Halal Research **4 (2)** (2024) 117-131

Chatbot Model Development Using BERT for West Sumatra Halal Tourism Information

Irmasari Hafidz[a*], Bayu Siddhi Mukti[a], Qudsiyah Zahra Ilham Naseela[a], Ahmadhian Daffa Yudistira[a], I Putu Adhitya Pratama Mangku Purnama[a], Nurul Fajrin Ariyani[b, c], Hanim Maria Astuti[a, d], Aris Tjahyanto[a]

[a]Information Systems Department, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

[b]Informatics Department, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

[c]Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, United Kingdom

[d]School of Information, Florida State University, Tallahassee, FL, 32304, United States

ABSTRACT

Halal tourism in Indonesia is growing rapidly, highlighting the need for Muslim halal tourism information that gives unique and relevant information for traveller. However, providing timely and reliable information, specifically related to halal tourism remains a challenge. This research aims to address this by developing a chatbot model using BERT for West Sumatra's halal tourism. A total of 1,125 questions were prepared, divided into nine categories or labels with 125 questions each. Eighty percent (900 questions) was used to fine-tune the BERT-base-multilingual-uncased model, while 20% (225 questions) was used for evaluation. The model was fine-tuned using BertForSequenceClassification for three epochs with a batch size of 32. The chatbot demonstrated high performance, with an overall accuracy of 0.96. However, the lowest precision value was 0.89 for "*budaya*" (or culture) and "*kuliner*" (or culinary) labels, and the lowest recall value was 0.64 for the "*belanja*" (or shopping) label, yielding an F1-score of 0.78. This study describes chatbot model development, from data collection and pre-processing to experimental setup and model training using a fine-tuned BERT-base-multilingual-uncased model. The chatbot model can group user queries into specific purposes and respond to a predefined list. However, one label (e.g "*belanja*" or shopping) may have the lowest recall due to a poor training dataset and query variation.

Keywords: *chatbot, BERT, natural language processing, halal tourism, West Sumatra*

* Corresponding author. Tel: ++62812-4999-8827;
Email address: irma@its.ac.id

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31th, 2024

1 Introduction

Halal tourism is a rapidly growing sector in Indonesia. This shows Indonesia's great potential to become a major player in the global halal tourism industry as proven by the prestigious award as the Best Muslim Friendly Destination of the Year in the 2023 Mastercard CrescentRating Global Muslim Travel Index (GMTI) held in Singapore [1]. With its cultural and natural wealth, Indonesia has great potential to become a popular halal tourist destination. According to data from the West Sumatra Central Statistics Agency (BPS), in 2023 there will be 48,933 foreign tourists and 13,341,025 domestic tourists visiting West Sumatra, and it is projected that 20% of them will be Muslim tourists [2]. The majority of the Muslim population, attractive cultural and natural riches, and government support are the driving factors. It is proven that tourism levies have a significant effect on the local income of Padang City, West Sumatra. This city has natural tourism potential with beautiful panoramas, interesting historical sites, unique culture and strategic location [3]. As information and communication technology advances, more people plan and make travel decisions online [4]. However, there are challenges in finding valid and reliable information. For instance, Expedia Group's Path to Purchase study reveals that travelers spend significant time researching before booking, often viewing hundreds of pages over several weeks [5]. This may result of wasting the user's time. Chatbot development plays a role in increasing customer satisfaction, especially in tourism. Chatbot response time is one of the focus study and has a major impact on service quality [6]. Chatbot technology is needed to improve tourism in West Sumatra, Indonesia, which promotes the halal industrial economy.

Previous research has highlighted the importance of chatbots in enhancing user experience in various sectors. However, the exploration specifically focusing on chatbot development tailored in halal tourism, particularly in the Indonesian context (i.e Bahasa Indonesia) still limited. Natural language processing techniques such as BERT with its multilingual context can help address the challenge. This research aims to fill this gap by developing a BERT-based chatbot model for West Sumatra halal tourism. This study also contributes in developing framework for chatbots, especially for halal tourism in Bahasa Indonesia and the Indonesian context.

2 Materials and methods

2.1 Materials

2.1.1 Halal Tourism

In 2000, the halal industry experienced rapid development and began to expand into the lifestyle sector, including tourism, hospitality, fashion, cosmetics, medical care, and others. This development is driven by the rapid growth of Muslim populations worldwide and their increasing purchasing power. According to PEW Research data, Muslims are now the largest population in the world, with 1.7 billion people. The term halal tourism is often equated with Islamic tourism, sharia tourism, halal travel, and others. Halal tourism is a business activity which follows halal business practices. In Islamic literature, "halal" refers to what the religion commands and can be consumed according to the Koran or the Hadith of the Prophet [7]. Tourism in Arabic is known as "al-Siyahah", "al-Rihlah", and "al-Safar" or in English it is called "tourism", which means activities or travel undertaken by individuals or groups, both within the country and abroad. By utilizing supporting services and facilities provided by the government and the community to fulfill the desires of tourists with certain goals [8].

According to DSN-MUI fatwa Number: 108/DSN-MUI/X/2016 concerning Guidelines for Organizing Tourism Based on Sharia Principles, halal tourist destinations are geographical areas located in one or more administrative regions in which there are tourist attractions, religious and public facilities, tourism facilities, accessibility, and society are interrelated and complement the realization of tourism following sharia principles [9]. The key components of halal tourism ensure that all aspects of a Muslim tourist's experience adhere to Islamic principles and provide comfort, convenience, and inclusivity. This is different from ordinary tourism that broadly includes various forms of travel for leisure, business, or other purposes, such as cultural, adventure, eco, and luxury experiences [10].

In 2023, the Muslim market experienced a notable increase, with around 145 million Muslim international arrivals, equating to about 90% of the pre-pandemic figures of 2019. This rebound demonstrates the sector's strong recovery and continued demand [11]. The development of shariah or halal tourism offers an alternative for Indonesia's travel industry, aligning with the global trend of halal tourism as part of the Islamic economy. Looking forward to 2024 the market continues to expand. Indonesia has succeeded to be of Top Muslim Friendly Destination of the Year 2024 in the Mastercard CrescentRating Global Muslim Travel Index (GMTI). This achievement led Indonesia to receive this title twice in a row, the year of 2023 and 2024. One of the tourist destinations being looked at is West Sumatra. The number of foreign tourists to West Sumatra in 2023 reached 56.645 visits compared to the number of visits the previous year which was 4.144 visits. This proves that West Sumatra is one of the main tourism destinations in Indonesia. It also consists of 12 cities and 7 districts. Each district and city has a destination that can be developed as a tourist destination.

The Regional Government of West Sumatra Province as a region won 4 categories, namely the best halal tourist destination, the best culinary destination, the best halal tourist travel service agency, and the best halal restaurant. Based on these various achievements, this area is designated by the Ministry of Tourism and Creative Economy as one of the world's halal tourist destinations in Indonesia. West Sumatra is also a region that, implementing the customs of "Adat Basandi Syara', Syara' Basandi Kitabullah" proclaimed development. This philosophy of life is held in Minangkabau society, which makes Islamic teachings the only foundation and/or guideline for behavioral patterns in life. Until 2017, there are 688 locations that can be visited across districts and cities throughout the West Sumatra Province.

2.1.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a set of methods that makes human languages accessible to computers [12]. This field of study is an interdisciplinary subfield of computer science, involving artificial intelligence, computer science, and linguistics [13]. NLP as a study began in the 1950s with Noam Chomsky theoretical study of language grammars in 1956 which influenced the creation of Backus-Naur Form (BNF) by scientist John Backus and Peter

Naur in 1963 [14]. The BNF is used to specify a context-free grammar (CFG) as defined by Chomsky . Early on, the grammar of a language was analyzed by scientists, who then described the grammar using the BNF [15]. This was time consuming and not representative of human language, which has many exceptions to the grammar rules. Later, more efficient methods were taken, such as machine learning and deep learning [16].

2.1.3 Deep Learning and BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language model introduced by researchers at Google in 2018 [17]. Initially, BERT was implemented and trained in the English language using Toronto BookCorpus and English Wikipedia. In 2019, Google rolled out BERT in 72 languages for Google Search [18]. BERT is designed to be able to train bidirectional representations of unlabeled text in depth by considering both left and right context for all text layers. BERT pre-trained models can be refined by providing an additional layer to create a state-of-the-art model for various tasks. BERT improvises using a Masked Language Model (MLM). MLM randomly masks some word tokens from its input and attempts to predict the real vocabulary token ids of the masked words based on the context in the input sentence. When tested using the GLUE benchmark, BERT managed to score better than OpenAI GPT in all metrics (MNLI, QQP, QNLI, SST-2, CoLA, STS-B, MRPC, RTE), having an average score of 79.6 for BERT base and 82.1 for BERT large, surpassing OpenAI GPT's average of 75.1. This model was chosen because it performs well in a variety of natural language processing (NLP) tasks and can handle text in a variety of languages. The use of BERT-base-multilingual-uncased in this research is expected to increase accuracy and efficiency in question classification tasks.

2.1.4 Chatbot

Chatbots, a portmanteau of the word "chat" and "robots" are computer programs designed to mimic natural conversations with human correspondents. A chatbot takes natural human language text as input and gives a return relevant text in human languange. Chatbots have various applications in education such as for learning assistance, information gathering, school administration, language learning, in healthcare such as for informing patients about existing products and services, recommending diagnosis and treatments based on patients' symptoms, reminding patients to take their medications, and in customer service such as for product support, answering customer question about product availability and specification, and helping potential buyer choose the appropriate product for their needs [19]. Major developments occurred in late 2022 as OpenAI announced and released ChatGPT, a chatbot based on generative pre-trained transformer that is capable of making a very convincing human-like conversation [20].

2.1.5 Doherty Threshold

According to the Doherty Threshold, a system should respond in under 400 ms (0.4 seconds) to ensure efficient interaction between the computer and the user. If a system takes longer than 400 ms to respond, it should provide an indication that the process is ongoing, such as a loading bar or spinning wheel animation [21]. User productivity increases when the interaction between the user and the computer is such that neither is waiting for the other

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31st, 2024

[22]. In the development of chatbots, this theory is applied by ensuring that the chatbot responds quickly, keeping users informed and preventing frustration.

2.2 Methods

The methodology used in this research is divided into five processes, as shown in Fig. 1 Research Methods. The process starts with (1) preparing the question dataset, (2) cleaning the dataset, (3) training the model, (4) predicting intent, and (5) evaluating the model. Each of these steps is crucial to ensure the accuracy and reliability of the chatbot's performance.

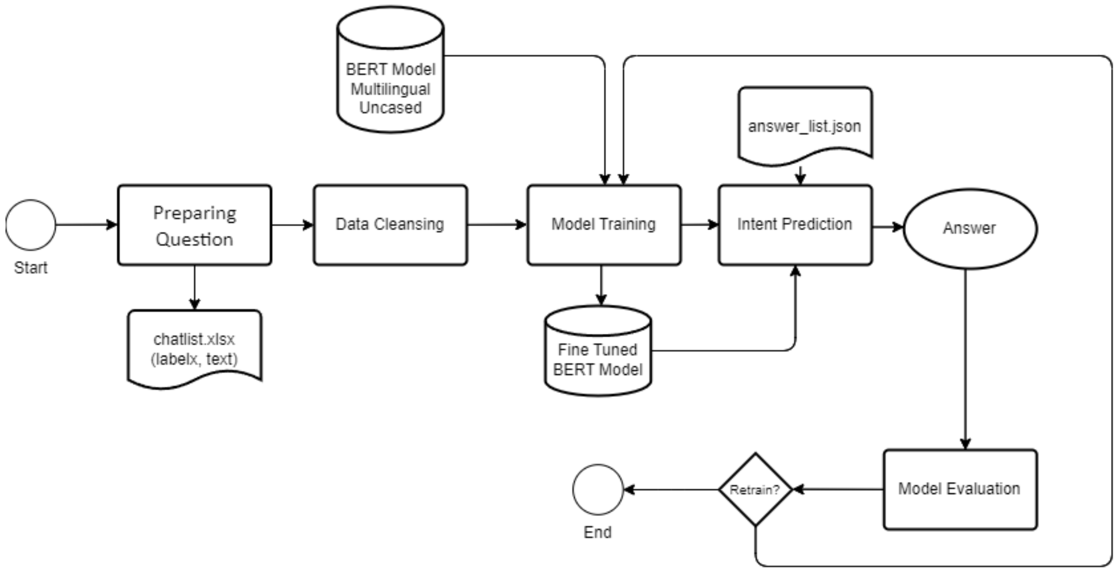

Fig. 1 Research Methods

2.2.1 Preparing Question

In the first stage, the process of creating a question dataset will be carried out. In this process, the types of questions (labels) that the chatbot can answer will be determined, along with the number of labels. The chatbot categorizes user queries into nine labels, which were created by the author based on their analysis of West Sumatra tourism articles in Bahasa Indonesia. Then from each type of question, many variations of questions that might be asked will be created. The questions also developed by the author based on these labels through their own identification. The dataset is created in the format as shown in Table 1.

Table 1. Dataset format

Label	Text
umum	Apa yang dimaksud dengan wisata halal?
umum	Apa definisi dari wisata halal?
…	…
kuliner	Makanan halal apa yang terkenal di Sumatra Barat?
kuliner	Apa saran kuliner halal di Sumatra Barat?
…	…

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31th, 2024

2.2.2 Data Cleaning

Data Cleaning is the process of removing noise from data, to maintain data integrity and uniformity before modeling. In this stage, four methods are applied. First, the text format is standardized to lowercase. Second, special characters, punctuation marks, extra spaces, and extra dots are removed. Third, words that do not have significant meaning (stopwords) are removed from the text. Finally, stemming is used to simplify word forms and change various word forms into their basic forms. These steps are crucial to preparing the dataset for subsequent analysis and modeling, ensuring that the dataset is clean and reliable. Some examples of data cleaning processing can be seen in Table 2 below.

Table 2. Data Cleaning Process

Text	Lowercase	No Punctuation	No Stopwords	Stemming
Apa saja pusat perbelanjaan di Sumatra Barat?	apa saja pusat perbelanjaan di Sumatra barat?	apa saja pusat perbelanjaan di Sumatra barat	pusat perbelanjaan Sumatra barat	pusat belanja Sumatra barat
Paket wisata halal yang bagus di Sumatra Barat apa saja?	paket wisata halal yang bagus di Sumatra barat apa saja?	paket wisata halal yang bagus di Sumatra barat apa saja	paket wisata halal bagus Sumatra barat	paket wisata halal bagus Sumatra barat
Apa saja aktivitas halal yang bisa dilakukan di Sumatra Barat?	apa saja aktivitas halal yang bisa dilakukan di Sumatra barat?	apa saja aktivitas halal yang bisa dilakukan di Sumatra barat	aktivitas halal Sumatra barat	aktivitas halal Sumatra barat
Apa pengertian dari wisata halal?	apa pengertian dari wisata halal?	apa pengertian dari wisata halal	pengertian wisata halal	erti wisata halal
Apa kuliner halal terpopuler di Sumatra Barat?	apa kuliner halal terpopuler di Sumatra barat?	apa kuliner halal terpopuler di Sumatra barat	kuliner halal terpopuler Sumatra barat	kuliner halal populer Sumatra barat

2.2.3 Model Training

Model training is the process of preparing a chatbot using Natural Language Processing (NLP) to interpret human questions. The primary goal is to fine-tune a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model specifically designed for natural language understanding tasks.

The fine-tuning method modified the pre-trained BERT model to the dataset using defined training arguments, allowing the model to learn and recognize data patterns for intent classification tasks. This involves setting training parameters such as batch size, number of epochs, and logging configurations. The model is trained over several epochs to understand and response accurately. Finally, the fine-tuned model is saved for use in the chatbot application, to interpret user input questions based on the patterns learned during the training.

2.2.4 Intent Prediction

Intent prediction is the process of identifying the purpose or intention of a user's statement or question. Intent prediction in chatbots systems seeks to understand user intent about what their text input. The main activities in intent prediction include (1) tokenization, (2) model inference, (3) logits extraction, (4) conversion to probabilities, and (5) determination of the predicted label. Below is the intent used to predict the answer.

Tokenization is the initial step where input text is transformed into tokens that can be understood by machine learning models. Model inference involves using the previous pre-trained model to predict intents based on the input questions from users. Logits extraction involves retrieving the model's output, which is then converted into probabilities using the softmax function. Determining the predicted label is done by selecting the index of the highest probability to determine the corresponding intent.

Responses are developed through intent results. The list of responses is used to generate responses to the chatbot by randomly selecting a variety of answers. The format of intent and list of responses can be seen in Table 3.

Table 3. List of Responses

Intent	List of Responses
umum	Wisata halal adalah konsep pariwisata yang dirancang khusus untuk memenuhi kebutuhan dan preferensi wisatawan Muslim. Konsep ini …
	Wisata halal adalah konsep pariwisata yang dirancang untuk memenuhi kebutuhan dan preferensi wisatawan Muslim, dengan menawarkan berbagai ...
	Wisata halal adalah konsep pariwisata yang menawarkan layanan dan fasilitas yang sesuai dengan syariat Islam, seperti makanan halal, tempat ibadah, …
…	…
kuliner	Sumatra Barat terkenal dengan kuliner Minangkabau yang khas dan tentunya halal. Rendang, misalnya, adalah hidangan daging yang dimasak dengan …
	Kuliner halal di Sumatra Barat sangat beragam dan menggugah selera. Salah satu yang wajib dicoba adalah Ayam Pop, ayam goreng khas Minangkabau …
	Saat berkunjung ke Sumatra Barat, Anda akan dimanjakan dengan berbagai kuliner halal yang kaya rasa. Salah satunya adalah Gulai Kepala Ikan, hidangan …
…	…

2.2.5 Model Evaluation

Model evaluation measures the intent prediction model performance. Its main task is comparing model-predicted intents to the actual intents of the test data. The evaluation

methods include computing metrics such as confusion matrix, accuracy, precision, recall, and f1-score. Table 4 explain chatbot model evaluation measures [26].

Table 4. Metrics in Evaluation Models [26]

Accuracy	Measures ratio of model's correct prediction: true positive (TP) and true negative (TN), to overall number of predictions made. $$a = \frac{TP + TN}{TP + TN + FP + FN}$$
Precision	Measures the ratio of true positive (TP) to the sum of true positive (TP) and false positive (FP). $$p = \frac{TP}{TP + FP}$$
Recall	Measures of the ratio of the true positive (TP) to the sum of true positive (TP) and false negative (FN). $$r = \frac{TP}{TP + FN}$$
f-1 measure	The F1-score is defined as the weighted harmonic mean of its precision and recall. $$F = \frac{1}{a \times \frac{1}{p} + (1 - a) \times \frac{1}{r}}$$

In addition to evaluating the prediction accuracy, the chatbot model reaction speed is also measured. This involves comparing the model's response times against the Doherty Threshold, a benchmark for acceptable system response times in human-computer interactions. Thus, we can ensure that the chatbot understands user intents accurately and replies quickly, improving the user experience and meeting usability criteria. Our experiment code and datasets are available at the GitHub repos: `https://github.com/irhafidz/2024chatbot_halaltourism_WestSumatra`. The implementation code was performed on Google Colab and can be found in the repos with the name `chatbot.ipynb`. All the datasets prepared: questions and pre-defined labels in `chatlist.csv`, answers of the questions in `answers.json`, cleaned dataset in `clean.csv`, testing datasets in `testing.csv`, are available in the Github repos mentioned above, in the folder named `data`.

3 Results and discussion

This chapter presents our BERT-based question classification model experiment findings with tasks to classify questions into various predetermined labels. The results show the model's performance in understanding and classifying questions and the effectiveness of response variations.

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31th, 2024

3.1 Results

A total of 1,125 questions have been prepared for this research. The questions are divided into 9 labels and has 125 questions for each label. Each label is explained in Table 5.

Table 5. List of Definitions

Labels	Definitions
akomodasi	Lodging, hotels, and places to stay in Sumatra Barat
aktivitas	Activities, events, and things to do in Sumatra Barat
belanja	Shopping options and retail experiences in Sumatra Barat
budaya	cultural aspects, traditions, and heritage of Sumatra Barat
kuliner	Related to food, dining, restaurants, and culinary experiences in Sumatra Barat
paket	Tourism packages, deals, or bundled services in Sumatra Barat
rekomendasi	Suggestions or recommendations about places in Sumatra Barat
transportasi	Transportation options, routes, and logistics in Sumatra Barat
umum	General overview and concepts related to halal tourism in Sumatra Barat

Approximately 80% of the data, or 900 questions, was used to fine tune the BERT language model, while the remaining 225 questions were utilized to evaluate it’s performance. This dataset is designed to ensure the model can provide more exact and relevant to the context of the question given.The pre-trained model used is BERT-base-multilingual-uncased. This pre-trained model uses masked language modeling (MLM) to trained on the top 102 languages with the largest Wikipedia corpus [23]. Since this chatbot works by classifying questions, the BertForSequenceClassification [24] transformer will be employed. The transformer model uses head-on sequence classification/regression. Since BERT-base-multilingual-uncased is trained for MLM purposes, using BERT-base-multilingual-uncased for classification purposes requires initializing new weights, which will be obtained during the fine-tuning process. The fine-tuning process is carried out by running 3 epochs using a batch size of 32. After fine-tuning process, the BERT model will predict labels on the testing data and be evaluated. Table 6 shows the classification results and the response given.

Evaluation of the performance of the classification model is carried out by evaluating the resulting precision, recall, f1-score and accuracy values. From the 225 test questions used, a classification report was obtained in the form of precision, recall, f1-score and accuracy values which can be seen in Table 7.

The evaluation results show a high accuracy rate for the classification model, reaching 0.96. The average precision value for all labels is 0.96, indicating consistent performance across most categories. Additionally, the average recall value for all labels is 0.96, showing the model's general effectiveness in retrieving relevant intents. The lowest value for precision of 0.89 belonging to the “budaya” and “kuliner” labels, the lowest value for recall of 0.64 belonging to the “belanja” label, and f1-score of 0.78 belonging to the “belanja” label.

The speed of the chatbot's responses is also a critical factor in evaluating its performance. The average time it takes a chatbot to answer a question, from predicting the label to providing an answer, is 0.074 seconds. The fastest time to answer a question was found to be 0.056 seconds, and the longest time was 0.32 seconds.

Table 6. Example of Classification and Response Results

Number	Results
1	**Question**: *Apa pengertian dari wisata halal?* **Intent**: *umum* **Response**: *Wisata halal adalah konsep pariwisata yang menawarkan layanan dan fasilitas yang sesuai dengan syariat Islam, seperti makanan halal, tempat ibadah, akomodasi ramah Muslim, dan aktivitas wisata yang sesuai dengan nilai-nilai Islam. Tujuannya adalah untuk menciptakan pengalaman wisata yang nyaman, aman, dan sesuai dengan agama bagi wisatawan Muslim.*
2	**Question**: *Bisa sebutkan aktivitas halal di Sumatra Barat?* **Intent**: *aktivitas* **Response**: *Melakukan perjalanan bersepeda di sekitar Danau Maninjau adalah salah satu cara terbaik untuk menikmati keindahan alam Sumatra Barat. Jalur sepeda yang membelah pemandangan alami dan desa-desa tradisional memberikan pengalaman yang mendalam dan menyegarkan. Wisatawan dapat menyewa sepeda di sekitar danau dan menikmati perjalanan santai sambil menikmati udara segar dan panorama yang menakjubkan. Aktivitas ini tidak hanya memberikan kesempatan untuk menjelajahi keindahan alam, tetapi juga memberikan kesempatan untuk berinteraksi dengan masyarakat setempat dan merasakan kedamaian desa-desa di sepanjang rute.*
3	**Question**: *Apa saja ritual adat di Sumatra Barat?* **Intent**: *budaya* **Response**: *Sumatra Barat memiliki beragam budaya dan adat istiadat yang kaya dan unik. Masyarakatnya merayakan peristiwa adat dengan penuh semangat dan kegembiraan, menjunjung tinggi nilai-nilai tradisional yang telah diwariskan dari leluhur mereka. Festival budaya dan seni, seperti Pawai Budaya atau Festival Randai, menjadi momen penting dalam mempertahankan dan mempromosikan keanekaragaman budaya daerah. Adat istiadat dalam pernikahan, upacara kematian, dan musim panen turut menambah warna dalam kehidupan sehari-hari masyarakat Sumatra Barat.*
...	...

3.2 Discussion

The tuned BERT-base-multilingual-uncased in our model achieved 0.96 question task accuracy. The precision, recall, and f1-score result also show good score with an average value of 0.96. This is because the BERT model excels at NLP tasks including classification. BERT model has been trained on a large text corpus and can understand words in context [25]. In Aloria et al. [25], the BERT-base-multilingual-uncased in the BERT-GRU model detected sarcasm in 5250 Hindi and English tweets with hashtag #sarcasm, and #irony, achieved 0.96 accuracy and f1-score. Manias et al. [26] use BERT-base-multilingual-uncased

for Keung's et al. [27] multilingual amazon reviews corpus (or MARC) dataset. The MARC dataset includes six languages: English, Japanese, German, French, Chinese, and Spanish, achieving accuracy and f1-score at 0.7002 and 0.7075 respectively.

Table 7. Classification Report

	Precision	Recall	F1-score	Support
akomodasi	0.93	1.00	0.96	25
aktivitas	1.00	1.00	1.00	25
belanja	1.00	0.64	0.78	25
budaya	0.89	1.00	0.94	25
kuliner	0.89	1.00	0.94	25
paket	1.00	1.00	1.00	25
rekomendasi	0.96	1.00	0.98	25
transportasi	1.00	1.00	1.00	25
umum	1.00	1.00	1.00	25
accuracy			0.96	225
macro avg	0.96	0.96	0.96	225
weighted avg	0.96	0.96	0.96	225

However, in our experiment, numerous labels have lower recall and f1-score values. The lowest recall and f1-score are 0.64 and 0.78 of "*belanja*" (or shopping) label. As shown in the confusion matrix in Table 5, from "belanja" label classifications, only 16 labels were predicted correctly. This table further demonstrates "belanja" label targets had the most classification errors. In this study, Bahasa Indonesia were employed for both model training and testing. This research also contributes to natural language processing use case, especially for underrepresented languages (i.e Bahasa Indonesia corpus) connected to tourism and halal-related text.

The low model performance on these labels is likely due to the insufficient training data and question variety. The chatbot model misclassified "belanja" label most frequently as "budaya" and "kuliner", indicating it may still struggle to distinguish between questions related to these three labels.

Besides accuracy, the system's response speed is an important factor for model quality. The average time needed to answer a question from label prediction to response is 0.074 seconds. Our response time is substantially faster than the Doherty Threshold of 0.4 seconds [21]. Additionally, the longest and fastest times, 0.32 and 0.056 seconds, are below this threshold

.

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31th, 2024

Table 8. Confusion Matrix Results

		Predicted Labels								
		akomodasi	aktivitas	belanja	budaya	kuliner	paket	rekomendasi	transportasi	umum
True Labels	akomodasi*	25	0	0	0	0	0	0	0	0
	aktivitas	0	25	0	0	0	0	0	0	0
	belanja	2	0	16	3	3	0	1	0	0
	budaya	0	0	0	25	0	0	0	0	0
	kuliner	0	0	0	0	25	0	0	0	0
	paket	0	0	0	0	0	25	0	0	0
	rekomendasi	0	0	0	0	0	0	25	0	0
	transportasi	0	0	0	0	0	0	0	25	0
	umum	0	0	0	0	0	0	0	0	25

*in Bahasa Indonesia: *akomodasi* (or accommodation), *aktivitas* (or activity), *belanja* (or shopping), *budaya* (or culture), *kuliner* (or culinary), *paket* (or package), *rekomendasi* (or recommendation), *transportasi* (or transportation), *umum* (or general)

This rapid response time ensures that users experience minimal delay when interacting with the chatbot, enhancing the overall user experience. According to the Doherty Threshold, a system response time of 0.4 seconds or less is necessary to keep the user's flow of thought uninterrupted, thereby maintaining engagement [21]. The chatbot system meets and exceeds this criterion by effectively recognizing and classifying user intents and delivering prompt responses. This combination of accuracy and speed is crucial for maintaining user engagement and satisfaction.

However, various limitations in the development model affect the results and classification system’s performance. One of the main limitations is the data used to train the model. Even though the dataset used consists of 1,125 questions and consists of 9 labels, the variety and complexity of questions still do not cover all the variations of questions that can arise in a real context. Each label has a difference in the expression of the question sentence. Additionally, the chatbot can classify user questions, but cannot deliver dynamic answers. Currently, chatbots can only give pre-written or static answers. This reduces its ability to adapt responses to context or changes in the conversation in real time.

Asking questions out of context is also a limitation of this chatbot implementation. Questions that do not match the topics the model has been trained on can lead to misclassification and irrelevant answers. This shows that although the model has a high level of accuracy, its success depends largely on the quality and suitability of the training dataset to the questions faced in actual use.

These constraints can be overcome by expanding the dataset by adding a greater variety of types of questions that may be asked by users. Dataset quality and representation greatly affect model performance. A more diversified and representative dataset can improve the model's ability to classify various types of questions. Increasing the amount and variety of training data can help the model understand new questions and perform well on labels with lower recall and f1-score values. Although BERT model performs well, exploration to newer or task-specific language models may increase performance. These models may offer specialized optimizations that could address BERT limitations.

To provide more dynamic answers, using an answer generator can be considered as a solution. Based on the context of the query, the chatbot can answer more dynamically and be relevant to user questions. Providing more precise and useful answers in various contexts, the chatbot can improve user interaction and experience. In addition, user feedback can help the model learn and increase accuracy. In this way, chatbots can become more adaptive and responsive to various types of questions, including those outside the initial context in which they were trained.

4 Conclusion

Halal tourism in West Sumatra, Indonesia will experience more growth and development in near future. With government assistance, the region's natural and cultural riches may make it a global halal tourism leader. With the fast development of technological innovation, more people are using web information for travel planning and decision-making.

In this paper, a chatbot model for halal tourism in West Sumatra, Indonesia, has been successfully developed using the BERT language model. The chatbot categorizes user queries into specific intents and responds from a predefined list corresponding to those intents. The chatbot model scored a high accuracy of 0.96, with an average response time of 0.074 seconds, based on 225 test queries. A fast response time under the Doherty Threshold of 0.4 seconds indicates real-time efficiency interactions.

However, four issues and vulnerabilities of the model have been found. First, classification under the “*belanja*” (or shopping) still has low recall and f1-scores. This can happen because the training data for questions is limited. Second, the lack of question diversity may cause numerous classification errors in the model. Third, this chatbot model still cannot handle out-of-context questions not included in its training data. Lastly, the chatbot still cannot provide dynamic responses to users.

Several recommendations for future research include increasing the quantity and variety of training questions. Second, using language models other than BERT may increase the chatbot model accuracy. Third, integrating an answer generator and user feedback mechanism would enable the chatbot to generate more dynamic responses, allowing the model to learn and improve its accuracy.

Received June 30th, 2024; Revised July 29th, 2024; Accepted July 31st, 2024

Acknowledgements

This research is done under the Laboratory of Akuisisi Data dan Diseminasi Informasi (ADDI Labs), Department of Information Systems, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya. We also thank our colleagues and fellow researchers for the knowledge and meaningful discussions, which greatly enriched this research.

References

[1] Redaksi, "Indonesia Kembali Raih Predikat Top Muslim Friendly Destination of The Year 2024 Versi GMTI", Amphuri. https://amphuri.org/indonesia-kembali-raih-predikat-top-muslim-friendly-destination-of-the-year-2024-versi-gmti/ (accessed June 25, 2024).

[2] B. Pusat Statistik, Wisatawan Sumatera Barat Tahun 2023. https://sumbar.bps.go.id/indicator/16/765/1/jumlah-kunjungan-wisatawan-nusantara-kabupaten-kota-di-provinsi-sumatera-barat.html (accessed April 28, 2024).

[3] L. Lusiana, M. Neldi, and S. Sanjaya, "Analisis Investasi Sektor Pariwisata, Jumlah Objek Wisata, Jumlah Kunjungan Wisatawan Dan Retribusi Kawasan Wisata Terhadap Pendapatan Asli Daerah Di Kota Padang", *Jurnal Riset Akuntansi dan Keuangan*. **9, No. 1** (2021) 25–34. doi: 10.17509/jrak.v9i1.28964.

[4] I. Abashidze, "The Influence of Online Platforms on Decision-Making Process and Behavioural Traits of International Travelers", *European Scientific Institute, ESJ*. **20, No. 37** (2024) 51. doi: 10.19044/esj.2024.v20n37p51.

[5] M. Swanson, "This Is How Much Time Travelers Spend Researching Before Booking a Trip, Travel Wires". https://www.travelwires.com/this-is-how-much-time-travelers-spend-researching-before-booking-a-trip (accessed July 31, 2024).

[6] M. R. Harisi and E. M. Hiwono, "Pengaruh Chatbot Usage terhadap Customer Satisfaction", *Jurnal Ilmiah Manajemen Ekonomi Dan Akuntansi*. **1, No. 2** (2024) 66–73. doi: 10.62017/jimea.v1i2.338.

[7] U. Suyatman, Ruminda, and I. Yatmika, "Pengembangan Muslim Friendly Tourism dalam Konsep Pariwisata Budaya di Pulau Dewata", *Pusat Penelitian dan Penerbitan UIN Sunan Gunung Djati Bandung*. (2019).

[8] L. Pusvisasari, "Tafsair Ahkam tentang Pariwisata Syariah", *Jurnal Peradaban dan Hukum Islam*, **3, No. 1** (2020) 39–58. doi: 10.29313/tahkim.v3i1.5642.

[9] N. A. Noviyani and R. T. Ratnasari, "Pengaruh Halal Destination Attributes di Sumatera Barat terhadap Wisatawan Muslim", *Jurnal Ekonomi Syariah Teori dan Terapan*. **8, No. 4** (2021) 401–412. doi: 10.20473/vol8iss20214pp401-412.

[10] R. S, "Exploring Halal Tourism: Current Trends and Future Directions", *Journal of Islamic Marketing*. **11, No. 4** (2020) 920–932.

[11] Mastercard & Crescentrating, "The Global Muslim Travel Index 2024", 9th ed. 2024.

[12] A. Shamshiri, K. R. Ryu, and J. Y. Park, "Text mining and natural language processing in construction", *Automation in Construction*. **158** (2024) 105200. doi: 10.1016/j.autcon.2023.105200.

[13] J. Einstein, "Introduction to Natural Language Processing". MIT Press, 2019.

[14] A. Magableh, "Context-Free Grammar for Aspect-Oriented UML Design Modeling Diagrams", *Journal of Theoretical and Applied Information Technology*. **97 No. 22** (2019) 3377–3385.

[15] P. Johri, S. K. Khatri, A. T. Al-Taani, M. Sabharwal, Suvanov, and A. Kumar, "Natural Language Processing: History, Evolution, Application, and Future Work", in *Proceedings of 3rd International Conference on Computing Informatics and Networks*, A. Abraham, O. Castillo, and D. Virmani, Eds., Singapore: Springer Singapore. **167** (2021) 365–375. doi: 10.1007/978-981-15-9712-1_31

[16] D. McAllester and A. Brown, "A Maximum Entropy Approach to the Hidden Markov Model", in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Association for Computational Linguistics. (2003) 154–161.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Cornell University*. (2019) doi: 10.48550/arXiv.1810.04805.

[18] R. Montti, "Google's BERT Rolls Out Worldwide", Search Engine Journal. https://www.searchenginejournal.com/google-bert-rolls-out-worldwide/339359/ (accessed June 29, 2024)

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

[20] K. Hines, "History Of ChatGPT: A Timeline Of The Meteoric Rise Of Generative AI Chatbots", Search Engine Journal. https://www.searchenginejournal.com/history-of-chatgpt-timeline/488370/ (accessed June 29, 2024).

[21] Doherty, W., & Thadhani, A. "The economic value of rapid response time" (ibm technical report ge20-0752-0). (1982) Zugriff via http://www. vm. ibm. com/devpages/jelliott.

[22] I. Pereyra, "Universal Principles of UX - 100 Timeless Strategies to Create Positive Interactions between People and Technology". Rockport Publishers, 2023.

[23] Hugging Face, BERT multilingual base model (uncased). 2024. https://huggingface.co/google-bert/bert-base-multilingual-uncased (accessed June 29, 2024).

[24] Hugging Face, BertForSequenceClassification. 2024. https://huggingface.co/docs/transformers/v4.41.3/en/model_doc/bert#transformers.BertForSequenceClassification (accessed June 29, 2024).

[25] S. Aloria, I. Aggarwal, N. Baliyan, and M. Ghosh, "Hilarious or Hidden? Detecting Sarcasm in Hinglish Tweets using BERT-GRU", in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). (2023) 1–6. doi: 10.1109/ICCCNT56998.2023.10307171.

[26] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, "Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data", Neural Comput & Applic**. 35, No. 29** (2023) 21415–21431. doi: 10.1007/s00521-023-08629-3.

[27] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The Multilingual Amazon Reviews Corpus", in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics. (2020) 4563–4568. doi: 10.18653/v1/2020.emnlp-main.369.