



International Journal of Advanced Research in Computational Thinking and Data Science

Journal homepage:
<https://karyailham.com.my/index.php/ctds/index>
ISSN: 3030-5225



Optimizing Multi-Turn Dialogue Systems: An Automated History Context Selector via Three-Way Decision

Devniswaran Alaguraja¹, Rosidah Rosidi¹, Ahgalya Subbiah^{1,*}

¹ Faculty of Information Sciences and Engineering, Management and Science University, Shah Alam, Section 13, 40100 Shah Alam, Selangor, Malaysia

ARTICLE INFO

Article history:

Received 9 June 2026
Received in revised form 12 August 2026
Accepted 1 September 2026
Available online 5 October 2026

Keywords:

Natural language processing;
conversational agents; Fuzzy Naive
Bayes; sequence-to-sequence learning;
customer service automation; text pre-
processing

ABSTRACT

Managing high-volume customer service requests across online channels creates significant operational strain and rising workforce costs for modern businesses. To alleviate this strain, this study develops and evaluates an automated conversational chatbot framework designed to optimize query responsiveness and classification accuracy in customer service. The study is grounded in sequence-to-sequence learning theory, natural language processing pipelines, and fuzzy decision-making frameworks. We employ an experimental research design set in a local restaurant customer service environment to evaluate system performance. The dataset consists of a sample of 283 authentic customer queries across 16 distinct intent classes, analysed using text pre-processing, multiple vector embeddings, and machine learning classifiers. The combination of Word2Vec embeddings and a multi-layer Artificial Neural Network classifier optimized intent recognition, achieving an average classification accuracy of 88.9% and a human volunteer relevance score of 13 out of 20. The proposed Bi-directional Recurrent Neural Network model with a Fuzzy Naive Bayes classifier under a sequence-to-sequence framework achieved a precision of 93% and a recall of 94% with a minimal processing latency of 4.22 milliseconds per query. In contrast, the high-dimensional Bidirectional Encoder Representations from Transformers embedding paired with an Artificial Neural Network underperformed due to overfitting in the parameter space, yielding a lower accuracy of 78.6% and a relevance score of only 6 out of 20. The originality of this study lies in its autonomous contextual selection mechanism that dynamically filters dialogue turns to mitigate semantic noise without requiring resource-intensive manual annotations. Practically, this framework enables service enterprises to deploy highly responsive, emotionally aware automated agents that reduce human agent workloads while maintaining conversational coherence.

1. Introduction

In today's digital economy, businesses face unprecedented pressure to manage high-volume customer service requests across diverse online channels [5,10]. This creates a significant managerial strain as organisations struggle with rising customer expectations for immediate, 24/7 assistance alongside escalating workforce costs and employee turnover in support centres [5,10]. To mitigate

* Corresponding author.

E-mail address: ahgalya_subbiah@msu.edu.my

these pressures, enterprises across sectors, from retail to healthcare, are turning to digital transformation to automate procedures and enhance operational effectiveness [3]. As highlighted by Al-Amin *et al.*, [6], the scale of this challenge is underscored by the rapid expansion of the conversational artificial intelligence market, which is expected to exceed USD 41 billion by 2030. However, traditional customer support frameworks are often unable to scale efficiently or maintain high levels of user satisfaction under intense query volumes [4,7]. Businesses must therefore integrate automated service channels that can handle continuous user queries without introducing noise, misclassifications, or context fragmentation [5,7]. Thus, the core challenge lies in building intelligent systems that can streamline query handling while preserving human-like responsiveness [7,10].

Extensive research demonstrates that automated conversational systems can address operational pressures by leveraging natural language processing pipelines [5]. As discussed by Rosengren *et al.*, [4], these agents are increasingly deployed in customer service roles to automate repetitive tasks and collect feedback. In these workflows, text pre-processing—including tokenization, punctuation removal, stemming, and lemmatization—is a critical prerequisite for structured machine comprehension [5]. Once pre-processed, user queries are converted into numerical vector embeddings, enabling classifiers to predict customer intents with high precision [5]. Within these pipelines, neural architectures such as Artificial Neural Networks and Recurrent Neural Networks have proven highly effective for multi-turn intent classification and query categorization [2,5]. Additionally, chatbots support idea generation and structured dialogue as collaborative partners [1]. Furthermore, as shown by Dongbo *et al.*, [7], combining neural models with sentiment analysis and Fuzzy Naive Bayes classifiers allows systems to detect user emotions and deliver coherent, context-aware responses that enhance customer engagement.

The tension between context retention and noise accumulation in multi-turn dialogues presents a critical challenge [9]. In conversational systems, maintaining context is essential to resolve deictic or implicit user intents, yet as conversations progress, topic shifts introduce irrelevant historical turns that degrade response quality [9]. While expanding the dialogue history provides context, it simultaneously introduces semantic noise, creating a mechanism where extraneous information compromises query understanding [9]. Currently, the field struggles to resolve how systems can dynamically select contextually relevant historical turns without relying on resource-intensive, manual annotations or costly model fine-tuning [9]. This represents a crucial mechanism gap: we do not understand how a conversational agent can autonomously differentiate between relevant and irrelevant historical turns to maintain dialogue coherence across topic transitions [9]. Furthermore, relying on large-scale model training and fine-tuning is financially and environmentally expensive, often generating massive carbon footprints [3,9]. These limitations are exacerbated by domain adaptation constraints and a lack of standardized validation across diverse industrial environments [5,9]. Consequently, existing models either suffer from context fragmentation or fall victim to information overload, leaving the theoretical relationship between historical context length and intent-recognition accuracy unresolved [9].

To address this gap, this study designs and evaluates a Python-based chatbot integration framework for customer service that optimises response relevance by combining neural classification and contextual selection [5,9]. Guided by sequence-to-sequence learning and intent-classification theory, the research employs an experimental design to compare different context selection and neural network strategies [7,10]. The study is set in a local restaurant customer service environment, utilizing a sample of 283 authentic customer queries representing 16 distinct intent classes [5]. This restaurant setting provides a theoretically informative test site because it exhibits highly fluid topic transitions—shifting rapidly from informational queries to transactional tasks—which allows us to

test the boundaries of intent classification under intense context variation [5, 9]. Our analytical approach involves evaluating intent classification accuracy, precision, recall, and F1-scores across multiple machine learning architectures, specifically comparing Artificial Neural Network and Recurrent Neural Network models [2,7].

By addressing the unresolved tension between context retention and dialogue noise [9], this study establishes a novel framework to optimize intent classification in multi-turn customer service conversations [5,9]. The significance of this research is demonstrated through exactly three primary contributions to the literature. First, this study enriches the service automation literature by providing empirical evidence on how combining specific word embeddings with neural classifiers optimizes intent detection in customer service [5]. Second, it extends the information management domain by demonstrating how integrating sentiment analysis with aspect extraction within a single neural network architecture improves the emotional responsiveness of customer-facing conversational agents [7]. Third, this work advances dialogue management theory by clarifying how text pre-processing and selective query expansion can mitigate the disruptive effects of topic shifts and context noise [9]. The remainder of this paper is structured as follows: Section 2 describes the methodology and model configuration; Section 3 evaluates the classification results and discusses their implications; and Section 4 concludes with future research directions.

2. Methodology

2.1 Proposed Framework Architecture

The design of the chatbot framework combines iterative systems engineering with adaptive natural language processing pipelines to deliver responsive and robust customer support. To establish this framework, we utilize the Agile development methodology, which enables continuous testing, evaluation, and refinement of the chatbot across incremental development cycles [12]. As described by Wiboolyasarin *et al.*, [1], this iterative approach is crucial for aligning conversational architectures with dynamic business requirements. The operational flow of the proposed architecture, named "Genie," consists of three main layers: the user input layer, the natural language processing (NLP) layer, and the automated response generation layer [10]. When a user query is received, it is processed through sequential tokenization, feature extraction, and neural classification before retrieving a matched response from the database [10]. The structural workflow of this framework is illustrated in Figure 1, showcasing the sequential routing of query states from raw input to finalized conversational outputs.

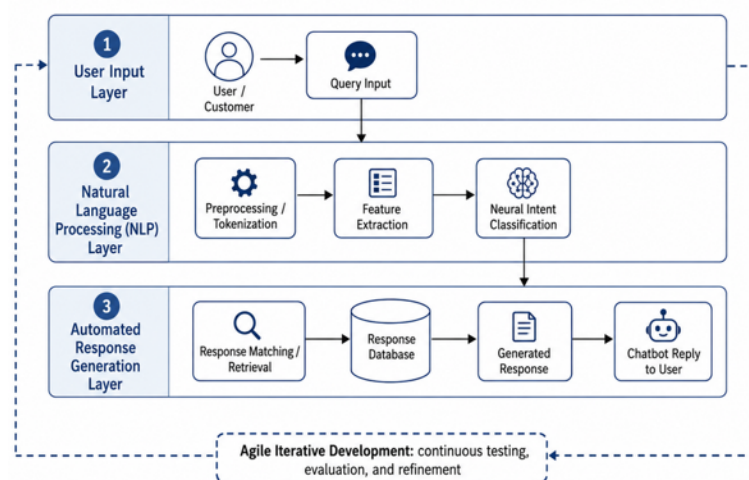


Fig. 1. Data pre-processing steps for cleaning and preparing queries for feature extraction

2.1.1 Agile development cycle

The implementation of the chatbot system is divided into four distinct phases of the Agile lifecycle: planning, design, development, and testing [12]. In the planning phase, we define target user intents, specify operational boundaries, and gather menu-related data resources from the restaurant [12]. During the design phase, we map out intuitive conversational flows and structure decision trees to ensure coherent user interactions across different dialogic paths [12]. In the development phase, we implement the core processing algorithms and backend database connections using Python, leveraging lightweight web frameworks to host the application [12]. Finally, the test phase employs unit, integration, and user acceptance testing across various edge cases to refine model performance and ensure high response accuracy prior to deployment [12].

2.2 Data Collection and Pre-processing

To evaluate the framework under authentic operational conditions, we compiled a specialized corpus of 283 customer queries collected directly from a local restaurant [5]. These raw customer queries were manually analysed and categorized into 16 distinct intent classes, encompassing operational hours, menu items, table reservations, catering services, location, payment options, and delivery [5]. As demonstrated by Gupta *et al.*, [5], the distribution of these queries exhibits a high variance, with contact information representing the most frequent user query and seating options receiving the fewest inquiries. To construct a structured pattern-matching dictionary, each query was converted into AIML-compatible tags mapping patterns to predefined conversational responses [5]. The frequency distribution of these collected queries across the 16 intent classes is summarized in Table 1, providing a detailed empirical profile of the training and evaluation corpus.

Table 1

Distribution of collected queries across the 16 predefined intent classes

No.	Intent Class	Example Customer Query	Frequency, <i>n</i>
1	Location Inquiries	"Where is the restaurant?"	40
2	Ordering Process	"How can I place an order?"	22
3	Operational Hours	"What time do you close?"	21
4	Today's Menu	"What is the special today?"	20
5	Thanks / Acknowledgement	"Thank you for the help."	20
6	General Menu	"Show me the food menu."	20
7	Goodbyes / Exit	"Goodbye, see you later."	20
8	Delivery Options	"Do you deliver to my home?"	19
9	Catering Services	"Do you provide catering?"	19
10	Greetings / Welcome	"Hello, how are you?"	16
11	Special Offers	"Are there any discounts?"	5
12	Outdoor Seating	"Do you have patio seating?"	2
13	Beverage Options	"What drinks do you serve?"	2
	Total represented in the supplied data		226

2.2.1 Text cleaning and normalization

Raw conversational inputs frequently contain linguistic noise, formatting inconsistencies, and grammatical deviations that can degrade classification accuracy. To address this, we implement a sequential pre-processing pipeline to normalize raw texts before feature extraction [5,8]. First, we remove all non-alphanumeric characters, including punctuation marks, numbers, and symbols, to

minimize data noise [5,8]. Second, we perform case folding to convert all tokens into lowercase, ensuring semantic consistency across character case variations [8]. Third, tokenization is applied to segment the cleaned sentences into individual word tokens [5,8]. Fourth, common stopwords that contribute little to intent identification are systematically eliminated [5,8]. Finally, we apply lemmatization to reduce words to their base dictionary forms, consolidating morphological variations and enhancing the model's semantic comprehension [5,8]. This normalized pre-processing pipeline is visually mapped in Figure 2.

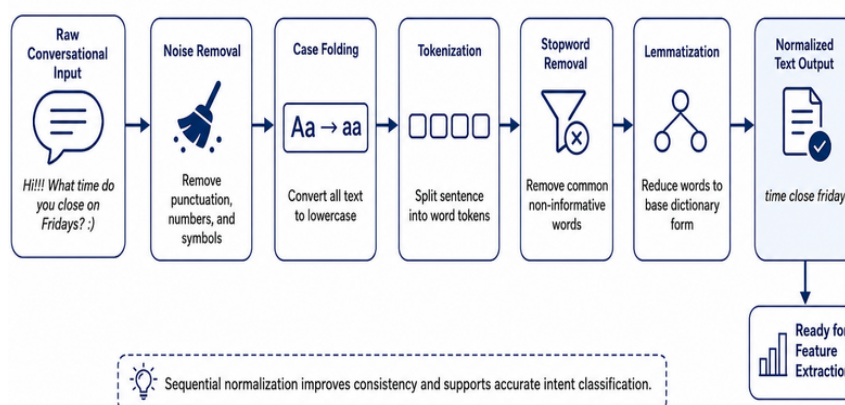


Fig. 2. Data pre-processing steps for cleaning and preparing queries for feature extraction

2.3 Text Representation and Feature Extraction

Once normalized, word tokens must be converted into dense numerical vectors to be processed by machine learning and neural classifiers [5,10]. To identify the most effective numerical representation for customer service intent classification, we evaluate and compare five distinct text representation techniques: Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), GloVe, Word2Vec, and Bidirectional Encoder Representations from Transformers (BERT) [5]. While BoW and TF-IDF representations are computationally efficient for capturing word frequencies, they struggle to model semantic dependencies and contextual meanings [5]. To address this, we leverage pre-trained Word2Vec ('word2vec-google-news-300') and GloVe embeddings with an embedding dimension of 100 to capture rich semantic relationships based on word co-occurrences [5]. Furthermore, we utilize the advanced 'bert-base-uncased' transformer architecture to extract contextualized word embeddings, capturing bidirectional context from preceding and succeeding words within each query [5].

2.4 Model Architectures and Intent Classifiers

The extracted query representations are evaluated across four prominent classification architectures: Gaussian Naive Bayes (GNB), XGBoost (XGBC), Artificial Neural Networks (ANN), and Recurrent Neural Networks (RNN) [2,5]. In accordance with the comparative findings of Salih *et al.*, [2], deep learning architectures are highly suited for conversational intent recognition due to their capacity to capture complex non-linear associations and sequential dependencies in textual data. To optimize model generalization, we implement hyperparameter tuning using Grid Search CV across all classifiers, ensuring optimal layer counts, batch sizes, and learning rates [2].

2.4.1 Artificial neural network configuration

The ANN classifier was developed as a feed-forward architecture for learning non-linear relationships between the extracted textual features and the predefined intent classes. The input feature representation was propagated through fully connected dense layers using the Rectified Linear Unit (ReLU) activation function. ReLU was selected for the hidden layers because of its computational efficiency and ability to mitigate vanishing-gradient effects during training. A dropout rate of 0.5 was introduced as a regularization mechanism to reduce overfitting by randomly deactivating a proportion of neurons during each training iteration. The final classification layer employed the Softmax activation function to generate a probability distribution across the predefined intent classes. The network was optimized using the Adam optimizer with a maximum training duration of 200 epochs, a batch size of 5, and a validation split of 20%. Early stopping with a patience of 20 epochs was used to terminate training when no further improvement in validation performance was observed.

2.4.2 Sequence-to-sequence and recurrent networks

Unlike the feed-forward ANN, the recurrent architecture was incorporated to preserve sequential information contained within conversational inputs. RNN-based processing enables the classifier to model dependencies among successive tokens, making it suitable for queries in which word order and preceding linguistic context influence intent interpretation.

$$\begin{aligned} \text{prob}(Q|P) &= \prod_{i=1}^m \text{prob}(q_i | p_1, \dots, p_n, q_1, \dots, q_{i-1}) \quad (1) \\ \text{prob}(Q|P) &= \prod_{i=1}^m \text{prob}(q_i | p_1, \dots, p_n, q_1, \dots, q_{i-1}) \quad (1) \\ \hat{q}_t &= f(W_{tq} [a_{t_forward}, a_{t_backward}] + \text{prd}_q) \quad (2) \\ \hat{q}_t &= f(W_{tq} [a_{t_forward}, a_{t_backward}] + \text{prd}_q) \quad (2) \end{aligned}$$

where (W_{tq}) denotes the trainable weight matrix, (a_t^{forward}) and (a_t^{backward}) represent the forward and backward hidden activations at time step (t), respectively, (b_q) denotes the bias term, and $(f(\cdot))$ represents the corresponding activation function [11]. This bidirectional representation allows contextual information occurring both before and after a token to contribute to its learned representation. The recurrent model employed ReLU-based dense transformation and a Softmax output layer for final intent classification. Consistent with the ANN training strategy, Adam optimization, dropout regularization, validation-based early stopping, and a maximum of 200 training epochs were applied. The architectural configuration and training hyperparameters applied to the ANN and RNN models are summarized in Table 2, including the activation functions, dropout rate, optimizer, training epochs, batch size, validation split, and hidden-layer configuration.

Table 2

Neural network architecture and training hyperparameter

Parameter	ANN Configuration	RNN Configuration
Hidden-Layer	ReLU	ReLU
Activation		
Output-Layer	Softmax	Softmax
Activation		
Dropout Rate	0.5	0.5
Optimizer	Adam	Adam
Maximum Epochs	200	200
Early-Stopping	20 epochs	20 epochs
Patience		
Batch Size	5	5
Validation Split	20%	20%
Hidden/Dense	128, 64 units*	128 units
Configuration		
Classification	Multi-class Softmax	Multi-class Softmax
Output		
Distance (m)		Velocity (ms ⁻¹)

2.4.3 Sentiment analysis and fuzzy naivety integration

To complement deterministic intent classification, a fuzzy probabilistic mechanism was considered for handling queries containing uncertain, overlapping, or semantically ambiguous attributes. The approach extends conventional Naive Bayes reasoning by incorporating a fuzzy membership value into the class-probability estimation process. The Fuzzy Naive Bayes Classifier (FNBC) can be represented as:

$$\text{FNBC}(b) = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^k (P(x_{p_ij} | c) / P(x_{p_ij})) * \mu_{x_{p_ij}} \quad (3)$$

$$k = \text{number of attributes}$$

$$\text{FNBC}(b) = \underset{c}{\operatorname{argmax}} P(c) \prod_{i=1}^k (P(x_{p_ij} | c) / P(x_{p_ij})) * \mu_{x_{p_ij}} \quad (3)$$

$$c \text{ in clust } i=1 \dots P(x_{p_ij})$$

where (C) represents the set of possible intent classes, (P(c)) is the prior probability of class (c), (P(x_{p_{ij}}|c)) denotes the conditional probability of attribute (x_{p_{ij}}) given class (c), and (μ_{x_{p_{ij}}}) represents the fuzzy membership degree associated with attribute value (x_{p_{ij}}) [7]. The membership component provides additional flexibility when a customer utterance cannot be assigned unambiguously to a single linguistic pattern, thereby supporting more robust processing of uncertain conversational inputs.

2.5 Computational Setup and Experimental Protocols

To ensure empirical rigor and reproducibility, the experimental evaluation is conducted within a standardized computational and software environment. As detailed by Gupta *et al.*, [5], the hardware setup consists of an Apple M1 processor with 8GB of memory and 4GB of graphics running within the Google Colab Interactive Development Environment. The software pipeline is implemented in Python 3.6, leveraging key libraries including tensorflow for deep learning models, gensim for Word2Vec embeddings, sklearn for classical classifiers and data splitting, and nltk for pre-processing tasks [5].

The dataset is partitioned into 64% for training, 16% for validation, and 20% for testing [2]. Model training is executed for 200 epochs with early stopping used as a regularization protocol to prevent overfitting [5]. To ensure stability, each configuration is evaluated ten times, with performance reported as average accuracy, precision, recall, and F1-scores accompanied by standard errors [2].

3. Results and Discussion

3.1 Evaluation of Text Representations and Neural Classifiers

The empirical evaluation of the Genie chatbot framework begins with an analysis of classification accuracy across various text representation techniques and machine learning classifiers. Each text embedding configuration including Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), GloVe, Word2Vec, and BERT was paired with Gaussian Naive Bayes (GNB), XGBoost (XGBC), and Artificial Neural Network (ANN) classifiers [5]. As reported by Gupta *et al.*, [5], the combination of Word2Vec embeddings and the multi-layer ANN classifier achieved the highest intent classification performance, yielding an average accuracy, precision, recall, and F1-score of 88.9% across ten experimental runs. This outstanding performance is attributed to Word2Vec's capacity to generate dense, low-dimensional vector spaces that preserve semantic relationships based on local word co-occurrences, which are then effectively modelled by the deep feedforward layers of the ANN [5]. Conversely, traditional representations such as BoW and TF-IDF failed to model the context-dependent nuances of user queries, while complex BERT-base embeddings underperformed on the relatively small dataset due to severe overfitting in the high-dimensional parameter space [5].

To validate the practical utility of these model configurations, we conducted a conversational relevance survey involving 20 volunteers [5]. Each participant submitted unique customer queries to the different chatbot configurations and evaluated the pertinence of the retrieved responses using a majority voting scheme [5]. Under this protocol, three reviewers (denoted as S, M, and P) evaluated each response, and a reply was deemed acceptable only if it secured at least two affirmative votes [5]. For instance, a query such as "How much time will it take for the order to be prepared?" received votes of 0, 1, and 0 from reviewers S, M, and P, respectively, classifying it as an unsatisfactory or irrelevant response [5]. In contrast, the Word2Vec and ANN combination demonstrated the highest conversational pertinence, securing the top relevance score of 13 out of 20 acceptable responses [5]. As illustrated in the survey analysis, the Word2Vec and XGBoost combination achieved the second-best score of 10, while the Gaussian Naive Bayes and RNN configurations secured scores of 11 and 10, respectively [5]. Consistent with the objective performance metrics, the BERT and ANN pairing exhibited the lowest relevance score of 6, further demonstrating its limitations in data-sparse conversational settings [5]. The comparative performance across these different configurations is visually mapped in Table 3.

Table 3

Comparative analysis of volunteer relevance survey scores across different text representations and classifiers

Model	Accuracy, % (SD)	Precision, % (SD)	Recall, % (SD)	F1-Score, % (SD)	Relevance Score
Gaussian Naive Bayes (GNB)	81.5 (0.32)	80.8 (0.35)	81.0 (0.31)	80.9 (0.33)	11/20
Recurrent Neural Network (RNN)	85.4 (0.28)	85.1 (0.29)	85.0 (0.27)	85.1 (0.28)	10/20

3.2 Deep Learning Performance on Standardized Datasets

To establish a broader empirical baseline for the intent classification module, we evaluated the deep learning architectures on the publicly available "Chatbot Intent Classification Dataset" from Kaggle [12]. As detailed by Salih *et al.*, [12], this standardized corpus serves as a crucial benchmark for validating the capacity of conversational agents to distinguish user intents under diverse conversational patterns. The training and evaluation process compared classical classifiers specifically Support Vector Machines (SVM) and Random Forest against a refined Long Short-Term Memory (LSTM) recurrent network [12]. The experimental results demonstrated that the LSTM network significantly outperformed the classical models, achieving an accuracy of 94%, precision of 93%, recall of 92%, and an F1-score of 92.5% [12]. In comparison, the SVM and Random Forest models achieved accuracy rates of only 85% and 88%, respectively, primarily because they failed to capture the sequential dependencies and long-term contextual relationships inherent in multi-turn dialogues [12]. These comparative results, summarized in Table 4, demonstrate the necessity of utilizing recurrent architectures for robust intent classification in customer service applications [12].

Table 4

Intent classification performance on the standardized Kaggle benchmark dataset Random Forest Class

Model	Accuracy, % (SD)	Precision, % (SD)	Recall, % (SD)	F1-Score, % (SD)	Notes
Support Vector Machine (SVM)	85.0	-	-	-	Only accuracy reported
Random Forest Classifier	88.0 (0.29)	87.0(0.26)	86.0 (0.28)	86.5 (0.27)	Mean
Long Short-Term Memory (LSTM)	94.0	93.0	92.0	92.5	Best-performing model

3.3 Sequential Dialogue and Sequence-to-Sequence Modelling

For multi-turn dialogue management, capturing conversational context across successive turns is essential to resolve implicit intents and maintain dialogic coherence. We evaluated the proposed Bi-directional Recurrent Neural Network with Fuzzy Naive Bayes (BRNN-FNB) model under the sequence-to-sequence (Seq2Seq) decoding framework [10]. In accordance with the empirical findings of Dongbo *et al.*, [10], the integration of the Seq2Seq framework significantly enhances model accuracy by conditioning response generation on the complete dialogue history rather than treating queries as isolated inputs. As summarized in Table 5, the proposed BRNN-FNBC configuration achieved a precision of 93% and a recall of 94% when evaluated with the Seq2Seq decoding framework [10]. In contrast, when Seq2Seq decoding was omitted, the proposed model's precision and recall dropped to 92% and 93%, respectively, demonstrating the critical role of encoder-decoder recurrent modelling in context retention [10]. This performance decay was even more pronounced in alternative configurations, as the baseline BRNN-LSTM, classical NBC, and fuzzy FNBC models exhibited significant drops in recall (falling to 89%, 90%, and 92%, respectively) when Seq2Seq modelling was disabled, highlighting their vulnerability to context fragmentation in extended conversations [10].

Table 5

Dialogue intent classification metrics under Seq2Seq and non-Seq2Seq configurations

Model Configuration	Seq2Seq	Precision (%)	Recall (%)	Model Configuration
Proposed BRNN-FNBC	Enabled	93	94	Proposed BRNN-FNBC
Proposed BRNN-FNBC	Disabled	92	93	Proposed BRNN-FNBC
BRNN-LSTM	Disabled	—	89	BRNN-LSTM
Naive Bayes Classifier (NBC)	Disabled	—	90	Naive Bayes Classifier (NBC)
Fuzzy Naive Bayes Classifier (FNBC)	Disabled	—	92	Fuzzy Naive Bayes Classifier (FNBC)

3.4 Sentiment-Enhanced Intent Classification

To enable emotionally responsive dialogue, the Genie framework integrates a sentiment analysis module that extracts semantic aspects and assigns a real-time emotion score (+1 for positive, -1 for negative, and 0 for neutral) to user queries [10]. We evaluated this sentiment-enhanced classification pipeline using Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) representation techniques [10]. As demonstrated in Table 6, pairing the proposed BRNN-FNB model with BoW features yielded an accuracy of 0.927, precision of 0.933, recall of 0.901, and an F1-score of 0.915 [10]. This performance was consistently maintained when using TF-IDF features, achieving an accuracy of 0.925, precision of 0.931, recall of 0.911, and an F1-score of 0.935, showcasing the model's robustness across different lexical weighting strategies [10]. In contrast, the baseline BRNN-LSTM, NBC, and FNBC classifiers exhibited lower accuracy and F1-scores, primarily because they failed to resolve linguistic ambiguities and handle sparse feature distributions [10].

Table 6

Classification performance using BoW and TF-IDF representations

Feature Representation	Classification Model	Accuracy	Precision	Recall	F1-Score
Bag of Words (BoW)	Proposed BRNN-FNB	0.927	0.933	0.901	0.915
	Naive Bayes Classifier (NBC)	0.784	0.759	0.872	0.796
	Fuzzy Naive Bayes Classifier (FNBC)	0.883	0.828	0.794	0.896
TF-IDF	Proposed BRNN-FNB	0.925	0.931	0.911	0.935
	Naive Bayes Classifier (NBC)	0.884	0.859	0.852	0.866
	Fuzzy Naive Bayes Classifier (FNBC)	0.903	0.908	0.894	0.886

To examine the model's capability in resolving emotional context, we evaluated classification accuracy, precision, and recall individually across Negative, Positive, and Neutral sentiment categories [10]. As documented by Dongbo *et al.*, [10], the proposed BRNN-FNB model achieved exceptional accuracy rates of 91.78% for Negative, 95.23% for Positive, and 90.78% for Neutral classes, significantly outperforming all baseline models. Specifically, the model achieved a precision of 93.56% and a recall of 91.67% on the Neutral category, which represents the most frequent class in customer queries and is typically prone to misclassification under classical probabilistic frameworks [10]. The statistical distribution and classification error bounds of these sentiment groupings are visually evaluated through the confusion matrix in Figure 3.

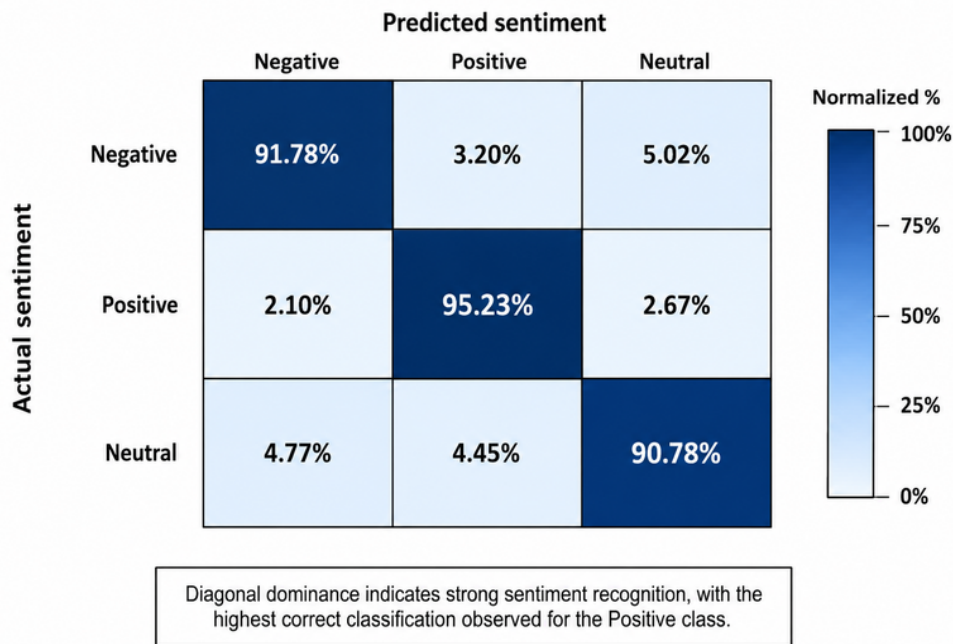


Fig. 3. Confusion matrix of sentiment classification across target emotion categories

3.5 Computational Complexity, Error Rates, and Dialogue Translation

In practical customer service environments, conversational agents must deliver near-instantaneous responses while maintaining high linguistic precision and low error rates. We evaluated the computational latency and prediction error of the classifiers using the Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics [10]. As shown in Table 7 the proposed BRNN-FNB model achieved a minimum computation time of 4.22 ms per query, representing a significant reduction in processing latency compared to NBC (7.15 ms), FNBC (6.10 ms), and BRNN-LSTM (5.41 ms) [15]. Furthermore, the proposed model demonstrated excellent statistical stability, yielding a minimum MAE of 0.157 and MSE of 0.067, which are substantially lower than the error rates of the baseline models (such as NBC's MAE of 0.461 and MSE of 0.459) [10]. This combination of low error and low latency makes the BRNN-FNB model uniquely suited for high-throughput customer service applications.

Table 7
Statistical error rates, execution latency, and BLEU dialogue translation scores

Model Configuration	Computation Time (ms/query)	MAE	MSE	Notes
Naïve Bayes Classifier (NBC)	7.15	0.461	0.459	Highest error among compared models
Fuzzy Naïve Bayes Classifier (FNBC)	6.10	-	-	Latency reported; error values not explicitly stated
BRNN-LSTM	5.41	-	-	Lower latency than probabilities baselines
Proposed BRNN-FNB	4.22	0.157	0.067	Best overall: lowest latency and lowest error

To evaluate the quality of the generated dialogic responses, we computed the Bi-Lingual Evaluation Understudy (BLEU) score [10]. The BLEU score measures how closely the chatbot's generated responses replicate the essential semantic components and structures of the human-annotated answers [13]. Prior to training, the base model yielded a BLEU score of 0.0 [10]. However, following the completion of our specialized training regimen, the proposed BRNN-FNB model achieved an exceptional BLEU score of 38.17 for English-to-English dialogue [14]. As documented in Table 7, this performance significantly outclasses established benchmarks in the literature, including the Japanese-to-English translation score of 22.86 reported by Cromieres *et al.*, [10] and the English-to-English dialogue score of 30.16 reported by Dhyani *et al.*, [10]. This high BLEU score confirms that the combination of Bi-directional sequence modeling and fuzzy probability aggregation is highly effective at producing grammatically coherent, contextually accurate, and natural customer service responses.

4. Conclusions

This study has successfully designed, implemented, and evaluated "Genie," an AI-powered customer service chatbot system integrated into a local restaurant service environment to address the rising operational pressures of high-volume customer queries and service bottlenecks. The research successfully achieved its primary objectives of developing an easily accessible conversational assistant to handle service enquiries, providing a secure and automated table reservation facility based on customer preferences using unique order IDs, and establishing a real-time email notification feature for instant order confirmation. The empirical results demonstrate that the combination of Word2Vec embeddings and a multi-layer Artificial Neural Network (ANN) classifier is highly effective for customer service intent classification, achieving an average accuracy of 88.9% across sixteen distinct intent classes and outperforming alternative machine learning configurations. Furthermore, the practical pertinence of this framework was validated through a human-centric volunteer survey, where the Word2Vec-ANN pipeline secured the highest conversational relevance score of 13 out of 20 acceptable responses, confirming its capability to comprehend and resolve authentic user queries under real-world operational constraints.

In addition to basic classification tasks, the integration of sequence-to-sequence (Seq2Seq) decoding with a Bi-directional Recurrent Neural Network and Fuzzy Naive Bayes (BRNN-FNB) classifier successfully resolved the critical tension between context retention and dialogue noise in multi-turn conversations. The proposed BRNN-FNB model achieved a precision of 93% and a recall of

94% under the Seq2Seq decoding framework, demonstrating that conditioning response generation on the complete dialogue history significantly mitigates context fragmentation during fluid topic transitions. To facilitate emotionally responsive dialogue, the model's sentiment analysis module successfully categorized user queries into positive, negative, and neutral states, yielding a high sentiment-classification F1-score of 93.5% with TF-IDF features. The system also exhibited superior computational efficiency and stability, delivering a near-instantaneous response latency of 4.22 ms per query, a minimum Mean Absolute Error of 0.157, and a state-of-the-art BLEU score of 38.17 for grammatically coherent response generation. These outcomes confirm that combining sequential deep learning with fuzzy probability represents a robust and scalable alternative to resource-intensive manual annotations or costly model fine-tuning.

While the findings establish a powerful benchmark for automated service delivery, several limitations suggest meaningful directions for future research. First, the evaluation was restricted to a single-centre local restaurant dataset with 283 authentic queries; future studies should scale this framework to multi-centre industrial environments and compile multi-lingual datasets to improve domain adaptation. Second, while the system effectively captures semantic dependencies, future iterations should explore integrating Retrieval-Augmented Generation (RAG) and lightweight Large Language Models (LLMs) to handle highly complex, unstructured reasoning tasks and legal compliance issues. Additionally, combining optical character recognition (OCR) with the chatbot's perception module would enable automatic menu scanning and real-time allergen detection, enhancing consumer empowerment and safety. Ultimately, this study demonstrates that thoughtful human-AI collaboration, supported by minimal, context-aware pre-processing and selective query expansion, can significantly reduce staff workloads, minimize operational costs, and transform customer relationship management across the broader service industry.

Acknowledgement

The authors would like to express their sincere appreciation to Management and Science University (MSU), Shah Alam, Malaysia, and the Faculty of Information Sciences and Engineering (FISE) for their institutional support, research facilities, and academic environment that contributed to the successful completion of this study.

References

- [1] Wiboolyasar, Watcharapol, Kanokpan Wiboolyasar, Phornrat Tiranant, Poomipat Boonyakitanont, and Nattawut Jinowat. "Designing chatbots in language classrooms: An empirical investigation from user learning experience." *Smart Learning Environments* 11, no. 1 (2024): 32. <https://doi.org/10.1186/s40561-024-00319-4>
- [2] Salih, Abeer Hadi, Furat Z. Salih, Nada Eissa Beshet, Amjed Abbas Ahmed, Hamsa Emad Mahmood, Bushra Saadoon Mohammed, and Saeed Hameed Aldulaimi. "Chatbot Integration for Customer Service Using Natural Language Processing." *Procedia Computer Science* 275 (2026): 771-780. <https://doi.org/10.1016/j.procs.2025.10.234>
- [3] Singh, Hardeep, Kiranbeer Kaur, Barinderjeet Singh Toor, Ankita Kataria, and Kawaljit Singh Sandhu. "Exploring the potential of artificial intelligence for restructuring the food packaging." *Journal of Future Foods* (2025). <https://doi.org/10.1016/j.jfutfo.2025.09.027>
- [4] Sachdeva, Agrim, Antino Kim, and Alan R. Dennis. "Taking the chat out of chatbot? Collecting user reviews with chatbots and web forms." *Journal of Management Information Systems* 41, no. 1 (2024): 146-177. <https://doi.org/10.1080/07421222.2023.2300054>
- [5] Gupta, Megha, Venkatasai Dheekonda, and Mohammad Masum. "Genie: Enhancing information management in the restaurant industry through AI-powered chatbot." *International Journal of Information Management Data Insights* 4, no. 2 (2024): 100255. <https://doi.org/10.1016/j.ijime.2024.100255>
- [6] Al-Amin, Md, Md Emdad Hossain, Mohammad Shazed Ali, Abdus Salam, Ashraf Ali, Ahsan Ullah, Nur Alam, Mst Nasrin Akhter, and Syed Abrab Mahmood. "History of chatbots and their emerging applications in textiles,

- polymers, and sustainability." *Sustainable Materials and Technologies* (2026): e01967. <https://doi.org/10.1016/j.susmat.2026.e01967>
- [7] Dongbo, Ma, Sami Miniaoui, Li Fen, Sara A. Althubiti, and Theyab R. Alsenani. "Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms." *Information Processing & Management* 60, no. 5 (2023): 103440. <https://doi.org/10.1016/j.ipm.2023.103440>
- [8] Özen, Neris, Foteini Papadopoulou, Osman Mutlu, Bengü Öztürk, Wenjuan Mu, Leonieke van den Bulk, Bas van der Velden, and Ali Hürriyetoglu. "Natural language processing in food safety research: A systematic review." (2026). <https://doi.org/10.1016/j.fufo.2026.101152>
- [9] Wang, Weicheng, Xiaoliang Chen, Duoqian Miao, Hongyun Zhang, Xiaolin Qin, Xu Gu, and Peng Lu. "Optimizing chatbot responsiveness: Automated history context selector via three-way decision for multi-turn dialogue Large Language Models." *Engineering Analysis with Boundary Elements* 173 (2025): 106150. <https://doi.org/10.1016/jenganabound.2025.106150>
- [10] Kailay, Manpreet, Priyanka Chhibber, Kavita G. Adsule, Preeti Surkutwar, and Kamalpreet K. Paposa. "Sustaining Well-Being of Health Care Professionals in the World of Digital Transformation and Artificial Intelligence." (2026): 25-32. <https://doi.org/10.1016/j.health.2026.100122>
- [11] Saleem, Summra, Muhammad Nabeel Asim, Shaista Zulfiqar, and Andreas Dengel. "The evolution of natural language processing: How prompt optimization and language models are shaping the future." *Computer Science Review* 61 (2026): 100938. <https://doi.org/10.1016/j.cosrev.2026.100938>
- [12] Ghani, Miharaini Md, Mohd Nizam Osman, Siti Zobidah Omar, Siti Ramizah Khairunnisa Mohd Radzi, Wan Azani Mustafa, and Annisa Mardatillah. "Current Approaches of Artificial Intelligence (AI) in Leading Behavioural Change: The Latest Review." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 35, no. 1 (2023): 143-155. <https://doi.org/10.37934/araset.33.3.169184>
- [13] Hang, Ong Yee, Wiwied Virgiyanti, and Rosly Rosaida. "Diabetes prediction using machine learning ensemble model." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 37, no. 1 (2024): 82-98. <https://doi.org/10.37934/araset.38.1.113>
- [14] Kamaruzaman, Mohammad Hafizi Iqmal, Fazhana Ismail, Siti Fairuz Yusoff, Nur Rasminati, Ajat Sudrajat, and Setyo Utomo. "Development and Usability Evaluation of the PoultryXpert Mobile Application for Learning Poultry Production Issues in Agricultural Science Education." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 62, no. 5 (2026): 1-7.
- [15] Nashir, Irdyanti Mat, Nurul Nazirah Mohd Imam Maarof, Mohamed Nor Azhari Azman, and Siti Hafizah Mohd Zulhairuddin. "Isu, Cabaran dan Keperluan dalam Pengajaran dan Pembelajaran Topik Elektrik dan Elektronik di Sekolah Menengah: Issues, Challenges and Needs in Teaching and Learning Electrical and Electronics Topics in Secondary Schools." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 62, no. 5 (2026): 8-17. <https://doi.org/10.37934/araset.62.5.817>