HIJM
HORIZON: Indonesian Journal of Multidisciplinary

p-ISSN: 2987-7776
e-ISSN: 2987-7180
Vol. 4, No. 3, 2026

IDENTIFICATION OF QUANTITY RELATIONSHIPS IN MATH STORY PROBLEMS USING BIDIRECTIONAL LONG SHORT-TERM MEMORY (Bi-LSTM)

Diana Ayu Puspita[1], Agung Prasetya[2], Yayak Kartika Sari[3]
[1, 2, 3]Universitas Bhinneka PGRI, Jl. Mayor Sujadi Timur No. 7, Tulungagung, Jawa Timur, Indonesia
Email: dianaayustudent@ubhi.ac.id

Article History

Received: 29-05-2026

Revision: 18-06-2026

Accepted: 23-06-2026

Published: 27-06-2026

Abstract. This research is motivated by the difficulty in automatically identifying and classifying quantity relationships in mathematical word problems due to variations in sentence structure, the use of natural language, and the complexity of the context contained in the text. These problems pose a challenge in the development of natural language processing systems to support mathematical problem understanding. This study aims to identify quantity relationships in mathematical word problems using a deep learning approach based on Bidirectional Long Short-Term Memory (BiLSTM). The research method used is a computational experimental study with stages of data collection, text preprocessing, vocabulary formation, model training, and model performance evaluation. The evaluation was carried out using accuracy, precision, recall, and F1-score metrics to measure the model's ability to classify quantity relationships. The results show that the BiLSTM model is able to identify quantity relationships well, indicated by an accuracy value of 0.9224 and high precision, recall, and F1-score values in each class category. These findings indicate that BiLSTM is effective in understanding the pattern of quantity relationships in mathematical word problems and has the potential to be used as an automated approach in natural language processing-based mathematical text analysis.

Keywords: BiLSTM, Text Classification, Mathematical Word Problems, Deep Learning, Quantity Relationships

Abstrak. Penelitian ini dilatarbelakangi oleh kesulitan dalam mengidentifikasi dan mengklasifikasikan hubungan kuantitas pada soal cerita matematika secara otomatis akibat variasi struktur kalimat, penggunaan bahasa alami, dan kompleksitas konteks yang terkandung dalam teks. Permasalahan tersebut menjadi tantangan dalam pengembangan sistem pemrosesan bahasa alami untuk mendukung pemahaman soal matematika. Penelitian ini bertujuan untuk mengidentifikasi hubungan kuantitas dalam soal cerita matematika menggunakan pendekatan deep learning berbasis Bidirectional Long Short-Term Memory (BiLSTM). Metode penelitian yang digunakan adalah penelitian eksperimen komputasional dengan tahapan pengumpulan data, prapemrosesan teks, pembentukan vocabulary, pelatihan model, serta evaluasi kinerja model. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, dan F1-score untuk mengukur kemampuan model dalam mengklasifikasikan hubungan kuantitas. Hasil penelitian menunjukkan bahwa model BiLSTM mampu mengidentifikasi hubungan kuantitas dengan baik, ditunjukkan oleh nilai accuracy sebesar 0,9224 serta nilai precision, recall, dan F1-score yang tinggi pada setiap kategori kelas. Temuan ini menunjukkan bahwa BiLSTM efektif dalam memahami pola hubungan kuantitas pada soal cerita matematika dan berpotensi digunakan sebagai pendekatan otomatis dalam analisis teks matematika berbasis pemrosesan bahasa alami.

Kata Kunci: BiLSTM, Klasifikasi Teks, Soal Cerita Matematika, *Deep Learning*, Hubungan Kuantitas

How to Cite: Puspita, D. A., Prasetya, A., & Sari, Y. K. (2026). Identification of Quantity Relationships in Math Story Problems Using Bidirectional Long Short-Term Memory (Bi-LSTM). *HORIZON: Indonesian Journal of Multidisciplinary*, 4 (3), 2407-2418. http://doi.org/10.54373/hijm.v4i3.6049

INTRODUCTION

Math Word Problem (MWP) is a mathematical problem presented in the form of a narrative that requires learners not only to perform calculations but also to understand the meaning of sentences and the context embedded in the story (Huang et al., 2018). Solving MWP involves interpreting quantitative information expressed through natural language and transforming it into appropriate mathematical representations. Therefore, successful problem solving depends on both mathematical reasoning and language comprehension. Because of this complexity, MWP has become one of the important benchmarks in the development of Intelligent Tutoring Systems and educational artificial intelligence, as it reflects the ability of a system to understand language and perform reasoning processes similar to human cognition (Son, 2024).

One of the most critical stages in solving MWP is identifying the relationships among quantities mentioned in the text. Numerical values in a problem are rarely presented as isolated entities; instead, they interact to form meaningful mathematical relationships. Mandal and Naskar (2021) classify quantity relationships into four major categories: Change, Compare, Combine, and Division–Multiplication. Accurate identification of these relationships is essential because it serves as the foundation for constructing mathematical models and equations. Errors in recognizing quantity relationships can lead to incorrect mathematical representations and ultimately produce incorrect solutions (Roy & Roth, 2015).

Despite its importance, automatic identification of quantity relationships remains a challenging task, particularly in natural language-based problems. Variations in sentence structure, lexical choices, synonym usage, and implicit information often make it difficult to consistently recognize relationships among quantities. Previous studies have attempted to address this challenge using rule-based approaches, linguistic features, and statistical methods. However, rule-based systems are generally effective only for simple and explicitly structured sentences and tend to fail when encountering complex sentence patterns or implicitly stated information (Koncel-Kedziorski et al., 2016; Mandal & Naskar, 2021). Consequently, there is still a significant gap between the need for flexible quantity relationship recognition and the capabilities of conventional approaches.

Recent studies further emphasize that solving MWP involves more than extracting numerical values from text. It requires understanding semantic relationships among events, entities, and quantities within a narrative context (Jie et al., 2022). When a system fails to capture these semantic connections, the resulting quantity relationships may be incorrectly identified, leading to inaccurate mathematical reasoning. This finding suggests that the primary

challenge in MWP lies not in mathematical computation itself but in the language understanding process that precedes mathematical modeling.

To address these limitations, contextual representation-based approaches have gained increasing attention in natural language processing research. Bidirectional Long Short-Term Memory (Bi-LSTM) is one of the deep learning architectures capable of capturing contextual information from both preceding and succeeding words simultaneously. This capability enables the model to learn semantic dependencies more effectively than traditional rule-based methods. Previous studies have demonstrated that Bi-LSTM performs well in various text classification and sequence labeling tasks because of its ability to extract contextual information from complex sentence structures (Prasetya et al., 2024). Furthermore, research in Indonesian language processing has shown that contextual neural models achieve superior performance in understanding semantic patterns and linguistic variations compared with conventional approaches (Willie et al., 2020).

Although numerous studies have explored automatic MWP solving, most research has focused on equation generation, answer prediction, or quantity extraction, while relatively few studies have specifically investigated quantity relationship classification in Indonesian mathematical texts. Existing studies also predominantly employ rule-based approaches, creating limitations in handling diverse linguistic patterns. Therefore, a research gap remains regarding the application of deep learning-based contextual models for identifying quantity relationships in Indonesian Math Word Problems.

The novelty of this study lies in the application of a Bidirectional Long Short-Term Memory (Bi-LSTM) model to automatically classify quantity relationships in Indonesian Math Word Problems into four categories: Change, Compare, Combine, and Division–Multiplication. Unlike previous studies that primarily relied on rule-based techniques, this research utilizes contextual semantic representations to capture complex linguistic patterns and quantity interactions. In addition, the study evaluates the effectiveness of the proposed Bi-LSTM model by comparing its performance with conventional rule-based approaches, thereby providing empirical evidence regarding the advantages of deep learning for quantity relationship identification. Based on these considerations, this study aims to develop and evaluate a Bi-LSTM model for identifying quantity relationships in Indonesian Math Word Problems. The findings are expected to contribute to the advancement of natural language understanding in educational applications and support the development of more intelligent, accurate, and adaptive tutoring systems for solving narrative-based mathematical problems.

METHODS

This study employed a quantitative experimental approach to develop and evaluate a Bidirectional Long Short-Term Memory (Bi-LSTM) model for identifying quantity relationships in Indonesian Math Word Problems. The research was conducted through several sequential stages, including data collection, preprocessing, model development, training, and evaluation. The dataset was compiled from various sources, including Electronic School Books (BSE), the Indonesian Book Information System (SIBI), e-books, and printed elementary mathematics textbooks for Grades I–VI. To increase linguistic diversity, data augmentation was performed using paraphrasing techniques while preserving the original quantity relationships. Each problem was then annotated into one of four quantity relationship categories, namely Change, Compare, Combine, and Division–Multiplication. The annotation process followed the Expression Tree representation and was validated through a human-in-the-loop procedure to ensure consistency and accuracy.

Prior to model training, the text data underwent preprocessing consisting of text normalization, tokenization, number mapping, and sequence padding. The processed text was transformed into word embeddings and subsequently fed into a Bi-LSTM architecture capable of capturing contextual information from both preceding and succeeding words within a sentence. The output generated by the Bi-LSTM layer was classified using a Softmax function to determine the corresponding quantity relationship category. The model was implemented using Python 3.10 with TensorFlow/Keras, NumPy, Pandas, and Scikit-learn libraries. Experiments were conducted using three train-test split scenarios, namely 70:30, 80:20, and 90:10. The training process employed an embedding dimension of 128, hidden dimension of 256, batch size of 32, learning rate of 0.001, and 30 epochs using the Adam optimizer. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics, while a confusion matrix was utilized to provide a more detailed analysis of classification performance for each quantity relationship category.

RESULTS AND DISCUSSION

Dataset Collection

Source and Amount of Data

The dataset used in this study is in the form of a mathematical story problem in Indonesian which contains one basic arithmetic operation, namely addition, subtraction, multiplication and division. Each question is arranged in the form of a narrative text consisting of one to several interrelated sentences.

Table 1. Distribution of data by category

Yes	Categories	Amount of Data
1	*Exchange*	114
2	*Combine*	73
3	*Compare*	8
4	*Division-Multiplication*	108
	Total	303

Based on these distributions, the dataset has covered all categories of quantity relationships used in the study. This variety of categories is important to train models to be able to recognize the various patterns of relationships that appear in math story problems. In the modeling stage, the labels of the quantity relationships are then mapped into the labels of arithmetic operations, namely OP_ADD, OP_SUB, OP_MUL, and OP_DIV. This mapping is carried out based on the meaning of the relationships contained in each question, so that the model can learn the relationship between the text and the corresponding arithmetic operations. The dataset that has been compiled is then divided into training data and validation data using three data sharing scenarios, namely 70:30, 80:20, and 90:10. This division was carried out to analyze the influence of the proportion of training data on the performance of the Bi-LSTM model.

Data Augmentation

The augmentation process in this study produced a variety of story questions from previously collected data. These variations still maintain numerical values, types of quantity relationships, and the direction of the questions in the question. Through this process, each question has several different sentence forms but has the same meaning. This aims to enrich language patterns so that the model can better recognize variations in sentence structure.

Table 2. Implementation of Augmentation on Questions

Steps	Results
Original Data	Pak Aji has 3 chickens, each of which lays 4 eggs. How many eggs are in total?
Paraphrasing	Pak Aji has 3 chickens. Each chicken produces 4 eggs. What is the total number of eggs?
Domain Word Changes	Pak Aji has 3 ducks, each one produces 4 eggs. What is the total number of eggs in all?
Expanding Equation Tree	Pak Aji has 3 chickens. If each chicken produces 4 eggs, what is the total number of eggs?

Based on the table, sentence variations can be generated without changing the quantity value or the structure of the operation used.

Data Annotation

The annotation results on the dataset show that each question has been represented in the form of a consistent mathematical structure. Annotation is done by converting the story problem into *an expression tree* in postfix format and a representation in the form of Python code. This representation is used to ensure that the quantity relationship of each problem corresponds to the resulting arithmetic operation.

Table 3. Data annotation results

Components	Results
Questions	Pak Aji has 3 chickens, each of which lays 4 eggs. How many eggs are all?
Types of Relationships	Div-Mul | *Multiplication*
Expression Tree	"3 4 [OP_MUL]"
Python Code	*"var_a = 3\nvar_b = 4\nprint(var_a * var_b)"*

Based on the annotation results, each question is successfully represented in the form of a mathematical structure that corresponds to its quantity relationship. This representation ensures that the calculation process can be carried out consistently and structured.

Dataset Structure in JSON Format

The dataset is stored in JSON format to facilitate the management and retrieval of data in the model training process. Each entry contains the identity of the question, question text, operation label, and supporting information such as relationship type, final answer, and operation annotation.


```
{
  "id_soal":"001",
  "teks_soal":"Adi punya 3 topi, diberi paman 2 topi lagi. Berapa topi adi sekarang?",
  "jenis_hubungan":"Change",
  "label_operasi":"[OP_ADD]",
  "jawaban_akhir":"5",
  "anotasi_operasi":{
  "expression_tree_postfix":"3 2 [OP_ADD]",
  "python_code":"var_a = 3\nvar_b = 2\nprint(var_a + var_b)"
  }
  },
{
  "id_soal":"002",
  "teks_soal":"Rani bermain sendiri di taman, kemudian datang Dina dan Toyib. Berapa banyak anak di taman?",
  "jenis_hubungan":"Combine",
  "label_operasi":"[OP_ADD]",
  "jawaban_akhir":"3",
  "anotasi_operasi":{
  "expression_tree_postfix":"1 2 [OP_ADD]",
  "python_code":"var_a = 1\nvar_b = 2\nprint(var_a + var_b)"
  }
  },
```


Figure 1. Example of a JSON Data Structure

Figure 1 is a sample of the question dataset. Based on this structure, each data has been compiled systematically and consistently, thus facilitating the process of pre-processing, training, and model evaluation.

Pre-Data Processing

The pre-processing results show that the question text data has been successfully converted into a form that can be processed by the model. The pre-processing stage produces a more structured and consistent representation of the text so that it can be used as input to the Bi-LSTM model**.**

Normalization of Text

The pre-processing stage produces a more structured and consistent representation of text so that it can be used as input to the Bi-LSTM model.

Table 4. Comparison of texts before and after normalization

Before	After
Pak aji has 3 chickens and each one lays 4 eggs	Pak Aji has 3 chickens, each of which lays 4 eggs.

Number Mapping

The number *mapping results* show that each numeric value in the question text has been replaced with special tokens such as <n1>, <n2>, and so on. This process aims to standardize the representation of numbers so that the model can focus more on the patterns of relationships between quantities.

Table 5. Number Mapping *Results*

Text	Results
Pak Aji has 3 chickens each laying 4 eggs	Pak Aji has <n1> chickens per egg <n2> eggs

The results *of number mapping* make the representation of numbers more consistent so that the model can focus more on studying the patterns of quantity relationships in the text of mathematical story problems.

Model Training Configuration

The configuration of the training model is determined based on the parameters used at the implementation stage.

Table 6. Model training configuration

Parameters	Value
Embedding Dimension	128
Hidden Dimension	256
Number of *Layers*	2
Dropout	0,3
Batch Size	32
Learning Rate	0,001
Optimizer	Adam
Loss Function	*CrossEntropyLoss*
Epoch	30

Model Evaluation Results

In this study, the classification of quantity relationships is represented in the form of mathematical operations, namely addition (OP_ADD), subtraction (OP_SUB), multiplication (OP_MUL), and division (OP_DIV). Each of these operations represents a specific type of quantity relationship in a mathematical story problem. Conceptually, quantity relationships such as *Change*, *Compare*, *Combine*, and *Division-Multiplication* are related to mathematical operations used in problem solving.

Change relationships are generally related to addition or subtraction operations (OP_ADD and OP_SUB), *Compare* is related to subtraction operations (OP_SUB), *Combine* is related to addition operations (OP_ADD), while *Division-Multiplication* is related to multiplication and division operations (OP_MUL and OP_DIV). Therefore, the results of the model classification in the form of operation labels are used as representations to identify the quantity relationships in the text of the question.

Evaluation was carried out on test data as many as 116 questions using *precision*, *recall*, and F1-score metrics per label, as well as aggregate metrics in the form of *accuracy*, *macro average*, and *weighted average*.

Table 7. Model test evaluation metrics per class

Operation Labels	*Precision*	*Recall*	F1 Score	Quantity
OP_ADD	0.8125	0.8667	0.8387	15
OP_SUB	0.9259	1.0000	0.9615	25
OP_MUL	1.0000	0.8205	0.9014	39
OP_DIV	0.9026	1.0000	0.9487	37

The results in table 7 show that the model performs well across labels, as seen from the F1-score values in the range of 0.83 to 0.96.

The OP_SUB and OP_DIV labels show the most stable performance with a *recall* value of 1,0000, which means that all data in that class is successfully detected. On the OP_MUL label, *the precision* value reaches 1.0000, but *a recall* of 0.8205 indicates that there is still some data that has not been detected. This shows that the model is still less sensitive to some of the patterns in the class. The OP_ADD label has an F1-score value of 0.8387, which indicates that there are still errors in the classification process in some of the data.

Table 8. Evaluation metrics

Metrics	Value
Accuracy	0.9224
Macro avg Precison	0.9102
Macro avg Recall	0.9218
Macro avg F1 *score*	0.9126

Weighted avg Precision	0.9287
Weighted avg Recall	0.9224
Weighted avg F1 *score*	0.9214

The model managed to obtain an accuracy value of 0.9224, which indicates that most of the predictions in the test data were correct. The *macro average* and *weighted average* values that are not much different show that the model's performance is relatively evenly distributed on each label. The *high recall* value in most classes indicates that the model is able to detect data well, although there are still some prediction errors in certain pattern questions.

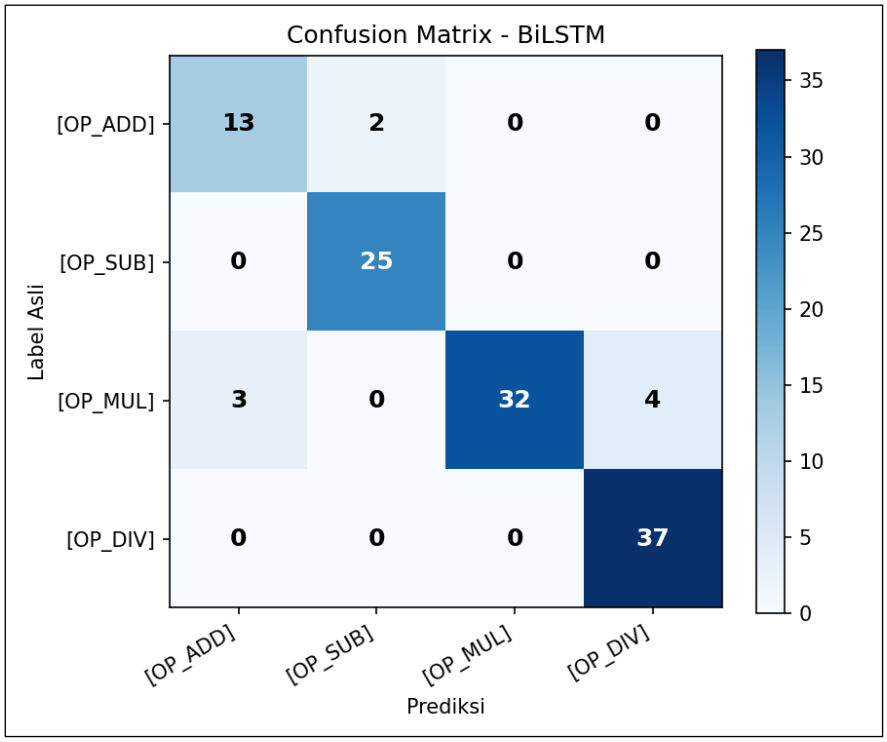

Figure 2. Confusion matrix results on test data

Based on Figure 2, the majority of quantity relationships were classified correctly, indicating that the proposed Bi-LSTM model was able to capture contextual information effectively from Indonesian Math Word Problems. The highest number of correct predictions was observed in the OP_DIV and OP_MUL classes, suggesting that multiplication and division relationships tend to contain more distinctive linguistic patterns that can be recognized consistently by the model. In contrast, a small number of misclassifications occurred, particularly between OP_ADD and OP_SUB classes. This finding indicates that addition and subtraction problems often share similar lexical and syntactic structures, making it more challenging for the model to distinguish between them accurately.

The occurrence of misclassification can be explained by the semantic similarity of mathematical story problems. In many cases, addition and subtraction relationships are expressed through comparable contextual cues, such as changes in quantity, possession, or transfer events. Although Bi-LSTM is capable of learning contextual dependencies from both directions of a sentence, it may still struggle when the semantic signals associated with different operations are highly overlapping. Similar findings have been reported by Roy and Roth (2015) and Mandal and Naskar (2021), who noted that quantity relationship identification becomes more difficult when numerical information is conveyed implicitly or through linguistically similar expressions.

Nevertheless, the relatively low number of classification errors demonstrates the effectiveness of contextual representation learning in capturing quantity relationships. Compared with rule-based approaches, which rely heavily on predefined linguistic patterns, the Bi-LSTM model is more flexible in handling variations in sentence structure and vocabulary. This result supports previous studies showing that deep learning models can better capture semantic information and contextual dependencies in natural language processing tasks (Willie et al., 2020; Prasetya et al., 2024). Therefore, the findings suggest that Bi-LSTM provides a promising approach for automatic quantity relationship identification in Indonesian Math Word Problems, although further improvements are still needed to address semantically similar classes.

Model Classification Results

The results of the model classification show that the system is able to identify the quantity relationship in the mathematical story problem based on the context of the given sentence. The prediction process is carried out using the Bi-LSTM model that has been trained beforehand. Figure 3 shows the results of the classification of the quantity relationship of *the Change label.*

```
=== PREDIKSI SOAL CERITA ===

Masukkan soal (ketik exit):
> Adi punya 3 topi, diberi paman 2 topi lagi. Berapa topi adi sekarang?

[HASIL]
Operasi       : Penjumlahan (+)
Tipe Soal     : Change
Confidence    : 99.9%

Probabilitas:
[OP_ADD] : 0.999
[OP_SUB] : 0.001
[OP_MUL] : 0.000
[OP_DIV] : 0.000
```

Figure 3. Label Change *classification results*

```
Masukkan soal (ketik exit):
> Rani bermain sendiri di taman, kemudian datang Dina dan Toyib. Berapa banyak anak di taman?

[HASIL]
Operasi       : Penjumlahan (+)
Tipe Soal     : Combine
Confidence    : 99.9%

Probabilitas:
[OP_ADD] : 0.999
[OP_SUB] : 0.001
[OP_MUL] : 0.000
[OP_DIV] : 0.000
```

Figure 4. Label Combine *classification results*

```
Masukkan soal (ketik exit):
> Pak Tani memiliki 156 ekor ayam. Setiap ayam bertelur 4 butir per minggu. Berapa telur yang dihasilkan dalam seminggu?

[HASIL]
Operasi       : Perkalian (×)
Tipe Soal     : Division-Multiplication
Confidence    : 97.5%

Probabilitas:
[OP_ADD] : 0.008
[OP_SUB] : 0.006
[OP_MUL] : 0.975
[OP_DIV] : 0.010
```

Figure 5. *Division-multiplication* label classification results

```
Masukkan soal (ketik exit):
> Pak Joni memelihara 327 ekor burung dara. Pak Edi memelihara 143 ekor burung dara. Berapa banyak selisih burung daranya?

[HASIL]
Operasi       : Pengurangan (-)
Tipe Soal     : Compare
Confidence    : 98.9%

Probabilitas:
[OP_ADD] : 0.011
[OP_SUB] : 0.989
[OP_MUL] : 0.000
[OP_DIV] : 0.000
```

Figure 6. Label Compare *classification results*

Based on the results of the classification, the model is able to identify the quantity relationship in each category with a high *level of confidence*. The prediction results show the match between the question type and the label generated by the model. In addition, the highest probability value in each prediction shows that the model is able to understand the pattern of quantity relationships based on the context of the sentence in the mathematical story problem.

CONCLUSION

Based on the results of the research that has been conducted, it can be concluded that the Bi-LSTM model can be used to identify the relationship between quantity in Indonesian elementary school math story problems. The built model showed good performance with an accuracy value of 92.24%, which indicates that most of the test data was correctly predicted. In addition, precision, recall, and F1-score values indicate that the model has relatively stable performance in each class. Although there are still some misclassifications, overall the model is able to understand the context of the story quite well. Thus, the Bi-LSTM approach is effectively used to classify quantity relationships in the text of mathematical story problems.

RECOMMENDATIONS

Based on the results of the research that has been carried out, there are several suggestions for the development of further research. The dataset can be expanded with larger amounts of data and more diverse sentence variations so that the model is able to learn more complex language patterns. In addition, future research can utilize more advanced models such as Transformer or BERT to improve understanding of language context. The addition of linguistic features, such as part-of-speech tagging and dependency parsing, can also be considered to help the model understand sentence structure more deeply. In addition, research can be developed into an end-to-end system that not only identifies quantity relationships, but is also able to generate mathematical solutions automatically from a given story problem.

REFERENCES

Huang, D., Liu, J., Lin, C. Y., & Yin, J. (2018). Neural math word problem solver with reinforcement learning. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 213–223.

Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., & Hajishirzi, H. (2016). MAWPS: A math word problem repository. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, January 2016*, 1152–1157. https://doi.org/10.18653/v1/n16-1136

Mandal, S., & Naskar, S. K. (2021). Classifying and Solving Arithmetic Math Word Problems - An Intelligent Math Solver. *IEEE Transactions on Learning Technologies*, *14*(1), 28–41. https://doi.org/10.1109/TLT.2021.3057805

Prasetya, A., Nugroho, A., & Wijaya, D. (2024). Contextual deep learning approaches for Indonesian natural language processing tasks. *Journal of Information and Communication Technology*, 23(1), 45–58.

Roy, S., & Roth, D. (2015). Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1743–1752). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1202

Willie, W., Purwarianti, A., & Surendro, K. (2020). Deep learning approaches for Indonesian text classification: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11(8), 451–458.