https://doi.org/10.31436/iiumej.v27i3.4191

A Hybrid Framework for Real-Time Sentiment Analysis and Adaptive Conversational Outcome Prediction in Domain-Specific Dialogue Systems

MEJDL SAFRAN*

Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

**Corresponding author: mejdl@ksu.edu.sa*

(Received: 04 February 2026; Accepted: 28 August 2026; Published online: 11 September 2026)

ABSTRACT: The lack of non-verbal signals in text-based communication forces dialogue systems to rely entirely on dynamic textual sentiment changes to infer user intent and emotional shifts. However, substantial class imbalance and the static, post-hoc nature of existing text analysis models prevent dialogue systems from recognizing these emotional dynamics and predicting conversational outcomes in real time. This research presents a hybrid machine learning framework to address these challenges through integrated sentiment-aware analysis and real-time conversation outcome prediction. We first utilize a pre-trained RoBERTa model to assign sentiment scores to individual messages, which is required for two subsequent tasks: (i) conversational outcome classification to classify conversations into Interested in Islam, Wants to Convert, and Accepted Islam, and (ii) message impact prediction to assess the emotional influence of agents' responses before deployment. In the initial task, our Enhanced Bi-LSTM model (enhanced with multi-head attention mechanisms and a boosted input capacity of 2500 tokens) achieves a macro-F1 score of 0.76, a balanced accuracy of 74.3%, and an overall accuracy of 90.0% on the full, unbalanced dataset when addressing the 3-class outcome classification task. By conducting an ablation study, we investigate the main source of error, the extremely ambiguous "Wants to Convert" class, and show the resilience of the model on well-defined classes, revealing that accuracy may exceed 99% by simplifying the task to a binary classification. In the second task, a fine-tuned BERT model achieves 80% accuracy in message-level sentiment prediction. The framework enables real-time feedback to refine adaptive responses and guide conversations toward desired outcomes. We assess our proposed framework using a case study of domain-specific conversations (such as religious outreach), supported by a dataset from a partner organization, to demonstrate applicability. The framework's flexibility makes it applicable to customer service, education, and healthcare, where effective communication depends on real-time sentiment analysis and outcome prediction.

ABSTRAK: Kekurangan isyarat bukan lisan dalam komunikasi berasaskan teks memaksa sistem dialog bergantung sepenuhnya pada perubahan sentimen tekstual dinamik membuat kesimpulan tentang niat pengguna dan perubahan emosi. Walau bagaimanapun, ketidakseimbangan kelas yang ketara dan sifat statik dan pasca-hoc model analisis teks sedia ada menghalang sistem dialog daripada mengenali dinamik emosi ini dan meramalkan hasil perbualan pada masa nyata. Kajian ini membentang rangka kerja pembelajaran mesin hibrid bagi menangani cabaran melalui analisis kesedaran sentimen bersepadu dan ramalan hasil perbualan masa nyata. Pada permulaan, model RoBERTa dilatih terlebih dahulu bagi memberi skor sentimen kepada mesej individu. Ini diperlukan untuk dua tugasan seterusnya: (i) klasifikasi hasil perbualan kepada Berminat tentang Islam, Ingin Beranut Islam dan Menerima Islam, dan (ii) ramalan impak mesej bagi menilai pengaruh emosi respons ejen sebelum penggunaan. Melalui tugasan awal, dapatan model Peningkatan Bi-LSTM

(dipertingkatkan dengan mekanisme perhatian berbilang kepala dan peningkatan kapasiti input sebanyak 2500 token) mencapai skor makro-F1 sebanyak 0.76, ketepatan seimbang sebanyak 74.3%, dan ketepatan keseluruhan sebanyak 90.0% pada set data penuh tidak seimbang apabila menangani tugasan pengelasan hasil 3 kelas. Dengan menjalankan kajian ablasi, kami menyiasat punca ralat utama, kelas "Ingin Menukar" yang sangat samar-samar, dan menunjukkan daya tahan model pada kelas yang ditakrif baik, mendedahkan bahawa ketepatan mungkin melebihi 99% dengan memudahkan tugasan kepada pengelasan binari. Menerusi tugasan kedua, model BERT yang ditala halus mencapai ketepatan 80% dalam ramalan sentimen peringkat mesej. Rangka kerja ini memudahkan maklum balas masa nyata bagi memperhalusi respons adaptif dan membimbing perbualan ke arah hasil yang diingini. Kami menilai rangka kerja yang dicadangkan menggunakan kajian kes yang mengandungi perbualan khusus domain (seperti jangkauan keagamaan). Ini disokong oleh set data daripada organisasi rakan kongsi, bagi menunjukkan kebolehgunaan. Fleksibiliti rangka kerja ini boleh digunakan untuk khidmat pelanggan, pendidikan dan penjagaan kesihatan, yang mana komunikasi berkesan bergantung pada analisis sentimen masa nyata dan ramalan hasil.

KEYWORDS: *Sentiment Analysis; Adaptive Conversational Outcome Prediction; Dialogue Systems; RoBERTa; Bi-LSTM.*

1. INTRODUCTION

Text-based dialogue systems are becoming crucial in areas that demand subtle, immediate adjustments to user emotions and intent, spanning from healthcare advice to customer support. These systems exploit advances in deep learning for sentiment analysis, such as models that analyze medical records to determine treatment quality, with high accuracy in classifying unstructured data [1]. These systems depend entirely on monitoring the temporal evolution of message sentiment to understand emotional state and assess interaction evolution, as text-based communication lacks non-verbal information (e.g., tone, gestures). Although sentiment analysis and conversational classification have been explored in static contexts, these techniques remain largely unintegrated into dynamic, real-time systems [2], especially in challenging situations where emotional shifts affect conversational outcomes.

The effectiveness of text-based conversational agents depends on their ability to understand emotional cues and adapt their responses in real time. Traditional approaches such as lexicon-based sentiment analysis struggle to capture context changes in multi-turn dialogues. Transformer-based models (such as BERT [3] and RoBERTa [4]) and sequential models (such as Bi-LSTM [5]) have their own merits but are seldom integrated end-to-end for dual-task optimization. Other problems, such as class imbalance and preprocessing latency, further plague adaptive dialogue systems, and no special datasets exist for outcome-driven dialogue.

In order to overcome these challenges, we propose a hybrid machine learning framework that combines three components: (1) a pre-trained RoBERTa model to assign sentiment scores to individual messages in dialogue conversations, utilizing its robust contextual embeddings; (2) an enhanced Bi-LSTM to classify conversations into Interested in Islam, Wants to Convert, and Accepted Islam by modeling sequential dependencies; and (3) a fine-tuned BERT model to predict the emotional impact of unsent responses, allowing for proactive adjustments. To ensure resilience in dynamic conversational scenarios, we evaluate this approach using metrics such as accuracy, precision, recall, F1-score, balanced accuracy, and macro F1-score.

The main contributions of this research are outlined as follows:

- We present a dawah conversation dataset of labeled conversations (Interested in Islam, Wants to Convert, and Accepted Islam), in which each conversation's message is assigned a sentiment score (positive/negative).
- We propose a hybrid machine learning framework that combines an Enhanced Bi-LSTM architecture for conversation outcome classification (which is enhanced with multi-head attention layers and an increased token limit) and a fine-tuned BERT for response impact prediction. The Enhanced Bi-LSTM achieves 90% accuracy in a challenging 3-class outcome classification, and BERT reaches 80% accuracy in response impact prediction.
- We comprehensively evaluate the proposed framework using accuracy, precision, recall, F1-score, balanced accuracy, and macro F1-score, along with data validation that includes statistical tests and confusion matrices.
- We validate the framework using a domain-specific use case (persuasive conversations), demonstrating its applicability to scenarios requiring sentiment-aware interaction management.

The paper is structured as follows: Section 2 reviews related work in conversational classification and sentiment analysis. Section 3 describes the framework pipeline, dataset, preparation methods, and the proposed methodology in depth. Section 4 presents the performance of the proposed framework, experimental results, and a real-world implementation prototype. Section 5 concludes the paper with future directions.

2. RELATED WORK

This section summarizes current developments in conversational systems and sentiment analysis, going from static text classification to dynamic, goal-oriented dialogue modeling. We divide the current literature into three main categories: dynamic sentiment analysis in conversational systems, conversational outcome prediction, and hybrid architecture and class imbalance.

2.1. Dynamic Sentiment Analysis in Conversational Systems

Sentiment analysis has progressed from lexicon-based approaches to more advanced deep learning models that can consider context. Early studies focused on static documents, such as social media posts and product reviews. Approaches proven effective in static domains, such as classical Support Vector Machines (SVM) and Naïve Bayes, have been shown to work well for tweets [6] and product review classification [7]. Sharin and Kartiwi [8] used sentiment analysis to determine the public's overall attitude toward Technical and Vocational Education and Training (TVET) on social media and showed the applicability of these methods in supporting policy decisions. Similarly, Buntoro et al. [9] showed that machine learning is reliable for gauging public sentiment in high-stakes political contexts by using SVM and Naïve Bayes algorithms in Indonesia's 2019 presidential election.

However, unlike these static applications, dialogue systems pose special challenges. Dialogues, unlike social media posts, require simulating emotional changes over time. Poria et al. [10] state that dialogue turns, as individual text units, fail to capture the "inter-speaker dependency," which is crucial to understanding context. While Recurrent Neural Networks (RNNs) have improved at capturing sequential dependencies, they can struggle to retain information over long-term contexts. The advent of Transformer-based models, especially BERT [3] and RoBERTa [4], marked a paradigm shift by enabling bidirectional context and capturing the subtle semantic nuances required for real-time dialogue.

2.2. Conversational Outcome Prediction

Predicting the outcome of a conversation, such as a sale, a resolution, or a conversion, differs from sentiment classification. It requires awareness of the strategic connection between goals and dialogue dynamics. In the business field, researchers have used machine learning to predict stock trends [11] and negotiation outcomes [12], primarily using aggregate sentiment features. More recent studies have explored bidirectional reasoning to track how emotional states change and how they affect resulting states [13]. These deep learning methods are superior to the traditional ones as they capture important conversational turning points. However, a critical issue with adaptive prediction systems is that most existing systems are post hoc, meaning that they are evaluated by analyzing completed dialogues [1]. One important limitation of previous studies is the absence of models designed to forecast what happens during the interaction (at the message level) so that real-time intervention can take place.

2.3. Hybrid Architectures and Class Imbalance

Hybrid solutions have been developed to balance the sequential nature of dialogues with the granular results of sentiment analysis. Pure transformer models (such as BERT) are great for semantic embedding but can be computationally expensive when used for real-time sequence modeling of long conversations. However, Bi-LSTMs are efficient sequence modelers [5] and lack the depth of contextual knowledge of pre-trained models. Recent studies have shown that combining these methods improves performance. Basiri et al. [7] showed that hybrid models, that is, attention models combined with Bi-LSTMs, show a significant improvement over standalone models in complex sentiment tasks. Additionally, the dataset is imbalanced, meaning the class “successful” is not often present, which is a problem that needs to be addressed. Extensive studies [14] show that to achieve good performance with deep neural networks trained on skewed data, dynamic sampling and class weighting are crucial. We build on this foundation by combining the strong embedding power of RoBERTa with our Enhanced Bi-LSTM, which is specifically designed to handle the high class imbalance in persuasive dialogue datasets.

2.4 Research Gaps and Novelty

While there is extensive literature on conversational AI, much work remains to be done on real-time, sentiment-aware dialogue systems. First, interaction is based on text only, without physical non-verbal cues like vocal tone and facial expressions; most high-accuracy sentiment models [2], [6], [8] and [9] are static, working on individual texts, such as reviews or tweets, rather than being dynamic and capturing the turn-taking interactions of live dialogue. Second, current outcome prediction models are mainly used as post-hoc analysis on finished logs [1] and cannot intervene in real time to guide agents turn by turn. Third, hybrid architectures are available [7], but few are optimized or evaluated on highly imbalanced, goal-oriented conversational datasets with rare goal outcomes.

This study directly bridges these gaps by shifting the paradigm from passive post-hoc evaluation to active real-time adaptation. Our novelty lies in integrating granular sentiment scoring and sequential outcome prediction into a unified feedback loop. The framework combines RoBERTa's message-level sentiment analysis with an enhanced Bi-LSTM for trajectory classification and a fine-tuned BERT for pre-send impact analysis, enabling dialogue systems to compensate for the lack of non-verbal cues and proactively guide dialogues toward desired outcomes.

https://doi.org/10.31436/iiumej.v27i3.4191

3. METHODOLOGY

This section details the proposed framework, first describing the pipeline and dataset. Then, we discuss the methodology of the main components of our proposed framework: sentiment scoring, conversation outcome classification, and message impact prediction, along with the proposed models for each task. Lastly, we define the evaluation metrics used in this study.

3.1. Framework Overview and Pipeline

Fig. 1 illustrates the proposed framework pipeline, which integrates conversation outcome classification and message impact prediction for adaptive dialogue management. The process starts with a common base, then splits into two interdependent tracks.

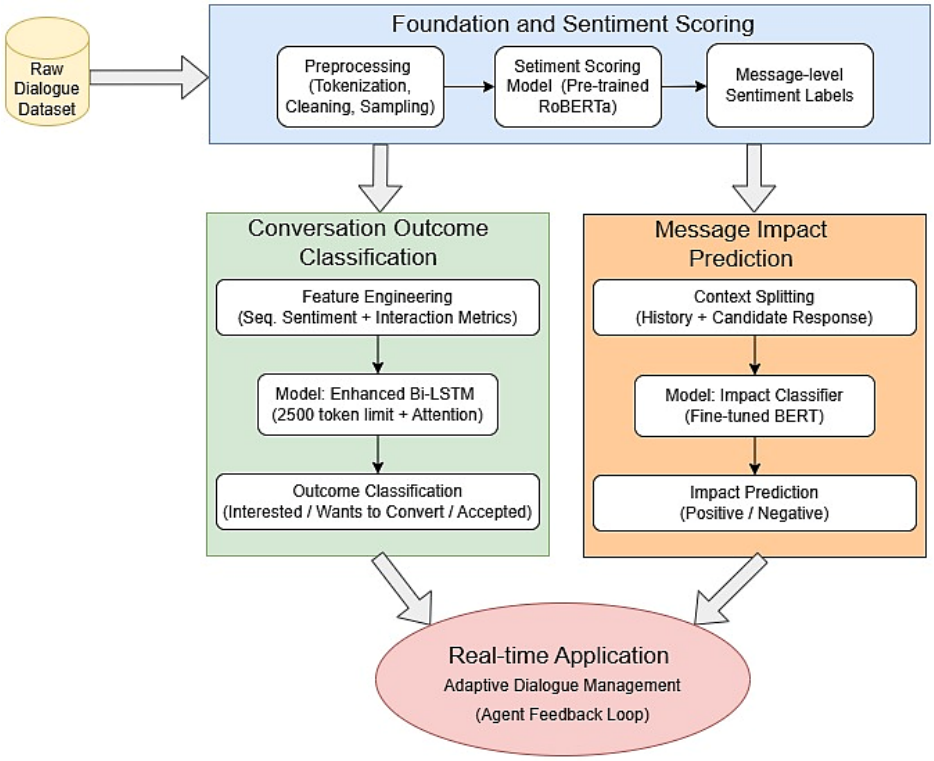

Figure 1. The proposed framework's overall pipeline.

1) Foundation and Sentiment Scoring

- Preprocessing: To guarantee data quality, the raw dialogue dataset is first subjected to tokenization, noise reduction, and text standardization.
- Sentiment Scoring: A pre-trained RoBERTa model assigns sentiment scores (positive/negative) to individual messages.
- Message-Level Sentiment: These granular scores are aggregated to track the emotional trajectory of the conversation turn by turn.

2) Track 1: Conversation Outcome Classification

- Engineered features (such as response latency and keyword density) and sequential sentiment scores are utilized in this track.

https://doi.org/10.31436/iiumej.v27i3.4191

- Our enhanced Bi-LSTM model is fed with these inputs to analyze the evolving context and classify the conversation into Interested in Islam, Wants to Convert, or Accepted Islam.

3) Track 2: Message Impact Prediction

- The agent's proposed response, its calculated sentiment score, the visitor's prior conversational context, and the outcome classification of the current ongoing conversation are considered inputs for this track to perform a predictive analysis before a response is sent.
- A fine-tuned BERT model is fed with these inputs to predict whether the proposed response will positively or negatively affect the direction of the conversation.

The pipeline dynamically links these tracks towards real-time adaptive dialogue management. Message-level sentiment tracking guides the conversation outcome classification, while message impact prediction enables agents to iteratively improve their proposed responses to steer that outcome positively.

3.2. Dataset Overview and Preparation

For this study, we use a unique dataset of 25,734 English conversations. The data were gathered from 2019-2022 from Rukn Al-hewar [15], an online platform for conversations about Islam. There are 5,677 (22.1%) conversations about "Accepted Islam", 17,553 (68.2%) conversations about "Interested in Islam", and 2,504 (9.7%) conversations about "Wants to Convert". This indicates class imbalance, as "Interested in Islam" represents 68.2% of conversations, posing challenges for model generalization.

Table 1. Types and numbers of conversations in the dataset

Type of conversation	Number of conversations
Accepted Islam	5,677
Interested in Islam	17,553
Wants to Convert	2,504
Total number of conversations	25,734

Table 2. Key statistical features of the dataset

Type of conversation	Number of conversations
Number of visitors	21,715
Number of agents	120
Number of words identified	26,762,112
Average number of words per conversation	967
Average number of words per question	9

Table 2 provides summary statistics for key features. The dataset contains 26.7 million words, 967 words per conversation, and an average of 9 words per question. The dataset includes 120 agents and 21,715 visitors engaged in conversations. Fig. 2 shows the differences in conversation length for each class. The longest dialogues concern Accepted Islam, while dialogues about Interested in Islam and Wants to Convert are shorter and more consistent. Therefore, pre-processing and sampling methods are needed to avoid training biases caused by this structural imbalance. Because of class imbalance and dialogue length, these factors must be addressed separately to ensure model robustness; however, the dataset size and annotations are domain-specific and can serve as a reliable resource for training a sentiment-aware dialogue system.

https://doi.org/10.31436/iiumej.v27i3.4191

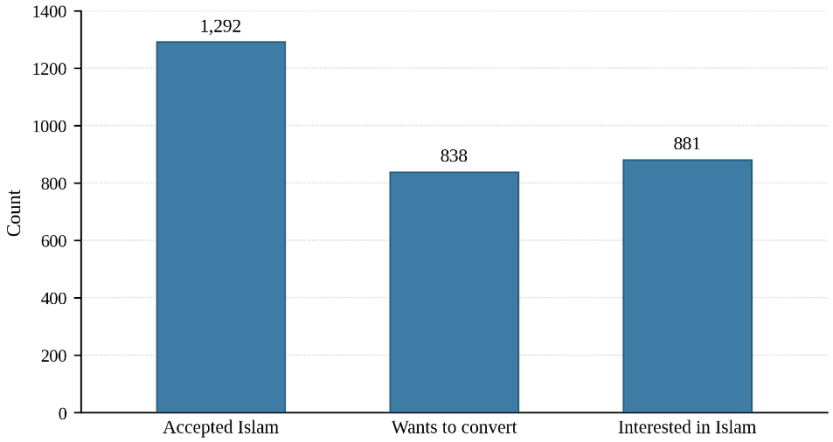

Figure 2. Average word count for each conversation class

To ensure analytical clarity and address potential semantic overlaps, explicit operational boundaries were defined for the three target categories:

- *Interested in Islam:* Exploratory dialogues where the visitor asks general, informational, or theological questions without expressing explicit intent to embrace the faith (e.g., *Can you explain how prayer works?*).
- *Wants to Convert:* Dialogues where the visitor explicitly expresses personal alignment, conviction, or a direct desire to become Muslim, but has not yet pronounced the testimony of faith during the session (e.g., *I believe Islam is the truth and I want to convert; what are the steps?*).
- *Accepted Islam:* Dialogues where the visitor formally completes the conversion process during the live chat by uttering or typing the testimony of faith (e.g., *I declare that there is no God but Allah and Muhammad is His messenger*).

The dataset's ground-truth labels were originally assigned by trained counselors at the partner organization (Rukn Al-hewar) after the session, using standard administrative protocols. The "Accepted Islam" label strictly required an explicit, logged declaration of the Shahada. To rigorously validate the quality of these primary labels and measure human consistency on ambiguous cases, a randomized sample of $N = 1{,}000$ conversations was independently re-annotated by two domain experts following a formal guidelines codebook. Inter-annotator agreement yielded a Fleiss' Kappa score of 0.82, indicating strong agreement and confirming the reliability of the ground-truth annotations.

To ensure analytical consistency and reduce noise, we prepared the data using standard preprocessing techniques. Text standardization involved lowercasing all characters and removing punctuation, special symbols, and non-alphanumeric elements. Lemmatization was applied to minimize lexical redundancy (e.g., "converting" → "convert"). Domain-specific stop-words (e.g., repetitive greetings such as "Assalamu Alaikum"), URLs, and emojis were removed to concentrate on meaningful dialogue content.

The dataset was partitioned into 70% training, 15% validation, and 15% test splits using a fixed global random seed (seed $= 42$). To address class imbalance during model development, we evaluated the Synthetic Minority Over-sampling Technique (SMOTE) and undersampling strategies. To strictly prevent data leakage between evaluation subsets, resampling techniques (including SMOTE) were applied exclusively to the 70% training split after partitioning. The

validation (15%) and held-out test split (15%, comprising 3,860 conversations) remained completely unmanipulated to evaluate true real-world operational performance.

3.3. Sentiment Scoring for Individual Messages

In this study, the two main tasks are conversation outcome classification (in Section 3.4) and message impact prediction (in Section 3.5); for both, each unannotated question and answer in the dataset is assigned a sentiment score (positive or negative) as a basis. The two transformer models tested were: (i) RoBERTa [4], an optimized version of BERT [3] known to capture subtle linguistic patterns, and (ii) DistilBERT, a transformer with minimal performance trade-off but drastically reduced computation time. As evaluated in Section 4, these empirical performance metrics determine which of RoBERTa and DistilBERT perform better. In our approach, we fully retain the original conversational context while keeping the pipeline as modular as possible, with sentiment scoring independent of the pre-processing step.

3.4. Conversation Outcome Classification

This subsection outlines the methodology of the conversation outcome classification. We first discuss the domain-specific features and then explain our Enhanced Bi-LSTM architecture. Finally, we highlight the importance of simulating real-time conversation, where classification happens dynamically rather than waiting for the dialogue to conclude.

Table 3. Domain-specific features and their descriptions

Feature	Description
Sentiment Score	Sentiment scores associated with all conversation messages, calculated using the RoBERTa model.
Average Sentiment Score	Average sentiment scores for questions, answers, and the overall conversation, offering insights into the general tone and engagement levels.
Number of Sessions	Tracks the total sessions within a conversation to measure the extent of repeated interactions.
Conversation Length	Represents the total number of words across all messages in a conversation.
Average Message Length	Calculates the average number of words per message, providing insights into the verbosity of participants.
Number of Messages	Counts the total number of messages in each conversation to reflect the level of engagement.
Duration	Measures the total time span of a conversation, from the first to the last message, to understand the temporal dynamics of interactions.

3.4.1. Feature Engineering

Three key aspects – sentiment metrics, interaction patterns, and temporal dynamics – were the focus of domain-specific features engineered to represent conversational dynamics, as shown in Table 3. Sentiment metrics include average sentiment scores for questions, answers, and the overall conversation, as well as per-message sentiment scores produced by the RoBERTa model. Interaction patterns such as conversation length and message frequency quantify engagement. Temporal dynamics such as response latency and keyword density track behavioral shifts.

3.4.2. Full-Conversation Classification Using Enhanced Bi-LSTM

To address the limitations of standard sequential models, we propose an Enhanced Bi-LSTM architecture. Our improved model incorporates multi-head attention mechanisms to dynamically weigh the significance of specific conversation turns, whereas standard Bi-LSTMs struggle with long-term dependencies. Additionally, we increased the input capacity

from the typical 512 or 1024 tokens to 2,500 tokens, enabling the model to analyze long, multi-turn conversations without losing any data. Our proposed model is optimized to classify full conversations into the three target classes using pre-trained GloVe embeddings.

3.4.3. Real-time Classification Simulation

Instead of waiting for the dialogue to end, the system in a live deployment must predict outcomes dynamically as the conversation progresses. When given partial context, models trained only on completed conversations often fall short. Therefore, we truncated the dataset at different completion milestones (50%, 60%, ... 100%) to simulate this real-time scenario and evaluate the robustness of our Enhanced Bi-LSTM. This guarantees that the model can act in advance and provide agents with actionable insights while the conversation is still active.

3.5. Message Impact Prediction

Effective dialogue systems should proactively analyze how individual messages impact interaction trajectories in addition to classifying full conversations into final outcomes. Traditional classification techniques evaluate completed conversations, which restricts their usefulness for real-time adaptation. Predicting whether a possible agent message will positively or negatively affect the conversation flow before it is sent is an essential challenge addressed in this subsection. This framework enables iterative improvement of agent strategies by isolating the causal role of specific responses and ensuring alignment with goals.

3.5.1. Target Label Definition

We created a new target label based on the conversational changes seen before and after a Daei message to understand the impact of these messages. This required examining two versions of the conversations:

1) **Pre-Daei Message Context:** this refers to the entire conversation up to the Daei message.
2) **Post-Daei Message Context:** this refers to the entire conversation (at most 5 messages from the visitor) after the Daei message.

Table 4. Daei message impact based on the transition of pre- and post-message context classes and their probabilities

Daei Message Impact	Pre-Daei Message Context	Post-Daei Message Context
Positive	Interested in Islam	Interested in Islam (probability increased)
	Interested in Islam	Wants to Convert
	Interested in Islam	Accepted Islam
	Wants to Convert	Wants to Convert (probability increased)
	Wants to Convert	Accepted Islam
	Accepted Islam	Accepted Islam (probability increased)
Negative	Interested in Islam	Interested in Islam (probability decreased)
	Wants to Convert	Interested in Islam
	Wants to Convert	Wants to Convert (probability decreased)
	Accepted Islam	Interested in Islam
	Accepted Islam	Wants to Convert
	Accepted Islam	Accepted Islam (probability decreased)

The predicted classes and their probabilities for both contexts stated above were calculated using the Enhanced Bi-LSTM as the classifier model. The sentiment score of the Daei message—which was computed using RoBERTa—was also included to evaluate how the tone and sentiment of the Daei message affected the classification. The Daei message's impact was

https://doi.org/10.31436/iiumej.v27i3.4191

assigned a positive or negative label based on the transition of the classes and their probabilities from the pre-Daei message context to the post-Daei message context, as illustrated in Table 4.

To ensure that deriving impact labels from model probability transitions does not introduce circularity, we subjected target labels to an independent human reliability audit. Domain experts manually evaluated a sample of dialogue transition segments to verify alignment between automated state shifts and true human-perceived message impact. Section 4.3 reports the detailed reliability findings.

Using the sentiment score of the Daei message and the classes and their probabilities of the pre- and post-Daei messages, we implemented and trained two comparative architectures - a fine-tuned BERT and a standard Bi-LSTM – to predict the Daei message impact. The performance of these two methods is thoroughly evaluated in Section 4.3.

3.5.2. Data Preparation

The following seed-based filtering method was used to prepare the dataset for the message impact prediction task:

- We split conversations into segments in which each segment contained a single Daei message as the focal point. We ensured that each segment had at least five messages from the visitor before the focal point and at least one message after it.
- We combined Daei messages into a single focal point if they appeared consecutively, and we averaged their sentiment scores.

This seed-based filtering method yields a dataset with 122,056 files in the "Accepted Islam" category, 259,720 files in the "Interested in Islam" category, and 32,045 files in the "Wants to Convert" category.

3.6. Evaluation Metrics

To rigorously evaluate model performance across classification and impact prediction tasks, particularly under class imbalance, we use standard quantitative performance metrics [16]. Given True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), the metrics are defined as follows:

- Overall accuracy measures the total proportion of correct predictions:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{1}$$

- Precision and recall measure positive predictive value and sensitivity, respectively [17]:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}},\ Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

- Per-class F1-score is a harmonic mean of precision and recall for class i [17]:

$$F1_i = \frac{2 \times \mathrm{Precision}_i \times \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i} \tag{3}$$

- Macro F1-score is the unweighted mean of F1-scores across all C classes, evaluating performance equally across minority and majority categories [16]:

$$Macro - F1 = \frac{1}{\mathrm{C}} \sum_{i=1}^{C} F1_i \tag{4}$$

- Balanced accuracy is the unweighted average of recall across all C classes to account for class imbalance [18]:

https://doi.org/10.31436/iiumej.v27i3.4191

$$Balanced\ Accuracy = \frac{1}{C}\sum_{i=1}^{C} Recall_i \quad (5)$$

3.7. Ethical Considerations and Privacy Protection

Because conversational datasets are typically text-based, they may include personal conversations; we implemented careful data privacy and ethical practices in data collection, preprocessing, and model execution. We thoroughly de-identified every raw conversation log submitted by the partner organization before analysis. We systematically removed Personally Identifiable Information (PII), including personal names, phone numbers, e-mail addresses, IP logs, and specific geographic locations, through automated regex filtering with manual spot-checking. Data collection followed the partner platform's terms of service, which allow interactions to be logged under administrative consent for quality assurance and research.

To address potential concerns regarding manipulative dialogue management, the proposed framework is explicitly architected as a Human-in-the-Loop (HITL) decision-support system rather than a fully autonomous agent. Real-time sentiment scores and message impact predictions serve as advisory feedback within the agent interface, ensuring that human counselors maintain full agency and ethical oversight over all generated responses before deployment.

4. Results and Discussion

This section presents the experimental findings for sentiment scoring, message impact prediction, and conversation outcome classification. Quantitative performance and comparative analyses validate our framework's ability to improve real-time dialogue.

4.1. Sentiment Scoring of Individual Messages

We analyzed two models, DistilBERT and RoBERTa, to predict the sentiment of specific questions and responses. Table 5 summarizes the performance of both models for each of the three categories based on average sentiment scores. Fig. 3 and Fig. 4 visually compare the average sentiment scores for RoBERTa and DistilBERT, respectively. To determine classification superiority, we manually annotated a randomized gold-standard evaluation subset of N=1,000 individual dialogue messages (Positive, Negative, and Neutral) with two domain experts. We used standard classification metrics for both models to assess this benchmark. The quantitative evaluation shows that RoBERTa outperforms DistilBERT in all metrics: accuracy of 88.4%, precision of 0.87, recall of 0.86, and Macro-F1 score of 0.86, compared with DistilBERT's 81.2%, 0.80, 0.79, and 0.79, respectively. The contextual embeddings generated by RoBERTa were especially good at capturing conversational nuances and domain-specific terminology. As a result, we chose RoBERTa as the backbone sentiment-scoring model for all downstream tasks.

Message sentiment scores range continuously between -1.0 (highly negative) and $+1.0$ (highly positive). Scores strictly above $+0.15$ are classified as positive, scores below -0.15 as negative, and scores within the interval as neutral. The neutral boundary was established empirically via grid validation on a representative sample of $N = 1000$ dialogue messages. In conversational contexts, neutral turns predominantly consist of factual inquiries, links, or procedural statements (e.g., asking for website links). A grid search across threshold intervals from to demonstrated that optimally separates purely procedural messages from subtle emotional shifts, preventing minor polarity fluctuations from misclassifying neutral inquiries.

https://doi.org/10.31436/iiumej.v27i3.4191

We performed sentiment analysis using the RoBERTa model and computed sentiment scores for each message in the text conversation. Rather than examining the entire conversation, we evaluated the sentiment of individual messages. This gave us a basic understanding of how people convey their views and feelings through messaging. Tables 6 and 7 illustrate some examples of sentiment scores for visitor messages and Daei messages using RoBERTa, respectively.

Table 5. Comparing DistilBERT and RoBERTa for predicting average sentiment scores

Item	RoBERTa	DistilBERT
Accepted Islam Answers	0.3625	0.3031
Accepted Islam Questions	0.2529	0.2535
Wants to Convert to Islam Answers	0.2872	0.2541
Wants to Convert to Islam Questions	0.1526	0.1918
Interested in Islam Answers	0.2131	0.2084
Interested in Islam Questions	0.0853	0.1386

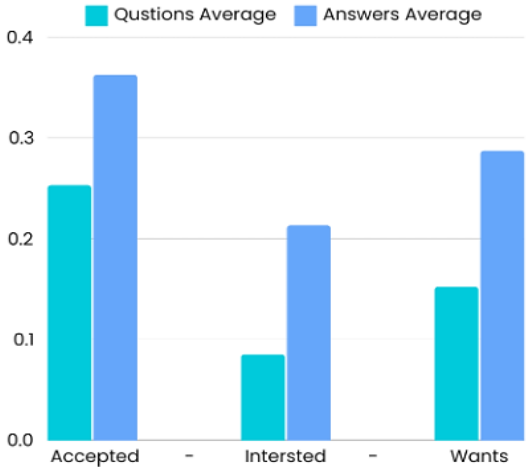

Figure 3. RoBERTa's average sentiment score performance

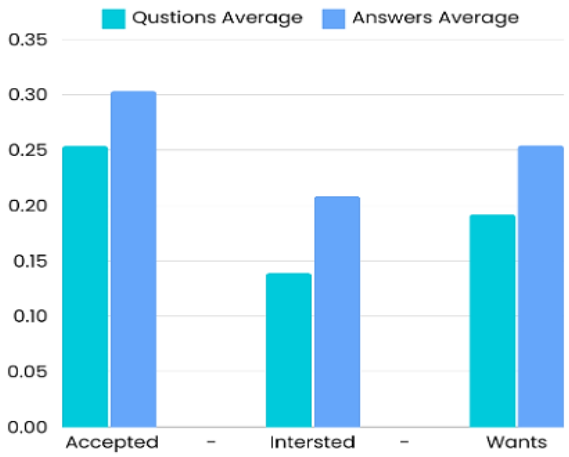

Figure 4. DistilBERT's average sentiment score performance

https://doi.org/10.31436/iiumej.v27i3.4191

Table 6. Sentiment Scores for visitor messages using RoBERTa

Visitor Message	RoBERTa Scores
"Yes I would love that."	0.9696
"Yeah, I studied Islam well and understand a lot of it."	0.7180
"So ur logic dont make sense."	-0.7479
" I'm not interested in converting."	-0.4887
"What is the importance of Salah"	0.107

Table 7. Sentiment Scores for Daei messages using RoBERTa

Visitor Message	RoBERTa Scores
"We are happy to help; keep coming to ask, please."	0.9435
"When you take shahadah, all that is forgiven by Allah, and you become new like the time you were born"	0.7180
"The vast universe itself, with its meticulous order and fine-tuning, is proof that God exists"	0.8500
"So would you like for me to share the basics of Islam with you?"	0.0411
"Correct, and that's what the Qur'an says"	0.053

We used scatter charts to display the sentiment scores for the three categories (Accepted Islam, Interested in Islam, and Wants to Convert) in order to better analyze the model's performance. Each scatter chart shows the sentiment scores for each question and response, along with their averages and detected outliers.

With very few outliers, the Accepted Islam scatter chart in Fig. 5 displays the most consistent results. This is expected because conversations in this category are usually clear and direct, as visitors have already accepted Islam, making it simpler for the algorithm to provide accurate sentiment scores.

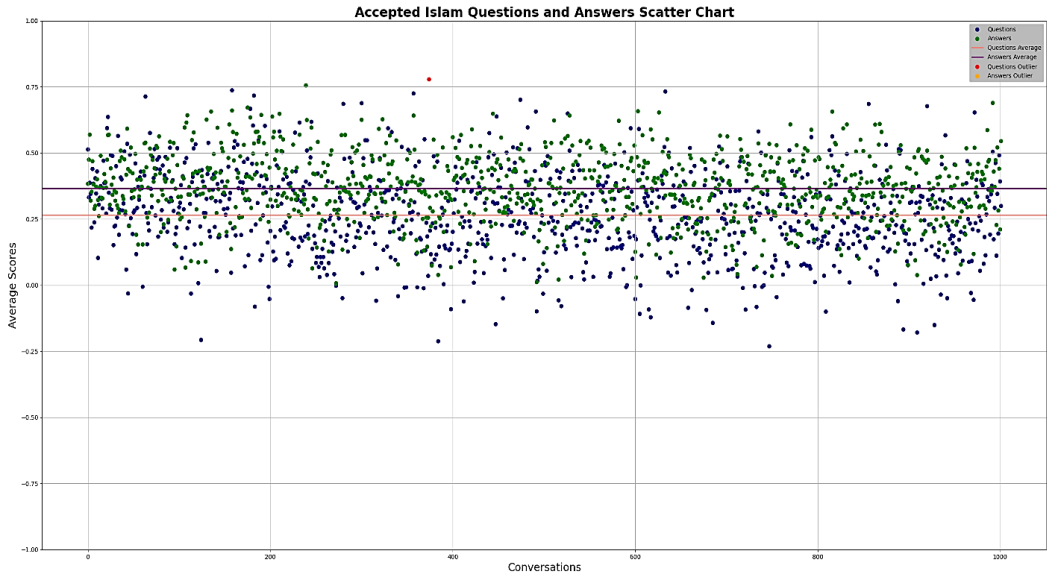

Figure 5. Scatter chart of Accepted Islam questions and answers

Compared to Accepted Islam, the Wants to Convert scatter chart in Fig. 6 shows slightly more variability, but it still performs well with few outliers. This is because, unlike the Accepted Islam category, interactions in this category often feature participants expressing a strong interest or intent to convert, resulting in more varied emotional expressions.

https://doi.org/10.31436/iiumej.v27i3.4191

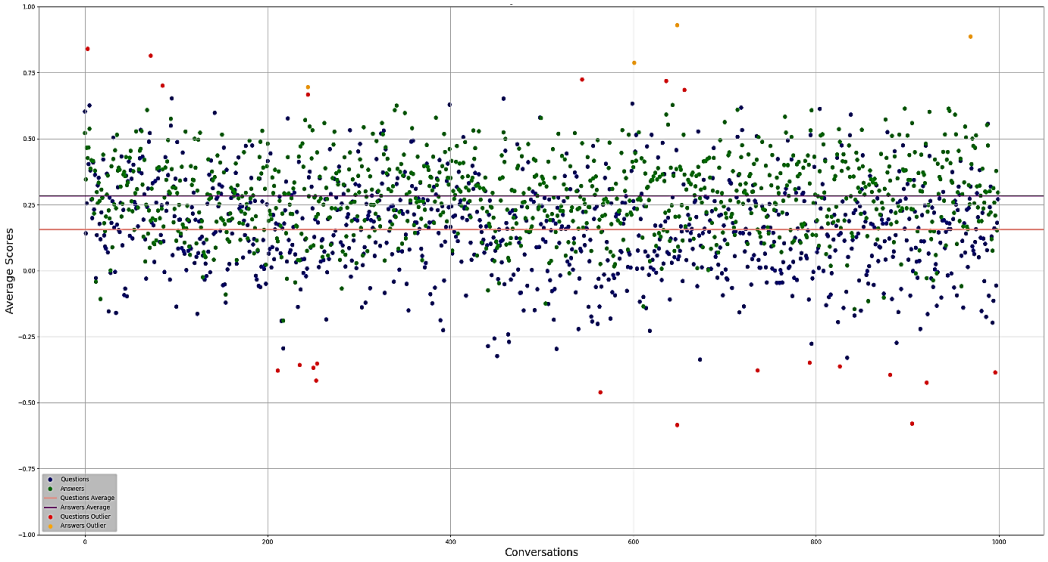

Figure 6. Scatter chart of Wants to Convert questions and answers

The scatter chart for Interested in Islam in Fig. 7 shows the most variability and outliers among the three categories. This is expected, as conversations in this category are more exploratory, with participants asking questions or expressing curiosity without clear intent or decisions. This natural variability in conversations results in a wider range of sentiment scores.

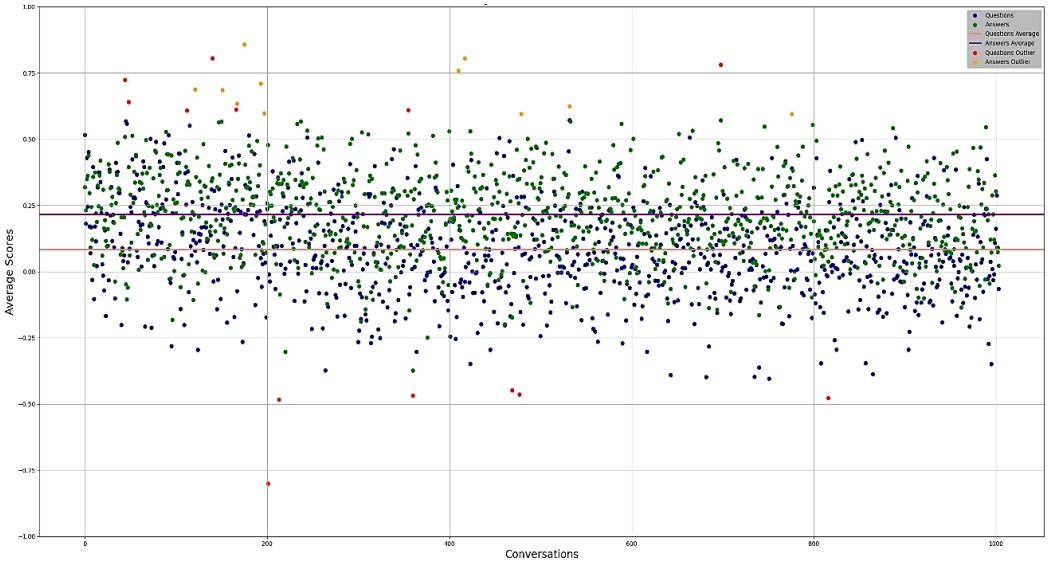

Figure 7. Scatter chart of Interested in Islam questions and answers

4.2. Conversation Outcome Classification

Comprehensive experimental results for conversation classification are shown in this subsection. Using the entire 3-class conversational dataset, we methodically assess the performance of three deep learning algorithms: BERT, Bi-LSTM, and our Enhanced Bi-LSTM. We then thoroughly validate our main model and analyze the effects of several approaches to resolving class imbalance. To confirm the cause of classification difficulty, we also perform an ablation study. Finally, we evaluate the models' adaptability and performance when trained on incrementally truncated conversation lengths, simulating real-time interaction scenarios.

https://doi.org/10.31436/iiumej.v27i3.4191

To categorize the conversation into one of the three groups, we initially examined three models on complete conversations without breaking them up into shorter segments.

- **BERT Model:** The accuracy of the pretrained BERT model was 74.8%. BERT's results highlighted its shortcomings for our dataset. In particular, 65% of our conversations were longer than the 512 tokens that BERT can handle. As illustrated in Fig. 8, this limitation reduced its effectiveness. To address this sequence length limitation, we also evaluated Longformer-base (supporting up to 4,096 tokens). However, Longformer exhibited severe training instability, excessive per-epoch latency, and early validation plateauing. Consequently, long-context transformers proved impractical for this task.

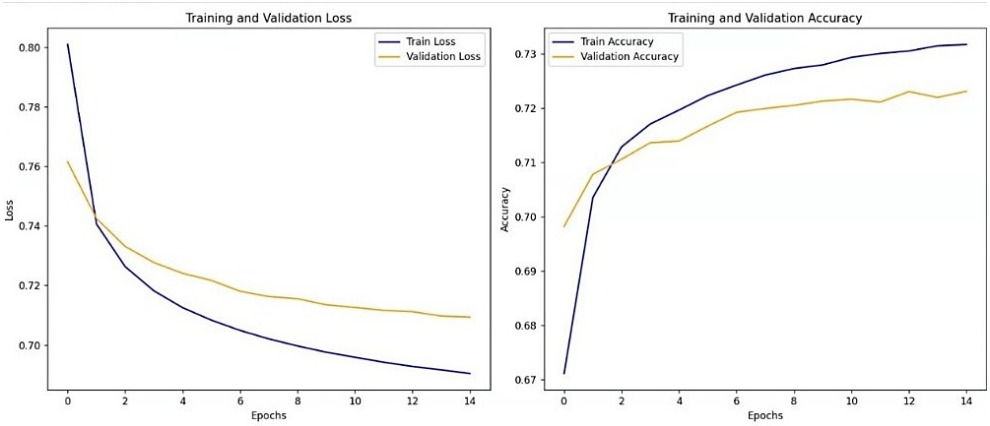

Figure 8. Performance of the pretrained BERT model, showing training and validation loss and accuracy over epochs.

- **Bi-LSTM Model:** A standard Bidirectional LSTM model with a 1024 token limitation demonstrated its ability to handle longer sequences with an accuracy of 80% on the test set (Fig. 9).

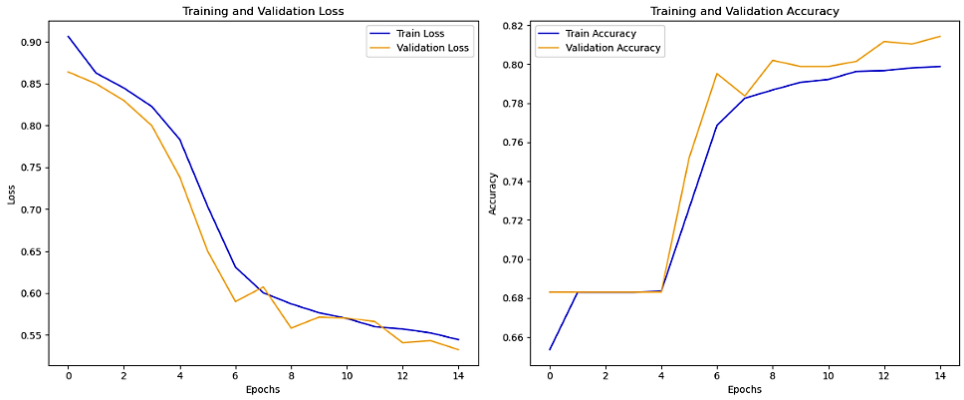

Figure 9. The pretrained Bi-LSTM model's performance on conversations with a token limit of 1024 demonstrates training and validation loss and accuracy over epochs.

- **Enhanced Bi-LSTM Model:** We improved the Bi-LSTM by adding attention layers, increasing the token limit to 2500, and adjusting the learning rate. With a test accuracy of 90%, this model achieved our best 3-class result. These adjustments led to significant improvements, as shown in Fig. 10, with accuracy remaining close to 90% and loss decreasing consistently.

https://doi.org/10.31436/iiumej.v27i3.4191

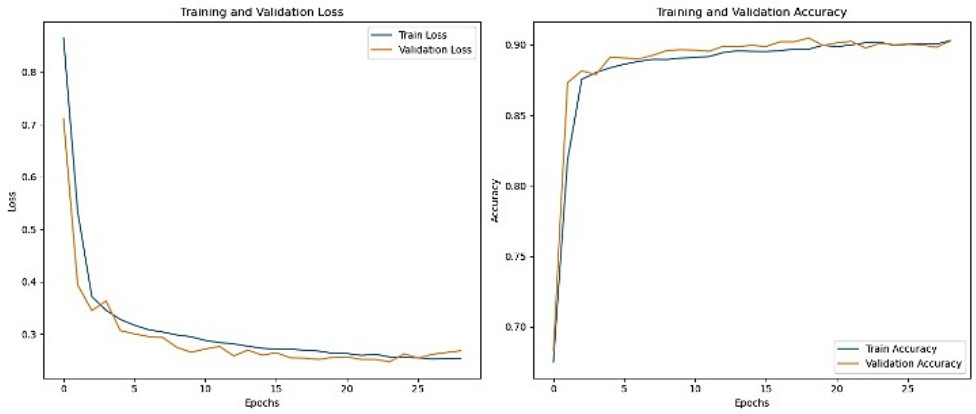

Figure 10. Performance of the Enhanced Bi-LSTM model with training and validation loss and accuracy over epochs after raising the token limit to 2500.

Table 8. Classification results of the Enhanced Bi-LSTM (3-Class Model)

Classes	Precision	Recall	F1-Score
Interested in Islam	0.90	0.96	0.93
Wants to Convert	0.57	0.30	0.39
Accepted Islam	0.96	0.97	0.97
Overall Accuracy	0.900		
Balanced Accuracy	0.743		
Macro-F1 Score	0.763		

The detailed classification results in Table 8 show that while the model achieves an overall accuracy of 90.0%, its balanced accuracy is 74.3% and its macro-F1 score is 0.763. This breakdown shows that overall accuracy is driven largely by strong performance on the majority classes ("Interested in Islam" with F1 = 0.93 and "Accepted Islam" with F1 = 0.97). Conversely, the model struggled with the minority "Wants to Convert" category, achieving a low recall of 0.30 and an F1-score of 0.39, indicating frequent misclassifications due to semantic overlap with "Interested in Islam". We further analyze this challenge in the following section.

Table 9. Hyperparameters, performance baseline comparison, and execution timing across models

Hyperparameters	BERT	Bi-LSTM	Enhanced Bi-LSTM
Learning rate	5.00e-05	5.00e-05	1.00e-04
Batch size	64	64	128
Number of layers	12 (transformers)	4 layers	4 + Attention layers
Hidden Units / Dimension	768	128 (per direction)	250 (per direction)
Token limit	512	1024	2500
Optimizer	AdamW	Adam	Adam
Number of epochs	15	15	50
Dropout Rate	0.1	0.2	0.3
Early Stopping	Patience = 3	Patience = 5	Patience = 5
Overall Accuracy	74.8%	80%	90%
Training Time/Epoch	~14 min	~3 min	~6 min
Avg. Inference Latency	42 ms	35 ms	85 ms
Additional features	Pretrained embeddings	GloVe embeddings	GloVe embeddings, multi-head attention

https://doi.org/10.31436/iiumej.v27i3.4191

Comparing these architectures demonstrates the component-wise contributions of our framework: extending token capacity from 512 (BERT) to 1,024 (Bi-LSTM) yields a 5.2% accuracy gain, while adding multi-head attention and expanding context to 2,500 tokens in the Enhanced Bi-LSTM provides an additional 10.0% accuracy increase (reaching 90.0% accuracy, 74.3% balanced accuracy, and 0.763 macro-F1).

Table 9 summarizes the hyperparameter configurations, architectural dimensions, performance baseline metrics, and execution timing across all evaluated models to establish a consistent experimental setup (with a detailed discussion of real-time latency and computational efficiency in Section 4.5).

4.2.1. Data Validation and Robustness Analysis

We validated the performance of our 90% accuracy 3-class model and transparently analyzed its limitations by presenting and discussing the confusion matrix and the statistical test of agreement (Cohen's Kappa). Then, we applied several techniques to address the imbalance identified in the confusion matrix.

To visualize the challenge presented in Table 7, we reported the confusion matrix (Fig. 11) for the Enhanced Bi-LSTM model's predictions on the 15% test split, which comprises 3860 conversations (2633 "Interested", 376 "Wants to Convert", and 851 "Accepted").

The confusion matrix in Fig. 11 shows that the model is very good at distinguishing between "Interested in Islam" (96.0% recall) and "Accepted Islam" (96.9% recall). Between these two major classes, there is relatively little uncertainty. But the major reason for the model's error is the class "Wants to Convert", where 66.5% of the conversations (250 conversations) are misclassified as "Interested in Islam". This strongly suggests that the language dimensions of "interested" and "wants to convert" are very similar, and this class is inherently ambiguous.

We computed Cohen's Kappa score, which evaluates inter-rater agreement while accounting for chance, to further validate the model's performance beyond simple accuracy. The model received a Cohen's Kappa score of 0.77 on this classification task. This indicates a high degree of agreement, demonstrating that the model's 90% accuracy is reliable and not a result of the class imbalance.

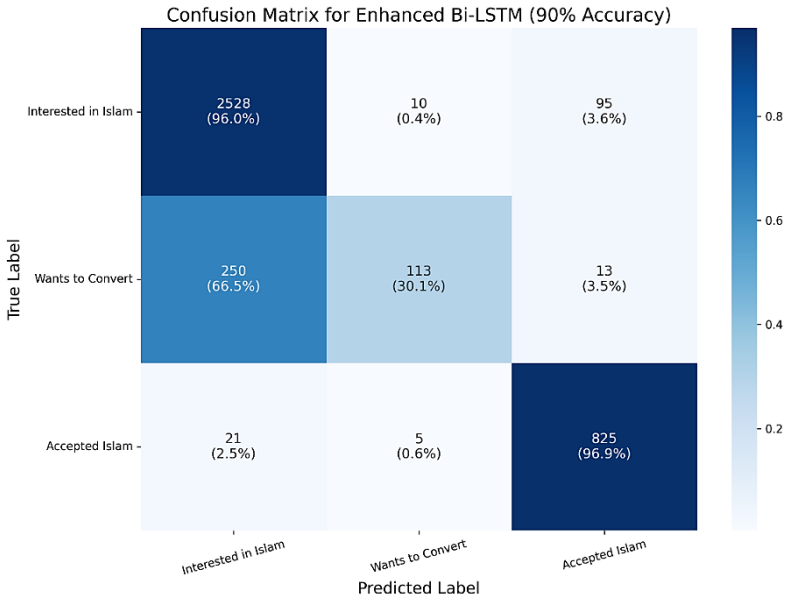

Figure 11. Confusion Matrix for 3-Class Enhanced Bi-LSTM (90% Accuracy)

https://doi.org/10.31436/iiumej.v27i3.4191

Additionally, we tested several strategies to address the imbalance found in the confusion matrix, including the following.

a) Using class weights:

During training, we assigned the minority class "Wants to Convert" higher weights. To ensure that the model focused more on the minority class, this adjustment updated the loss function to penalize misclassification of this class more heavily. The performance of this method is shown in Table 10. The model's recall for "Wants to Convert" improved to 0.71, but its precision for this class stayed low at 0.38, resulting in an F1-score of 0.50. The model's overall accuracy of 86% shows moderate effectiveness in balancing performance across classes.

Table 10. Enhanced Bi-LSTM's performance after applying class weights

Classes	Precision	Recall	F1-Score
Interested in Islam	0.96	0.84	0.89
Wants to Convert	0.38	0.71	0.50
Accepted Islam	0.97	0.97	0.97
Accuracy		0.86	

b) Under-sampling the dominant classes:

To balance the dataset, the number of samples in the main categories ("Interested in Islam" and "Accepted Islam") was reduced. This strategy improved the model's ability to classify the minority category, "Wants to Convert." As a result, the model attained an overall accuracy of 84%. Table 11 shows that the precision and recall for "Wants to Convert" were both 0.76, yielding an F1-score of 0.76. F1-scores for "Interested in Islam" and "Accepted Islam" were 0.77 and 0.97, respectively. These findings indicate that while undersampling reduced the model's bias toward dominant categories, the smaller dataset size marginally limited the model's learning potential.

Table 11. Performance of the Enhanced Bi-LSTM when the dominant classes were undersampled.

Classes	Precision	Recall	F1-Score
Interested in Islam	0.77	0.78	0.77
Wants to Convert	0.77	0.76	0.76
Accepted Islam	0.97	0.97	0.97
Accuracy		0.84	

c) Oversampling the minority class:

To increase the representation of the "Wants to Convert" category, we employed oversampling by duplicating its data and generating synthetic samples. This method ensured the model was exposed to more minority-class samples during training. According to Table 12, the recall for "Wants to Convert" increased to 0.59, but the precision stayed low at 0.37, yielding an F1-score of 0.45. High F1-scores of 0.90 and 0.96 were attained by the "Interested in Islam" and "Accepted Islam" categories, respectively. Although the precision-recall tradeoff for this category still presented difficulties, these results demonstrate that oversampling improved the model's capacity to identify the minority class.

Table 12. Performance of Enhanced Bi-LSTM after oversampling the "wants to convert" class.

Classes	Precision	Recall	F1-Score
Interested in Islam	0.94	0.86	0.90
Wants to Convert	0.37	0.59	0.45
Accepted Islam	0.97	0.95	0.96
Accuracy	0.86		

d) Merging classes:

In this technique, we consolidated the "Accepted Islam" and "Wants to Convert" categories into a single class. The model performed well on the entire dataset while focusing more on underrepresented data. The model attained an accuracy of 91%, with detailed results shown in Table 13.

Table 13. Performance of Enhanced Bi-LSTM after merging "accepted Islam" and "wants to convert" classes.

Classes	Precision	Recall	F1-Score
Interested in Islam	0.90	0.97	0.94
Merged (Wants to Convert & accepted Islam)	0.93	0.77	0.85
Accuracy	0.91		

The results of these experiments demonstrate that while typical imbalance strategies can increase minority-class recall, they tend to reduce overall model accuracy and do not address the fundamental problem of linguistic ambiguity.

4.2.2. Ablation Study: Removing the Ambiguous Class

Because of its substantial linguistic overlap with the "Interested" class, the study in the preceding section clearly confirms that the "Wants to Convert" class is the main source of model error. We conducted an ablation study to test this hypothesis by eliminating this one problematic class and reducing the task to a binary classification (Interested vs. Accepted).

Table 14. Performance of Enhanced Bi-LSTM (2-Class Ablation Study)

Classes	Precision	Recall	F1-Score
Interested in Islam	0.99	0.99	0.99
Accepted Islam	0.97	0.98	0.98
Overall Accuracy	0.990		
Balanced Accuracy	0.985		
Macro-F1 Score	0.985		

After removing the "Wants to Convert" class, we retrained our Enhanced Bi-LSTM model on the remaining two categories (2-class ablation study). Table 14 summarizes the findings, and Fig. 12 shows the results. The model attained an overall accuracy of 99%, a balanced accuracy of 98.5%, and a macro-F1 score of 0.985. While this high performance demonstrates the Enhanced Bi-LSTM's capacity to discriminate between distinct conversational endpoints (Interested vs. Accepted), this binary result must be interpreted cautiously. Simplifying the task to a two-class problem removes the primary source of linguistic overlap but does not resolve the real-world three-class classification task. Thus, this ablation study is not presented as evidence of strong performance for the full system, but rather as a diagnostic experiment confirming that semantic ambiguity in the "Wants to Convert" class is the primary bottleneck of the 3-class architecture. This diagnostic ablation isolates the contribution of architectural

https://doi.org/10.31436/iiumej.v27i3.4191

components from dataset characteristics, quantitatively proving that our sequence length expansion and attention layers effectively solved trajectory modeling, leaving semantic boundary overlap as the sole bottleneck.

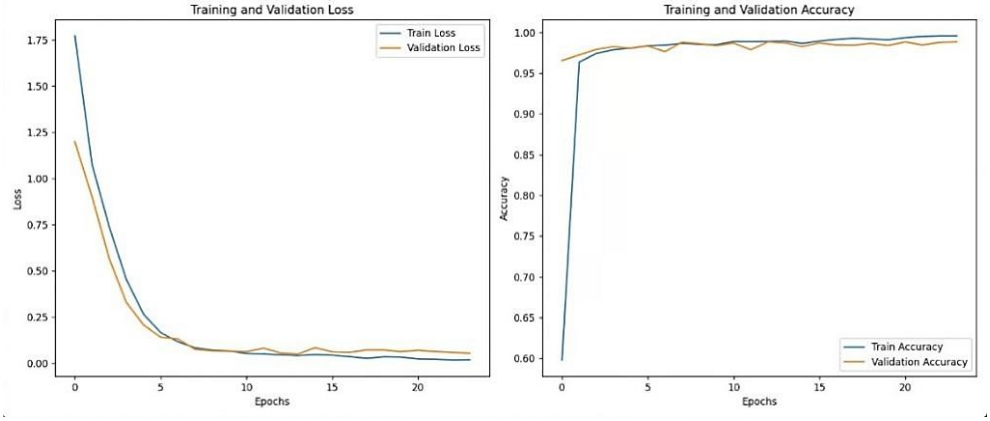

Figure 12. Loss and accuracy curves for training and validation after the removal of the "wants to convert" class (2-Class Ablation Study).

4.2.3. Real-time Classification Simulation: Effect of Conversation Length

To evaluate the model's performance across conversations of varying lengths (i.e., to more realistically represent real-world chats), we used a second approach to train the model with split data (50%, 60%, up to 100%). We utilized this splitting method exclusively for the Enhanced Bi-LSTM model.

Table 15. Enhanced Bi-LSTM's performance (training with split conversations in the 3-class classification task).

Classes	Precision	Recall	F1-Score
Interested in Islam	0.90	0.94	0.92
Wants to Convert	0.49	0.41	0.45
Accepted Islam	0.89	0.83	0.86
Overall Accuracy	0.860		
Balanced Accuracy	0.727		
Macro-F1 Score	0.743		

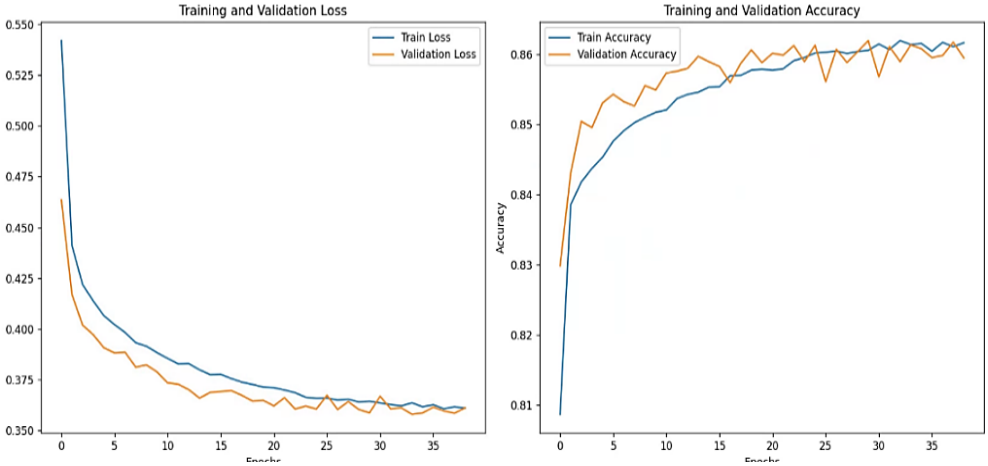

Figure 13. Performance of Enhanced Bi-LSTM (second approach: training on split conversations, 3-class classification task).

When evaluated on split conversation milestones to simulate real-time interaction, the model achieved an overall accuracy of 86%, a balanced accuracy of 72.7%, and a macro-F1 score of 0.743 on the 3-class task. As detailed in Table 15, while the model maintains solid dynamic classification across varying dialogue lengths for the majority classes (Interested F1 = 0.92, Accepted F1 = 0.86), performance on the minority "Wants to Convert" category remains lower (Recall = 0.41, F1 = 0.45) due to mid-conversation semantic ambiguity. Fig. 13's training and validation plots demonstrate consistent training progress.

As with our full-conversation model, we applied the same ablation study to the split-conversation setup by eliminating the "Wants to Convert" class. As shown in Table 16 and Fig. 14, the model attained an overall accuracy of 99%, a balanced accuracy of 98.5%, and a macro-F1 score of 0.985. This confirms that dynamic truncated sequence modeling remains robust across varying conversation lengths when class definitions are distinct, though addressing the full three-class ambiguity in live settings remains a target for future work.

Table 16. Performance of Enhanced Bi-LSTM (training on split conversations, 2-Class Ablation).

Classes	Precision	Recall	F1-Score
Interested in Islam	0.99	0.99	0.99
Accepted Islam	0.98	0.98	0.98
Overall Accuracy	0.990		
Balanced Accuracy	0.985		
Macro-F1 Score	0.985		

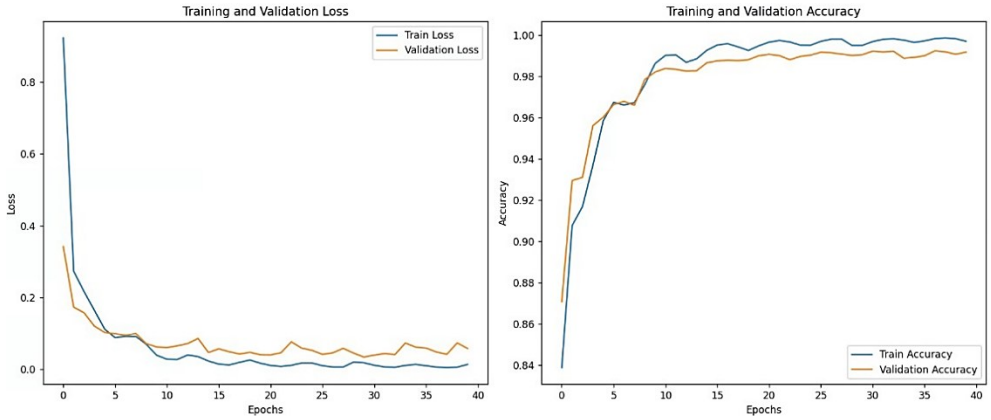

Figure 14. Performance of Enhanced Bi-LSTM (training on split conversations, 2-Class Ablation).

4.3. Message Impact Prediction

In this section, we evaluate the performance of Bi-LSTM and BERT models in predicting the impact of agent (Daei) messages prior to deployment (Positive vs. Negative), as defined in Section 3.5. Before assessing the predictive architectures, we performed a formal reliability analysis to validate the target labels and remove possible circularity in label generation. Two qualified dialogue counselors independently audited a randomized sample of N=500 agent message segments, which included pre-message context, agent message, and visitor turns. Observed agreement between human impact ratings and our automated transition rules was 84.6%, and Cohen's Kappa was 0.71. This large interrater agreement demonstrates the empirical ground-truth validity of the target impact labels.

The class imbalance in the impact dataset was addressed by embedding class weights in the loss function during training to prevent bias toward the dominant negative class.

https://doi.org/10.31436/iiumej.v27i3.4191

Table 17 shows that the Bi-LSTM model achieved an accuracy of 78%, which is a good score for the negative class, with an F1-score of 0.85, and a moderate score for the positive class, with an F1-score of 0.59. The weighted average of 0.79 shows balanced overall performance, with weights used to help the model focus on both classes.

The BERT model achieved an accuracy of 80% (as shown in Table 18), which is higher than the Bi-LSTM model. It performed well on the negative class (F1-score of 0.86) and improved on the positive class (F1-score of 0.62). This illustrates BERT's ability to address class imbalance effectively.

Table 17. Performance of Bi-LSTM for predicting message impact.

Classes	Precision	Recall	F1-Score
Negative	0.87	0.83	0.85
Positive	0.56	0.64	0.59
Accuracy		0.78	

Table 18. Performance of BERT for predicting message impact.

Classes	Precision	Recall	F1-Score
Negative	0.87	0.85	0.86
Positive	0.59	0.64	0.62
Accuracy		0.80	

4.4. Prototype Implementation and Real-world Application

We developed a working prototype to validate the practicality and real-time capabilities of our proposed framework. This prototype integrates the best-performing models and techniques from our research, demonstrating their effectiveness in a real-world operational setting. The prototype demonstrates real-time sentiment analysis and conversational outcome prediction. It has been effectively expanded beyond the original domain to a new application where the key classification task is to determine whether a visitor's overall sentiment regarding a conversation is positive or negative.

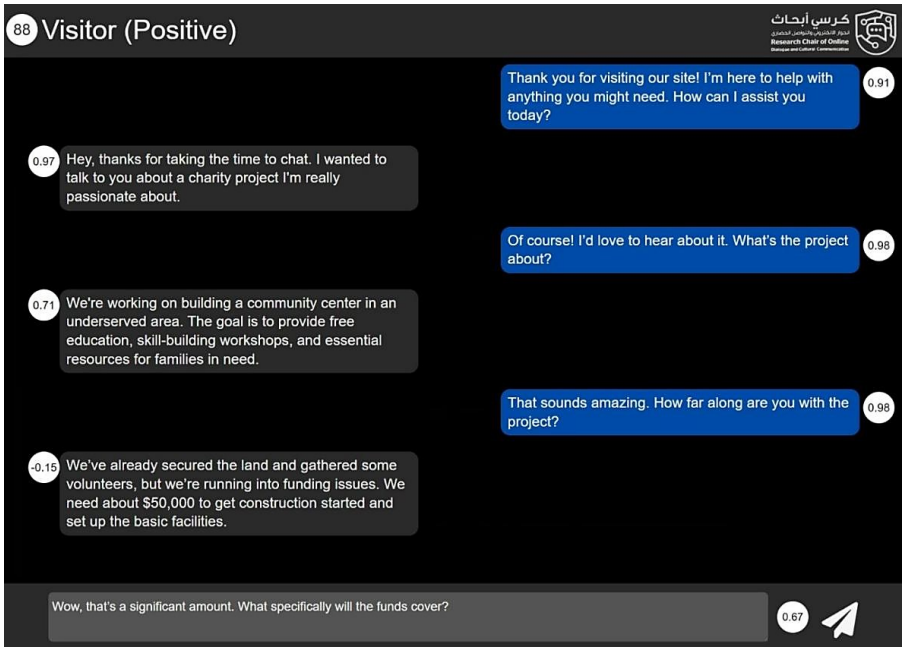

Figure 15. A prototype agent interface demonstrates real-time sentiment scoring, conversation outcome prediction, and pre-send message impact evaluation.

Fig. 15 depicts a screenshot of the agent's UI within this prototype. As shown, a real-time sentiment score is given to each message sent during the conversation, offering detailed emotional insights. Importantly, the expected outcome of the entire conversation (such as "Positive" in the upper-left corner of Fig. 15) is constantly shown, giving the agent an immediate sense of the interaction's trajectory. Additionally, the system predicts and shows the predicted impact of the agent's unsent message right next to the submit button to promote proactive communication. Agents can iteratively improve their responses through this real-time feedback mechanism, ensuring their communication is strategically matched to steer the conversation toward desired outcomes. The successful development and demonstration of these features within the prototype prove the feasibility of real-time deployment and underline the proposed approach's robust adaptability to diverse domains involving customer-agent interactions.

4.5. Real-time Performance and Latency Analysis

We tested the inference latency of RoBERTa sentiment scoring and Enhanced Bi-LSTM outcome prediction on a test set of 1,000 messages to verify the proposed framework's real-time capacity. The experiments were run on a local workstation with an NVIDIA GeForce RTX 4080 GPU. With a 95th percentile latency of 110 milliseconds, the average end-to-end processing time per message was 85 milliseconds. This latency is negligible because turns in human-agent text chat typically occur between seconds and minutes. This confirms that the system can produce sentiment scores and impact predictions instantly without disrupting the conversation, as it falls well within the acceptable threshold for real-time text chat interactions.

5. CONCLUSION

Without non-verbal cues, text-based dialogue systems struggle to infer users' intentions and emotional states; thus, they have limited ability to adapt dynamically and guide the conversation appropriately. This study proposed a machine learning model to address these pressing issues by performing sentiment analysis and predicting conversational outcomes in real time. The main goals of this research were to correctly categorize conversation outcomes and to assess the potential effect of agent messages in real time before deployment.

Our Enhanced Bi-LSTM model performed strongly on the full 3-class challenge for conversational outcome classification, achieving 90% accuracy. The class with the most language ambiguity, and the main source of error, was the "Wants to Convert" class, as shown by the statistical tests and data validation results. We removed this class in an ablation study, and the model's accuracy improved to 99%, showing its effectiveness on well-defined classification problems. Further, our meticulously trained BERT model outperformed the Bi-LSTM (78%), achieving 80% accuracy in real-time evaluation of message impact. This feature ensures high-quality dialog trajectories, allowing agents to proactively improve their answers. We further validated our approach's practical applicability and extensibility to other customer interaction domains by developing and deploying a prototype demonstrated in a real-time agent interface. This framework represents a significant advance for dialogue systems, enabling more intelligent, adaptive, and outcome-driven communication.

REFERENCES

[1] Waheeb SA, Khan NA, Shang X. (2022) An Efficient Sentiment Analysis Based Deep Learning Classification Model to Evaluate Treatment Quality. Malaysian Journal of Computer Science, 35(1):1-20.

[2] Al-Qablan TA, Mohd Noor MH, Al-Betar MA, Khader AT. (2023) A survey on sentiment analysis and its applications. Neural Comput & Applic, 35:21567-21601.

[3] Devlin J, Chang MW, Lee K, Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers): 2019; Minneapolis, Minnesota. pp 4171-4186.

[4] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692v1.

[5] Siami-Namini S, Tavakoli N, Namin AS. (2019) The Performance of LSTM and Bi-LSTM in Forecasting Time Series Data. In *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 3285–3292. doi: 10.1109/BigData47090.2019.9005997.

[6] Naresh A, Krishna PV. (2021) An efficient approach for sentiment analysis using machine learning algorithm. Evol Intell, 14(2):725-731. doi: 10.1007/s12065-020-00429-1.

[7] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., Acharya, U. R. (2021) ABCD: Attention-based bidirectional CNN-RNN-deep model for sentiment analysis. Future Generation Computer Systems, 115, 279-294, doi: 10.1016/j.future.2020.08.005.

[8] Sharin, N. H., Kartiwi, M. (2025) Identifying Technical and Vocational Education and Training (TVET) Sentiment from Social Media Using a Machine Learning Approach. *IIUM Engineering Journal*, 26(3), 295-303. doi: 10.31436/iiumej.v26i3.3739

[9] Buntoro, G., Arifin, R., Syaifuddiin, G., Selamat, A., Krejcar, O., Hamido, F. (2020) The Implementation of the Machine Learning Algorithm for the Sentiment Analysis of Indonesia's 2019 Presidential Election. *IIUM Engineering Journal*, 22(1), 78-92. doi: 10.31436/iiumej.v22i1.1532

[10] Poria, S., Majumder, N., Mihalcea, R., Hovy, E. (2019) Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7, 100943-100953, doi: 10.1109/ACCESS.2019.2929050

[11] Bouktif S, Fiaz A, Awad M. (2019) Stock Market Movement Prediction using Disparate Text Features with Machine Learning. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE; pp 1-6. doi: 10.1109/ICDS47004.2019.8942303.

[12] Malhotra P, Patil M, Anchan T, Ganatra K, Gaglani J. (2021) Negotiation Closer Bot: Negotiation Outcome Predictor based on Sentiment Analysis. In 2021 5th International Conference on Information Systems and Computer Networks (ISCON). IEEE; pp 1-5. doi: 10.1109/ISCON52037.2021.9702347.

[13] Li, W., Shao, W., Ji, S., Cambria, E. (2022) BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis. *Neurocomputing*, 467, 73-82, doi: 10.1016/j.neucom.2021.09.057.

[14] Johnson, J. M., Khoshgoftaar, T. M. (2019) Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 27, doi: 10.1186/s40537-019-0192-5

[15] Rukn Al-hewar non-profit Organization. (2025). Home. [https://edialoguec.org.sa/]

[16] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

[17] Van Rijsbergen, C. J. (1979). Information Retrieval (2nd ed.). Butterworths.

[18] Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In the 2010 20th International Conference on Pattern Recognition (ICPR), IEEE, pp. 3121- 3124.