSTM model for classifying Chatbot messages, Lhasiw, N., Sanglerdsinlapachai, N., & Tanantong, T 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing, IEEE
[13] Artificial Intelligence Approaches In Database Management Systems. Годишник на Стопанския факултет на СУ „Св. Климент Охридски, SCRS conference proceedings on intelligent system, DOI: 10.52458/978-93-91842-08-6-12
[14] Chatbots: History, technology, and applications, Eleni Adamopoulou, Lefteris Moussiades
[15] Billion-scale similarity search with GPUs, Johnson, Jeff, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs." IEEE Transactions on Big Data 7, no. 3 (2019): 535-547. IEEE Transactions on Big Data 7, no. 3 (2019): 535-547
[16] Bert: Pre-training of deep bidirectional transformers for language understanding, Devlin, Jacob, arXiv preprint arXiv:1810.04805.
[17] Attention is all you need, Vaswani, A, Advances in Neural Information Processing Systems (2017).
[18] Chatbots and conversational agents: A bibliometric analysis, Io, H. N., and C. B. Lee, 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 215-219). IEEE
[19] https://toloka.ai/blog/transformer-architecture/
[20]https://medium.com/@infinitylearnings1201/databricks-gen-ai-rag-in-detail-79da2299cb97
[21]https://www.digitalocean.com/community/tutorials/bidirectional-rnn-keras