ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 2, July 2023

SIMILAR QUESTIONS IDENTIFICATION ON INDONESIAN LANGUAGE SUBJECTS USING MACHINE LEARNING

Hasmawati[1], Ade Romadhony[2]

[1,2]School of Computing, Telkom University

email: Hasmawati@telkomuniversity.ac.id[1], aderomadhony@telkomuniversity.ac.id[2]

Abstract

Question similarity is carried out to evaluate similarities between questions in a collection of questions in the question and answer forum and on other platforms. This is done to improve the performance of the question-and-answer forum so that new questions submitted by users can be identified as similar to existing questions in the database. Currently, research related to question similarity is still being carried out on foreign language datasets. The purpose of this research is to identify the similarity of questions in a collection of questions in Indonesian. The method used is Support Vector Machine and IndoBERT. For feature extraction, we evaluate the lexical features and syntax features of each question. For lexical feature extraction, we use the cosine similarity algorithm to calculate the distance between two objects which are represented as vectors. For syntax feature extraction we use the Indonesian part of speech tagger (POS Tag). The dataset used is a collection of questions on Indonesian subjects at the primary and secondary school levels. The results of this study show that the best performance of the Support Vector Machine is obtained from the use of the cosine similarity feature with an accuracy of 85%. While the use of the POS Tag feature or the combination of POS Tag and cosine similarity causes the model to be overfitted and the accuracy decreases to 77%. Meanwhile, for the IndoBERT model, an accuracy of 95% was obtained.

Keywords: Question Similarity, Support Vector Machine, IndoBERT, Cosine Similarity, POS Tag

Received: 02-06-2023 | **Revised:** 23-06-2023 | **Accepted:** 24-07-2023
DOI: https://doi.org/10.23887/janapati.v12i2.62582

INTRODUCTION

In the question and answer forums such as Stack Overflow, Quora, Yahoo, etc., often new questions submitted have already been asked, so the answers given should be the same as answers to similar questions already stored in the database. To improve the performance of the question and answer forum, the identification of similar questions is one possible solution. The aim is to identify whether the new questions given by the user are similar to the existing questions in the database so that the question-and-answer system can provide answers quickly.

Research related to question similarity has been developed using various approaches, one of which is using machine learning with various types of feature extraction methods according to the characteristics of the language being evaluated. One of them is the research conducted by Muntaka Al-asa'd et al [1], where they proposed a method of predicting the similarity of questions by extracting morphological, syntactic, semantic, and lexical features in a dataset of questions in Arabic. The approach taken involves several processes including preprocessing for Arabic text, feature extraction, and text classification. The dataset used is a collection of questions in Arabic with a total of 4000 pairs of questions. The method used in this research is Extreme Gradient Boosting (XGB) and feature selection with Random Forest. The performance of the method is evaluated by calculating the values for accuracy, precision, recall, and F1 score. This study succeeded in classifying questions with an accuracy of 78.2%.

Another study was conducted by [2]. In this research, they implemented the convolutional neural network to measure the similarity of questions on community question-answering systems. In this research, they used SemEval 2016 dataset and implement different feature

extraction. The results of this study indicate that the combination of CNN with external knowledge gives the best results.

Another approach using deep learning was conducted by [3], [4]. In research [4], they combine 2 methods, namely Versatile Global T-max pooling which is used to predict the subsequent word in the data collection, and DeepLSTM which is used for predicting the best answers. The combination of these methods gives good performance.

Another technique for solving the similarity task question was also proposed by [3], [5]–[8]. In research conducted by [8], they took an approach by utilizing the answers from the questions as a bridge between the 2 questions. They compared the patterns of the 2 question-answer pairs to identify similar questions. The dataset used in this research is a collection of questions from the Q&A forum CQADupStack and QuoraQP-a. In its implementation, they made 3 modules, the representation-based similarity module to predict the similarity vectors of 2 questions. The second is the matching pattern module which uses the Siamese Network to compare the matching patterns of 2 questions based on the same answers. The third module is the aggregation module which combines the similarity vectors of the two previous modules. Their experiments show that the proposed model works significantly and outperforms previous models.

The difference between this research and previous research is the previous research was carried out on a dataset of questions in foreign languages such as Arabic and English [9]–[11] and also the different methods used. Therefore, the purpose of this research is to identify the similarity of questions in a collection of questions in Indonesian. Another contribution to this research is that we built a labeled dataset for pairs of questions in Indonesian. In this study, we used a machine learning approach including Support Vector Machine (SVM) and pre-trained IndoBERT to predict the similarity of new questions to existing questions. For the question features, we evaluate the lexical features and syntax of each question. The reason for selecting lexical and syntactic features is based on previous research [1] which obtained good performance when using these features. In addition, the selection of this feature is also based on the availability of Indonesian language processing tools that are available and open-source accessible. The dataset used in this study is a collection of questions on Indonesian subjects at the elementary and secondary school levels.

METHOD

The flowchart of this research can be seen In Figure 1.

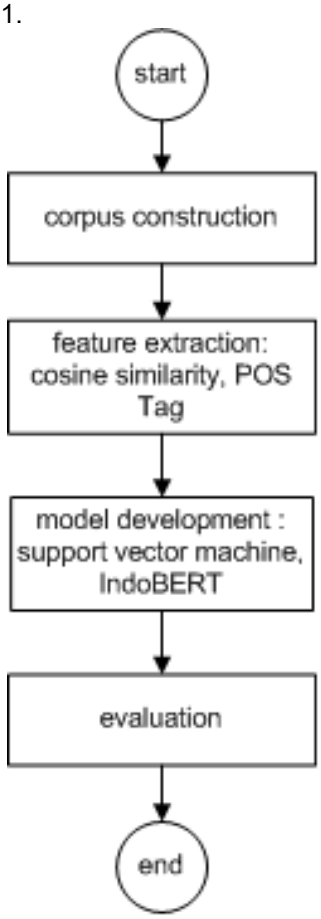

Figure 1. Research flowchart diagram

Corpus Construction

The first step in this research is collecting datasets and labeling datasets. The data collected is a collection of Indonesian language subject questions at the primary and secondary school levels. An example of the dataset used can be seen in Table 1.

The next step is to label the dataset. For each pair of questions, we label "yes" if the pair of questions are similar and label "no" otherwise. There were 622 pairs of questions consisting of 407 pairs of questions labeled "yes" and 215 pairs of questions labeled "no". To do the dataset labeling, we involved 3 undergraduate students. An example of a dataset of questions

that have been labeled based on their similarity can be seen in Table 2.

Table 1. Sample of dataset

NO	Text	Question	Answer
1	Adi:Benar dalam liburan ini sekolah kita akan berdarma wisata, Pak? Kepala Sekolah:Benar! Mengapa Adi bertanya? Adi:Untuk meyakinkan diri. Darmawisata kemana, Pak? Kepala Sekolah:Belum dipastikan. Mungkin ke Kebun Raya Bogor. Mungkin pula, ke Pantai Pangandaran. Adi:Mudah-mudahan ke Kebun Raya Bogor. Saya, belum pernah kesana. Kepala Sekolah:Itu hasil rapat yang menentukan.	Watak kepala sekolah berdasarkan penggalan drama di atas adalah	Bijaksana
2	Mak Dahlia: (Menghela napas panjang) Memangnya kamu mau kemana? Mengapa kamu merias diri? Cantika:Tidak kemana-mana, tetapi aku suka berias saja. Lihatlah, Mak. Bukankah aku ini cantik? Ah, bukan. Aku bukannya cantik. Tapi aku cantik sekali! (sambil terus mengedip-ngedipkan mata di depan cermin	Kalimat yang ada di dalam kurung pada cuplikan naskah drama "Batu Menangis" di atas disebut	Kramagung
3		' Ide yang sangat bagus. Mengapa kamu tidak bercerita terlebih dahulu.Kesalahan penulisan kalimat langsung tersebut, adalah …	Tidak ada tanda petik di akhir kalimat dan tidak membubuhkan tanda tanya di akhir kalimat ke-2.
4		Salah satu alat komunikasi yang berkembang di zaman sekarang adalah penggunaan WA dan instagram. WA dan instagram termasuk media	Sosial
5		Yang termasuk potongan struktrur teks Laporan Hasil Observasi adalah....	Definisi umum; deskripsi bagian; simpulan

In this study, we did not preprocess the dataset, such as stopword removal, stemming, etc., because removing common words in a question would change the context of the question. So that it can affect the prediction of the similarity of the questions.

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 12, Issue 2, July 2023

Table 2. Randomly selected pair of questions

NO	Question 1	Question 2	Label
1	Yang merupakan kalimat utama dari paragraf tersebut adalah ...	Manakah kalimat utama dalam paragraf tersebut?	yes
2	Dibawah ini yang merupakan makanan khas daerah Yogyakarta adalah ...	Contoh makanan khas Yogyakarta adalah ..	yes
3	Berikut ini yang bukan merupakan unsur unsur pada peta pikiran adalah..	Berikut kalimat iklan yang sesuai untuk iklan jenis pengumuman ..	no
4	Berikut yang bukan merupakan ciri ciri iklan yang baik dan benar adalah..	Berikut yang bukan merupakan media untuk pengiklanan adalah..	no

In Table 2, it can be seen that the pair of questions in line 1 have similarities, both ask about the main sentence in a paragraph. While the pairs of questions in line 4 are not similar because question 1 asks about the characteristics of good advertising, while question 2 asks about advertising media.

Feature Extraction

We perform lexical and syntax feature extraction. To extract the lexical features we calculate the cosine similarity between the two pairs of questions using the cosine similarity algorithm from scikit-learn. Cosine Similarity is used to measure the similarity between pairs of questions represented in vectors by calculating the cosine value. The formula for calculating the cosine similarity value of two pairs of questions is shown in Equation 1 [1].

$$\cos(\mathbf{A},\mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i\mathbf{B}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{A}_i)^2}\sqrt{\sum_{i=1}^{n}(\mathbf{B}_i)^2}} \quad (1)$$

Where A and B are question pairs.

To extract the syntax features we use Indonesian language POS tagging by adopting the POS Tag extraction stages in the research of Rani Aulia et al [12], [13]. Table 3 shows a sample of feature extraction result.

Table 3. sample of feature extraction result

NO	Question 1	Question 2	Cosine Similarity	POS Tag of Question 1	POS Tag of Question 2	Same Postag	Match Postag	Percentage of Match Postag
1	sikap rio sebaiknya	bagaimanakah sebaiknya sikap rio?	0.796904	['NN', 'NN', 'RB']	['NN', 'RB', 'VB', 'NNP']	2	1	0.25
2	rambu lalu lintas memiliki arti	makna dari rambu lalu lintas tersebut yaitu	0.783669	['NN', 'CC', 'NN', 'VB', 'NN']	['NN', 'IN', 'NN', 'CC', 'NN', 'PR', 'SC']	4	3	0.428571
3	saat pergi ke ciwidey, fahri menemukan banyak ...	fahri menemukan banyak rambu lalu lintas ini s...	0.984628	['NN', 'VB', 'IN', 'NN', 'NN', 'VB', 'CD', 'NN...	['NN', 'VB', 'CD', 'NN', 'CC', 'NN', 'PR', 'NN...	17	16	0.888889
4	bendera ini dikenal dengan nama bendera	bendera ini disebut bendera	0.881748	['NN', 'PR', 'VB', 'IN', 'NN', 'JJ']	['NN', 'PR', 'VB', 'NN']	4	3	0.75
5	rambu lalu	arti rambu	0.880561	['NN',	['NN',	4	4	0.8

lintas ini menunjukkan	lalu lintas ini	'CC', 'NN', 'PR', 'VB']	'NN', 'CC', 'NN', 'PR']

Question Similarity Model

To build a question similarity identification model, we used two algorithms; a pre-trained model indoBERT and Support Vector Machine.

1) Support Vector Machine (SVM) is a supervised algorithm. In the case of text classification, the algorithm divides the data into two classes using a vector line called a hyperplane. For implementing the SVM algorithms we use Python and the library scikit-learn. For the tuning parameter, we used kernel rbf, regularization parameter C is 1.0 and gamma is auto.
2) IndoBERT is a monolingual BERT model for Indonesian. IndoBERT has 3 models; IndoBERT-liteBase, IndoBERTBase, IndoBERTLarge. In this research, we implement the pre-trained indoBERT BASE p1 proposed by B. Wilie [14] and Fajri [15] that was pre-trained on Indo4B Indonesian corpus.

Evaluation

To evaluate the model performance, we measure accuracy, precision, recall, and f1 score. We conducted an experiment to see the performance of the SVM and indoBERT algorithms with a combination of feature extraction;

1) Syntax features only (POS Tag),
2) Lexical features only (Cosine Similarity),
3) A combination of POS Tag features and cosine similarity.

In implementing the algorithm, we split the dataset into 3 parts; 80% data for training, 10% for validation, and 10% for testing.

RESULT AND DISCUSSION

Based on the experiment scenario mentioned above, we conducted an experiment to assess the performance of the SVM and indoBERT classification algorithms with a combination of features 1) syntax features only (POS Tag), 2) lexical features only (Cosine Similarity), 3) a combination of POS Tag and cosine similarity features.

To evaluate the classification model, we calculate the accuracy, precision, recall, and F1 score using the confusion matrix.

1) Accuracy is the ratio of Correct predictions (positive and negative) to the entire data. Accuracy, in this case, is "how many percent of the questions are predicted to be similar and not similar from all questions".
2) Precision is the ratio of correctly positive predictions compared to the overall positive predicted outcome. Precision in this case is the percentage of questions that are similar to the total questions that are predicted to be similar.
3) Recall (sensitivity) is the ratio of correctly positive predictions compared to all true positive data. Recall in this case what percentage of questions are predicted to be similar compared to all questions that are similar.
4) F1 Score is a weighted average comparison of precision and recall.

Table 4 Performance of the classification algorithm based on the feature extraction used

No	Model	Accuracy			Recall			Precision			F-1		
		Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
1	SVM POSTag	0.86	0.77	0.80	0.83	0.71	0.74	0.86	0.76	0.82	0.84	0.72	0.76
2	SVM Cosine Similarity	0.81	0.84	0.85	0.76	0.76	0.83	0.80	0.90	0.84	0.77	0.79	0.83
3	SVM POSTag + Cosine Similarity	**0.88**	0.75	0.77	**0.85**	0.68	0.69	**0.87**	0.75	0.79	**0.86**	0.69	0.70
4	indoBERT-base-p1	0.54	**0.98**	**0.95**	0.49	**0.98**	**0.94**	0.49	**0.99**	**0.95**	0.49	**0.98**	**0.94**

Based on Table 4, it can be seen that in the training data, the highest accuracy is obtained from the combination of using the POS Tag and Cosine similarity features, which is 88%. However, in testing data, the accuracy decreased to 77%. The same condition was when using the POS Tag feature only where the accuracy of the training data was obtained by 86%, but decreased to 80% in the test data. In using the cosine similarity feature, better results are obtained, the accuracy of the training data is obtained by 81% and increases to 85% in the test data. As for the indoBERT model, an accuracy of 54% was obtained on the training data, and 95% on the test data.

Based on the results of this performance it was concluded that on the training data, the use of a combination of POS Tag and cosine similarity features improves the performance of the algorithm. However, in the testing data, the highest accuracy is obtained when only using the cosine similarity feature. In other words, the model is overfitting to the dataset. Based on our evaluations and observations, this condition occurs because the data used is less varied when compared to the complexity of the model. In this case, it can be seen from the words used in the Indonesian questions at the primary and secondary school levels that are still less varied.

CONCLUSION

A model has been built to identify the similarity of questions at primary and secondary school levels using the SVM and indoBERT algorithms. In the implementation, we extract the lexical and syntactic features of each question. The experiment results show that the best model performance is obtained from the use of the cosine similarity feature in the SVM algorithm. Meanwhile, the use of the POS Tag feature or the combination of POS Tag and cosine similarity causes the model to become overfitting to the dataset and the model accuracy decreases. To improve the performance of the model in future studies, we propose the use of other feature extraction, such as TF-IDF, a count vectorizer that focuses on the frequency of occurrence of words in a document, and also evaluate the semantic features with various feature extraction approaches. In addition, improvements can also be made to the dataset by increasing the number and variety of words used.

REFERENCES

[1] M. Al-Asa'd, N. Al-Khdour, M. B. Younes, E. Khwaileh, M. Hammad, and M. AL-Smadi, "Question to Question Similarity Analysis using Morphological, Syntactic, Semantic, and Lexical Features," in *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/AICCSA47632.2019.9035248.

[2] V.-T. Nguyen, A.-C. Le, and H.-N. Nguyen, "A Model of Convolutional Neural Network Combined with External Knowledge to Measure the Question Similarity for Community Question Answering Systems," *Int J Mach Learn Comput*, vol. 11, no. 3, pp. 194–201, May 2021, doi: 10.18178/ijmlc.2021.11.3.1035.

[3] Y. Yulin and Z. Guiyun, "High school math text similarity studies based on CNN and BiLSTM," in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, IEEE, Dec. 2020, pp. 1982–1986. doi: 10.1109/ICMCCE51767.2020.00434.

[4] D. V. Vekariya and N. R. Limbasiya, "A Novel Approach for Semantic Similarity Measurement for High Quality Answer Selection in Question Answering using Deep Learning Methods," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020, pp. 518–522. doi: 10.1109/ICACCS48705.2020.9074471.

[5] Y. Chali and R. Islam, "Question-Question Similarity in Online Forums," in *Proceedings of the 10th Annual Meeting of the Forum for Information Retrieval Evaluation*, New York, NY, USA: ACM, Dec. 2018, pp. 21–28. doi: 10.1145/3293339.3293345.

[6] T.-T. Ha, V.-N. Nguyen, K.-H. Nguyen, K.-A. Nguyen, and Q.-K. Than, "Utilizing SBERT For Finding Similar Questions in Community Question Answering," in *2021 13th International Conference on*

Knowledge and Systems Engineering (KSE), IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/KSE53942.2021.9648830.

[7] K. M. Shivani and M. R. Aswathy, "Study on Techniques for Analyzing Semantic Similarity in Question Answering System," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, May 2018, pp. 633–636. doi: 10.1109/ICOEI.2018.8553832.

[8] Z. Wang *et al.*, "Match^2," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2020, pp. 559–568. doi: 10.1145/3397271.3401143.

[9] W. T. Alshammari and S. AlHumoud, "TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT With BiLSTM," *IEEE Access*, vol. 10, pp. 91509–91523, 2022, doi: 10.1109/ACCESS.2022.3198955.

[10] F. Kunneman, T. Castro Ferreira, E. Krahmer, and A. van den Bosch, "Question Similarity in Community Question Answering: A Systematic Exploration of Preprocessing Methods and Models," in *Proceedings - Natural Language Processing in a Deep Learning World*, Incoma Ltd., Shoumen, Bulgaria, Oct. 2019, pp. 593–601. doi: 10.26615/978-954-452-056-4_070.

[11] N. Othman, R. Faiz, and K. Smaïli, "Learning English and Arabic question similarity with Siamese Neural Networks in community question answering services," *Data Knowl Eng*, vol. 138, p. 101962, Mar. 2022, doi: 10.1016/j.datak.2021.101962.

[12] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," in *2014 International Conference on Asian Language Processing (IALP)*, IEEE, Oct. 2014, pp. 70–73. doi: 10.1109/IALP.2014.6973521.

[13] R. A. Hidayat, I. N. Khasanah, W. C. Putri, and R. Mahendra, "Feature-Rich Classifiers for Recognizing Textual Entailment in Indonesian," *Procedia Comput Sci*, vol. 189, pp. 148–155, 2021, doi: 10.1016/j.procs.2021.05.094.

[14] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 843–857, Dec. 2020.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.