Khazanah Journal of Religion and Technology
Sunan Gunung Djati State Islamic University, Bandung, Indonesia
https://journal.azkia.id/index.php/kjrt
Online ISSN: 2987-6060
Volume 4 Issue 1 (2026)
DOI: https://doi.org/10.66948/kjrt.v4i1.51

Evaluation of DistilBERT and BiLSTM Models for the Development of Islamic Chatbots Based on Tag Classification

Muhammad Rizki Al-Fathir
Tricada Intronik Ltd.
Bandung, Indonesia
alfthr378@gmail.com

Muhammad Saifurridwani 'Ijazi
State Treasury Service Office (KPPN) I
Bandung, Indonesia
m.saifurridwani@gmail.com

Nabila Lailatanzila
Dicoding Akademi Indonesia Ltd.
Bandung, Indonesia
nabila5des@gmail.com

Nirwan Rasyid Ridlo
LEN Persero Ltd.
Bandung, Indonesia
nirwanrasyidridlo@gmail.com

Riza Anwar Fadil
Kreasi Bali Sasmita Ltd.
Bandung, Indonesia
rizaafadil@gmail.com

***Abstract*—** This study evaluates the performance of DistilBERT and Bidirectional Long Short-Term Memory (BiLSTM) models for intent classification in Islamic chatbots, with the main challenge being a highly imbalanced dataset containing 2,031 unique intents. Following the CRISP-DM methodology, the DistilBERT model was fine-tuned using Focal Loss to address class imbalance, while the BiLSTM model was built from scratch with a standard loss function. The evaluation results demonstrated the absolute superiority of DistilBERT, achieving an accuracy of 65.15%, far surpassing BiLSTM, which achieved only 34.50% due to severe overfitting. Although the final model sizes of both were similar, DistilBERT training proved to be significantly more efficient. These findings demonstrate that a Transformer-based architecture combined with an appropriate strategy, such as Focal Loss, is a much more robust and effective solution for large-scale, imbalanced text classification in specific domains. The practical feasibility of this approach was validated through its successful implementation in a publicly accessible, functional chatbot prototype.

Keywords- BiLSTM, Chatbot, Deep Learning, DistilBERT, NLP

I. INTRODUCTION

Chatbots have permeated various aspects of modern life, transforming them from simple tools into intelligent and interactive virtual assistants. Their ability to provide instant information, automate customer service, and personalize user experiences has made them a crucial solution for individuals and organizations. This evolution is driven by rapid advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP) [1]. Modern chatbots rely heavily on advances in Deep Learning (DL), which enables them to understand user intent, extract important entities from queries, and generate coherent and relevant responses.

As digitalization increases, Muslims around the world are also seeking easier and faster ways to access religious

information. However, currently available digital platforms often struggle to provide accurate, contextual, and accessible Islamic information, especially for complex questions or those requiring a deep understanding of primary sources such as the Quran and Hadith [2]. Arabic, an official language in 22 countries and spoken by over 400 million speakers, is recognized as the fourth most widely used language on the internet [3]. Muslims use Classical Arabic in their daily prayers, while Modern Standard Arabic (MSA) is used in formal settings such as the media and classrooms, and Arabic dialects (AD) are used in everyday communication [2]. This gap creates a need for intelligent and reliable systems, such as Islamic chatbots, that can bridge this information gap.

Modern chatbot development relies heavily on advances in Deep Learning (DL), particularly in NLP. DL models are capable of learning complex patterns in text data. Two prominent model architectures in this context are Long Short-Term Memory (LSTM), specifically Bidirectional LSTM (BiLSTM), and Transformer-based models like BERT [1][3]. LSTM and Bi-LSTM networks have been used to produce promising results on a variety of tasks, including language modeling and speech recognition. BiLSTM excels at capturing long-term dependencies in text sequences from both forward and backward directions, making it effective for context understanding tasks. Bi-LSTM-CRF, for example, can efficiently leverage past and future input features [3]. Meanwhile, BERT, with its multi-head *attention mechanism*, has revolutionized NLP with its ability to learn rich and contextual language representations through *pre - training* on large data corpora [1].

While BERT is powerful, its large size and high computational requirements can be a bottleneck, especially for applications requiring fast responses or running on resource-constrained devices [1]. A study by Alkahtani et al. (2021) showed that DistilBERT, improved through fine-tuning, can achieve competitive and efficient text classification performance compared to other large models [4]. However, applying these state-of-the-art models, including DistilBERT and BiLSTM, in the context of Islamic chatbots presents unique challenges. Islamic source texts such as the Quran and Hadith often use Classical Arabic, which is rich in semantic nuances, complex morphology, and specific religious idioms, unlike modern Arabic or other common languages. Arabic is also characterized by its non-concatenative morphology and has issues with ambiguity due to the absence of short vowels and clear case markers in contemporary texts. The development of Transformer-based models specifically for Arabic, such as AraBERT, has shown significant progress in understanding Arabic across various dialects and formal contexts [5].

Furthermore, a hybrid approach combining CNN and BiLSTM has been successfully applied to classify Arabic medical queries, demonstrating the potential for similar models to be applied in the religious domain [6]. It is crucial to ensure that the responses provided are not only linguistically accurate but also sharia-compliant and contextually appropriate, avoiding misinterpretation or the spread of misinformation. The availability of high-quality datasets specifically annotated for Islamic NLP tasks remains a challenge. While some efforts have been made to build corpora and resources for Arabic and MSA dialects, much work remains, especially for Classical Arabic [7].

Thus, this study aims to fill this gap by in-depth analyzing the ability of DistilBERT and BiLSTM to understand and process Islamic questions, considering the unique challenges of Arabic and religious texts. We will also evaluate the performance of these two models (and their potential combinations) using relevant metrics and datasets curated specifically for the Islamic context, and provide practical insights into which model architecture is best suited for building efficient, accurate, and reliable Islamic chatbots, given the need for rapid responses and high contextual accuracy. This research is expected to significantly contribute to the development of more intelligent and culturally sensitive AI systems in the religious domain, while paving the way for broader applications of NLP in the Islamic context.

II. RELATED WORKS

Various studies have been conducted to explore the effectiveness of deep learning models in chatbot development, particularly in the Arabic language and Islamic domains. One early study in the Islamic chatbot context was SoulS scripture, which utilized BERT to understand questions related to the Quran and Hadith in a contextual and relevant manner [8]. Another study adopted a fine-tuning approach to DistilBERT to analyze sentiment from marketplace reviews with an Islamic perspective, demonstrating the efficiency and accuracy of DistilBERT in the religious domain [9].

In developing a chatbot for halal tourism, Hafidz et al. used multilingual BERT to handle questions related to Sharia tourism in West Sumatra, demonstrating the advantages of transformers in complex natural language processing. BiLSTM was used in a study by Anki et al. to produce a chatbot with high accuracy (~99.5%), demonstrating the strong potential of this architecture for deep conversational tasks [10][11].

Anki and Bustamam then conducted a comparative evaluation between LSTM and BiLSTM for chatbots and found that BiLSTM consistently outperformed in understanding the context of two-way conversations [12]. Another study compared transformers and BiLSTM in Arabic poetry meter classification, where BiLSTM achieved an accuracy of up to 90.53%, demonstrating the effectiveness of this model for understanding complex Arabic language structures [13]. Furthermore, Abdul-Mageed et al. developed Arabic-specific models named ARBERT and MARBERT, which have been used in various NLP tasks, including sentiment analysis and topic classification in the Arabic-Islamic context [14].

Research on Thai conversational chatbots has explored the use of BiLSTM combined with data augmentation techniques to improve classification performance [15]. Deep Recurrent Neural Networks (DNNs), specifically BiLSTM models, have been shown to produce conversational AI chatbots with high accuracy [11]. Modern Natural Language Processing techniques have been utilized to develop chatbots such as InfoGenie, specifically designed to improve the process of extracting information from users [16]. In the context of Islamic education, Natural Language Processing (NLP) has been implemented to build chatbot applications

aimed at assisting the learning of classical Islamic texts in Islamic boarding schools [17].

Chatbots are becoming increasingly widespread in modern life, evolving from simple tools to intelligent and interactive virtual assistants. Chatbots are now capable of providing instant information, automating customer service, and personalizing the user experience, making them essential solutions for both individuals and organizations. This evolution is driven by rapid advances in Artificial Intelligence (AI), particularly Natural Language Processing (NLP). Modern chatbots rely heavily on Deep Learning (DL) to understand user intent, extract critical information, and generate relevant responses [18].

III. RESEARCH METHODS

This research uses the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to ensure a systematic and structured approach [19]. This methodology guides the project from goal definition to implementation of the final results. Given the research's focus on comparative evaluation, each stage will be geared toward supporting the objective of comparing the DistilBERT and BiLSTM models.

3.1 Business Understanding

1. Project Objective: The main objective of this research is to evaluate and compare the performance of two deep learning architectures, namely DistilBERT and Bidirectional Long Short-Term Memory (BiLSTM), in the intent classification task for Islamic-themed chatbots.
2. Functional Requirements: The system developed must be able to receive text input from the user, classify it into one of the predetermined Islamic tags (intents), and provide an appropriate response from a list of prepared responses.
3. Success Criteria: The primary success is measured by the classification performance of each model. The model with the higher accuracy evaluation metric on the test set will be considered superior for this use case.

3.2 Data Sourcing (Data Collecting)

The data source used in this study is the Islamic Supervised Chatbot Dataset obtained from the Kaggle platform. Because the original dataset is in English, it was translated into Indonesian using the MarianMT model to align with the objectives of developing an Indonesian-language chatbot. This dataset was selected due to its relevance for the tag classification task and the development of an Islamic conversation-based chatbot application. However, initial identification revealed a significant weakness, namely the lack of clear and verified sources or references for the existing data.

To mitigate this risk, adding an external service for document retrieval of hadith *is* a strategic move. This mechanism serves as a validation layer that addresses the dataset's weaknesses by providing concrete evidence. By including authoritative hadith references with each relevant response, *the chatbot* not only provides answers but also demonstrates its knowledge base, hopefully building trust and enabling users to verify their own beliefs.

3.3 Data Understanding

This stage focuses on exploratory analysis to understand the composition and characteristics of the dataset.

1. Intent Distribution Analysis: Check the number of patterns for each tag to ensure the dataset is sufficiently balanced. Imbalanced data can introduce bias into the model.
2. Vocabulary Analysis: Identifying the number of unique words and word frequency distribution in corpus patterns.
3. Text Statistics: Analyzes the average, minimum, and maximum lengths of sentence patterns to determine parameters such as sequence length when padding.

3.4 Data Preparation

This stage is crucial to convert raw text data into a numeric format that can be processed by a machine learning model.

1. Data Extraction: Read a JSON file and extract all patterns and their corresponding tags into a dataframe or list.
2. Cleaning and Normalization: Converts all text patterns to lowercase and removes punctuation and irrelevant characters.
3. Label Encoding: Converting text tags into numeric (integer) representation.
4. Tokenization and Sequencing: This process is done differently for both models:
   a. For BiLSTM: Text is tokenized into words. Each word is then mapped to an integer index. Each sequence (sentence) is then padded to a uniform length.
   b. For DistilBERT: Text is tokenized using the WordPiece tokenizer specific to DistilBERT. This tokenization includes adding special tokens such as [CLS] and [SEP].
5. Data Distribution: The dataset is divided into three parts: training set, validation set, and test set in an 80:10:10 ratio. The test set will not be used until the final evaluation stage to ensure objective assessment.

3.5 Modeling

At this stage, two different text classification models are built and trained separately.

3.5.1 Modeling with BiLSTM

Model Architecture: The BiLSTM model is built with several layers:

1. Embedding Layer: Converts integer sequences into dense vectors. These vectors can be trained from scratch.
2. Bidirectional LSTM Layer: Processes data sequences from two directions (front to back and back to front) to better capture context.

3. Flatten layer
4. Dense Output Layer: A fully-connected layer with a softmax activation function to generate probabilities for each class (intent).
5. Training Process: The model is compiled using the Adam optimizer and the categorical_crossentropy loss function, then trained on the training data.

3.5.2 Modeling with DistilBERT

Model Architecture (Fine-tuning): The approach used is fine-tuning of the pre-trained DistilBERT model.

1. DistilBERT Base Model: The DistilBERT base model is used as an encoder to generate contextual representations of the input text.
2. Pooling: The output of the [CLS] token is taken as a representation of the entire sentence.
3. Dense Classification Layer: A fully-connected layer with softmax activation is added on top of the DistilBERT model to perform intent classification.
4. Training Process: All models (including the weights from DistilBERT) were retrained (fine-tuned) on the Islamic dataset. The models were compiled using the Adam optimizer with a low learning rate. Unlike the BiLSTM model, which uses categorical cross-entropy, this model uses Focal Loss as its loss function. The focal loss was chosen to address the significant class imbalance in the dataset by giving more weight to difficult-to-classify samples.

3.6 Evaluation

This stage focuses on comparing the performance of the two models using a test set that has never been seen before.

1. Evaluation Objective: Determine which model (DistilBERT or BiLSTM) provides the best intent classification performance on the test data.
2. Classification Metrics: Model performance will be measured using standard metrics for multi-class classification problems, namely accuracy, precision, recall, and F1-Score.

3.7 Deployment

Although the main focus is evaluation, a prototype can be implemented for demonstration.

1. Model Selection: The best performing model from the evaluation phase (e.g., DistilBERT) will be selected for implementation, but the user can still choose which model to use.
2. Application Logic:
    a. The application accepts text input from the user.
    b. The text is processed and tokenized according to the format required by the selected model.
    c. The model selected by the user predicts the tag (intent) of the input.
    d. The application looks for predicted tags within the JSON file.
    e. One answer from the list of appropriate responses is randomly selected and displayed to the user.
    f. Additional references originating from hadith document retrieval are displayed to the user as hadiths that are relevant to the user's query other than the greeting query.
3. Platform: This prototype can be built as a simple web application using Streamlit deployed on Streamlit Cloud.

IV. ANALYSIS AND RESULTS

This section presents the results of a comparative evaluation between the DistilBERT and BiLSTM models and provides an in-depth discussion of the findings. The presentation of the results follows the *Cross-Industry Standard Process for Data Mining* (CRISP-DM) methodology described previously, starting with exploratory data analysis, modeling results, and final evaluation.

4.1 Results of Exploration Data Analysis

The first step in this research was to thoroughly understand the composition and characteristics of the Islamic Supervised Chatbot Dataset used. This step is a crucial part of the CRISP-DM methodology, as understanding the data will influence the entire subsequent modeling process. This dataset consists of a total of 8,542 input sentences (patterns) classified into 2,031 unique intents (tags).

Analysis of the intent distribution reveals significant data imbalance. On average, each intent has only about 4.2 input sentences, with a minimum of 1 and a maximum of 35. This is further exacerbated by the fact that 50% of the intents (median) have only 3 or fewer input sentences. For example, the intent “not good” has 35 sentences, while an intent like “Why do we wear Ihram?” has only one input sentence. This class imbalance is a common challenge in text classification tasks, as models can potentially be biased towards the majority class and perform poorly on the minority class [20]. A visualization of the distribution of the 20 intents with the highest number of sentences is presented in Figure 1.

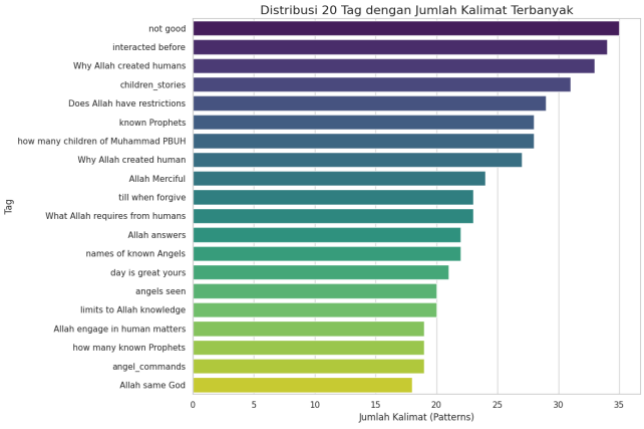

Fig. 1. Bar graph showing the distribution of the top 20 intents.

From a lexical perspective, vocabulary analysis identified 4,821 unique words in the entire data corpus. Meanwhile, statistical analysis of the text showed that the length of input sentences varied considerably. The average sentence length was approximately 6.6 words. Most sentences were relatively short, with 75% of the data containing 8 words or less. However, outliers were also found, consisting of very long sentences, up to 97 words, while the shortest sentence consisted of only 1 word. The overall distribution of sentence lengths can be seen in Figure 2.

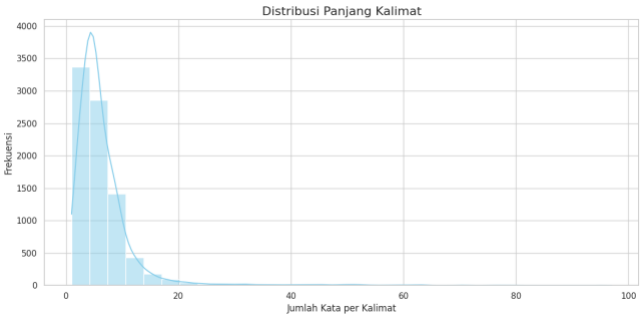

Fig. 2. The histogram shows the distribution of the number of words per input sentence.

These data characteristics, especially class imbalance and sentence length variation, are important considerations in the modeling and evaluation stages to see how reliably the DistilBERT and BiLSTM models handle non-ideal data conditions.

4.2 Model Training Results

After understanding the characteristics of the dataset, the next step is to train the DistilBERT and BiLSTM models. This section presents the results of the training process, focusing on analyzing the learning stability (loss and accuracy curves) and comparing the computational efficiency of the two models.

4.2.1 DistilBERT Training Performance

The DistilBERT model was trained using a fine-tuning approach. In terms of efficiency, the training process took 12.22 minutes (732.92 seconds) and produced a model file measuring 279 MB. The learning curve presented in Figure 3 shows how the model's performance changed during the training process. Training accuracy continues to increase, reaching above 90%, while evaluation accuracy tends to stagnate around 70% after the first few epochs. This gap between the training and validation performance curves indicates that the DistilBERT model is experiencing a tendency towards overfitting, where the model begins to "memorize" the training data too much. A similar phenomenon is also reflected in the loss curve, where the training loss continues to decrease towards zero, while the evaluation loss stops decreasing after the 10th epoch.

Fig. 3. DistilBERT model training and validation graph per epoch.

4.2.2 BiLSTM Training Performance

The BiLSTM model, built from scratch, exhibited distinct training characteristics. The final model size was 289 MB, slightly larger than DistilBERT. The training time was 1 hour. Based on the learning curve in Figure 4, the BiLSTM model exhibited significant overfitting symptoms. A wide performance gap between the training and validation data is evident. Training accuracy reached nearly 100%, indicating the model perfectly memorized the training data. However, performance on the validation data was very low, with accuracy barely exceeding 35%. The loss curve also confirmed this, with training loss dropping dramatically, while validation loss remained high and fluctuating.

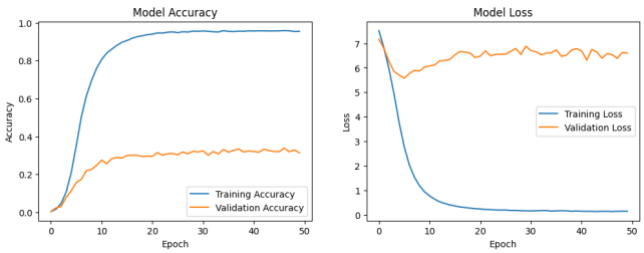

Fig. 4. BiLSTM model training and validation graph per epoch.

Overall, although both models exhibited *overfitting symptoms*, the DistilBERT model demonstrated significantly better generalization ability during training than BiLSTM. Its validation performance was higher and more stable. On the other hand, the BiLSTM model struggled to capture general patterns from the data and tended to memorize, a serious problem for real-world applications. These results provide an initial indication that a *transformer- based architecture* is better suited to the complexity of this task. DistilBERT's superior performance is not only due to its more sophisticated *transformer architecture* but also to the strategic choice of the focal loss. This function effectively suppresses the negative impact of data imbalance, a problem identified during the data analysis stage, allowing the model to focus on learning from the informative minority class.

4.3 Model Evaluation Results

The evaluation phase is the final test to determine which model performs best in classifying intents on the test data set. This data has never been processed by either model during

the training or validation phases, providing the most objective picture of each model's generalization ability. A performance comparison between DistilBERT and BiLSTM is presented in Table 1.

Table 1. Evaluation results

Metric	DistilBERT Model	BiLSTM Model
Accuracy	0.6515	0.3450
Precision	0.5319	0.3561
Recall	0.5273	0.3450
F1-Score	0.5123	0.3387

Based on the table, the DistilBERT model consistently demonstrates significantly superior performance across all evaluation metrics compared to the BiLSTM model. DistilBERT's accuracy is nearly twofold higher than BiLSTM's.

4.4 Deployment Stage Implications

The results of this study directly informed the deployment phase of the Islamic chatbot prototype. As planned, the best-performing model, DistilBERT, was selected as the primary engine for intent classification. With 65.15% accuracy on the test data, this model provided a reliable basis for the application logic to receive user input, predict relevant tags, and retrieve appropriate responses from JSON files. Users could still select the BiLSTM model as a benchmark. Figure 5 shows how the model selection interface is presented to users.

Fig. 5. Display of the selectable model interface

Hadith document retrieval, designed as a validation layer, is crucial here. Although the DistilBERT model provides the best tag predictions, presenting authoritative hadith references to users significantly increases the chatbot's trustworthiness and credibility. This is a strategic move to address the weaknesses of unverified initial data sources, ensuring users can perform independent verification. Figure 6 shows a screenshot of the application providing hadith *retrieval results* that match user input.

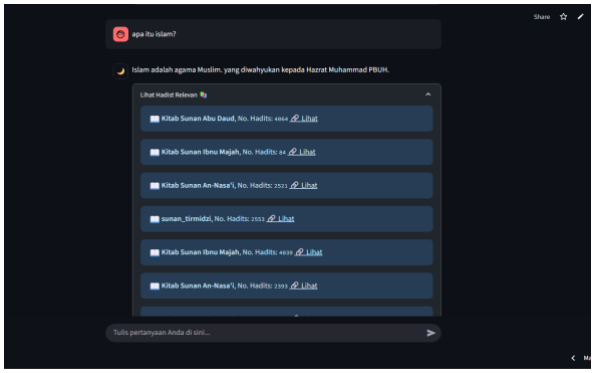

Fig. 6. Display of the interface for presenting relevant hadith references according to user input.

V. DISCUSSION

5.1 Absolute Superiority of DistilBERT in Predictive Performance

The final evaluation results left no doubt: the DistilBERT model was the clear winner. With an accuracy of 65.15% on the test data, its performance significantly outperformed BiLSTM, which achieved only 34.50%. This superiority can be attributed to two main factors. First, DistilBERT's underlying Transformer architecture allows it to more effectively capture contextual relationships in text through self-attention mechanisms. This is crucial for understanding the nuances of diverse Islamic questions.

Second, and no less important, is the methodological decision to use Focal Loss. As identified during the data analysis phase, this dataset has a severe class imbalance problem. Strategically using Focal Loss helps the model focus more on samples from the difficult-to-learn minority class, rather than being dominated by the majority class. This combination of a sophisticated architecture and a well-targeted loss function is key to DistilBERT's success.

5.2 Generalization Failure Analysis on BiLSTM

On the other hand, the BiLSTM model's poor performance on the test data (34.50% accuracy) confirms what was already observed during the training phase: severe generalization failure. The BiLSTM model exhibits symptoms of extreme overfitting, where it perfectly memorizes the training data (near 100% accuracy) but fails miserably when presented with new data. This failure is likely due to a combination of its simpler architecture (compared to transformers) and its inability to handle the highly imbalanced dataset with 2,031 classes without special handling mechanisms. The BiLSTM model, trained with the standard categorical_crossentropy loss function, tends to be biased and only predicts the most frequently occurring classes, resulting in poor overall performance.

5.3 Efficiency Considerations and Practical Implications

In terms of efficiency, the model size comparison yields interesting results. The DistilBERT (279 MB) and BiLSTM (289 MB) model sizes are not significantly different. This

contradicts the common assumption that LSTM-based models are always much lighter than transformer models. The training time for DistilBERT with a GPU accelerator was approximately 12.22 minutes. Given the significant performance difference with comparable model sizes, DistilBERT is clearly a superior and more practical choice for implementing this chatbot application, in line with the deployment plan.

VI. CONCLUSION

This study aims to evaluate and compare the performance of two deep learning architectures, DistilBERT and BiLSTM, in the intent classification task for developing chatbots in the specific Islamic domain with a dataset containing a very large number of classes and imbalance. The results show that the DistilBERT model (accuracy of 0.6515) significantly outperforms the BiLSTM model (accuracy of 0.3450) across all evaluation metrics. DistilBERT's success is supported by its Transformer architecture, which is capable of capturing context superiorly, and the use of Focal Loss, which effectively addresses the problem of data imbalance. In contrast, the BiLSTM model experiences generalization failure due to severe overfitting.

The primary contribution of this research is the demonstration of an effective and practical methodology for building custom chatbots under challenging data conditions. This research goes beyond theoretical evaluation and successfully demonstrates its implementation feasibility through the deployment of a publicly accessible functional prototype through IslamQ Streamlit . The use of strategies such as model hosting on Hugging Face Hub also demonstrates an efficient workflow for real-world applications.

However, this study has limitations centered on the unverified quality and source of the dataset and the chatbot's poor performance compared to chatbots using a generative approach. Therefore, future research can be directed at several areas: first, manually refining and verifying the dataset to improve the quality of the training data. Second, exploring more sophisticated Transformer models pre-trained specifically on Arabic corpora, such as MARBERT, for potential performance improvements. Third, implementing more advanced data augmentation or sampling techniques to address the class imbalance issue more comprehensively.

REFERENCES

[1] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistillBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” Oct. 2019.

[2] “Arabic natural language processing: An overview,” *Journal of King Saud University - Computer and Information Sciences* , vol. 33, no. 5, pp. 497–507.

[3] Z. Huang, W. Xu, and K. Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging,” Aug. 2015.

[4] S. Shah, S. Manzoni, F. Zaman, F. Es-sabery, F. Epifania, and I. Zoppis, “Fine-Tuning of Distil-BERT for Continual Learning in Text Classification: An Experimental Analysis,” *IEEE Access* , vol. PP, p. 1, Jul. 2024, doi: 10.1109/ACCESS.2024.3435537.

[5] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based Model for Arabic Language Understanding,” Jul. 2020.

[6] M. Bahbib, M. Yakhlef, and L. Tamym, “CNN-BILSTM Based-Hybrid Automated Model for Arabic Medical Question Categorization,” *Operations Research Forum* , vol. 6, Jul. 2025, doi: 10.1007/s43069-025-00436-x.

[7] A. Farghaly and K. Shaalan, “Arabic Natural Language Processing,” *ACM Transactions on Asian Language Information Processing* , vol. 8, no. 4, pp. 1–22, Jul. 2009.

[8] A. Malik, AP Gefadri, E. Sidik, and AP Syadrina, "SoulScripture: Chatbot using Bidirectional Encoder Representations from Transformers as a Medium of Spiritual Guidance," *Khazanah Journal of Religion and Technology* , vol. 2, no. 1, pp. 23–27, Aug. 2024.

[9] RF Reza, Muhmmad Thoriq, and Rd. Imam Saepul Millah, "Sentiment Analysis of Marketplace Review with Islamic Perspective using Fine-Tuning DistilBERT," *Khazanah Journal of Religion and Technology* , vol. 2, no. 2, pp. 45–54, Jan. 2025.

[10] I. Hafidz *et al.* , “Chatbot Model Development Using BERT for West Sumatra Halal Tourism Information,” *Halal Research Journal* , vol. 4, no. 2, pp. 117–131, Jul. 2024.

[11] P. Anki, A. Bustamam, HS Al-Ash, and D. Sarwinda, “High Accuracy Conversational AI Chatbot Using Deep Recurrent Neural Networks Based on BiLSTM Model,” in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)* , Nov. 2020, pp. 382–387.

[12] P. Anki and A. Bustamam, "Measuring the accuracy of LSTM and BiLSTM models in the application of artificial intelligence by applying chatbot program," *Indonesian Journal of Electrical Engineering and Computer Science* , vol. 23, no. 1, p. 197, Jul. 2021.

[13] AM Mutawa and S. Sruthi, “A Comparative Evaluation of Transformers and Deep Learning Models for Arabic Meter Classification,” Mar. 2025.

[14] M. Abdul-Mageed, A. Elmadany, and EMB Nagoudi, “ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* , Stroudsburg, PA, USA, 2021.

[15] N. Lhasiw, T. Tanantong, and N. Sanglerdsinlapachai, “Thai Conversational Chatbot Classification Using BiLSTM and Data Augmentation,” in *Communications in Computer and Information Science* , Singapore: Springer Nature Singapore, 2023, pp. 127–141.

[16] YD Kumar, MP Lahkar, AK Singh, B. Dey, and U. Sharma, “InfoGenie: A Chatbot that Enhances Information Extraction Using Modern Natural Language Processing Techniques,” in *Proceedings of the 1st International Conference on Cognitive & Cloud Computing* , 2024, pp. 239–247.

[17] Y. Sofyan and AFI Arroyan, "Implementation of Natural Language Processing (NLP) in Developing a Chatbot Application for Classical Islamic Text Learning at Pesantren El-Huda El-Islamy," *Journal TIFDA (Technology Information and Data Analytics)* , vol. 2, no. 1, pp. 34–41, June. 2025.

[18] N. Sandu and E. Gide, “Adoption of AI-Chatbots to Enhance Student Learning Experience in Higher Education in India,” in *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)* , Jul. 2019.

[19] D. Ruswanti, D. Susilo, and R. Riani, “Implementation of CRISP-DM in Data Mining to Predict Income with the C.45 Algorithm,” *Go Infotech: STMIK AUB Scientific Journal* , vol. 30, no. 1, pp. 111–121, Jun. 2024.

[20] C. Padurariu and M.E. Breaban, “Dealing with Data Imbalance in Text Classification,” *Procedia Comput Sci* , vol. 159, pp. 736–745, 2019.