Journal of Applied Data Sciences
Vol. 4, No. 4, December 2023, pp. 414-430

Long Short-Term Memory-Based Chatbot for Vocational Registration Information Services

Yudo Sembodo Hastoro Langgeng [1], Esther Irawati Setiawan [2, *], Syaiful Imron [3], Joan Santoso [4]

[1,2,4] Information Technology, Institut Sains dan Teknologi Terpadu Surabaya, Ngagel Jaya Tengah 73, Surabaya 60284, Indonesia
[3] PT Mulya Adhi Paramita, Kompleks Wonokitri Indah S24-26, Surabaya, Indonesia

(Received: September 27, 2023; Revised: October 20, 2023; Accepted: November 16, 2023; Available online: December 6, 2023)

Abstract

The development of chatbots can communicate fluently like humans thanks to the Natural Language Processing (NLP) technology. Using this technology, chatbots can provide more accurate and natural responses, providing an almost the same experience as human interaction. Therefore, chatbot technology is in great demand by companies and government agencies as a cost-effective solution for information and administrative services that require little human effort and can operate 24/7. The registration information service at BLK Surabaya still uses an operator who serves prospective trainees and answers questions via social media or chat. However, these operators have limitations in terms of time and effort. The purpose of this study is to examine how to use chatbots to answer questions about registration information training at BLK Surabaya using the Long Short Term Memory (LSTM) algorithm with a dataset of questions collected in the form of Frequently Asked Questions (FAQ) in Indonesian. The dataset consists of 2,636 labeled samples of questions, which were divided into three sets: 60% for training (1,581 pieces), 20% for validation (527 samples), and 20% for testing (528 samples) to evaluate the model's performance. Several steps were taken in implementing this research, including changing the list of questions and answers into the JSON data format, preprocessing, creating LSTM modeling, data training, and data testing. The test results show that Chatbot can provide accurate solutions related to training registration questions with Precision of 88.4%, Accuracy of 87.6%, and Recall of 87.3%.

Keywords: Vocational Training Centre information, AI Chatbot, LSTM, Natural Language Processing

1. Introduction

The Job Training Center (Balai Latihan Kerja / BLK), one of the institutions tasked with providing job training to the general public, also feels the positive impact of using Chat. Before the pandemic in 2020, information and registration services related to job training were carried out conventionally by coming directly to the BLK office to get information and services through customer service, which operators and services carried out were limited to office hours.

After the pandemic, community activities continued with social interaction restrictions, forcing BLK to change conventional information services to online. Online information services are carried out with the help of social media such as websites, Telegram, Instagram, Whatsapp, and Facebook, which can be accessed from anywhere but still have shortcomings. Information services with social media still require operator assistance to answer community questions. Still, limited service time and the existence of similar repetitive questions related to training information make this ineffective. Therefore, an information technology-based system is needed to assist operators in processing question data and displaying information to users. Artificial intelligence (AI)-based chatbots that can be embedded on various platforms commonly used by the wider community, such as websites, Telegram, and Whatsapp, can be used as a solution to information service problems[1], [2].

Chatbot[3], [4] is helpful in its utilization in the fields of education, information search, business, and e-commerce[5]–[7]. Some studies also explain the use of chatbots, including the use of chatbots to serve conversations with older people conducted by Wang, et al.[8] to encourage older people to communicate and increase affective interaction and reduce loneliness and isolation to slow down the level of disability of older people and reduce the financing and care burden of home workers. A multi-layer LSTM embedding mechanism is used to extract semantic information from long

***Corresponding author: Esther Irawati Setiawan (esther@stts.edu)**
DOI: https://doi.org/10.47738/jads.v4i4.128

ISSN 2723-6471

sentences written by older people. In addition to daily conversations, the chatbot system can provide recommendations for activities, restaurants, and health information for older people[9], [10]. The results of this study focus on not only the development of the chatbot system but also the influence of the number of training Epochs on the system's performance. The system cannot learn well if the number of training Epochs is insufficient. If the number of training epochs is also significant, it causes overfitting and affects the training time.

Based on the conducted research, the employed modeling approach demonstrates impressive performance. It achieves an accuracy of 79.96% for top 1 predictions, 93.14% for top 5 predictions, and 94.85% for message matching. These results showcase the model's effectiveness in accurately predicting the desired outcomes and matching messages. In another research conducted by P. Muangkammuen et al. [11], the focus was on designing and developing a Thai-language chatbot tailored to answer frequently asked questions (FAQs) in a specific domain. The research used deep learning modeling, using the Long Short-Term Memory (LSTM) algorithm to classify phrases from questions and answers within each class. The dataset used in this research consisted of 2,636 labeled samples of questions, which were divided into three sets: 60% for training (1,581 pieces), 20% for validation (527 samples), and 20% for testing (528 samples) to evaluate the model's performance. The research results indicated that the chatbot achieved a question processing rate of 86.36% with a correct answer accuracy of 93.2%.

In the field of government, chatbots can serve as automated assistants to provide support and answer various citizen inquiries, as highlighted in government[12]–[14]. The knowledge base for such chatbots can be built by collecting frequently asked questions (FAQs) related to common problems, and additional domains can be created from various eDistrict documents to develop a knowledge base consisting of topics on multiple services provided by the government. These chatbots are designed to communicate with users in everyday language and provide relevant responses based on the intent detected in the user's query. The chatbot communicates with the user and provides answers related to the user's query based on the purpose of the detected message using their daily language[15], [16]; this chatbot supports 11 Indian languages.

While research conducted by Y.D. Prabowo et al. [17] explained that LSTM is recognized as a very effective model in acknowledging and understanding data with a sequential structure. This model has a unique ability to understand the meaning and context of sentences or data sequences, producing output classes with a high degree of accuracy.

Several mobile messenger applications are also increasingly popping up with all their advantages; research [18] explains the use of telegram for learning needs. In this study, several messenger applications were selected, and it was found that Telegram was the right messenger.

Based on the background that has been described, this research will raise the topic of chatbot to answer questions about training registration information at the Job Training Center using the Long Short Term Memory (LSTM) algorithm and the FAQ dataset from a collection of questions that Indonesian-speaking people often ask; the chatbot answer response will be examined for accuracy, precision and recall values[19], [20]. With the growing development of chatbot research and LSTM architectures, the novelty of chatbot development research using Long Short-Term Memory (LSTM) networks lies in the application of Recurrent Neural Networks (RNN) with memory cells specifically designed to capture and maintain long-term dependencies in sequential data.

2. Research Methodology

From the general description previously explained, this research will explore the chatbot domain in the context of providing answers regarding training registration information at work training centers. The approach taken is to use the Long Short-Term Memory (LSTM) algorithm to improve the chatbot's understanding ability and responsiveness. The dataset used consists of a list of FAQs (Frequently Ask Questions/FAQ) of general questions regarding training registration submitted by Indonesian-speaking people. Chatbot performance evaluation will be carried out by measuring accuracy, precision and recall values. The following visual depiction through Figure 1 explores the theoretical framework of this research.

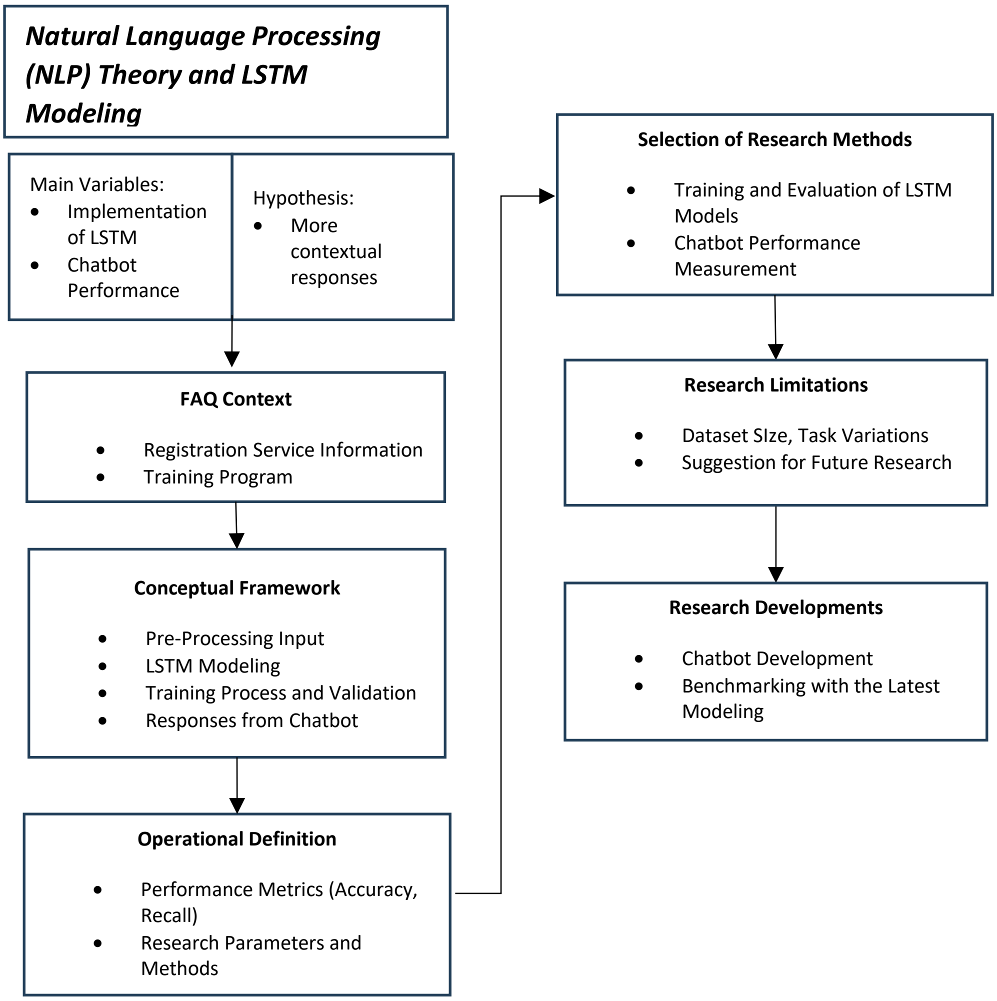


Figure 1. Research Theoretical Framework Chart

We will describe the structure and techniques used in the training enrollment information chatbot research[21]. The procedures carried out in this research are as follows:

1) Create a dataset from a list of questions (FAQ).
2) Perform data pre-processing techniques.
3) Providing tokenization and Padding.
4) Carry out training and validation of the dataset and responses using LSTM modeling, as well as testing and evaluating model performance using K-Fold Cross Validation by calculating Accuracy, Precision and Recall values as well as loss values from chatbot responses.

The chatbot agent will be named ICA (Interactive Chat Assistant), and the chat interface will use the Telegram application. The following system modeling architecture is shown in Figure 2.

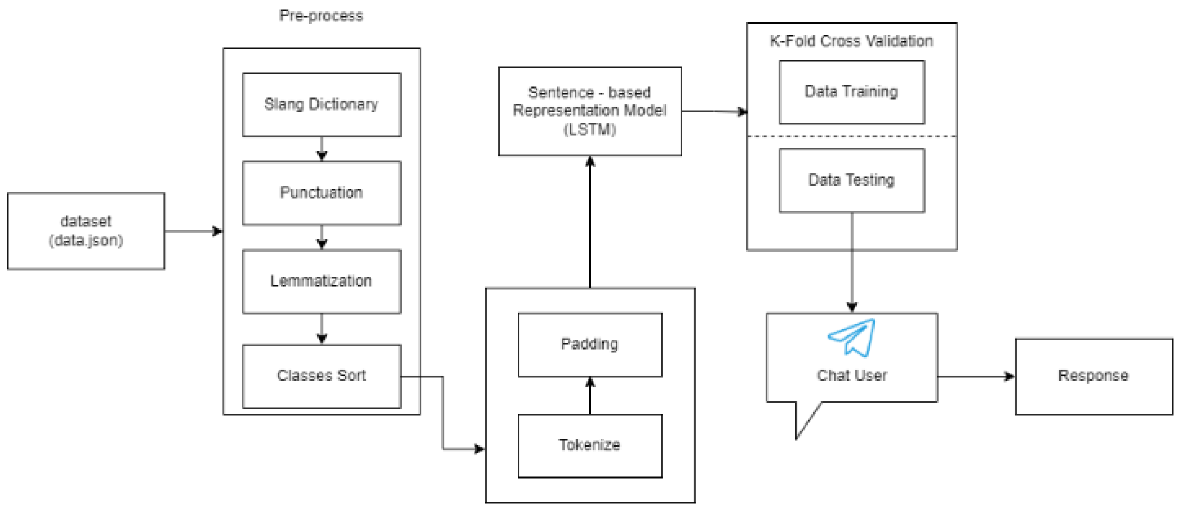


Figure 2. System Modelling Architecture

An Explanation of the system modelling architecture in Figure 2 is as follows:

1) Data.json is a collection of Frequently Asked Questions (FAQ) which have been collected from frequently asked questions, arranged in the form of a file containing a list of questions and answers.
2) Data or text pre-processing techniques focus more on converting raw data into an understandable structure, where primary attention is paid to keywords present in the text that highlight the context of the sentence or paragraph.
3) After the pre-processing stage, the next step is to change the text to numeric (tokenize) the words and adjust the length of the sentences by adding padding so that the length of the sentences is the same as one another, following the longest sentence.
4) The dataset will be trained and validated using LSTM modeling, the dataset will be divided into 2 parts, training data and testing data.
5) Using K-Fold Cross Validation where the dataset will be trained 5 fold times, and the accuracy and loss values are recorded.
6) After obtaining the ideal model, it is then saved into h5 format, to be run on the messenger platform.

2.1. Natural Language Processing

Natural language processing (NLP) [22], [23] is one part of the AI field that focuses on linguistic computing. NLP[24], [25] is the ability of computer programs to understand and comprehend human language as it is spoken and written - referred to as natural language. Defines NLP [26], [27] as a theoretical field regarding a computational technique used to analyze and represent text written in human language. It aims to obtain language processing as humans do so that it can be utilized in various fields. Liddy divides NLP [28], [29] into seven levels NLP in analyzing and interpreting natural language input, namely:

1) Phonology is related to the interpretation of sounds in speech in a word.
2) Morphology is related to the interpretation of the meaning of a word (related to prefixes and suffixes).
3) Lexical deals with the interpretation of the meaning of each word analyzed individually.
4) Syntactic deals with analyzing words in a sentence to find the grammatical structure.
5) Semantics deals with the interpretation of the meaning of words formed due to the interaction of meaning in words in a sentence.
6) Discourse deals with analyzing the meaning of a text consisting of several sentences.
7) Pragmatic, related to analyzing the selection of word usage using the context in the text.

NLP provides the theory and implementation to be applied in various fields[30], [31]. But in fact, any area that deals with text processing is a candidate for NLP. Some application fields that utilize NLP include Information Retrieval, Information Extraction, Question Answering, Summarization, Machine Translation, and Dialogue Systems.

2.2. Long Short Term Memory

Long Short Term Memory[32], [33] or commonly abbreviated as LSTM, is a unique form of RNN that can perform learning on long-term dependencies. This model was introduced by Hochreiter and Schmidhuber in 1997. The architecture of LSTM [34] is shown in Figure 3.

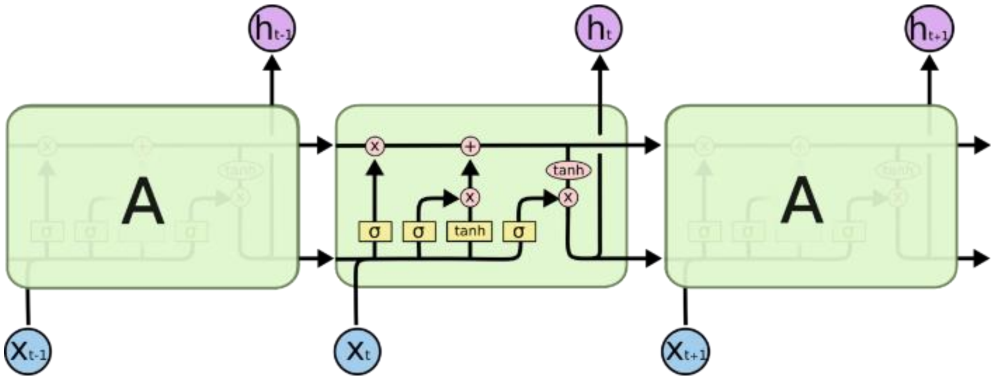


Figure 3. The Architecture of LSTM

The architecture of LSTM All recurrent neural networks have repeating series of neural network modules. LSTM also has the same structure but has the additional gates feature on the cells. LTSM will determine what information to discard from the cell. This decision is made by the forget gate layer. This layer will consider ht-1 and xt so that it will produce an output between 0 and 1. Output 0 represents that the information will be forgotten, while output 1 represents that the information will be noticed.

$$F_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \quad (1)$$

A sigmoid "input gate layer" determines which values to update. Next, a tanh layer creates a vector of new candidate values, Ct, that can be added to the state.

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i \quad (2)$$

$$C_t = tan\ tan\ h\ (W_c.[h_{t-1}, x_t] + b_c) \quad (3)$$

These two layers will be combined in the next step to update the state. Next, the old state will be corrected, Ct-1, to the new cell state Ct. Then, it will be multiplied with the old state by ignoring the information that has been forgotten before. Then, it is added with Ct.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

The last step is to determine what the output is. First, the sigmoid layer will determine which part of the cell to output. Then, the cell will be passed to the tanh Layer (to force the output value between -1 and 1) and multiplied by the output of the sigmoid gate.

$$O_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tan\ tan\ h(C_t) \quad (6)$$

2.3. K-Fold Cross Validation

Cross-validation is a technique in data analysis used to evaluate the performance of a prediction model by dividing the data into subsets and then testing the model on different subsets repeatedly, making it possible to obtain more accurate evaluation results and reduce the risk of overfitting the model. In other words, cross-validation is a method to test how well a model can be generalized to data that has never been seen before[35].

This technique separates the data into two subsets: the training and validation sets. The algorithm will be trained by the training set and evaluated by the validation set.

Cross-validation can be used to determine the error rate or error of the error in predicting the output. The error rate is obtained by calculating the average of the model error rate at each cross-validation iteration. The error rate can be measured by various metrics, such as accuracy, precision, recall, and F1 score, depending on the problem being addressed.

One of the cross-validation techniques is K-Fold Cross Validation [29], [30] where the dataset is divided into K equal subset partitions or parts, where each subset will become validation (testing) data in turn. At the same time, the rest is used as training data. The K-fold CV architecture is shown in Figure 4.

Figure 4. The Architecture of Cross Validation

The dataset is divided into five equal subsets. At each iteration, one subset becomes the validation data, while the other four subsets become the training data. The model is trained on the training data and tested on the validation data, and the error score is calculated. After five iterations, the average error score is estimated to get a more accurate model error score.

K-Fold Cross Validation will generate K different error scores (e.g., accuracy, precision, recall, or F1 score) that can be measured to evaluate the model's overall performance. These error scores can then be averaged to produce a more accurate model error score that represents the model's ability to generalize to data that has never been seen before.

2.4. Text Pre-processing

Text preprocessing is an essential step in natural language processing that helps improve the quality and accuracy of text analysis in various NLP applications. With proper text processing, text can be processed better, and NLP models can better understand the context.

Data or text pre-processing techniques focus more on transforming the raw data into an understandable structure, where primary attention is paid to the keywords present in the text that highlight the context of the sentence or paragraph. Simply, text pre-processing means converting documents into a format machine can easily understand, predict, and analyze through machine learning algorithms. Naseem et al. [36] explain several techniques for doing text pre-processing:

1) Removing Noise
2) Normalizing Text
3) Stopwords Removal
4) Stemming
5) Lemmatization

3. Results and Discussions

This chapter will discuss the process and trials related to the research, including datasets, preprocessing, LSTM Training model, and Testing model.

3.1. Dataset

Collecting data about registration information is the initial stage in the development of Chatbot. This research collects data through interview techniques with registration operators and registrants who come to BLK Surabaya. By using this technique, it can obtain the data needed for research purposes wholly and accurately.

In addition to interviews, questions that often arise related to registration information will be collected. Questions are collected from social media sources (Facebook and IG) and chat messages, then a list of answers is made, which will become a list of FAQs used as datasets in the research.

Eight hundred forty lists of paired questions were collected, then converted into .JSON and used as a database. Examples of some of the paired questions and answers are in Table 1.

Table 1. Paired Questions

Question	Category
Q: What time does the registration service open? (Jam berapa pelayanan pendaftaran buka?) A: Services are open from 08.00 - 15.00 Western Indonesia Time (Pelayanan buka mulai jam 08.00 – 15.00 WIB)	Service Hours
Q: Where is the address of UPT BLK Surabaya? (Dimana alamat UPT BLK Surabaya?) A: UPT BLK Surabaya is located at Jl. Dukuh Menanggal III/29 Gayungan Surabaya. (UPT BLK Surabaya beralamat di Jl. Dukuh Menanggal III/29 Gayungan Surabaya)	Address of BLK Surabaya
Q: What are the training registration requirements? (Apa persyaratan pendaftaran pelatihan?) A: Register requirements : Copy of Identity Card / Family Card (1 sheet), Copy of Vocational / Senior High School Certificate (1 sheet), 3x4 Photo (2 Sheets) (Persyaratan pendaftaran : Fotocopy KTP/KK (1 lembar), Fotocopy Ijasah SMK/SMA (1 lembar), Pas Foto 3x4 (2 lembar))	Registration Requirements
Q: How long does the training take? (Berapa lama waktu pelaksanaan pelatihan?) A: for training time 260 lesson hours or 33 working days (untuk waktu pelaksanaan pelatihan 260 JP (33 hari kerja))	Training Time
Q: How do I register for the training? (Bagaimana cara mendaftar pelatihan?) A: For registration, you can come directly to BLK Surabaya or online through the website (untuk pendaftaran bisa datang langsung ke BLK Surabaya atau secara online melalui website)	How to apply

3.2. Converting Dataset to .JSON Format

Data.json is a file containing a collection of questions and answers from messages often asked by registrants in the form of a Frequently Ask Question (FAQ) list. The data structure consists of Intents, each labeled/tagged according to its category. Within each tag are patterns, sentences of questions often asked by users, and responses, which are answers to information providers have provided. Therefore, Data.json makes it easy for the system to recognize patterns and determine responses to frequently asked questions.

The following is the syntax of data.json

```
{
"intents": [
  {
   "tag": "How_to_apply (Cara_daftar)",
```

Journal of Applied Data Sciences
Vol. 4, No. 4, December 2023, pp. 414-430

"patterns": [

"how to apply for training? (*bagaimana cara mendaftar pelatihan?*)",

"how do I join the training? (*bagaimana cara ikut pelatihan?*)",

"how to register? (*cara daftarnya gimana?*)",

"where do you register? *(Daftarnya lwat mana kak?*)",

"how to register for BLK? (*Cara daftar mengikuti BLK gmn ya?*)",

"I register BLK (*Kak daftar BLK*)",

"how to register? (*Cara daftar gmn ya?*)",

"how to register? (*cara daftar?*)",

"How to register in training program BLK? (*Bagaimana cara mendaftar program pelatihan di BLK Surabaya?*)"

],

"responses": [

"You can register online or offline (*Untuk pendaftarannya bisa secara online atau offline kak*)",

" You can choose to register online or offline (*Kamu dapat memilih untuk mendaftar secara online ataupun offline.*)",

"There are two registration options available, namely via the internet or conventionally. (*Ada dua opsi pendaftaran yang tersedia, yaitu melalui internet atau secara konvensional.*)",

"Registration can be done through online or manually. (*Pendaftaran dapat dilaksanakan melalui sistem online maupun dengan cara manual.*)"

]

},

]

}

3.3. Text Pre-processing

Preprocessing is the initial stage in data processing, where raw or unstructured data is tidied, formatted, and prepared to be processed further. Preprocessing is to prepare the raw data to be processed more efficiently and accurately and remove unnecessary noise and interference. The preprocessing stage includes several steps: duplicate data removal, data normalization, missing data filling, data format conversion, and data cleaning from unnecessary characters or symbols. Once preprocessing is complete, the data is ready for further processing using data analysis or machine learning techniques. There are four stages of preprocessing performed, namely:

3.3.1. Change the Abbreviation to Standard Form

When recognizing a sentence, the system automatically converts all found abbreviations into the standard or standard form. Therefore, a slang dictionary is provided, which contains a collection of non-standard words, including abbreviations. The slang dictionary structure consists of 2 columns, namely non-standard words and standard words, shown in Table 2. For standard word dictionaries, refer to research [37], [38].

Table 2. Slang Dictionary Lists

Slang Word	Formal Word
woww	wow
Where (*dimn*)	Where (*dimana*)

Slang Word	Formal Word
Happy (*met*)	Happy (*selamat*)
Hatch (*netas*)	Hatch (*menetas*)
Which Number (*kebrp*)	Which Number (*keberapa*)
Very (*bgt*)	Very (*banget*)
Sby	Surabaya
Cuteness (*gemess*)	Cuteness (*gemes*)
I (*aku*)	I (*saya*)

3.3.2. Punctuation

Punctuation is one of the stages in natural language processing that involves removing or replacing punctuation marks in text. Punctuation includes punctuation marks such as periods, commas, question marks, exclamation marks, and quotation marks. In natural language processing, punctuation is often considered irrelevant to the meaning of the text and can affect the model's performance. Therefore, removing or replacing punctuation marks can improve data quality and model performance. Punctuation processing can be done using libraries or algorithms available in natural language processing. Figure 5 displays the process of punctuation removal.

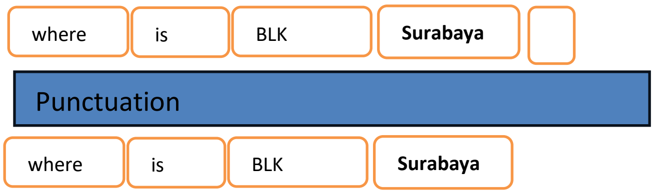


Figure 5. Punctuation Removal

3.3.3. Lemmatization

Lemmatization is a stage in text processing that aims to convert words to their base form. This is done by removing affixes and returning the word to its base form. Lemmatizing is usually done after the tokenizing and stop word removal stages. The lemmatizing process can use linguistic rules or a dictionary of base words for a particular language, then results in text data that has been processed and simplified, making it easier to process and analyze.

3.3.4. Grouping By Class

Grouping by class is a technique used in natural language processing to obtain sentences processed and adjusted according to the desired criteria. This process involves several stages, such as tokenization, removal of unwanted words, and conversion of words into their base form. The goal is to simplify text processing and make it more understandable to the machines or programs. This method can be used for applications like chatbots, sentiment analysis, or text classification. In natural language processing, preprocessing sentence base classes is essential to ensure the data obtained can be appropriately processed and provide accurate results. Figure 6 displays the list of classes from the data.

78 Classes

[‘Address_BLKSurabaya (*Alamat_BLKSurabaya*)’, ‘Dormitory (*Asrama*)’, ‘How_to_register (*Cara_daftar*)’, ‘Register_offline (*Daftar_offline*)’, ‘Register_online (*Daftar_online*)’, ‘Training_information (*Informasi_pelatihan*)’, ‘Business_and_management_training (*Pelatihan_bisman*)’, ‘Electronic_training (*Pelatihan_elektronika*)’, ‘Muslim_student_association_training (*Pelatihan_HMI*)’,’Electrical_training (*Pelatihan_listrik*)’, ‘Manufacture_training (*Pelatihan_manufaktur*)’, ‘Automotive_training (*Pelatihan_otomotif*)’, ‘Building_training (*Pelatihan_bangunan*)’, ‘No_answer (*tidakada_jawab*)’, ‘Pocket_money (*uang_saku*)’, ‘Competence_test (*uji_kompetensi*)’,’minimum_age (*umur_min*)’, ‘instructor_interview (*wawancara_instruktur*)’, ‘who_made_it_official’ (*yang_meresmikan*), ….]

Figure 6. List of Class/Category Groups

3.4. Padding Stage

After the data has been pre-processed, it enters the padding stage, a step for text data processing that aims to adjust the length of each sentence in the dataset. This is necessary to ensure that each input has a uniform size and meets the needs of the deep learning model used. This step can be done by adding unique tokens such as <PAD> at the end of sentences shorter than the specified maximum length. Thus, every sentence in the dataset will be the same size and ready to be used in model training. Padding preprocessing can also make it easier to process large amounts of text data efficiently, as the size of each input has been adjusted before being fed into the model. Figure 7 shows the structure of padding stage.

```
[154, 13, 4, 5, 68],
[30, 4, 5, 136, 309],
[310, 13, 4, 5, 68],
[91, 154, 13, 4, 5, 68],
[30, 4, 5, 68, 7, 311,
312],
[76, 6, 200, 4, 5],
[4, 5, 68, 73, 76],
```

(a)

```
[[  0   0   0 ...   0   0 149]
 [  0   0   0 ...   0   0 179]
 [  0   0   0 ...   0   0 149]
 ...
 [  0   0   0 ...   0   0   0]
 [  0   0   0 ...   0   0 509]
 [  0   0   0 ...   0   0 510]]
```

(b)

Figure 7. Padding Stage: (a) Tokenization Results and (b) Padding Results

3.5. Label Encoder

Currently, the datasets have the same data length; the next step is to convert categorical data into numeric data. In this step, the method used is fit_transform() from LabelEncoder to perform label encoding on the 'tags' data in the DataFrame data. This method will process the category data from the 'tags' column and replace it with the appropriate numeric value. The encoding results are shown in Figure 8.

```
[36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 35 35 35 35 35 35 35
 35 35 68 68 68 68 68 68 21 21 21 21 42 42 57 57 57 57 57 57 57 58 58 58
 58 58 23 23 23 23 23 67 67 67 67 67 67 67 77 77 77 46 46 46 22 22 22 22
 22 22 22 22 22 22 22 22 22 29 29 29 29 29 29 20 20 20 20 20 20 20 20 20
 20 20 20  2  2  2  2  2  2  2  2 66 66 66 66 66 66 66 66 66 66 66 66
 40 40 40 40 40 40 40 40 40  3  3  3  3  3  3  3  3  3  3  3  4  4  4  4
  4  4  4  0  0  0  0  0  0  0  0  0  0  0  0  0  5  5  5  5  5  5  5  5
  5  5  5  5  5  5 41 41 41 41 41 41 41 41 41 41 41 41 41 41 32 32 32 59
 59 59 59 59 59 59 14 14 14 14 14 14 14 14 14 10 10 10 10 10 10 10 10 10
 10 11 11 11 11 11 11 11 11 11 11 11  9  9  9  9  9  9  8  8  8  8  8  7
  7  7  7  7  7 18 18 18 18 18  6  6  6  6  6  6  6  6 16 16 16 16 16 16
 16 16 15 15 15 15 15 13 13 13 13 13 13 12 12 12 12 12 12 28 28 28 28 28
 28 28 28 28 28 49 49 49 49 49 49 49 43 43 43 43 43 27 27 53 53 75 75 54
 54 54 54 54 54 61 61 61 26 26 26 26 26 25 25 25 25 25 39 39 45 45 45 45
 45 63 63 63 63 63 63 62 62 62 62 62 62 19 19 19 19 19 19 19 17 17 17 17
 17 17 17 17 74 74 31 31 31 31 50 50 50 50 50 50  1  1  1  1  1  1  1  1
  1  1 64 64 64 64 64 51 51 51 60 60 60 60 60 60 60 60 60 69 69 69 69 69
 69 70 70 70 70 70 70 70 65 65 65 65 65 65 65 65 65 65 65 76 76 38 38 30]
```

Figure 8. Label Encoding Results

3.6. Cross Validation

To test the performance of LSTM modeling on the dataset, an evaluation of the modeling used is carried out, one of which uses cross-validation. Cross-validation is a model evaluation technique that measures model performance by dividing data into several parts or partitions. At each iteration, one data partition will be used as a test or validation set, while the other will be used as training data. The model will be trained on the training data and tested on the test data, then the performance of the model will be calculated and recorded. This process is repeated k times, with each partition acting as test data, so all data is used as test and training data. The model performance at each iteration is calculated and averaged for a more general and reliable model performance. This method allows maximum use of data without compromising the quality of model performance.

The cross-validation stage in the chatbot model development process using LSTM. First, the dataset will be divided into several subsets or folds. Next, train the model on each fold and test its performance on other subsets not used for training. After all the folds have been trained and tested, we can calculate the average performance of all the folds as an estimate of the model performance. By using cross-validation, we can obtain a more stable and reliable estimation of model performance and avoid the problem of overfitting the training data. In addition, with cross-validation, we can also select the optimal hyperparameters for the LSTM model in the chatbot.

3.7. LSTM Modeling

To get a model that meets expectations, repeated experiments are carried out to obtain parameters based on training and evaluation needs, along with the parameters and model summary used in the LSTM model.

Table 3. LSTM Parameter

Parameter	Value
Embedding	19
LSTM	64

ISSN 2723-6471

Parameter	**Value**
Regulizer L1 & L2	1 e-3
Dropout	0.5
Batch size	8
Optimizer	Adam
Learning Rate	1 e-3
Epoch	200

Table 4. Model Summary

Layer (type)	**Output Shape**	**Param #**
Input_1 (inputLayer)	[(None, 23)]	0
Embedding (Embedding)	(None, 23, 10)	5460
Lstm (LSTM)	(None, 23, 10)	840
Dropout (Dropout)	(None, 23, 10)	0
Flatten (Flatten)	(None, 230)	0
Dense (Dense)	(None, 78)	18018
Total params : 24318 (94.99 KB) Trainable params : 24318 (94.99 KB) Non-trainable params : 0 (0.00 Byte)		

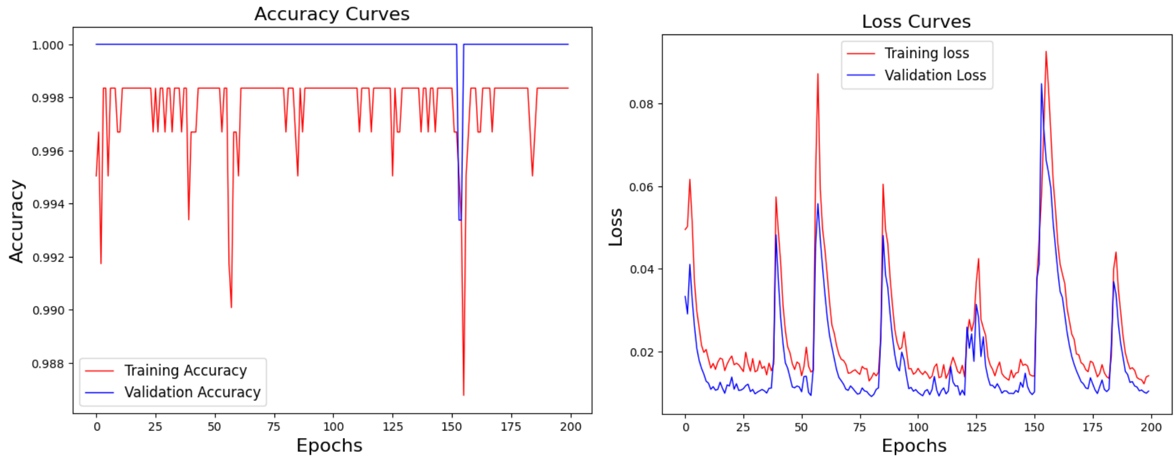
3.8. Model Testing

After conducting a series of test datasets with the LSTM model and validation using K-Fold with total K=5 folds, and each fold goes through 200 epochs, is obtained 80% dataset training results in the 5th fold with Accuracy 1.000 and Loss 0.006, and at 20% testing data, with mark Accuracy: 0.863, Precision: 0.885, Recall: 0.863 and F1-Score: 0.852 (weighted average), testing on Fold 1 has low accuracy 0.727, and loss 1.6410, for results testing presented in table 5.

Table 5. Testing Result

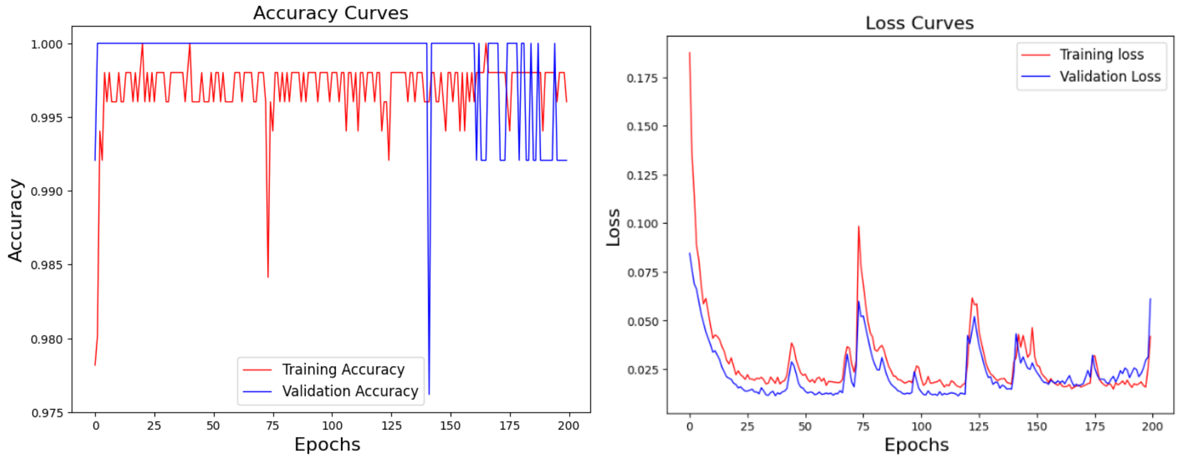
Dataset	**Accuracy**	**Precision**	**Recall**	**F1-Score**
90%	0.869	0.850	0.860	0.845
80%	0.876	0.884	0.873	0.863
70%	0.873	0.881	0.869	0.856

90% Graph Dataset



80% Graph Dataset

ISSN 2723-6471



70% Graph Dataset

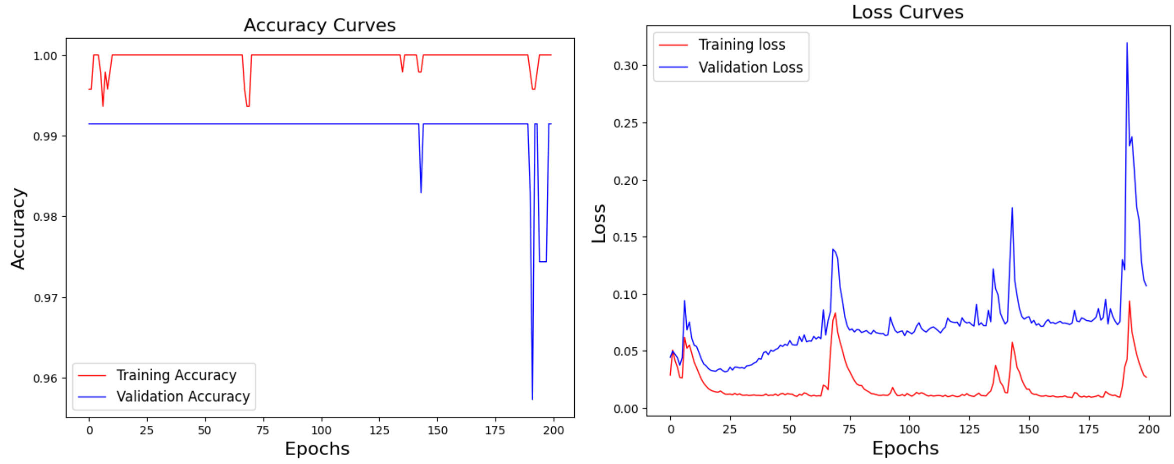


Figure 9. Label Encoding Results

In the Figure 9 comparison dividing the dataset, the dataset shows a 90% accuracy high followed by high loss. While the dataset is 70% accuracy small, the following loss is also tiny.

In this study, the authors also conducted trials using other methods, including GRU, CNN, and LSTM-CNN. The dataset used is 80% training data and 20% testing data. The test results can be seen in Table 6.

Table 6. Testing Result

Method	Accuracy	Precision	Recall	F1-Score
GRU	0.845	0.846	0.845	0.858
LSTM	0.876	0.884	0.873	0.863
CNN	0.922	0.908	0.922	0.904
LSTM-CNN	0.904	0.880	0.905	0.885

Table 7. Testing Accuracy Answer (20% Testing Data)

Question	Answer	Accuracy
Assalamualaikum	Walaikum salam wr wb, Introduce I am ICA, Anyone can I help?	99.72%
Where is BLK Surabaya? (*Dimn alamat blk Surabaya ?*)	BLK Surabaya is located on Jalan Dukuh Menanggal III number 29, Surabaya, in the Gayungan area	99.98%
How to register? (*Cara daftarnya gmna ?*)	Registration Requirements	99.99%

Question	Answer	Accuracy
What are the training registration requirements? (Apa persyaratan pendaftaran pelatihan?)	Register requirements : Copy of Identity Card / Family Card (1 sheet), Copy of Vocational / Senior High School Certificate (1 sheet), 3x4 Photo (2 Sheets) (Persyaratan pendaftaran : Fotocopy KTP/KK (1 lembar), Fotocopy Ijasah SMK/SMA (1 lembar), Pas Foto 3x4 (2 lembar))	74.69%
@^%#^&%$	Sorry, Ica doesn't understand, can you repeat it again? (*Maaf kak, Ica ngga ngerti bisa diulang pertanyaannya?*)	79.82%
Thank you (*terima kasih*)	Your welcome (*sama-sama kak*)	99.99%

In Table 7. a direct trial was conducted by giving six random questions. Of the several answers given, 4 with 99% accuracy, and 2 with 76% accuracy, chatbot agents can recognize different question models, some questions are abbreviated, there is a symbol, and answers are given with an accuracy of 90% and above.

3.9. Interface Design

For chatbots to be used by a wide range of people and to support training registration services at BLK Surabaya, it is necessary to design an easy and inexpensive interface. This study will use telegram messenger as the interface. After LSTM modeling, it takes four files which will be forming files on telegram, namely:

1) The LSTM learning model is saved with file name model.h5.
2) A collection of classes is stored with the file name classes.pkl.
3) Tokenization results with file name tokenizer.pkl.
4) The result of converting label encoder mapping source sequence with the variable length to vector with file name le.pkl.
5) A unique collection of words from lemmatized results with the name word.pkl.

Besides telegrams being able to run via handheld devices, they can also be run in web/desktop mode. So that chatbots can communicate via telegram. Several steps will be prepared:

1) Create a bot agent via Botfather.
2) After the agent is created, then submit a token that will be used to communicate with the chatbot.
3) From the token input into the chatbot application that has been prepared.
4) On the chatbot application side, python-telegram-bot must first be installed to connect.
5) After the chatbot is connected to the Telegram, it will try to give questions and answers and then record the resulting accuracy.

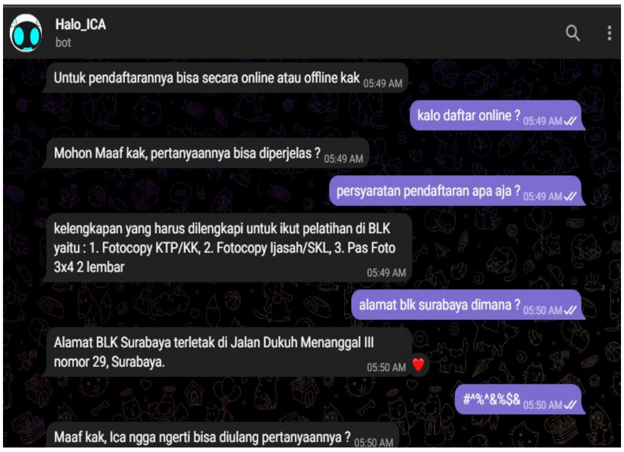


Figure 10. Label Encoding Results

After the program is run and the chatbot and telegram are connected, the operator tries to provide questions and answers given by the chatbot that have 90% accuracy, as shown in Figure 10.

4. Conclusions

After testing and evaluating, chatbot research using LSTM has obtained quite good results. The 70% dataset has the highest loss, while the 90% loss dataset is unstable. The number of datasets is still 840 paired questions, so it affects the accuracy of the modeling. For future research, the dataset will be added to 1000 questions to improve the accuracy obtained. In addition, it will compare with models other than LSTM so that a more varied Accuracy value will be received.

Besides that, comparison trials have been carried out using other methods, including GRU, LSTM, CNN and LSTM-CNN. It was found that the CNN method had the best results, namely an accuracy of 92.2%. It is acknowledged that the CNN method has advantages in extracting features from text, such as word patterns or short phrases, and can identify specific features in groups of words, while LSTM only focuses on understanding long sequence relationships which may be less accurate for capturing features between texts. From the research trials, it can be concluded that chatbot research using LSTM obtained quite good results, using a dataset of 80%:20%, namely Accuracy of 0.876, Precision of 0.884 and Recall of 0.873.

In this LSTM-based chatbot research, there are still many limitations experienced, one of which is related to the dataset which is only 840 paired questions, so it affects the accuracy and loss results. To get high accuracy results requires a large and varied dataset. Apart from that, the chat context is still text-based and the topic only serves training registration information. Currently, text-based chat can still handle information needs. Hopefully the results of this research in the future can be useful and perfected, especially in developing Chatbots using Artificial Intelligence.

To improve the accuracy, the dataset should be expanded to 1000 paired questions for future research. The modeling will be compared with the latest modeling such as Bidirectional LSTM, Bert Transformers so that more varied accuracy values will be obtained. The chatbot discussion topic will be expanded not only around training registration information but can also serve the training implementation process in each profession. Moreover, the text-based chat interface will be changed to voice-based, making conversation easier.

5. Declarations

5.1. Author Contributions

Conceptualization, E.I.S., S.I., J.S.; methodology, E.I.S., Y.S.H.L. and J.S.; software, Y.S.H.L.; validation, J.S.; formal analysis, E.I.S.; writing—original draft preparation, Y.S.H.L., E.I.S. and J.S.; writing—review and editing, E.I.S.; visualization, Y.S.H.L. and S.I; supervision, E.I.S.; project administration, E.I.S; funding acquisition, E.I.S. and J.S.

5.2. Data Availability Statement

All dataset for machine learning model are available from the website database (https://www.estherirawati.web.id/datasets/).

5.3. Funding

This project was supported by a research grant from the DIPA Academic Directorate of Vocational Higher Education Directorate General of Vocational Education, Ministry of Education, Culture, Research, and Technology with contract number SP DIPA-023. I 8.1.690524/2023. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.4. Institutional Review Board Statement

This research project has been reviewed and approved by the Research, Innovation, Publication Quality Assurance, and Curriculum Development Center of Institut Sains dan Teknologi Terpadu Surabaya (ISTTS). The IRB has determined that the study meets the ethical standards and guidelines for the protection of human subjects in research.

5.5. Informed Consent Statement

The research project, entitled "Long Short-Term Memory-Based Chatbot for Vocational Registration Information Services," is being conducted by Yudo Sembodo, Esther Irawati Setiawan, Syaiful Imron, and Joan Santoso from Institut Sains dan Teknologi Terpadu Surabaya (ISTTS). The primary objective of this research is to develop and evaluate a chatbot for vocational registration information services using Long Short-Term Memory technology.

5.6. Declaration of Competing Interest

The authors declare no competing of interest.

References

[1] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, “A new chatbot for customer service on social media,” in *Proceedings of the 2017 CHI conference on human factors in computing systems*, vol. 1, no.1, pp. 3506–3510, 2017.

[2] E. A. J. Croes and M. L. Antheunis, “Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot,” *J Soc Pers Relat*, vol. 38, no. 1, pp. 279–300, 2021.

[3] E. Adamopoulou and L. Moussiades, “An overview of chatbot technology,” in *IFIP international conference on artificial intelligence applications and innovations*, vol. 1, no.1, pp. 373–383, 2020.

[4] V. Vijayaraghavan, J. B. Cooper, and others, “Algorithm inspection for chatbot performance evaluation,” *Procedia Comput Sci*, vol. 171, no. 1, pp. 2267–2274, 2020.

[5] C. Nalini, R. S. Kumari, G. K. Parteban, T. N. Priyaa, and A. S. Sanchay, “AI based chatbot in food industry,” in *AIP Conference Proceedings*, vol. 1, no. 1, pp. 1-8, 2021.

[6] J.-S. Chen, T.-T.-Y. Le, and D. Florence, “Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing,” *International Journal of Retail & Distribution Management*, vol. 49, no. 11, pp. 1512–1531, 2021.

[7] B. A. Shawar and E. Atwell, “Chatbots: are they really useful?,” *Journal for Language Technology and Computational Linguistics*, vol. 22, no. 1, pp. 29–49, 2007.

[8] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, and H.-M. Wang, “A chatbot using LSTM-based multi-layer embedding for elderly care,” in *2017 International Conference on Orange Technologies (ICOT)*, v0ol. 1, no.1, pp. 70–74, 2017.

[9] J. Zhang, Y. J. Oh, P. Lange, Z. Yu, and Y. Fukuoka, “Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet,” *J Med Internet Res*, vol. 22, no. 9, pp. e22845-e22852, 2020.

[10] S. Abuse and H. I. V Prevention, “Department of Health and Human Services Substance Abuse and Mental Health Services Administration”, 2018.

[11] P. Muangkammuen, N. Intiruk, and K. R. Saikaew, “Automated thai-faq chatbot using rnn-lstm,” in *2018 22nd International Computer Science and Engineering Conference (ICSEC)*, vol. 1, no. 1, pp. 1–4, 2018.

[12] C. V. Misischia, F. Poecze, and C. Strauss, “Chatbots in customer service: Their relevance and impact on service quality,” *Procedia Comput Sci*, vol. 201, no. 1, pp. 421–428, 2022.

[13] E. W. T. Ngai, M. C. M. Lee, M. Luo, P. S. L. Chan, and T. Liang, “An intelligent knowledge-based chatbot for customer service,” *Electron Commer Res Appl*, vol. 50, no. 1, pp. 101098-101123, 2021.

[14] I. Hasan, S. Rizvi, S. Jain, and S. Huria, “The AI enabled Chatbot Framework for Intelligent Citizen-Government Interaction for Delivery of Services,” in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, vol. 1, no. 1, pp. 601–606, 2021.

[15] I. M. De Andrade and C. Tumelero, “Increasing customer service efficiency through artificial intelligence chatbot,” *Revista de Gestão*, vol. 29, no. 3, pp. 238–251, 2022.

[16] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, “My chatbot companion-a study of human-chatbot relationships,” *Int J Hum Comput Stud*, vol. 149, no. 1, pp. 102601-102608, 2021.

[17] Y. D. Prabowo, H. L. H. S. Warnars, W. Budiharto, A. I. Kistijantoro, Y. Heryadi, and others, "Lstm and simple rnn comparison in the problem of sequence to sequence on conversation data using bahasa indonesia," in *2018 Indonesian association for pattern recognition international conference (INAPR)*, vol. 1, no. 1, pp. 51–56, 2018.

[18] I. Tsidylo, S. Samborskiy, S.-I. Mazur, and M. Zamoroz, "Designing a Chat Bot for Learning a Subject in a Telegram Messenger.," in *ICTERI Workshops*, vol. 1, no. 1, pp. 1329–1340, 2020.

[19] B. Horn, *Robot vision*. MIT press, vol. 1, no. 1, pp. 1-7, 1986.

[20] G. Veruggio, "The euron roboethics roadmap," in *2006 6th IEEE-RAS international conference on humanoid robots*, vol 1, no. 1, pp. 612–617, 2006.

[21] N. Darapaneni *et al.*, "Customer support Chatbot for electronic components," in *2022 Interdisciplinary Research in Technology and Management (IRTM)*, vol. 1, no. 1, pp. 1–7, 2022.

[22] S. Naveen, "AUTOMATED CHATBOT IMPLEMENTED USING NATURAL LANGUAGE PROCESSING", *International Research Journal of Modernization in Engineering Technology and Science,* vol. 2, no. 8, pp. 619-624, 2020.

[23] K. Sarakhman, R. Kempnyk, and V. Chyhura, "ChatBot using NLP," in *Computational linguistics and intelligent systems: proceedings of the 4nd International conference (2), 2020*, vol. 1, no. 1, pp. 429–432, 2020.

[24] D. Sai Pranav, M. Mutreja, D. Punj, and P. Chawla, "Natural Language Processing in Chatbots," in *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 1*, Springer, vol 1, no. 1, pp. 87–98, 2022.

[25] E. Kasthuri and S. Balaji, "Natural language processing and deep learning chatbot using long short term memory algorithm," *Mater Today Proc*, vol. 1, no. 1, pp. 1-12, 2021.

[26] M. Aleedy, H. Shaiba, and M. Bezbradica, "Generating and analyzing chatbot responses using natural language processing," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, pp. 1-12, 2019.

[27] V. Tiwari, L. K. Verma, P. Sharma, R. Jain, and P. Nagrath, "Neural network and NLP based chatbot for answering COVID-19 queries," *International Journal of Intelligent Engineering Informatics*, vol. 9, no. 2, pp. 161–175, 2021.

[28] S. Ayanouz, B. A. Abdelhakim, and M. Benhmed, "A smart chatbot architecture based NLP and machine learning for health care assistance," in *Proceedings of the 3rd international conference on networking, information systems & security*, vol. 1, no. 1, pp. 1–6, 2020.

[29] R. Regin, S. S. Rajest, T. Shynu, and others, "An automated conversation system using natural language processing (nlp) chatbot in python," *Central Asian Journal of Medical and Natural Science*, vol. 3, no. 4, pp. 314–336, 2022.

[30] M. Mnasri, "Recent advances in conversational NLP: Towards the standardization of Chatbot building," *arXiv preprint arXiv:1903.09025*, vol. 1, no. 1, pp. 1-13, 2019.

[31] T. Lalwani, S. Bhalotia, A. Pal, V. Rathod, and S. Bisen, "Implementation of a Chatbot System using AI and NLP," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, vol.6, no-3, pp. 1-13, 2018.

[32] P. Anki, A. Bustamam, H. S. Al-Ash, and D. Sarwinda, "Intelligent chatbot adapted from question and answer system using RNN-LSTM model," in *Journal of Physics: Conference Series*, vol. 1, no. 1, pp. 12001-12007, 2021.

[33] K. Smagulova and A. P. James, "A survey on LSTM memristive neural network architectures and applications," *Eur Phys J Spec Top*, vol. 228, no. 10, pp. 2313–2324, 2019.

[34] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Comput*, vol. 31, no. 7, pp. 1235–1270, 2019.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter," *Multimed Tools Appl*, vol. 80, no. 1, pp. 35239–35266, 2021.

[37] S. Imron, E. I. Setiawan, and J. Santoso, "Deteksi Aspek Review E-Commerce Menggunakan IndoBERT Embedding dan CNN," *INSYST: Journal of Intelligent System and Computation*, vol. 5, no. 1, pp. 10–16, 2023.

[38] S. Imron, E. I. Setiawan, J. Santoso, M. H. Purnomo, and others, "Aspect Based Sentiment Analysis Marketplace Product Reviews Using BERT, LSTM, and CNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 3, pp. 586–591, 2023.