Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

People Entity Recognition in Indonesian Alquran Translation using Roberta

Aufa Mutia*, Moch Arif Bijaksana

School of Computing, Bachelor of Informatics, Telkom University, Bandung
Jl. Telekomunikasi. 1, Terusan Buahbatu - Bojongsoang, Telkom University, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat, Indonesia
Email: [1,*]aufamutia@student.telkomuniversity.ac.id, [2]arifbijaksana@telkomuniversity.ac.id
Correspondence Author Email: aufamutia@student.telkomuniversity.ac.id
Submitted: **16/01/2024;** Accepted: **26/01/2024;** Published: **27/01/2024**

Abstract−The Quran was revealed in Arabic, which has a complex linguistic structure, a unique writing system, and intricate grammar, making it challenging to understand. Therefore, understanding and interpreting the Quran is a primary goal for Muslims. To comprehend the teachings contained in the Quran, Muslims need an understanding of the human entities mentioned in it. However, manually labeling human entities in the Quran can be a complex task prone to errors. The aim of this research is to facilitate the process of labeling human entities in Quranic texts by building a model with good performance. RoBERTa is a Named Entity Recognition (NER) model that is an extension of BERT, trained with enhanced training methodologies. This study focuses on the use of the RoBERTa model to identify human entities in the translated text of the Quran in Bahasa Indonesia. The input to this system consists of translated Quranic sentences, which are then processed by the model to generate output in the form of predicted labels for those sentence entities. The model is constructed by utilizing a dataset from the Tanzil Quran corpus, covering chapters 1 to 6. Data preprocessing involves punctuation removal, tokenization, and case folding. The dataset is divided into training data (80%) and testing data (20%). The RoBERTa model is trained with hyperparameters such as epochs, learning rate, and batch size. Evaluation is performed using metrics such as Precision, Recall, and F-Score on the testing data. The evaluation results of the constructed RoBERTa model show an F-Score value of 52%. This score is not better compared to the BERT model, indicating that the RoBERTa model tends to have inferior performance in identifying human entities in the translated text of the Quran.

Keywords: People Entity; Quran; Indonesian; NER; RoBERTa

1. INTRODUCTION

As the holy book of Islam, the Quran was revealed in the Arabic language, which has a complex linguistic structure, including different writing systems, intricate grammar, and a phonetic system distinct from commonly studied languages[1]. The Quranic text consists of more than 6200 verses and 114 chapters[2]. Additionally, its rich linguistic style and profound meanings[3] make it challenging to comprehend. Understanding and interpreting the verses of the Quran involve exegesis through translations[4], which is the primary goal for Muslims. Muslims seek to understand the human entities within the Quran to grasp the teachings it contains. It is crucial to elucidate the messages and values embedded in each verse using human entities. However, manually labeling human entities in the Quran can be a complex task prone to errors. An effective approach is needed to identify and label objects due to the large size and linguistic complexity of the text.

Natural Language Processing (NLP) is a branch of computer science and linguistics that studies the interaction between computers and human natural language[5]. The main goal of NLP is to enable computers to work with text and human speech in the same way as humans. NLP enables computers to read text, listen to speech, interpret it, measure sentiment, and determine which parts are significant[6]. Some complex tasks that NLP can perform include language understanding, automatic translation, entity recognition, information extraction, sentiment analysis, and natural language generation. One way to understand and identify human entities is through Information Extraction (IE). Information Extraction is a document processing type that captures and produces factual information available in the document[5]. Unstructured information is a common issue in IE. Named Entity Recognition (NER) is an approach used to address this problem. Named Entity Recognition (NER) is an area of Natural Language Processing (NLP) that deals with the identification and classification of entities in written text[8]. Additionally, NER is the process of extracting named entities considered significant in a text, categorizing them into predefined categories[9]. In Figure 1, there is an example input for Named Entity Recognition (NER) in the form of text, and the output shows the identification of entities by the system. The input for Named Entity Recognition (NER) consists of sentences or text, with the output being words or tokens that are identified and categorized into several entity types. The NER task is to tag noun phrases with the following four classes, person (PER), organization (ORG), location (LOC), and miscellaneous (MISC[10] . In the Quran, NER can be employed to identify and label human entities such as character or group names.

In recent years, deep learning-based NER models have become dominant and achieved sophisticated results[11]. Deep learning involves multiple layers of processing to learn data representations with varying levels of abstraction. The main advantage of deep learning lies in its ability to learn semantic representations and compositions supported by vector representations and neural processing[11]. This allows machines to automatically discover the processing and latent representations needed for classification or detection.

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

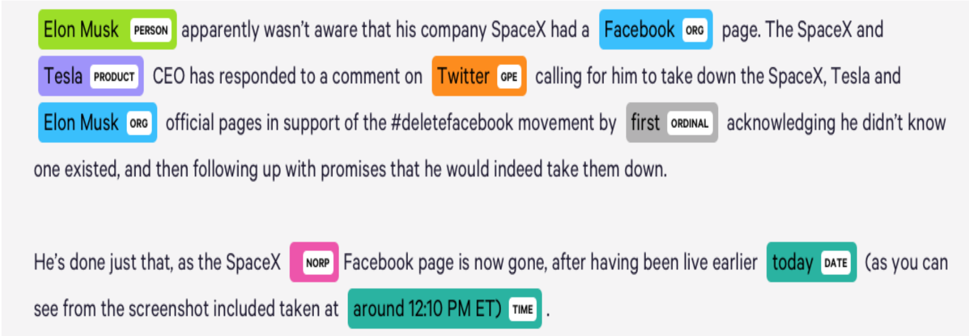

Figure 1. Example text along with detected entities

BERT (Bidirectional Encoder Representation from Transformers) is a novel language representation model designed to train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers[12]. In a previous study by Retno Diah et al., human entity extraction from the English-translated Quran was performed using BERT. The model was built using datasets from Quran Surah Al-Fatihah, Al-Baqarah, and Ali Imran, achieving an F1-Score of 53%. The study revealed challenges in classifying hierarchical phrases due to insufficient training data, resulting in low extraction performance[12]. Meanwhile, the use of RoBERTa for extracting human entities in the translation of the Indonesian-language Quran has not been done before.

RoBERTa (A Robustly Optimized BERT Pretraining Approach) is a BERT development released in 2019 by researchers from the University of Washington and Facebook AI[13]. RoBERTa is BERT retrained with an enhanced training methodology. Research results indicate that, when pretrained with the same dataset as BERT, RoBERTa is 1000% more accurate on SQUAD, MNLI-m, and SST-2 datasets and requires ten times fewer steps than BERT. An overview of the RoBERTa model for Named Entity Recognition (NER) can be seen in Figure 2. In enhancing training performance, RoBERTa eliminates the Next Sentence Prediction (NSP) training method from BERT's training method and introduces dynamic masking[14]. RoBERTa is improved by optimizing training hyperparameters. With this optimization, RoBERTa outperforms BERT in almost all tasks designed to be solved by BERT. Meanwhile, the use of RoBERTa for extracting human entities in the translation of the Indonesian language Quran has not been conducted.

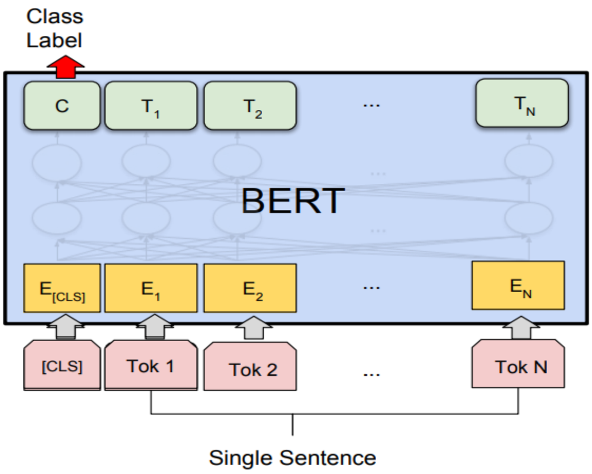

Figure 2. RoBERTa Model for Named Entity Recognition (NER)

This research aims to facilitate the labeling of human entities in the Quranic text using the RoBERTa model. RoBERTa exhibits superior language processing capabilities due to training on a larger and more diverse dataset compared to previous models such as BERT and XLNet[13]. RoBERTa also outperforms other models, including BERT and XLNet, in NER tasks[13]. Therefore, RoBERTa can expedite the process with more accurate results in labeling human entities in the Indonesian-translated Quran.

In creating the RoBERTa model, the dataset used is from the Tanzil Quran corpus, consisting of the Indonesian-translated Quran text. This research labels human entities using the BIO format, where Begin (B) indicates the beginning of a label phrase, Inside (I) indicates a part of a label phrase, and Outside (O) indicates non-label phrases[15]. Table 1 shows the dataset structure used in this study to build the RoBERTa model. The input to the constructed human entity extraction system is sentences or text of the Indonesian language translation of Quranic verses, and the system's output is the human entities extracted by the system. Figure 3 illustrates an example of input taken from a passage of Surah Al-Baqarah verse 275 of the Quran. The system takes input in the

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

form of sentences or texts of the Indonesian-language Quran translation and produces output consisting of detected human entities within those sentences.

Surah Al-Baqarah : 275

Input :

Orang-orang yang makan mengambil riba tidak dapat berdiri melainkan seperti berdirinya orang yang kemasukan syaitan lantaran tekanan penyakit gila. Keadaan mereka yang demikian itu adalah disebabkan mereka berkata berpendapat sesungguhnya jual beli itu sama dengan riba padahal Allah telah menghalalkan jual beli dan mengharamkan riba. Orang-orang yang telah sampai kepadanya larangan dari Tuhannya lalu terus berhenti dari mengambil riba maka baginya apa yang telah diambilnya dahulu sebelum datang larangan dan urusannya terserah kepada Allah. Orang yang kembali mengambil riba maka orang itu adalah penghuni-penghuni neraka mereka kekal di dalamnya.

Output :

1. Orang-orang yang makan
2. orang yang kemasukan syaitan
3. Orang-orang yang telah sampai kepadanya larangan dari Tuhannya
4. Orang yang kembali mengambil riba

Figure 3. Example input and output of the human entity extraction system.

This research is focused on labeling human entities in the translated text of the Indonesian-language Quran, with the aim of improving understanding of human entities in the Quran. In its application, the labeling process often involves complex challenges. Identifying and categorizing human entities such as character names or groups in the text requires precision to understand the context of Quranic verses. This study restricts the dataset used to train the model by only using labeled translations of Quranic verses from Juz 1 to 6. Additionally, labeling for the entities used consists of only one level, as seen in Table 1. The entities labeled in this research are specifically related to human entities.

The research is conducted by building a RoBERTa model to identify human entities in the translated text of the Indonesian-language Quran. RoBERTa is expected to construct a model with better performance in extracting human entities. From the constructed model, performance evaluation is carried out using the F-Score evaluation metric calculated based on Precision and Recall.

2. RESEARCH METHODOLOGY

2.1 System Flow

In this research, the RoBERTa model is employed to extract human entities in the translated text of the Indonesian-language Quran. The model is trained using the Tanzil Quran corpus dataset. Before training this dataset, data preprocessing is carried out in the pre-processing stage. Subsequently, the dataset is divided into two parts, namely training data and testing data. Once trained, the model is tested to predict labels on testing data different from the training data. The results of the testing are analyzed to determine the model's performance in identifying human entities. Figure 4 illustrates the flow of the research process constructed for this study.

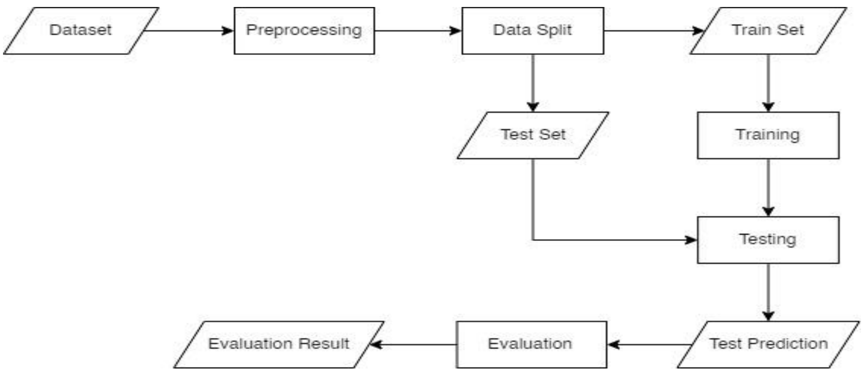

Figure 4. System flow

2.2 Dataset

The dataset used in this research is extracted from the Tanzil Quran corpus which can be accessed at the link https://tanzil.net. The Quranic verses utilized are in the form of translated text in Bahasa Indonesia, covering Juz 1 to Juz 6, with a total of 36,752 tokens and 3,139 entities. This study employs the entity tags PER (Person) to represent human entities and O (Other) for other entities. The entity labeling format used in this research is BIO, aligning with the dataset format of the NER CoNLL-2003. The dataset comprises entities (tokens) and BIO Format (Begin, Inside, and Outside) taken from excerpts of verses in the translated text of the Indonesian-language Quran. An example dataset structure used in this research is shown in Table 1.

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Table 1. The dataset structure table

Juz	Ayat	Token	Label
1	7	yaitu	O
1	7	jalan	O
1	7	orang-orang	B-PER
1	7	yang	I-PER
1	7	telah	I-PER
1	7	engkau	I-PER
1	7	beri	I-PER
1	7	nikmat	I-PER
1	7	kepada	O
1	7	mereka	O
1	7	bukan	O
1	7	jalan	O
1	7	mereka	B-PER
1	7	yang	I-PER
1	7	dimurkai	I-PER
1	7	dan	O
1	7	bukan	O
1	7	pula	O
1	7	jalan	O
1	7	mereka	B-PER
1	7	yang	I-PER
1	7	sesat	I-PER

2.3 Data Preprocessing

The data used is in CSV (Comma Separated Values) format, where the text in the dataset is separated by commas. This data can be read using the pandas library. Before testing and training on the data, a preprocessing stage is conducted. The preprocessing steps are as follows:

1. Punctuation Removal
   Punctuation removal involves the action of eliminating or not adding punctuation marks to the text. Characters such as periods (.), commas (,), semicolons (;), colons (:), exclamation marks (!), question marks (?), and others are used in punctuation to convey meaning and clarify sentence structure. The process at this stage is illustrated in Table 2, where punctuation marks still exist in the text and are then removed.

Table 2. Example Punctuation Removal

Input	Output
Rasul telah beriman kepada Al-Quran yang diturunkan kepadanya dari Tuhannya, demikian pula orang-orang yang beriman. Semuanya beriman kepada Allah, malaikat-malaikat-Nya, kitab-kitab-Nya, dan rasul-rasul-Nya. Mereka mengatakan “Kami tidak membeda-bedakan antara seseorangpun dengan yang lain dari rasul-rasul-Nya.” Dan mereka mengatakan, “Kami dengar dan kami taat.” Mereka berdoa “Ampunilah kami, ya Tuhan kami, dan kepada Engkaulah tempat kembali.”	Rasul telah beriman kepada Al-Quran yang diturunkan kepadanya dari Tuhannya demikian pula orang-orang yang beriman Semuanya beriman kepada Allah malaikat-malaikat-Nya kitab-kitab-Nya dan rasul-rasul-Nya Mereka mengatakan Kami tidak membeda-bedakan antara seseorangpun dengan yang lain dari rasul-rasul-Nya Dan mereka mengatakan Kami dengar dan kami taat Mereka berdoa Ampunilah kami ya Tuhan kami dan kepada Engkaulah tempat kembali

2. Tokenization
   Tokenization is the process of converting text into a collection of smaller units called "tokens." Tokens can be words, phrases, or characters. The process at this stage is shown in Table 3, where in this process, each word in the text is decoded into words marked with a comma (,) followed by a space.

Table 3. Example Tokenization

Input	Output
Rasul telah beriman kepada Al-Quran yang diturunkan kepadanya dari Tuhannya, demikian pula orang-orang yang beriman. Semuanya beriman kepada Allah, malaikat-malaikat-Nya, kitab-kitab-Nya, dan rasul-rasul-Nya. Mereka mengatakan “Kami tidak membeda-bedakan antara seseorangpun	‘Rasul’, ‘telah’, ‘beriman’, ‘kepada, ‘Al-Quran’, ‘yang’, ‘diturunkan’, ‘kepadanya’, ‘dari’, ‘Tuhannya’, ‘demikian’, ‘pula’, ‘orang-orang’, ‘yang’, ‘beriman’, ‘Semuanya’, ‘beriman’, ‘kepada’, ‘Allah’, ‘malaikat-malaikat-Nya’, ‘kitab-kitab-Nya’, ‘dan’, ‘rasul-rasul-Nya’, ‘Mereka’, ‘mengatakan’, ‘Kami’, ‘tidak’,

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Input	Output
dengan yang lain dari rasul-rasul-Nya.” Dan mereka mengatakan, “Kami dengar dan kami taat.” Mereka berdoa “Ampunilah kami, ya Tuhan kami, dan kepada Engkaulah tempat kembali.”	‘membeda-bedakan’, ‘antara’, ‘seseorangpun’, ‘dengan’, ‘yang’, ‘lain’, ‘dari’, ‘rasul-rasul-Nya’, ‘Dan’, ‘mereka’, ‘mengatakan’, ‘Kami’, ‘dengar’, ‘dan’, ‘kami’, ‘taat’, ‘Mereka’, ‘berdoa’, ‘Ampunilah’, ‘kami’, ‘ya’, ‘Tuhan’, ‘kami’, ‘dan’, ‘kepada’, ‘Engkaulah’, ‘tempat’, ‘kembali’

3. Case Folding
   Case Folding involves converting letters in the label column to uppercase. The process at this stage is shown in the table, where letters in words that were previously lowercase or a mixture of uppercase and lowercase become all uppercase in the output. This process is important so that labels on tokens can be classified correctly.

Table 4. Example Case Folding

Input	Output
B-Per, I-Per, o	B-PER, I-PER, O

2.4 Data Splitting

Two separate datasets are required for training and testing the model. The existing dataset is divided into two parts: training data (80% of the total dataset) and testing data (20% of the total dataset).

2.5 RoBERTa Model Development

In this stage, the model is trained using the training data from the previous process. Model development is carried out with hyperparameter settings. Hyperparameters are parameters used to adjust the behavior of a machine learning model or algorithm during the training process. The hyperparameter optimization technique employs Bayesian methods to select optimal values for parameters such as learning rate, batch size, and epoch[16]. Learning rate determines how quickly weight updates are made during the training process[17]. Batch size refers to the number of data points in one batch[18]. Meanwhile, an epoch is one cycle of processing data in deep learning, particularly in image processing[19].

2.6 Evaluation System

The trained model is tested using testing data different from the training data and unseen by the model. The model then predicts label tags for each testing data (B-PER, I-PER, O). NER is typically evaluated by comparing its output to human annotations[13]. The comparison can be measured using Exact-Match evaluation. With Exact-Match evaluation, named entities are considered correctly identified only if their boundaries and types match the actual values[13]. Precision, Recall, and F-Score can be calculated based on the counts of True Positive (TP), False Positive (FP), and False Negative (FN)[13]. True Positive represents entities recognized by NER that match the actual values. False Positive indicates entities recognized by NER that do not match the actual values. False Negative refers to entities described in the actual values but not recognized by NER. The F Score is used as a performance metric for the model[20]. The following is the method to calculate the F1 Score:

$$\text{F Score} = \frac{2 \times \text{Recall} \times \text{Precission}}{\text{Recall} + \text{Precission}} \quad (1)$$

Precision is the percentage of correctly identified named entities in the system[20]. The following is the method to calculate Precision:

$$\text{Precission} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (2)$$

Meanwhile, Recall is the percentage of named entities successfully identified by the system[20]. The following is the method to calculate Recall:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (3)$$

3. RESULT AND DISCUSSION

3.1 Results

The testing phase began with the Tanzil Quran corpus dataset, which underwent preprocessing to generate a ready-to-use dataset for model construction. This dataset was initially divided into training and testing sets. The model was built using the training dataset, employing the RoBERTa model. The model was utilized to detect training data classified into human entities. Evaluation of the predictions was then conducted using metric calculations. The results of this evaluation served as a reference for analysis, ultimately producing the evaluation outcomes.

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Testing was conducted using the translated text of the Indonesian-language Quran from Juz 1 to Juz 6, extracted from the Tanzil Quran corpus, which was then preprocessed to generate a dataset ready for model construction. The dataset can be accessed at the following link: https://github.com/aufamutia/RoBERTa. Figure 5 displays the statistical distribution of token counts in the dataset based on labels. The dataset is then divided into training and testing datasets. For the training dataset, a RoBERTa model for Named Entity Recognition (NER) was built using the SimpleTransformers library with the 'roberta' model type. Hyperparameters such as learning rate, epochs, and batch size were adjusted. Meanwhile, the testing dataset was used to evaluate the performance of the constructed model.

The implementation utilized T4 GPU hardware with 166 GB of memory designed for high-performance deep learning. PyTorch, a deep learning framework, was used for software implementation. The implementation also involved hyperparameters, including learning rate, epochs, and batch size. These three hyperparameters serve as the basis for the comparisons conducted in this study. Model performance evaluation was carried out using three evaluation metrics: Precision, Recall, and F-Score. F-Score served as the main metric to measure the model's ability to identify entities.

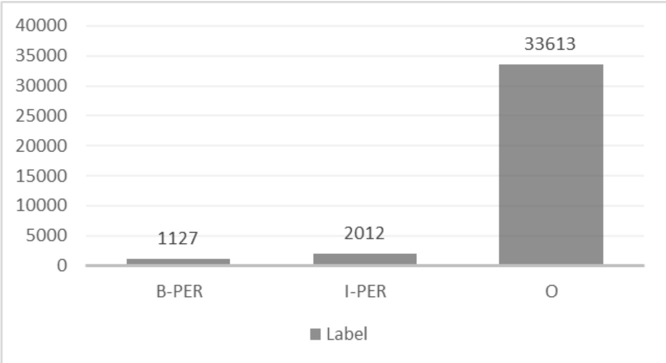

Figure 5. Label distribution in the dataset

A comparison was made between the RoBERTa and BERT models in identifying entities, as shown in Table 5. The comparison was done with the same hyperparameter settings: learning rate 0.0001, epochs 85, and batch size 16. Both models used the same dataset. The results indicated that the best Precision value was achieved by the RoBERTa model at 0.598. However, the best Recall and F-Score values were achieved by the BERT model, with Recall at 0.483 and F-Score at 0.527. These results suggest that the RoBERTa model did not outperform the BERT model when considering F-Score as the primary metric for performance measurement.

Table 5. Comparison of performance results between RoBERTa and BERT models

Model	Hyperparameter Value			Evaluation Metric		
	Learning Rate	Epoch	Batch Size	Precision	Recall	F-Score
RoBERTa	0,0001	85	16	**0,598**	0,463	0,522
BERT	0,0001	85	16	0,58	**0,483**	**0,527**

Table 6 presents a comparison of experiments conducted on the RoBERTa model with different batch sizes. The results show that the model with a batch size of 16 achieved the highest F-Score at 0.522.

Table 6. Comparison of experiments on the RoBERTa model

Model	Hyperparameter Value			Evaluation Metric		
	Learning Rate	Epoch	Batch Size	Precision	Recall	F-Score
RoBERTa	0,0001	85	16	**0,598**	0,463	**0,522**
RoBERTa	0,0001	85	32	0,563	**0,473**	0,514
RoBERTa	0,0001	85	64	0,555	0,465	0,506

The testing results of the RoBERTa model for classifying Indonesian-language Quran translation words into BIO labels (B-PER, I-PER, and O) can be seen in Table 7.

Table 7. Example of testing results from the RoBERTa model.

Kata	True Label	Model Prediction
		RoBERTa
Dan	O	O
sesungguhnya	O	I-PER
yang	O	O

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Kata	True Label	Model Prediction
		RoBERTa
demikian	O	O
itu	O	O
sungguh	O	O
berat	O	I-PER
kecuali	O	O
bagi	O	O
orang-orang	B-PER	B-PER
yang	I-PER	O
khusyu	I-PER	O
yaitu	O	O
orang-orang	B-PER	B-PER
yang	I-PER	O
meyakini	I-PER	O
bahwa	I-PER	O
mereka	I-PER	O
akan	I-PER	O
menemui	I-PER	O
Tuhannya	I-PER	O
dan	O	O
bahwa	O	O
mereka	O	O
akan	O	O
kembali	O	O
kepada-Nya	O	O

3.2 Discussion

A comparison of the RoBERTa model was made against the BERT model using the same hyperparameter and dataset settings as shown in Table 2. RoBERTa had a lower F-Score (0.527) compared to BERT (0.522). This suggests that in this experiment, BERT tended to perform better in identifying entities. RoBERTa had a Precision of 0.598, while BERT had a Precision of 0.580. Although the difference between these two values is small, RoBERTa can provide more accurate results in classifying certain entities as positive. On the other hand, BERT had a higher Recall of 0.483 compared to RoBERTa (0.463). This indicates that BERT is better at identifying entities that should be predicted as positive.

Despite both models having the same number of epochs and learning rate, differences in architecture and internal parameters can affect the approach and internal understanding formation of the model. In such situations, further exploration is needed to determine the most suitable combination of hyperparameters for the RoBERTa model. Additionally, batch size can significantly affect model performance. By conducting further experiments on batch size variations, F-Score and other metrics can be improved or optimized.

Internal comparisons within the RoBERTa model were made by varying the batch size (16, 32, and 64), while keeping learning rate and epochs constant at 0.0001 and 85, respectively. Table 3 shows the results of the internal comparison in the RoBERTa model. Model performance tended to decrease with an increase in batch size, indicating that a batch size of 16 provided better performance compared to larger batch sizes. In this case, the model with a batch size of 16 had a higher F-Score compared to batch sizes 32 and 64.

Based on Table 4, entities correctly and incorrectly predicted by the RoBERTa model are identified. Predictable entities are those that frequently appear in the training data, such as the faithful, the fortunate, and the unbelievers. The model's low performance in labeling human entities suggests that the model struggles to extract entities, but the RoBERTa model built in this study can still be used to extract human entities.

This study has limitations in determining hyperparameters due to hardware memory constraints. The results indicate that RoBERTa is slightly inferior to BERT in this experiment with an F-Score of 0.522. To enhance model performance in further research, additional dataset incorporation is recommended to enable the model to understand variations in the data. Additionally, fine-tuning hyperparameters more optimally for the model and increasing hardware memory are suggested to avoid limitations in hyperparameter adjustments.

4. CONCLUSION

Based on the analysis of the comparison between the RoBERTa and BERT models, it was found that BERT demonstrated better performance with an evaluated F-Score of 0.527 compared to RoBERTa with an F-Score of 0.522. Although the difference is relatively small, BERT tends to excel in identifying human entities in this case. In the internal analysis of the RoBERTa model, variations in batch size showed a significant impact on model performance. A batch size hyperparameter of 16 resulted in a higher F-Score compared to batch sizes 32 and 64, indicating that the model is better at identifying entities with a smaller batch size. Despite its lower performance, the RoBERTa model can be used to extract human entities in the translated text of the Indonesian-language Quran. Evaluation of entities shows that the model can predict entities that frequently appear in the training data accurately. For further research, the addition of datasets can be considered to generate a better model for extracting human entities in the translated text of the Indonesian-language Quran. Researchers also suggest conducting more diverse experiments, especially regarding the use of more optimal hyperparameters in model development. Additionally, the addition of memory to the hardware is recommended to avoid limitations in hyperparameter adjustments.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude and thanks to Allah SWT for His blessings and guidance, providing me with unwavering strength throughout the implementation of this research. Without Him, I am nothing. I would like to extend my boundless thanks to Mr. Moch Arif Bijaksana for the continuous guidance and directions during the preparation of this research. Your patience and knowledge have provided me with the guidance to stay spirited and optimistic in completing this research. Thanks also to my family for their prayers and meaningful support. And to my friends who supported and encouraged me during my most challenging moments. I acknowledge that the valuable participation from various parties is crucial to the success of this research. Regardless of the magnitude of the support provided, it is highly significant to me and the success of this research. I hope that many benefits can be derived from what I have written here.

REFERENCES

[1] M. N. Annisa and R. Safii, “Analisis Kebutuhan Belajar Bahasa Arab sebagai Bahasa Asing dalam Konteks Pendidikan Tinggi,” ELOQUENCE : Journal of Foreign Language, vol. 2, no. 2, pp. 313–328, Aug. 2023, doi: 10.58194/eloquence.v2i2.861.

[2] C. K. Dagli, M. M. Dakake, and J. E. B. Lumbard, “The Study Quran : A New Translation with Notes and Commentary,” 2017. [Online]. Available: http://www.harpercollins.com

[3] Permana G and Naelin Najihah B, “A Comparative Study of the Methods of Interpretation of Verses about Satan in Tafsir Al-Manar,” Journal of ‘Ulūm al-Qur’ān and Tafsīr Studies, vol. 2, no. 2, pp. 51–74, 2023, doi: 10.54801/juquts.v2i2.213.

[4] Bayan A, “Studi Kritis Terjemah Al-Qur’an Depag RI (Tela’ah Terhadap Ayat-Ayat Mutasayabihat dan Aqidah Dalam Perspektif NU),” Jurnal Syntax Admiration, vol. 1, pp. 638–655, 2020.

[5] K. Andesa, “Super Agent Chatbot ‘3S’ Sebagai Media Informasi Menggunakan Metoda Natural Language Processing (NLP),” Jurnal Teknologi dan Open Source, pp. 53–64, 2019, [Online]. Available: http://www.sar.ac.id

[6] H. Eka Rosyadi, F. Amrullah, R. David Marcus, and R. Rahman Affandi, “Rancang Bangun Chatbot Informasi Lowongan Pekerjaan Berbasis Whatsapp dengan Metode NLP (Natural Language Processing),” BRILIANT: Jurnal Riset dan Konseptual, vol. 5, 2020, doi: 10.28926/briliant.

[7] A. H. Tantri and N. A. Rakhmawati, “Ekstraksi Informasi Semantik dan Spatiotemporal pada Artikel Online Terkait Bencana di Indonesia,” JURNAL SISTEM INFORMASI BISNIS, vol. 10, no. 1, pp. 114–121, Aug. 2020, doi: 10.21456/vol10iss1pp114-121.

[8] P. J. Gorinski et al., “Named Entity Recognition for Electronic Health Records: A Comparison of Rule-based and Machine Learning Approaches,” ArXiv, 2019, [Online]. Available: http://arxiv.org/abs/1903.03985

[9] Giarsyani N, Hidayatullah A, and Rahmadi R, “Komparasi Algoritma Machine Learning dan Deep Learning Untuk Named Entity Recognition : Studi Kasus Data Kebencanaan,” Jurnal Informatika & Rekayasa Elektronika, vol. 3, pp. 48–57, 2020.

[10] H. L. Chieu and H. T. Ng, “Named Entity Recognition with a Maximum Entropy Approach,” 2003.

[11] J. Li, A. Sun, J. Han, and C. Li, “A Survey on Deep Learning for Named Entity Recognition,” IEEE Trans Knowl Data Eng, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[12] R. Diah, A. Ningtias, and M. Arif Bijaksana, “People Entity Recognition for the English Quran Translation using BERT,” Jurnal Media Informatika Budidarma, 2023, doi: 10.30865/mib.v7i1.5586.

[13] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” ArXiv, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[14] Haidir M and Purwarianti A, “Short Answer Grading Using Contextual Word Embedding and Linear Regression,” 2020.

[15] Setiawan M, Romadhony A, and Hasmawati, “Ekstraksi Informasi Beasiswa dari Media Sosial menggunakan BiLSTM-CRF,” 2023.

[16] S. Akrah, “DuluthNLP at SemEval-2021 Task 7: Fine-Tuning RoBERTa Model for Humor Detection and Offense Rating,” 2021. [Online]. Available: https://github.com/akrahdan/

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 648−656
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

[17] Rismayati and Luthfiarta A, “VGG16 Transfer Learning Architecture for Salak Fruit Quality Classification,” Jurnal Informatika dan Teknologi Informasi, vol. 18, no. 1, pp. 37–48, 2021, doi: 10.31515/telematika.v18i1.4025.
[18] Y. A. Suwitono and F. J. Kaunang, “Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras,” Jurnal Komtika (Komputasi dan Informatika), vol. 6, no. 2, pp. 109–121, Nov. 2022, doi: 10.31603/komtika.v6i2.8054.
[19] L. Marifatul Azizah, S. Fadillah Umayah, and F. Fajar, “Deteksi Kecacatan Permukaan Buah Manggis Menggunakan Metode Deep Learning dengan Konvolusi Multilayer,” Semesta Teknika, vol. 21, no. 2, 2018, doi: 10.18196/st.212229.
[20] D. I. Dzidny, M. A. Bijaksana, and K. M. Lhaksmana, “Supervised Learning Approaches for Nested People Entity Extraction in Indonesian Translated Quran,” Building of Informatics, Technology and Science (BITS), vol. 4, no. 1, Jun. 2022, doi: 10.47065/bits.v4i1.1758.