JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 8 Number 2 | December 2023: 213-221
DOI: 10.15575/join.v8i2.1138

Implementation of Recurrent Neural Network (RNN) for Question Similarity Identification in Indonesian Language

Muhammad Iqbal[1*], Hasmawati[2], Ade Romadhony[3]
[1,2,3]School of Computing, Informatics, Telkom University, Bandung, Indonesia

Article Info

Article history:

Received July 17, 2023
Revised August 18, 2023
Accepted September 23, 2023
Published December 28, 2023

Keywords:

Manhattan distance
Question in Indonesian
Similarity question
RNN

ABSTRACT

In a question-and-answer forum, the identification of question similarity is used to determine how similar two questions are. This procedure makes sure that user-submitted questions are compared to the questions in a database for matches to improve system performance on the online Q&A platform. Currently, question similarity is mostly done in foreign languages. The purpose of this research is to identify question similarities and evaluate the effectiveness of the methods used in Indonesian language questions. The data used is a public dataset with labeled pairs of questions as 0 and 1 where label 0 for different pairs of questions and label 1 for the same pairs of questions. The method used is a Recurrent Neural Network (RNN) with the Manhattan Distance approach to calculate the similarity distance between two questions. The question pairs are taken as two inputs with a reference label to identify the similarity distance between the two question inputs. We evaluated the model using three different optimizers namely RMSprop, Adam, and Adagrad. The best results were obtained using the Adam optimizer with 80:20 ratio split-data and overall accuracy is 76%, precision is 74%, recall is 98.8%, and F1-score is 85.1%.

Corresponding Author:

Muhammad Iqbal
School of Computing, Informatics, Telkom University, Bandung
Jl. Telekomunikasi Dayeuhkolot Bandung 40257
Email: muhammadiqbaaall@student.telkomuniversity.ac.id

1. INTRODUCTION

A sentence or text in a question can have the same meaning but with different wording [1]. In this case, a system requires semantic similarity measurement in text. Semantic measurement is a common method used in Natural Language Processing (NLP) [2], [3]. NLP is a research field that explores how computers can be used to understand and manipulate text and natural language processing [4]. Semantic similarity measurement of question similarity is often used in identifying question similarities. Identification of question similarity is crucial in information retrieval, online question-answering systems, machine translation, dialogue systems with Artificial Intelligence (AI), and document matching [5]. With semantic text measurement, a system can easily determine the answer to a query. Moreover, it can make a system work efficiently. For example, in an online question-answering system, the system stores each new question from the user and uses it as a reference when there is a new question with the same meaning but different wording [6]–[8]. Consequently, the system can quickly provide an answer to the question.

Several previous studies have been conducted on text similarity. One of them is the research conducted by Borui Ye and colleagues in 2017, using Chinese language Community Question Answering (cQA) data. In their study, they designed an Encoder-Decoder pre-trained framework RNN. The model was developed using a two-step implementation scheme [9]. In this research, two input questions were provided to be compared with a reference label. The data was automatically labeled into three categories: questions with the same, relevant, and different labels. In the final step, they used manually

labeled data with a smaller scale and achieved an accuracy of 83.1% [9]. Another study was conducted by Jiapeng Wang and Yihong Dong in 2020. In their research, Wang and colleagues measured the semantic similarity distance between two texts. The distance measurement was performed using string-based and corpus-based text representations. After going through various steps, they calculated the semantic similarity distance and obtained good performance, although it required high computation and resources [5].

In 2019, Muntaha Al-Asa'd analyzed question similarity on an online forum using Arabic language data. The proposed approach to detecting question similarity involved examining the syntactic, morphological, semantic, and lexical features of question pairs. The steps applied in the detection process included Arabic text processing, feature extraction, and text classification. Multiple algorithms such as Support Vector Machine (SVM) and Decision Trees were used in the text classification stage [10]. The method employed for the detection process was Random Forest with XGBoost feature extraction, resulting in a final accuracy of 78.2% in detecting question similarity in the Arabic language [10]. In another study in 2018, Zongkui Zhu identified duplicate questions with a computational model of semantic equations based on the Siamese Network and combined them with the BI LSTM approach with the result of an accuracy of 84.5%[8].

In 2022, Wiwin Surwanigsih and their research team carried out a study to assess the effectiveness of the A Lite BERT (ALBERT) model, Efficient Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), and the Robust Optimized BERT Pretraining Approach (RoBERTa) in the context of enhancing the creation of an Indonesian-language question and answer system. The languages spoken are Indonesian, Malay, and Esperanto. Wiwin and her team employed information from Indonesian Wikipedia and the Open Super-large Crawled ALMAnaCH coRpus (OSCAR) for Esperanto. They utilized the SentencePiece method with a dictionary size of approximately 3000 subtokens and implemented byte-level byte pair encoding (ByteLevelBPE). In this study, the paper on Bidirectional Encoder Representations from Transformers (BERT) tested learning rates of 1E-5 and 5e-5 for both languages, as per the reference. The results obtained are an accuracy of 91.7% and an F1-score of 86.2% with the conclusion RoBERTa model was better to implement the Indonesian question-and-answer system[11].

The research in this paper aims to identify similarities between two Indonesian language questions. The data used consists of Indonesian language questions from elementary school, junior high school, and senior high school levels. The identification process is conducted using an RNN model with the Manhattan Distance approach, where the RNN layer serves as the input layer for the two questions, and the distance is calculated using the Manhattan distance approach. Before processing the question pairs to determine their similarity scores, the question pairs are first labeled as "same" and "different". The next step involves taking the two questions as input and comparing whether the questions are the same or different based on the given labels. The purpose of this study is to see the performance of the RNN method to identify similarities of questions in the Indonesian language. This research is expected to serve as a reference for future related studies.

2. METHOD

The method used in this study is RNN with the Manhattan Distance approach. The pair of questions will be used as two inputs that will be processed by each layer in the model built. the pair of questions are processed by the RNN layer and then the similarity distance of the two questions is calculated using the Manhattan Distance equation. The flow of question similarity identification process in the Indonesian language using the RNN method is divided into five stages, as shown in Figure 1. The process begins with preparing the dataset to be processed into the preprocessing stage. Then, the dataset is transformed into vector form using feature extraction and split for further processing in the model training and model testing stages.

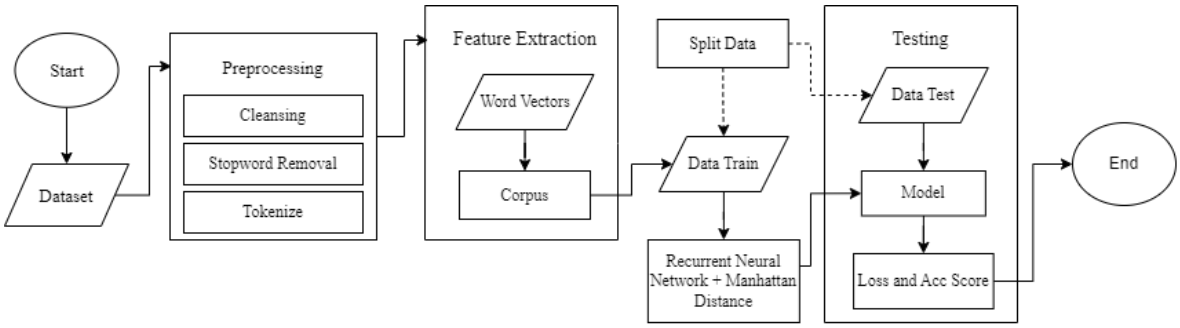

Figure 1. System Overview

2.1. Data Preparation

The data used in this study consists of question pairs sourced from Indonesian language questions at the elementary school, junior high school, and senior high school levels. In this study, we use little data because of limited data questions on the subject matter of Indonesian. The total number of collected question pairs is approximately 621 pairs. Each question pair is labeled with the value "same" for identical question pairs and "different" for question pairs that are different. The distinguishing factor between the same question pairs is the use of different sentence structures and contextual information. An example of the question pair data can be seen in the following Table 1.

Table 1. Example of Labeled Question Pairs

No	Question 1	Question 2	Label
1	Sikap Rio sebaiknya? Rio's attitude should?	Rambu lalu lintas memiliki arti? Do traffic signs have meaning?	Beda Different
2	Dari bacaan diatas tuliskan 3 kegunaan denah! From the reading above, write 3 uses of a map!	Yang merupakan kalimat utama dari paragraf tersebut adalah? What is the main sentence of the paragraph?	Beda Different
3	Bahan alam yang dapat menghasilkan benang diantaranya? Natural materials that can produce yarn are?	Berikut manakah tahapan pembuatan pakaian yang benar? What are the correct steps in making clothes?	Beda Different
4	Bendera ini dikenal dengan nama bendera? This flag is known by the name of the flag?	Bendera ini disebut dengan bendera? Is this called a flag?	Sama Same
5	jika Zio hendak pergi kepasar minggu zio harus? what if Zio wants to go to the market Sunday zio should?	Zio perlu .... apabila zio ingin pergi ke pasar minggu Zio needs .... if he wants to go to the Sunday market.	Sama Same
6	Saat ingin pergi ke Ciwidey, Fahri banyak menemukan rambu lalu lintas. Ini fahri harus berhati hati karena? When he wanted to go to Ciwidey, Fahri found many traffic signs. Do you have to be careful because?	Fahri menemukan banyak rambu lalu lintas di jalan, Fahri harus berhati hari karena? Fahri found a lot of traffic signs on the road, Fahri must be careful because?	Sama Same

From the total number of collected question pairs, the labels are changed to 0 and 1, where 0 represents different question pairs and 1 represents the same question pairs. The comparison between the same and different question pairs can be seen in Figure 2.

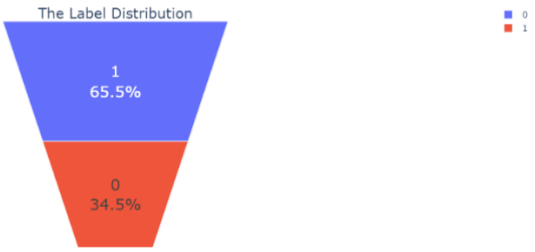

Figure 2. Visualization of Data Distribution

In Figure 2 it can be seen that the pair of questions with label 1 is 65.5% and the pair of questions with label 0 is 34.5%. We have not used the resampling data function because the subject data used in questions about the Indonesian language levels of elementary, junior, and senior high school students consists of recurring question types, such as those involving the use of the words "how" and "what". Consequently, a significant amount of data would be duplicated if data resampling were to be conducted.

2.2. Preprocessing

This process is performed to improve the sentence structure of the data, thereby reducing any issues that may arise during the analysis. Additionally, preprocessing can also improve the accuracy of data analysis results [12]. In general, an overview of the data preprocessing can be observed in Figure 3.

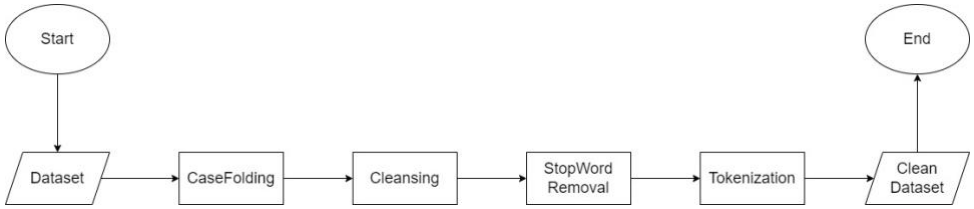

Figure 3. Preprocessing Overview

In Figure 3 the explanation of the preprocessing stages can be explained as follows:

1. Case Folding which is the process of standardizing uppercase or lowercase letters into all lowercase letters [13].
2. Cleansing is the process of cleaning data from characters other than alphanumeric characters such as punctuation marks and other characters.
3. Stopword Removal is a process of removing words or general terms that are not useful and have no meaning in the hope of research[14]
4. Tokenization is the process of breaking down sentences into words. After passing the previous stage of preprocessing, the sentences in the dataset are broken down and separated by a space.

Examples of the results of each of these stages are presented in Table 2.

Table 2. Preprocessing result

Process Name	Input	Output
Case Folding	Dari bacaan diatas tuliskan tiga kegunaan denah! From the reading above, write tiga uses of a map!	dari bacaan diatas tuliskan three kegunaan denah! from the reading above, write three uses of a map!
Cleansing	Saat ingin pergi ke Ciwidey, Fahri banyak menemukan rambu lalu lintas. Ini fahri harus berhati hati karena? When he wanted to go to Ciwidey, Fahri found many traffic signs. Do you have to be careful because?	Saat ingin pergi ke Ciwidey Fahri banyak menemukan rambu lalu lintas ini fahri harus berhati hati karena when he wanted to go to Ciwidey, Fahri found many traffic signs Do you have to be careful because
Stopword-Removal	Sikap Rio sebaiknya ? Rio's attitude should.	Sikap Rio. Rio's attitude.
Tokenize	Sikap Rio sebaiknya ? "Rio's", "attitude", "should"	"Sikap", "Rio" , "sebaiknya" "Rio's", "attitude", "should"

2.3. Feature Extraction

The feature extraction works by mapping each word into a vector [15], [16]. Each word assigned with a vector value represents the word's projection in a vector space [17]–[19]. Feature extraction is widely used in the field of NLP because it can capture the semantic similarities between words. In this analysis process, the feature extraction used is Word2Vec. Word2Vec feature works by assigning weights in the form of vectors to each word that can carry the semantic meaning of the word [20]. In this research, we imported the Word2Vec feature from the gensim library. The output produced is the representation of words in vector form. The results of the feature extraction process can be observed in Figure 4.

```
Hasil Vektor pertanyaan:

[[-0.00474142  0.00949607  0.00836911 ... -0.00040556 -0.00047918
  -0.00077861]
 [-0.00875505  0.02066677  0.0041516  ... -0.0133285   0.00532449
  -0.00855689]
 [-0.00363592  0.01052917  0.00385878 ... -0.00647272  0.00692675
  -0.00533451]
 ...
 [-0.00374862  0.00945855  0.00131877 ... -0.01117217  0.00488459
  -0.00720125]
 [-0.0048888   0.0134483   0.0055307  ... -0.0082364   0.00687727
  -0.00753349]
 [-0.00298248  0.00664337  0.00990108 ... -0.00406571  0.00128709
  -0.00227961]]
```

Figure 4. Extraction Feature Result

2.4. Split Data

Data splitting involves the separation of data into two distinct sets: one for training the model and the other for evaluating its performance. This process is conducted in two stages, with the data initially divided into a 70:30 ratio, allocating 70% for training and the remaining 30% for testing. Then, from the 70% training data, it is further divided with a ratio of 70:30 to separate a portion of the data as validation data. This process aims to assess the accuracy_validation during the model training process.

2.5. Recurrent Neural Network

Recurrent Neural Network or RNN is an effective machine learning technique designed for handling sequential data [15]. Generally, RNN can be described as a repetitive relationship [21]. An RNN leverages feedback generated by prior nodes as its input and constructs a recursive framework for mapping input and parameters to compute output and assess loss. It engages in iterative computations to establish a flexible network structure. The structure of an RNN can be likened to a sequence comprising an input layer, a hidden layer, and an output layer, functioning like a chain reaction [15]. Overall, the calculations of the architecture can be seen in equation (1).

$$h^t = func(h^{t-1}, x^t;\ \theta) \tag{1}$$

h^t represents the hidden layer of the RNN, and the calculation results are determined based on the previous neural node's h, h^{t-1}.

The model to be built in this study is a simple RNN model with the Manhattan Distance approach. The question pairs are applied to the model with an initial layer that receives two input questions, namely question 1 and question 2. After the pair of questions is processed on the RNN layer, the similarity distance is calculated using the Manhattan Distance equation so that the resulting output is a similarity score between two question objects. The architecture of the model to be built can be observed in Figure 5.

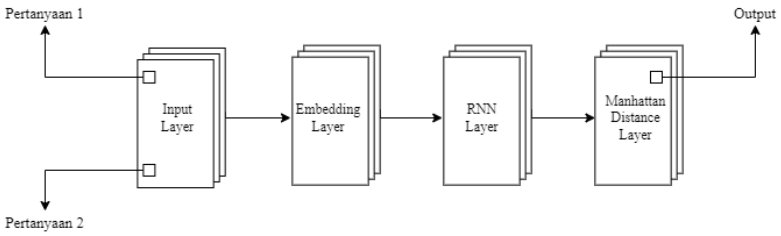

Figure 5. RNN Architecture

Figure 5 shows the RNN model with Manhattan Distance consisting of 4 layers. The first layer is the input layer which receives input in the form of two questions. In this layer, the questions are passed to the embedding layer. In the embedding layer, the input question pairs are transformed into continuous vector representations in a numerical space, enabling the calculation of similarity between the question pairs. The output from the embedding layer is then passed to the RNN layer, where the question pairs are processed. After going through the RNN layer, the similarity distance between the question pairs is calculated using the Manhattan Distance approach, resulting in an output that

represents the accuracy score of the question pairs' similarity. The calculation performed in the Manhattan Distance layer can be seen in equation (2).

$$Sim(x, y) = |x1 - x2| + |y1 - y2| \quad (2)$$

In equation (2), x represents the question and y represents the label reference in the processed data [5]. The distance between the two questions is calculated after going through the preceding layers.

2.6. Performance of System

In this research, we evaluate the model performance using the following metrics: accuracy, precision, recall, and F1-Score. The result can be easily seen in the confusion matrix. Confusion Matrix is an assessment technique that helps determine how effectively a system or model can accurately identify data. True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are the four primary parts of the Confusion Matrix. Several evaluation metrics, such as accuracy, precision, recall, and F1-score, can be computed from these constituents. These metrics help to provide a detailed understanding of the model's performance in classification tasks. The results of each component can be obtained using the calculations in equations 3, 4, 5, and 6.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (3)$$

$$Precision = \frac{TP}{(TN+FP)} \quad (4)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (5)$$

$$F1 - score = \frac{2TP}{2(TP+FP+FN)} \quad (6)$$

3. RESULT AND DISCUSSION

3.1. Result

The testing in this study was conducted to evaluate the performance of the RNN model in identifying question similarity in Indonesian language questions. at the initial stage, we tried to train the model using three different optimizers: Adam, Adagrad, and RMSprop. Then, the models trained with these optimizers were tested using various testing schemes, including different ratios of data splitting and the usage of stopword removal on the data. In this case, the tested data splitting ratios were 60:40, 70:30, and 80:20. We used data_train to train the model and data_test to validate the result. The accuracy matrix is used to evaluate the performance of the RNN model in the model training process. The results of the training on question data without stopword removal are presented in Table 3.

Table 3. Test Result without Stopword Removal Data

Optimizer	Ratio					
	60:40		70:30		80:20	
	Training	Validation	Training	Validation	Training	Validation
Adam	**0,6768**	**0,6533**	0,6542	0,7011	0,6616	0,6400
Adagrad	0,6591	0,6500	**0,6542**	**0,7162**	**0,6801**	**0,6533**
RMSProp	0,6734	0,6533	0,6599	0,6869	0,6616	0,6400

Table 3 shows the results of testing question pairs using the RNN model with different optimizers and data-splitting ratios. The optimizer Adagrad achieved the highest training accuracy score of 0.6801 with a data splitting ratio of 80:20. On the other hand, the optimizer Adagrad achieved the highest validation accuracy score of 0.7162 with a data splitting ratio of 70:30. The testing was then continued by incorporating Stopword removal on the data. The results of testing question data with Stopword removal can be observed in Table 4.

Table 4. Test Result with Stopword Removal Data

Optimizer	Ratio					
	60:40		70:30		80:20	
	Training	Validation	Training	Validation	Training	Validation
Adam	**0,8956**	**0,7867**	0,9261	0,7011	0,9337	0,7400
Adagrad	0,5923	0,5625	0,623	0,5714	**0,9671**	**0,7500**
RMSProp	0.8956	0,7333	**0,9406**	**0,7176**	0,9444	0,7500

Table 4 shows the results of testing question pairs with the addition of the stopwords removal function increase. In this testing, the highest training score was achieved using the Adagrad optimizer with a data splitting ratio of 80:20, resulting in an accuracy of 0.9671. On the other hand, the highest validation score was obtained using the Adam optimizer with a data splitting ratio of 60:40, with an accuracy of 0.7867. in NLP research, the preparation of data largely determines the results of classification by the model built including the use of stopword removal, and this study conducted a test scheme using and not using stopword removal on the data used. the result is that data using additional stopword removal is processed more optimally by the built model. This is because the use of stopword removal in the data reduces the distribution of words that do not have a contribution and meaning relevant to the question, thus making the model more focused on question keywords and efficient in the identification process.

Then, the final results of the accuracy for the identification of the similarity of questions in the Indonesian language using the RNN model are calculated with the confusion matrix function. The data selected for model evaluation was the data with stopword removal, as it demonstrated improved training accuracy and validation accuracy during the model training process.

3.2. Result Testing Analysis

In the analysis phase of the test results, the test evaluation of the performance of the model using data test with stopwords removal which had previously been shared in the split data process. The Data test used is new data and not yet recognized by the model. The total data used for testing is 125 pairs of questions. The results of the evaluation of the model are calculated by using the confusion matrix to obtain accuracy, recall, precision, and F1-score.

Table 5. Testing Result

Optimizer	Data Split Ratio			
	60:40			
	Accuracy	Precision	Recall	F1-Score
Adam	**0.6746**	**0.7277**	**0.8036**	**0.7638**
Adagrad	0.5702	0.6555	0.7239	0.6880
RMSprop	0.6144	0.6488	0.8957	0.7525
	70:30			
	Accuracy	Precision	Recall	F1-Score
Adam	0.7219	0.7358	**0.9212**	0.8181
Adagrad	0.5582	**0.8441**	0.3987	0.5416
RMSprop	**0.7433**	0.7548	0.9212	**0.8297**
	80:20			
	Accuracy	Precision	Recall	F1-Score
Adam	**0.7600**	**0.7478**	**0.9885**	**0.8514**
Adagrad	0.7440	0.7722	0.8965	0.8297
RMSprop	0.7040	0.7314	0.9080	0.8102

Table 5 shows the results when using the Adam optimizer on data that includes additional stopwords removal. The results are as follows: 76% accuracy, 74.7% precision, 98.8% recall, and 85.1% F1 score. The results were obtained using a data separation ratio of 80: 20. The results were achieved with a ratio of 80: 20 where 80% of the data was used to train the built model. it can be concluded that the RNN model with the Manhattan distance approach to identify the similarity of questions in Indonesian requires a lot of data to train the model so that the model can produce maximum performance. The selection of an appropriate optimizer and the careful preprocessing of data play crucial roles in determining the ultimate performance of a text classification model. These factors significantly impact the model's ability to achieve accurate and reliable results.

3.3 DISCUSSION

The results of the identification of the question similarity on the Indonesian subject matter that has been done were obtained after conducting a series of studies. The research includes data

preparation, model development, model training, and model evaluation. The test that has been carried out includes using different optimizers and different data split ratios (60:40, 70:30, 80:20). In addition, the data preparation process also determines the results obtained by the model. The results in Table 3 show the difference in scores obtained when the model was trained using optimizers and different data split ratios. The selection of different optimizers in the classification of texts can affect the results obtained. The process carried out in the data preparation section also determines the final results obtained by the model. In Table 4, the test is done by adding a stopword removal function to the data. Adding a stopword removal function to the data increases the effectiveness of the question similarity identification process by the model. This is because the use of stopword removal eliminates irrelevant words from the questions and enables the model to concentrate on the keywords within each question sentence. As a result, the model executes the identification process more efficiently. The results obtained are a training accuracy of 96% and a testing accuracy of 78%. Then to validate the accuracy of testing used separate data that has not been recognized by the model. The process is done by choosing an Adam optimizer with a split-data ratio of 80: 20 and using data that has been added to the stopword removal function in the preprocessing section. The result was calculated using the confusion matrix equation and the overall results are accuracy of 76%, precision of 74.7%, recall of 98.8%, and F1-Score of 85.1%.

4. CONCLUSION

Based on the testing, the RNN method with the Manhattan Distance approach can identify similarities between the two questions. After several test schemes against the model used, it can be concluded that the preprocessing process can improve the accuracy of the model, especially for limited data. The accuracy results are also determined by the optimizer, the size of the dataset used, and the data splitting ratio for model training. In this study, the dataset consisted of approximately 621 question pairs in the Indonesian language. Due to limited data on questions about the Indonesian language, the dataset size can be considered relatively small, as it typically requires a substantial amount of data to effectively train the constructed model. The highest accuracy was achieved using the Adam optimizer with the addition of stopword removal in the data preprocessing. The final evaluation of the model using the confusion matrix achieved the following results: an accuracy of 76%, precision of 74.7%, recall of 98.8%, and F1-score of 85.1%.

For future research, we must do further research using data on a larger scale because the application of deep learning for the classification process requires large data to be optimal in classifying. So that the model can produce better accuracy than previous research.

REFERENCES

[1] I. M. S. Putra, Putu Jhonarendra, and Ni Kadek Dwi Rusjayanthi, "Deteksi Kesamaan Teks Jawaban pada Sistem Test Essay Online dengan Pendekatan Neural Network," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1070–1082, Dec. 2021, doi: 10.29207/resti.v5i6.3544.

[2] R. E. Setiawan, T. Fabrianti Kusumasari, and M. A. Hasibuan, "Penerapan Deep Learning, NLP(Natural Language Processing) dan Data Visualization untuk Customer Research Digital Marketing Instagram," Bandung, 2019.

[3] E. D. Liddy, "Natural Language Processing Natural Language Processing Natural Language Processing 1," 2001. [Online]. Available: https://surface.syr.edu/istpub

[4] G. G. Chowdhury, "Natural Language Processing Dept. of Computer and Information Sciences University of Strathclyde," *Glasgow G1 1XH, UK*, 2003.

[5] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9. MDPI AG, pp. 1–17, Sep. 01, 2020. doi: 10.3390/info11090421.

[6] F. Kunneman, T. C. Ferreira, E. Krahmer, and A. Van Den Bosch, "Question similarity in community question answering: A systematic exploration of preprocessing methods and models," in *International Conference Recent Advances in Natural Language Processing, RANLP*, Incoma Ltd, 2019, pp. 593–601. doi: 10.26615/978-954-452-056-4_070.

[7] D. Bogdanova, C. dos Santos, L. Barbosa, and B. Zadrozny, "Detecting semantically equivalent questions in online user forums," in *CoNLL 2015 - 19th Conference on Computational Natural Language Learning, Proceedings*, Association for Computational Linguistics (ACL), 2015, pp. 123–131. doi: 10.18653/v1/k15-1013.

[8] Z. Zhu, Z. He, Z. Tang, B. Wang, and W. Chen, "A Semantic Similarity Computing Model based on Siamese Network for Duplicate Questions Identification," School of Computer Science and Technology, Soochow University, 2018.

[9] B. Ye, G. Feng, A. Cui, and M. Li, "Learning Question Similarity with Recurrent Neural Networks," in *Proceedings - 2017 IEEE International Conference on Big Knowledge, ICBK 2017*, Institute of Electrical and Electronics Engineers Inc., Aug. 2017, pp. 111–118. doi: 10.1109/ICBK.2017.46.

[10] N. A. M. B. Y. E. K. M. H. M. A.-S. Muntaha Al-asa'd, *Question to Question Similarity Analysis Using Morphological, Syntactic, Semantic, and Lexical Features*. 2019.

[11] W. Suwarningsih, R. A. Pratama, F. Y. Rahadika, and M. H. A. Purnomo, "RoBERTa: language modeling in building Indonesian question-answering systems," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 20, no. 6, pp. 1248–1255, Dec. 2022, doi: 10.12928/TELKOMNIKA.v20i6.24248.

[12] N. Adani Setyadi, M. Nasrun, and C. Setianingsih, *Text Analysis For Hate Speech Detection Using Backpropagation Neural Network*. 2018.
[13] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Compl aint Document Classification Using Sastrawi," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2020. doi: 10.1088/1757-899X/874/1/012017.
[14] M. Anandarajan, C. Hill, and T. Nolan, "Text Preprocessing," 2019, pp. 45–59. doi: 10.1007/978-3-319-95663-3_4.
[15] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very Deep Convolutional Networks for Text Classification," Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.01781
[16] M. Mansoor, Z. Ur Rehman, M. Shaheen, M. A. Khan, and M. Habib, "Deep learning based semantic similarity detection using text data," *Information Technology and Control*, vol. 49, no. 4, pp. 495–510, 2020, doi: 10.5755/j01.itc.49.4.27118.
[17] L. Efrizoni, S. Defit, M. Tajuddin, and A. Anggrawan, "Komparasi Ekstraksi Fitur dalam Klasifikasi Teks Multilabel Menggunakan Algoritma Machine Learning," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 653–666, Jul. 2022, doi: 10.30812/matrik.v21i3.1851.
[18] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information (Switzerland)*, vol. 10, no. 4. MDPI AG, 2019. doi: 10.3390/info10040150.
[19] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, "Comparative study of word embedding methods in topic segmentation," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 340–349. doi: 10.1016/j.procs.2017.08.009.
[20] E. M. Dharma, F. Lumban Gaol, H. Leslie, H. S. Warnars, and B. Soewito, "The Accuracy Comparison Among Word2Vec, Glove, and FastTexr Towards Convolution Neural Network (CNN) Text Classification," *J Theor Appl Inf Technol*, vol. 31, no. 2, 2022, [Online]. Available: www.jatit.org
[21] X. Zhang, M. H. Chen, and Y. Qin, "NLP-QA Framework Based on LSTM-RNN," in *Proceedings - 2nd International Conference on Data Science and Business Analytics, ICDSBA 2018*, Institute of Electrical and Electronics Engineers Inc., Dec. 2018, pp. 307–311. doi: 10.1109/ICDSBA.2018.00065.