Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AI-DRIVEN CHATBOT IMPLEMENTATION FOR ENHANCING CUSTOMER SERVICE IN HIGHER EDUCATION: A CASE STUDY FROM UNIVERSITAS NEGERI SEMARANG

MUHAMAD ANBIYA NUR ISLAM[1], BUDI WARSITO[2], OKY DWI NURHAYATI[3]

[1,3]Master of Information System, School of Postgraduate, Diponegoro University, Indonesia

[2]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Indonesia

E-mail: [1]mail.anbiya@gmail.com, [2]budiwrst2@gmail.com, [3]okydwin@gmail.com

ABSTRACT

Given the limited human resources and the needs of service users at Universitas Negeri Semarang (UNNES) helpdesk, there is a need for a solution regarding service problems. This study aimed to implement and evaluate an integrated chatbot system using similarity-based and generative-based response generation models at UNNES' helpdesk. The primary contribution is enhancing response efficiency and user satisfaction through automated, context-aware responses, which is a novel approach in higher education institutions. The primary objective was to enhance response efficiency and user satisfaction using automated and context-aware response generation. The methodology involved deploying the TF-IDF model for initial query handling to quickly retrieve relevant Frequently Asked Questions (FAQ) responses. Additionally, a generative model, Llama RAG, was employed for generating nuanced answers when queries fell below a defined similarity threshold. The steps included data collection, preprocessing, model training, and performance evaluation using precision, recall, F1 score, and BLEU score metrics. The TF-IDF model effectively handled 78% of queries, while the Llama RAG model addressed the remaining 22%. The average similarity score of TF-IDF responses was 0.85, and the BLEU score for generative responses was 0.61, demonstrating high relevance and linguistic coherence, respectively. These findings underscore the potential of integrating advanced AI models to improve helpdesk operations, suggesting that such systems can significantly enhance user interaction and operational efficiency.

Keywords: *AI-Driven Chatbots, Customer Service Automation, Natural Language Processing, Higher Education Helpdesk, Hybrid Chatbot Systems*

1. INTRODUCTION

Universitas Negeri Semarang (UNNES) operates a helpdesk designed to address inquiries from students, staff, and external users. The increasing volume of inquiries over the past three years (2021-2023) has placed significant strain on the limited customer service personnel. Addressing this issue is critical because efficient helpdesk operations are essential for maintaining high levels of user satisfaction and operational effectiveness, which directly impact the institution's reputation and service quality. This growing volume has resulted in longer wait times and a substantial number of unresolved complaints, highlighting the need for an efficient and scalable solution. The rapid transition from in-person to online services due to the COVID-19 pandemic has further exacerbated these challenges, leading to increased user dissatisfaction and many unresolved inquiries. This study aims to address these issues by implementing an AI-driven chatbot system to enhance response efficiency and user satisfaction at UNNES. However, this work does not cover the integration of the chatbot with other university services or explore its application in non-academic settings.

The helpdesk faces several specific challenges, including a limited number of customer service personnel, increasing inquiry volumes, and the transition from in-person to online services due to the COVID-19 pandemic. The average number of daily complaints has risen from 29 in 2021 to 105 in 2023, resulting in many unresolved inquiries and increased user dissatisfaction. The limited capacity

Journal of Theoretical and Applied Information Technology
31st July 2024. Vol.102. No. 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

of the customer service representative personnel has turned this growing volume into a potential threat, characterized by a significant number of unresolved complaints.

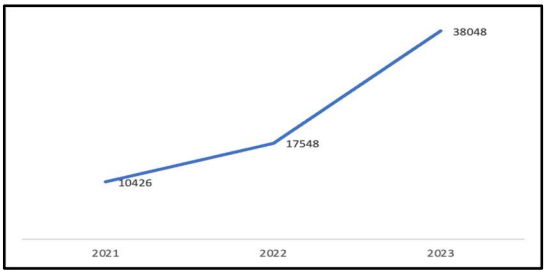

Figure 1: Number of UNNES services from 2021 to 2023

Prior to the COVID-19 pandemic, UNNES provided in-person help desk services. However, following the onset of the pandemic, the help desk transitioned to online services, utilizing email and live chat. Live chat services allow for real-time question-and-answer sessions with customer service representatives. Over the past decade, live chat has emerged as a preferred method for user engagement due to its direct and immediate interaction capabilities [1]. This feature requires personnel to individually address each user's information or service requests.

Further analysis of the live chat service duration at UNNES is detailed in Figure 2, which shows the average duration of a live chat session is approximately 16 minutes and 20 seconds, with the longest recorded session lasting 115 minutes. Figure 3 reveals that the average response time by customer service personnel to incoming requests is 5 minutes. This leads to user perceptions of the integrated service, as depicted in Figure 4, which shows that out of 4004 responses, or approximately 99% of the customers, rated the service as neutral, while the remaining 1% provided positive feedback. These findings indicate the need for a thorough evaluation and in-depth analysis of the service data provided.

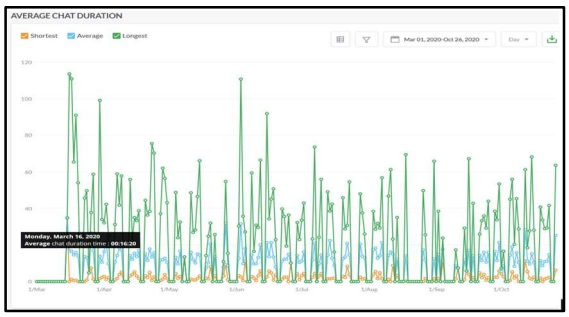

Figure 2: Average service duration at UNNES help desk

A significant issue arises when the volume of service requests increases but is met with a limited number of customer service personnel, leading to many user inquiries remaining unaddressed or causing longer wait times, thereby potentially reducing user satisfaction.

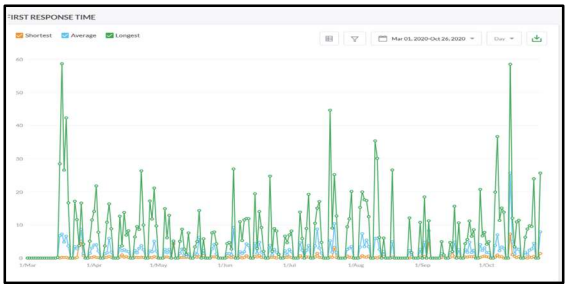

Figure 3: First response time trends at UNNES help desk

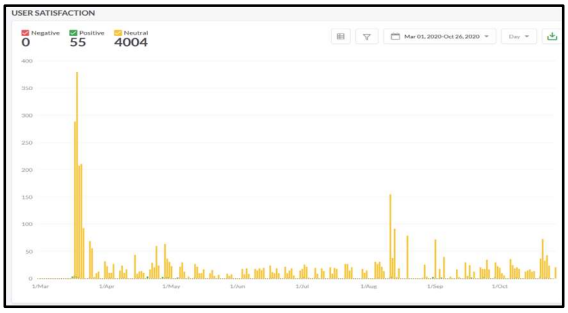

Figure 4. User satisfaction score at UNNES help desk

An analysis of the UNNES help desk service revealed that 10% of users submit the same complaint multiple times per day, and 30% submit questions that are already answered in the Frequently Asked Questions (FAQ) section. This indicates a low engagement with the FAQ literature by users. This scenario presents two main challenges for the UNNES help desk. Firstly, the workload for customer service personnel increases as many users submit the same queries repeatedly. Secondly, a higher volume of incoming queries directly correlates with the time required by customer service personnel to address all inquiries.

The limited customer support personnel at UNNES helpdesk have made it imperative to utilize Artificial Intelligence (AI) as a pivotal alternative to human intervention for primary customer service tasks. AI technology facilitates machines to operate in a manner akin to humans, thereby potentially substituting human labor [2]. This not only reduces the workload on customer service representative but also improves the response time and satisfaction for users seeking information. An automated system ensures that users receive consistent and accurate information promptly, enhancing overall service quality.

The primary objective of this study is to explore and evaluate various methods for building an

effective FAQ-based chatbot tailored to the needs of higher education institutions like UNNES. This problem needs attention because the increasing volume of inquiries and limited human resources lead to longer wait times and user dissatisfaction, which can negatively affect the institution's operational efficiency and reputation. Through a detailed comparative analysis of different natural language processing (NLP) techniques, this study aims to identify the most suitable methods for developing such a system. The contributions of this research are threefold:

Firstly, the study provides quantitative findings on the performance of various methods, including TF-IDF, BERT, Universal Sentence Encoder (USE), Sentence-BERT (SBERT), and GloVe, in the context of an FAQ-based chatbot. By evaluating these methods using metrics such as Precision, Recall, F1 Score, and Accuracy, the research identifies the strengths and weaknesses of each approach. As known from research before, SBERT provides promising results for limited number FAQ dataset [3].

Secondly, the research offers practical recommendations for implementing a chatbot that combines both retrieval-based and generative components [4]. The retrieval-based component addresses common and repetitive questions by pulling relevant answers from the FAQ database. In contrast, the generative component, utilizing models like LLaMA3, is triggered when the similarity score of a query does not meet the predefined threshold, indicating that the chatbot is not confident in its ability to provide an accurate response. In such cases, the generative model, trained on general knowledge about the organization, steps in to create appropriate responses. This hybrid approach ensures that users receive accurate and helpful information, even for queries that fall outside the predefined FAQ scope.

Lastly, the study presents a detailed case study on the implementation of the chatbot at UNNES, addressing specific challenges and proposing solutions. Compared to existing literature, such as the works by Adamopoulou and Moussiades [1] and Labadze et al.[5], our findings demonstrate a significant improvement in response efficiency and user satisfaction. Unlike previous studies that primarily focused on either retrieval-based or generative models, our hybrid approach combines both, resulting in a more comprehensive solution. This contribution aligns with and extends the current understanding of AI applications in higher education settings. This case study highlights the real-world application of the research findings, showcasing the potential benefits and impact of an AI-driven FAQ-based chatbot in an academic setting. The insights gained from this implementation can guide other institutions in adopting similar technologies to enhance their customer service operations.

By addressing these objectives, the study aims to contribute to the broader understanding of AI and chatbot technologies in higher education, providing a valuable resource for both academic researchers and practitioners. The findings of this research are expected to significantly improve the efficiency and effectiveness of helpdesk services, leading to higher user satisfaction and better resource management within educational institutions.

2. LITERATURE REVIEW

Recent studies have highlighted the growing incorporation of AI technologies within university settings, focusing on the potential for improved efficiency and the challenges associated with their adoption [6]. The theoretical framework for customer service enhancement through AI is well-established, with researchers emphasizing the importance of seamless integration, personalization, and context-aware responses [2]. Within the domain of educational settings, the adoption of chatbots has shown varied results, with some studies highlighting the benefits of student interaction enhancement and support service [7], or educators, the main advantages are the time-saving assistance and improved pedagogy. However, our research also emphasizes significant challenges and critical factors that educators need to handle diligently. These include concerns related to AI applications such as reliability, accuracy, and ethical considerations. This underscores the need for a comprehensive and tailored approach to [5] chatbot development and deployment within the unique context of higher education institutions.

Existing methods for building FAQ-based chatbots include both traditional and modern NLP techniques. Traditional methods, such as TF-IDF and bag-of-words models, rely on statistical representations of text data and have been widely used due to their simplicity and efficiency. However, these methods often lack the ability to capture semantic meaning and context, which limits their effectiveness in providing accurate responses to user queries. Modern NLP techniques, such as BERT, Universal Sentence Encoder (USE), and SBERT, leverage deep learning and transformer architectures to understand the context and semantics of text. These models have demonstrated significant

improvements in various NLP tasks, including question answering and text classification (IEEE Access). BERT, for instance, uses bidirectional context to achieve a deeper understanding of language, which enhances its performance in FAQ-based chatbots.

Generative models, such as LLaMA3 and ChatGPT, represent a different approach to building chatbots. These models generate responses based on learned knowledge from large datasets, enabling them to produce more human-like and contextually relevant answers. While generative models excel in creating diverse and coherent responses, they can struggle with providing specific answers to domain-specific queries unless they are fine-tuned with relevant data.s

Email, Live Chat, and other Chat services implemented in several major universities in Indonesia have not been able to sufficiently replace or assist customer service in handling numerous and repetitive queries, or those with identical responses. In the industrial realm, for example, choki, an AI used by the e-commerce platform Shopee, employs a predefined question-and-answer design. Users select from available questions, which may complicate the process of finding a query that matches their specific issue. Based on the impact of AI application in specific research of customer satisfaction, shopee's choki provides 48% contribution to customer satisfaction [8].

3. METHODOLOGY

The methodology section outlines the systematic approach used to develop and evaluate the FAQ-based chatbot for Universitas Negeri Semarang (UNNES). This research employs a mixed-method design, combining qualitative and quantitative approaches to provide a comprehensive evaluation of the chatbot system. The methodology includes detailed steps of data collection from various sources, data preprocessing, model training, and performance evaluation using metrics such as precision, recall, F1 score, and BLEU score.

3.1 Data Collection

The data for chatbot development were collected from various sources, including historical chat logs, email inquiries, and WhatsApp chat history from the UNNES helpdesk. The data collection spanned from January 2021 to December 2023, capturing a comprehensive range of user queries and responses. In total, the dataset comprised over 4,000 historical user queries, which were refined into 200 unique and frequently asked question (FAQ) pairs. The dataset is organized in a tabular format with columns for questions and answers, examples of which are shown in Table 1. This comprehensive dataset ensures that the chatbot is trained on a wide range of queries and responses, capturing the most common issues and their resolutions.

Table 1: Sample of FAQ Dataset from helpdesk UNNES

Question	Answer
-I am going to register for an exam/seminar/scholar ship. I need proof of UKT payment. -How can I print proof of UKT payment?	We can process the printing of UKT payment proof only for undergraduate students. For postgraduate students, you can request it directly from the postgraduate finance department.
-The Virtual Account number for UKT payment does not appear -Why does my Virtual Account number for UKT payment not appear?	During the administrative registration payment period (UKT payment), we will regenerate your VA number. Please recheck the payment menu for UKT/SPP/SPI. If it is not during the UKT payment period, please write a complaint on the ult.unnes.ac.id page (create a new ticket).
-How to reset student email password? -I forgot my student email password	We can reset the email password for students and staff. Please write a complaint "request to reset email password" on the ult.unnes.ac.id page (create a new ticket) by attaching your student ID card (KTM) or national ID card (KTP).
-My thesis/dissertation title is wrong, how can I correct it?	The title can be directly changed by the department supervisor in the SBVT menu on Sikadu.

3.2 Data Preprocessing

Preprocessing is a crucial step to prepare the raw data for analysis and model training. The preprocessing steps include:

3.2.1 Data cleaning

The data cleaning process involves the removal of duplicate entries, irrelevant data, and missing values by discarding incomplete or irrelevant records. This step ensures that the dataset is free from noise and inconsistencies, which can negatively impact the performance of the model.

3.2.2 Data normalization

Data normalization includes lowercasing and punctuation removal, standardizing the text format for consistent processing. Sample of data normalization process is shown in table 2.

Table 2: Lowercasing and punctuation removal sample

Before	After
-I am going to register for an exam/seminar/scholarship. I need proof of UKT payment.	i am going to register for an exam seminar scholarship i need proof of ukt payment

3.2.3 Tokenization

Splitting the text into tokens (words or sub words) for further processing. Tokenization is performed using NLTK and the tokenizers provided with transformer models, such as BERT tokenizer for BERT encoder.

3.2.4 Stop words removal

Removing common stop words that do not contribute significant meaning. This step helps in reducing the dimensionality of the dataset while retaining the essential information.

3.2.5 Vectorization

Converting text data into numerical format using methods like TF-IDF, word embeddings (Word2Vec, GloVe), or contextual embeddings (BERT, USE, SBERT). This transformation is necessary for feeding the text data into machine learning models for training and evaluation

3.3 Template Based Response Generation

The system development phase involves designing and implementing the FAQ-based chatbot using various NLP techniques. Different models such as TF-IDF, BERT, USE, SBERT, and GloVe are employed to generate vector representations of the queries and responses. The system is designed to handle both retrieval-based and generative responses, ensuring that it can provide accurate answers to frequently asked questions while also generating appropriate responses for less common queries. The development process includes training the models on the preprocessed dataset, fine-tuning them for improved performance, and integrating them into the chatbot system.

3.3.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) method is a widely used algorithm for measuring the similarity between texts. It represents a text as a vector, where each dimension corresponds to a word term that appears in the text [9]. It helps identify words that are distinctive within individual articles but relatively uncommon across all articles [10]. It combines the frequency of a word in a document (TF) with the inverse document frequency (IDF) across the corpus

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

where t is the term, d is the document, N is the total number of documents, and $DF(t)$ is the number of documents containing the term t.

3.3.2 BERT (Bidirectional Encoder Representations from Transformers)

Since its introduction by Devlin et al. in 2018, BERT (Bidirectional Encoder Representations from Transformers) has established itself as a foundational technology for various Natural Language Processing (NLP) tasks [11]. In recent developments, BERT has significantly influenced the advancement of methods for assessing text similarity, an application crucial in the fields of information retrieval, document clustering, and question-answering systems. The bidirectional nature of BERT allows it to understand the context of words within a sentence more deeply than unidirectional models, enabling BERT to capture nuances of meaning that are essential for determining text similarity [12].

Recent studies have focused on leveraging pre-trained BERT models to compute semantic similarity between texts. By fine-tuning BERT with specific datasets, researchers have been able to enhance the model's performance in identifying similar texts, even when they employ varied words or structures. For instance, a 2022 study demonstrated that fine-tuned BERT models outperform traditional cosine similarity metrics, particularly in tasks requiring a deep understanding of sentence context. These models are not only more accurate but also significantly faster at processing large volumes of text, which is crucial for real-time applications.

3.3.3 USE

USE (Universal Sentence Encoder) is a technique involves converting textual data into high-dimensional vector representations. These vector embeddings can then be utilized for a variety of natural language processing tasks, such as categorizing and classifying text, grouping related texts through clustering algorithms, measuring semantic similarity between texts, and other text analysis applications [13]. In contrast to traditional word embedding methods that represent individual words as vectors, the Universal Sentence Encoder (USE) aims to encode entire sentences into vector representations. This approach enables USE to capture not only the semantic meanings of words, but also the syntactic relationships and sentence-level context. By representing sentences as vectors that incorporate both semantic and syntactic information,

USE is particularly well-suited for natural language processing tasks that require an understanding of the contextual meaning within full sentences [13]. Research comparing different sentence embedding approaches has demonstrated that the Universal Sentence Encoder (USE) frequently achieves superior performance over other techniques like Sentence-BERT and InferSent [14].

3.3.4 SBERT

Sentence-BERT (SBERT) is a modified version of the BERT model that incorporates siamese and triplet network architectures. It aims to enhance the performance of ALBERT models, more so than standard BERT models, specifically for tasks involving semantic textual similarity benchmarks [15]. SBERT is optimized for efficient semantic similarity search, making it significantly faster than BERT. Unlike BERT which requires inputting both sentences together, leading to massive computational overhead, SBERT reduces the computation time drastically [16].

3.3.5 GloVe

GloVe (Global Vectors for Word Representation), created by Stanford NLP Group, is a word embedding model that maps words to vectors based on their co-occurrence in large text datasets. It uses global corpus statistics to capture semantic relationships, differing from models like Word2Vec that rely on local context [17]. GloVe can be fine-tuned for specific fields like medicine or finance, allowing it to capture domain-specific word meanings. This customization helps GloVe perform better than generic embeddings on specialized tasks within these fields [18].

3.4 Generative Response Generation

The generative retrieval technique involves creating responses based on the knowledge stored in large-scale models. Unlike traditional retrieval methods that rely solely on matching existing responses to user queries, generative retrieval techniques generate new responses by leveraging the vast amounts of information encoded within the models.

3.4.1 Large Language Model

Large Language Models (LLMs) have transformed the field of natural language processing (NLP) with their exceptional ability to comprehend and produce human language. Pre-trained on extensive datasets, these models have achieved notable advancements in numerous NLP tasks, capturing the attention of both academic researchers and industry professionals. Figure 1 illustrates the rapid advancement and increasing complexity of LLMs over time. The graph shows significant milestones in the development of LLMs, beginning with T5 and progressing through notable models such as GPT-3, Codex, InstructGPT, ChatGPT, LLAMA, and GPT-4. This exponential growth underscores the transformative impact these models have had on the field of natural language processing [19].

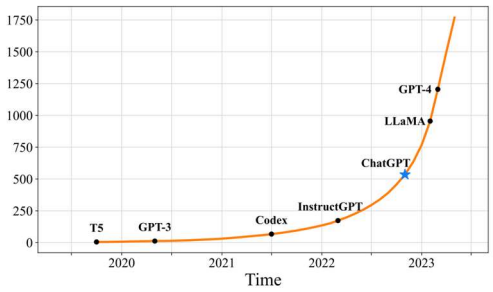

Figure 1 Growth and Development of Large Language Models over Time[20]

LLMs provide significant insights into language, comprehension, intelligence, social interaction, and the concept of personhood. They demonstrate that statistical methods can lead to understanding and that complex sequence learning combined with social interaction may be adequate for achieving general intelligence [20].

3.4.2 Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is a method that improves the performance of LLMs by combining information retrieval with text generation. By utilizing external knowledge sources, RAG enhances the precision and pertinence of the generated content, addressing common problems like hallucinations and factual errors typically associated with LLMs. Figure 2 provides an overview of the Retrieval-Augmented Generation (RAG) process. The retrieval-augmented generation framework comprises three key elements: the retrieval source, retrieval metric, and integration methods [21].

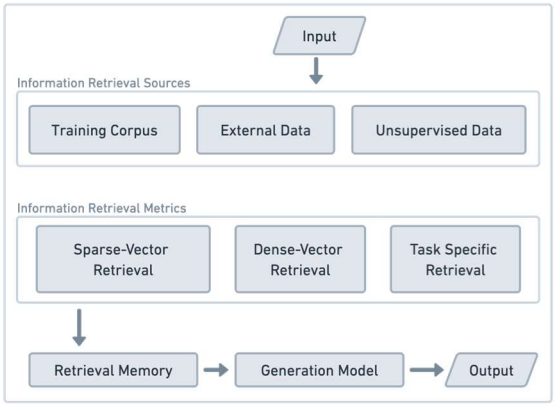

Figure 2 Retrieval Augmented Generation process [21]

RAG enhances the accuracy of open-domain question answering by creating diverse contexts and integrating them with dense representations, surpassing the performance of other approaches [22].

3.5 Chatbot Response Generation Architecture

The chatbot's response generation architecture combines multiple components from the model development phase into a unified system, adept at managing user queries efficiently. This design facilitates smooth user-chatbot interactions, utilizing both retrieval-based and generative models to deliver precise and contextually appropriate responses. Figure 3 illustrates the overall architecture of the chatbot system.

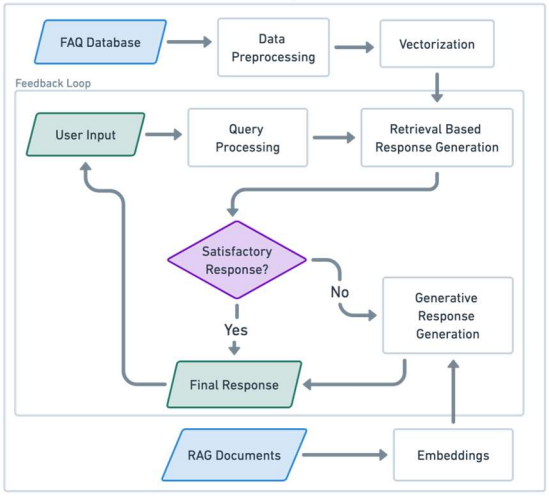

Figure 3 Chatbot Response Generation Flowchart

This architecture manages user queries efficiently and effectively, delivering accurate, relevant, and contextually suitable responses. The satisfactory response decision is designed to differentiate user queries by determining whether they pertain to common specific complaints or general organizational knowledge. This decision point ensures that the chatbot directs queries either to a retrieval-based response generation model, when the query matches known FAQs, or to a generative response model, when the query requires a broader or more contextually nuanced answer. By integrating both retrieval-based and generative models, the system is equipped to handle a broad spectrum of queries, ranging from straightforward FAQs to more intricate and novel questions.

3.6 Evaluation Metrics

To assess the chatbot's performance, several key metrics are employed to ensure the model's accuracy, relevance, and response quality. These metrics offer a thorough evaluation of the chatbot's effectiveness in managing user queries.

3.6.1 Precision and Recall

Precision measures the accuracy of the positive predictions made by the model. It is a crucial metric for understanding how many of the retrieved documents or responses are relevant. Recall measures the ability of the model to capture all relevant instances. It is an essential metric for evaluating the model's effectiveness in retrieving all relevant documents or responses.

$$Precision = \frac{True\ Positives}{True\ Positives+Fals\ \ Positives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives+Fals\ \ Negatives} \quad (3)$$

3.6.2 F1 Score

The F1 Score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance, considering both precision and recall. This metric is particularly useful when the dataset is imbalanced, as it combines the benefits of both precision and recall into a single metric.

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Reca} \quad (4)$$

3.6.3 Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances. It is a straightforward metric that provides an overall measure of the model's performance. High accuracy indicates that the model is generally performing well across both relevant and irrelevant instances.

$$Accuracy = \frac{True\ Positives+Tru\ \ Negatives}{Total\ Instances} \quad (5)$$

3.6.4 BLEU Score

The BLEU (Bilingual Evaluation Understudy) Score evaluates the quality of text generated by the model by comparing it to a reference text. It is commonly used for evaluating the performance of machine translation models but is also applicable for assessing generative models in chatbot applications. A higher BLEU score indicates that the generated text closely matches the reference text, suggesting better quality and relevance of the generated responses.

4. RESULTS AND DISCUSSION

The Results and Discussion section presents the findings from the evaluation of the hybrid chatbot, analyses the performance metrics,

and discusses the implications of these results in the context of improving customer service at UNNES.

4.1 Performance Evaluation of Similarity-Based Models

The study evaluated the effectiveness of various similarity-based models including TF-IDF, BERT, USE, SBERT, and GloVe in processing user queries within the context of a university helpdesk system. Performance metrics focused on F1 Score and Accuracy, which are crucial for assessing the precision and reliability of each model in retrieving relevant FAQ responses.

Table 3 Precision, recall and F1 score for each model

Model	Precision	Recall	F1 Score	Accuracy
TF-IDF	**0.67**	**0.68**	**0.67**	**0.8**
BERT	0.48	0.5	0.49	0.62
USE	0.36	0.37	0.36	0.48
SBERT	0.61	0.62	0.61	0.74
GloVe	0.34	0.38	0.35	0.46

The initial analysis of the similarity-based models reveals varied effectiveness across the metrics of Precision, Recall, F1 Score, and Accuracy (see Table 3). TF-IDF demonstrated the highest accuracy (0.80) and F1 Score (0.67), indicating its strength in handling straightforward keyword-matching queries. In contrast, BERT and SBERT, while achieving lower accuracy (0.62 and 0.74, respectively), exhibited competitive F1 Scores (0.49 and 0.61), reflecting their capability in understanding and processing context. However, models like USE and GloVe showed limited effectiveness with lower accuracy and F1 Scores, highlighting their struggles with complex language queries. A notable limitation of this study is the exclusion of other advanced generative models that could potentially enhance response quality further.

Diversity in performance underscores the importance of selecting the right model based on the specific needs of the helpdesk system, balancing between precision in retrieval and depth of language comprehension. Figure 2 illustrates this diversity by displaying the performance metrics of Precision, Recall, F1 Score, and Accuracy for each model alongside the processing time required. The graph clearly shows the trade-offs between accuracy and efficiency, highlighting how different models can cater to varying operational requirements. The comparative analysis provides a visual representation of each model's strengths and limitations, assisting in making an informed choice about the most appropriate model for UNNES's helpdesk operations based on specific query complexities and response time constraints.

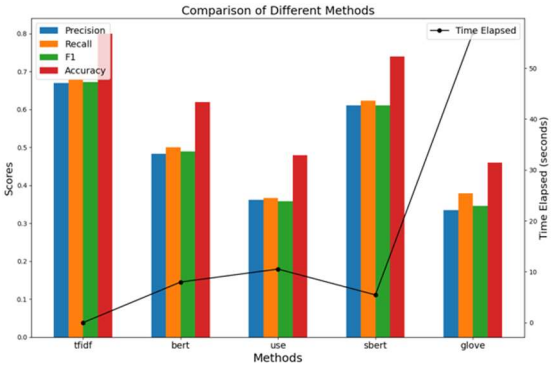

Figure 4 Comparison of Similarity-based methods

The graph presents a multi-dimensional comparison of the models TF-IDF, BERT, USE, SBERT, and GloVe. Each bar represents a different metric, Precision, Recall, F1 Score, and Accuracy which are critical for assessing the model's ability to accurately and reliably handle user queries. The black line graph overlay shows the time elapsed for processing, adding an additional layer of evaluation regarding the operational efficiency of each model.

TF-IDF shows high scores across all performance metrics and has the lowest processing time, making it the most efficient model in terms of both speed and accuracy. This suggests that while TF-IDF may lack some of the deeper contextual understanding capabilities of more advanced models, its speed and precision make it highly effective for tasks where response time is critical.

BERT and SBERT display moderate to high scores in F1 and Accuracy but at a cost of increased processing time. SBERT, in particular, shows a significant increase in time elapsed, which may be a consideration for real-time applications. Despite this, its high Accuracy and F1 Score indicate that it provides valuable deep semantic understanding, which could be crucial for complex query handling.

USE and GloVe are the least effective according to the metrics considered. Both show lower performance across the board and also take longer to process than TF-IDF. This indicates that these models might not be the best fit for environments where both accuracy and speed are required.

There is a noticeable trade-off between time efficiency and accuracy. Models that perform better in understanding the context of queries (SBERT)

require more processing time, which might be a critical factor depending on the application's requirements.

4.2 Determining Optimal Threshold Based on Similarity Scores

In the integration of the TF-IDF model with the Llama RAG generative model, determining an optimal threshold is critical for deciding when to transition from retrieval-based to generative responses. This threshold is strategically set based on the distribution of similarity scores for correctly identified answers by the TF-IDF model, ensuring an efficient and effective use of both models.

The determination of the threshold was grounded in a statistical analysis of the similarity scores obtained from the TF-IDF model's correctly guessed answers. Observations were compiled into a histogram to visually assess the distribution and identify a statistically significant cutoff point. The mean of the similarity scores, depicted by the red dashed line in Figure 3, was calculated at 0.81. To adopt a conservative yet effective approach in transitioning to the generative model, the threshold was set at one standard deviation below the mean, equating to 0.69, as illustrated by the green dashed line.

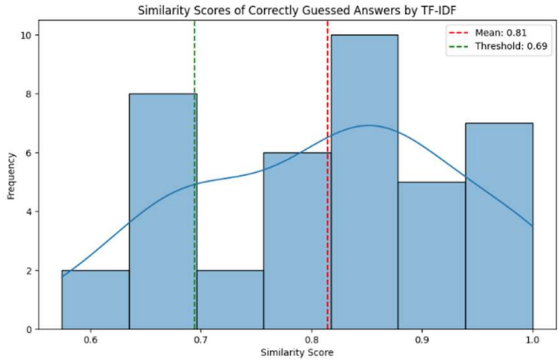

Figure 5 Histogram of Similarity Scores for Correctly Guessed Answers by TF-IDF

Setting the threshold at 0.69 strikes a deliberate balance, leveraging the rapid retrieval capabilities of TF-IDF for the majority of queries and reserving the sophisticated, computationally intensive generative capacity of the Llama RAG model for queries that fall below this confidence interval. This strategic placement ensures that the generative model is utilized in scenarios where its advanced processing capabilities are most needed, thereby optimizing resource use and maintaining response quality.

4.3 Development and Performance Evaluation of the Integrated TF-IDF and Llama RAG Model

The integration of the TF-IDF and Llama RAG models into our helpdesk system is designed to enhance both efficiency and the quality of responses. This combined model operates in a two-stage process:

1. Initial query handling, the TF-IDF model first processes incoming queries, quickly searching the FAQ database to find the most relevant existing answers based on similarity scores.
2. Advanced response generation, if the TF-IDF model's similarity score for the best-matching FAQ response falls below the threshold of 0.69, the query is then routed to the Llama RAG model. This advanced model generates contextually relevant responses by leveraging both retrieved content and dynamic embeddings.

The architecture strategically utilizes the generative capabilities of the Llama RAG model only when necessary, conserving computational resources while ensuring responses are both timely and contextually appropriate. Regular adjustments and monitoring are conducted to optimize the interaction between the retrieval and generative components. After implementation, a performance analysis was conducted to assess the system's effectiveness. The results are as follows:

- Total Responses: 50 queries processed.
- TF-IDF Responses: 39 responses (78%), demonstrating its ability to quickly and accurately handle the majority of queries.
- Llama RAG Responses: 11 responses (22%), employed for more complex queries that required nuanced answers.
- Average Similarity Score: 0.85 for TF-IDF responses, indicating high relevance and accuracy.
- BLEU Score: The generative responses from the Llama RAG model recorded a BLEU score of 0.61. This score reflects the linguistic quality and coherence of the generated responses compared to human-like references, underscoring the model's capability to produce contextually appropriate and fluent text.

While the metrics confirm the TF-IDF model's effectiveness in addressing straightforward queries and the Llama RAG model's utility in providing detailed, generative responses for more complex questions, the study has certain limitations. First, the dataset is primarily based on FAQs, which

may not cover all potential user queries. Second, the evaluation metrics, while comprehensive, do not fully capture user satisfaction and the long-term impact on helpdesk operations. Additionally, the integration of more advanced generative models like GPT-4 or other emerging technologies could provide further improvements in response quality and user engagement. Future work should explore the integration of these models and incorporate real-time user feedback mechanisms to continuously refine and enhance the chatbot system.

4.4 Reporting the Implementation Results at Universitas Negeri Semarang

The implementation of the integrated TF-IDF and Llama RAG model at Universitas Negeri Semarang has been rigorously evaluated based on user satisfaction and the efficiency in handling queries. This evaluation is crucial in assessing the impact of the model on user experience and operational effectiveness.

4.4.1 User Satisfaction Analysis

We collected user satisfaction data to gauge the effectiveness of the helpdesk system enhanced by the integrated model. The graph below represents the distribution of user satisfaction ratings.

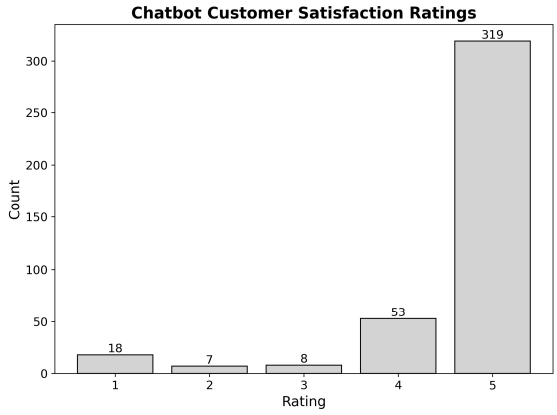

Figure 6 Chatbot Customer Satisfaction Ratings

A substantial majority of the users rated their experience as excellent, which signifies strong approval of the system's performance. Ratings of 4 and below, which are less frequent, suggest specific areas where improvements could potentially enhance user satisfaction further.

4.4.2 Operational Efficiency

Prior to the integration of the chatbot, the average chat duration displayed significant variability, with spikes indicating prolonged interactions on certain days (see Figure 2). The average chat durations varied widely, occasionally extending beyond several minutes, which often resulted in delayed responses and potential user dissatisfaction.

In contrast, after the chatbot was implemented, the average handling time for customer sessions was significantly reduced to approximately 1 minute and 42 seconds. This marked reduction in chat duration not only improved the efficiency of the helpdesk operations but also contributed positively to user satisfaction, as reflected in the high ratings.

The large amount of positive satisfaction ratings, along with the notably quick average handling time, clearly shows that the integrated TF-IDF and Llama RAG model has greatly enhanced the helpdesk's operations at Universitas Negeri Semarang. The successful rollout and favorable user feedback highlight the model's capability to manage a diverse array of inquiries effectively. This has not only improved interactions at the helpdesk but has also significantly boosted both the user experience and the overall efficiency of operations.

5. CONCLUSION

The integration of advanced NLP techniques and generative retrieval models at the UNNES helpdesk has significantly improved operational efficiency and user satisfaction. The TF-IDF model effectively handled 78% of queries, providing quick and relevant responses to common issues. For instance, when students inquired about the procedure for printing proof of UKT payment, the TF-IDF model retrieved accurate FAQ responses promptly. The Llama RAG model addressed the remaining 22%, generating detailed answers for complex inquiries such as procedural clarifications for specific administrative tasks. This combination ensured high relevance and linguistic coherence, as demonstrated by the average similarity score of 0.85 for every correctly guessed TF-IDF responses and a BLEU score of 0.61 for generative responses. This hybrid approach ensures timely and contextually appropriate responses, leading to higher user satisfaction and reduced handling times. Key findings from the implementation include:

1. High user satisfaction, the deployment of the chatbot has resulted in overwhelmingly positive user satisfaction ratings, with the majority of users rating their experience as excellent. This indicates that the chatbot's responses meet or exceed user expectations in terms of relevance and timeliness.
2. Reduced handling times, the average handling time of user interactions has been significantly

reduced. This improvement demonstrates the chatbot's ability to provide quick and accurate answers, which is essential for maintaining a high level of user satisfaction and operational efficiency.

3. Operational improvements, the introduction of the chatbot has stabilized the variability in interaction durations previously observed, leading to more predictable and efficient helpdesk operations.

These results highlight the significant impact of integrating advanced AI models in enhancing the efficiency of educational support services. By combining retrieval-based and generative models, this study demonstrates a comprehensive approach to addressing common and complex inquiries, thereby setting a precedent for future implementations in similar educational settings. The hybrid model's ability to handle a diverse range of queries not only improves immediate response times but also ensures the delivery of contextually appropriate information, which is crucial for user satisfaction.

However, this study raises several questions that warrant further investigation. For instance, how will the chatbot system integrate with other digital services within the university? What are the long-term impacts on user behavior and satisfaction? Additionally, how can ethical and privacy concerns be effectively addressed in the deployment of AI-driven systems in educational settings? These questions highlight areas for future research to further refine and expand the application of AI in higher education.

In conclusion, the implementation of the integrated TF-IDF and Llama RAG chatbot at Universitas Negeri Semarang represents a forward-thinking approach to enhancing educational support services. This project serves as a model for other institutions aiming to harness the power of AI to improve user engagement and service delivery in an academic environment.

REFERENCE

[1] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, p. 100006, Dec. 2020, doi: 10.1016/j.mlwa.2020.100006.

[2] J. Reis, M. Amorim, Y. Cohen, and M. Rodrigues, "Artificial Intelligence in Service Delivery Systems: A Systematic Literature Review," 2020, pp. 222–233. doi: 10.1007/978-3-030-45688-7_23.

[3] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results in Engineering*, vol. 17, p. 100856, Mar. 2023, doi: 10.1016/j.rineng.2022.100856.

[4] L. Yang *et al.*, "A Hybrid Retrieval-Generation Neural Conversation Model," *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, [Online]. Available: https://api.semanticscholar.org/CorpusID:126187120

[5] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *International Journal of Educational Technology in Higher Education*, vol. 20, Jun. 2023, doi: 10.1186/s41239-023-00426-1.

[6] S. Akinwalere and V. Ivanov, "Artificial Intelligence in Higher Education: Challenges and Opportunities," *Border Crossing*, vol. 12, pp. 1–15, Jun. 2022, doi: 10.33182/bc.v12i1.2015.

[7] P. F. Oliveira and P. Matos, "Introducing a Chatbot to the Web Portal of a Higher Education Institution to Enhance Student Interaction," in *ASEC 2023*, Basel Switzerland: MDPI, Dec. 2023, p. 128. doi: 10.3390/ASEC2023-16621.

[8] A. M. Alghaniy, "Pengaruh Teknologi Artificial Intelligence Pada Layanan Chatbot Shopee Terhadap Kepuasan Pelanggan di Bandung Raya, Indonesia," *International Journal Administration, Business & Organization*, vol. 5, no. 1, pp. 48–55, May 2024, doi: 10.61242/ijabo.24.337.

[9] F. Lan, "Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method," *Advances in Multimedia*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7923262.

[10] E. Aljohani, "Hybrid Feature-Driven Ensemble Learning In Arabic Nlp: Fusing Sequential Neural Networks With Advanced Text Analysis Techniques," *J Theor Appl Inf Technol*, vol. 102, no. 5, pp. 1686–1700, Mar. 2024.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019. [Online]. Available:

https://api.semanticscholar.org/CorpusID:52967399

[12] H. G. F. W. D. Y. Dengyun Zhu, "Semantic Similarity Caculating based on BERT," *Journal of Electrical Systems*, vol. 20, no. 2, pp. 73–79, Apr. 2024, doi: 10.52783/jes.1099.

[13] G. Deepthi and A. M. Sowjanya, "Query-Based Retrieval Using Universal Sentence Encoder," *Revue d'Intelligence Artificielle*, vol. 35, no. 4, pp. 301–306, Aug. 2021, doi: 10.18280/ria.350404.

[14] A. Kruspe, "A simple method for domain adaptation of sentence embeddings," *ArXiv*, vol. abs/2008.11228, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:221319659

[15] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of BERT and ALBERT Sentence Embedding Performance on Downstream NLP Tasks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, Jan. 2021, pp. 5482–5487. doi: 10.1109/ICPR48806.2021.9412102.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[17] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[18] A. Patel, "Word Embeddings for Banking Industry," *ArXiv*, vol. abs/2306.01807, 2023, doi: 10.48550/arXiv.2306.01807.

[19] W. X. Zhao *et al.*, "A Survey of Large Language Models," Mar. 2023.

[20] B. A. y Arcas, "Do Large Language Models Understand Us?," *Daedalus*, vol. 151, no. 2, pp. 183–197, May 2022, doi: 10.1162/daed_a_01909.

[21] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A Survey on Retrieval-Augmented Text Generation," Feb. 2022.

[22] Y. Mao *et al.*, "Generation-Augmented Retrieval for Open-Domain Question Answering," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 4089–4100. doi: 10.18653/v1/2021.acl-long.316.