ICIC Express Letters
Volume **17**, Number **4**, April **2023**

 ISSN 1881-803X
pp. **389–396**

SENTENCES SIMILARITY DETECTION IN INDONESIAN POETRY COMPARISON USING SIAMESE MALSTM

Egi Anggriawan, Farhan Farid and Riri Fitri Sari

Department of Electrical Engineering
Faculty of Engineering
University of Indonesia
Jl. Margonda Raya, Pondok Cina, Beji, Depok 16424, Indonesia
{ egi.anggriawan; farhan.farid; riri }@ui.ac.id

Received April 2022; accepted July 2022

Abstract. *Poetry is one of the fashions for someone to express their thoughts and feelings beautifully and imaginatively. To preserve a better transformation in the evolution of poetry, there must be a critical thinking for something new and original. Comparative literature is one of the methods often used to examine the originality of poetry. In comparative literature, the similarities and differences in several poetry become the object of research. One that can be compared is the semantic meaning present in each line of poetry. In this study, the originality verification will be carried out by the model of machine learning which is trained to understand the semantic meaning and similar wording of each line of poetry pairs. This study uses the Siamese MaLSTM algorithm to create the model needed. The identifier model is built using Indonesian poetry collection and goes through several stages such as data preprocessing, word embeddings, training and testing. The result of the evaluation model provides a good accuracy in recognizing semantic similarity in sentences pairs. This study also presents the description of system application which needed to gather Indonesian poetry collection, identify the similarity of poetry from the prior and manage Indonesian poetry community. The machine learning model that has been built will be used in the system application as a cloud service that will check the originality of registered poems with the poetry collection in the database.*
Keywords: Poetry, Originality, Semantic similarity, Siamese MaLSTM, Application development

1. **Introduction.** Poetry is a literary form that conveys the poet's feelings and thinking imaginatively, aesthetically, concise and rich in meaning [1]. Poetry are written with words selection and set carefully to improve people's awareness about life experience. The poetry presentation is not only through the manuscript but also the arrangement of pronunciations and rhythm, so that it can evoke distinctive impression for the listeners. Poetry tends to be easily distinguished from other literature seen from the typographic which contains repetition and rhyme in each line [2]. Poets often use various style to make the rhymes more aesthetic and exude much meaning.

Poetry has always encountered transformation and evolution throughout the ages. This happens considering as a work of art always creates tension between conventions and innovation [3]. Poetry is constantly transforming according to the evolution of palates and changes in aesthetic concepts. Changes with the intent of thriving need to be pursued in the context of poetry creation. This signifies the advancement of human intellectual in creating ideas that present novelty. Originality is an essential factor that must be pursued to achieve novelty in poetry creations. With originality, the poet will elevate an imaginative side which is full of uniqueness. Spontaneity of the unique imaginative side is usually presage of an evolutionary work [4].

DOI: 10.24507/icicel.17.04.389

To examine the originality of poetry, the approach that can be used is to do comparative literature. In comparative literature, the differences and similarities that exist in two literary works are objects that will be compared [5]. The more similarities found implies a high influence from other literature. The more differences, the more creative in the context of novelty in the literature. However, to declare an originality claim on a created poetry, comparing two poems alone is not enough. Comparisons should be made on a number of pre-existing poems. An expert might not be able to compare all the poems within a limited time, and the machine learning approach is needed to automate the originality checking process.

In recent decades, the machine learning algorithm has been widely used to solve problems in various areas of life, not least in the field of literature. Some researchers try to utilize the ability of the machine learning algorithm to create a model that can make a series of poems with a certain style like Haiku style [6] and Chinese style [7]. Other research discusses the detection of sentiment and emotion in poetry [8]. However, no one has conducted research to examine the comparison of two poetries in order to verify the originality of the created poetry. Therefore, the problem raised in this study is how the process of examining originality in poetry can be done automatically using the machine learning algorithm. Several machine learning approaches can be used to examine originality in poetry works. Most of these approaches utilize text similarity detection methods. The method that is quite good to use is the semantic-based, fuzzy-based, or citation-based method; the three methods are able to overcome several types of plagiarism such as copy, near copy, restructuring and paraphrasing. Based on other studies, semantic-based is the approach that has the best performance to detect text similarity [9]. In this research, the algorithm to be used is Siamese MaLSTM; the reason is because of the good accuracy ability in detecting sematic similarity of two texts [10].

1.1. **Contribution.** The purpose of this study is to propose the method to measure originality in the creation of poetry. There are two key contributions of this paper: first, presenting stages of building the machine learning model using dataset which is gathered and extracted from the collection of Indonesian poetries; second, designing a business process from the application of the poetry community. The application is intended to preserve original poems which uses a pre-trained model to check the semantic similarities in poetry pairs.

1.2. **Organization.** Section 2 describes the previous studies related to this study. Section 3 describes the stages of building the machine learning model starting from the dataset setup, preprocessing, the learning and testing process. Section 4 describes the process of designing the Indonesian poetry community application. Section 5 discusses conclusions and recommendations for further research.

2. **Originality in Poetry.** The discussion of originality is often related with detecting plagiarism in a written work. Plagiarism is born from the problem solving in various situations that consider the value of authenticity and originality of a literary work (problem-posing approach). The situations in question are for example such as forgery, copy paste, metaphor, language translating, paraphrasing writing and others. Detection of plagiarism is more practical to measure originality, because it departs from cases of forgery that have occurred before.

Plagiarism detection is generally divided into two approaches, namely intrinsic plagiarism detection and extrinsic plagiarism detection [11]. Detection of extrinsic plagiarism is done by comparing a work with a collection of literature from other authors. While the intrinsic approach is done by looking at the uniqueness of the writing style of an author. Each author is considered to have peculiarities or patterns in his writing; these

patterns which will then be extracted into features will be compared. Some of the features used are for example 1) stylometry features: average number of punctuation marks, n-gram frequency, word frequency, etc. [12], 2) rhetorical relation: elaboration, contrast, negation, etc. [13], and 3) metrical feature: meter of ductile, length of section, elision, etc. [14]. Several methods that can be used to determine plagiarism in the intrinsic approach are analysis outlier [12], relative frequency analysis [13], analysis information entropy [15], classification machine learning such as Support Vector Machine (SVM) [16], Random Trees, Naïve Bayes, and Logistic Regression [14]. Intrinsic detection method has several shortcomings; the most prominent is that the methods are not reliable enough for practical applications. The precision values resulting from this method are sufficient for raising suspicion but not for proving plagiarism [11].

Detection of extrinsic plagiarism uses an approach that is similar to comparative literature. In comparative literature, the differences and similarities that exist in two literary works are objects that will be compared [5]. Several studies on comparative literature take the steps by examine on each line in poetry, inspect the selection of words, and extract what meaning is presented [17-19]. The research in this paper uses steps similar to that conducted by Rahman et al. [20]. The procedure applied by Rahman is to 1) identify the themes of the two poems to be compared, 2) deconstruct the poetic text, so that each line of poetry can be aligned with another line of poetry, 3) decontextualize each line in the poem to examine the potential for adaptation in it, and 4) underline words, phrases or sentences that have contextual closeness [20]. The analysis process applied by Rahman is carried out manually; therefore, this study will take a different approach by automating the process with the help of the machine learning method.

3. Sentences Similarity Detection Model.

3.1. Dataset and preprocessing.

Dataset is an important component in building a machine learning model, and improperly preparing the dataset will produce the mistaken output. In this case, the data needed is a collection of sentences deconstructed from Indonesian poetries. The poems were collected from open sources and entered into the corpus database. The goal to be achieved is that the model is able to decide whether two sentences in poetry have semantic similarity or not; therefore, the dataset format required is that some sentences that have been tested have semantic similarity, as shown in Table 1. To meet these needs, some sentences in poetry will be paraphrased and labeled according to their semantic similarity. The complete arrangement of the dataset to be used is shown in Table 1.

Table 1. Description of the attributes in the dataset

Attribute name	Description
id	Primary id of poetry pair
poetry1	The sentence extracted from the first poetry
poetry2	The sentence extracted from the second poetry
is_duplicated	The target label, if it is 1 then both sentences have semantic similarity, 0 for the opposite.

Preprocessing is necessary to reduce the possibility of bias or the presence of outliers that may interfere with the performance of the learning process results. Some preprocessing techniques to be used are regular expression (regexes), tokenizing, stemming and stopword removal.

3.2. **Word embedding.** Word embeddings are used to extract the quality features of sentences. Each word in poetry will be converted to numeric representation as vector feature. In this research, we used the FastText as a pre-trained embedding model. FastText is an efficient word representation learning library provided by the Facebook research team. It contains 2 million common crawl words with 300 dimensions, providing 600 billion word-vectors. FastText is different from Google word embedding because it does the analysis in the sub word context by providing the n-gram character level representation of words [21]. FastText represents a word by the sum of the vector representation of its n-grams. It uses the same scoring function as skipgram model which objective is to maximize log-likelihood to the word context probability. With this method it allows the model to learn reliable representations for rare words [21]. Therefore, FastText is used for this work considering the enormous amount of diction in poetry.

3.3. **Siamese MaLSTM.** The preprocessed features are fed into the Siamese MaLSTM architecture for label prediction. The Manhattan Long Short Term Memory (MaLSTM) [22], as shown in Figure 1, is a sequence modeling technique which generates long term sequences by using its multiple layers inside. There are two networks, $LSTM_a$ and $LSTM_b$ in which each processes one of the sentences in a given pair. Each network consists of four components: o_t output gate, c_t cell memory block (current state determines which information will be fed to the next neuron), I_t input gate and the forget gate f_t. The input I_t feeds the LSTM layer in the form of real valued vectors. The hidden state representations h_t are updated sequentially between the gates. The update steps purely depend on memory cell block c_t. These four components decide which information is used and which information is omitted from the model for final prediction.

$$I_t = sigmoid(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$
$$f_t = sigmoid(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$
$$c_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3}$$
$$c_t = i_t \odot x_t + f_i \odot c_{t-1} \tag{4}$$
$$o_t = sigmoid(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$
$$h_t = o_t \odot tanh(c_t) \tag{6}$$
$$M_a = |x_1 - x_2| + |y_1 - y_2| \tag{7}$$

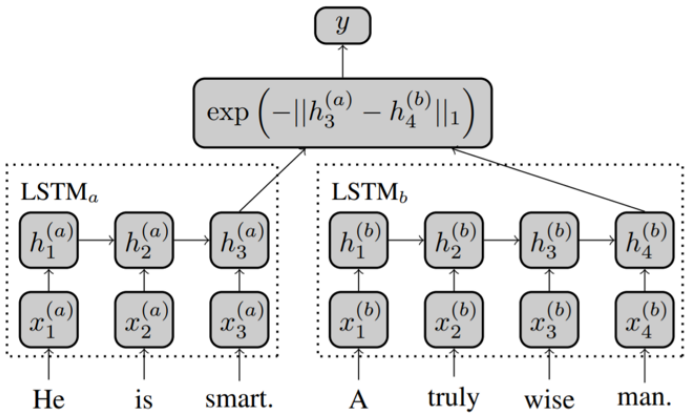

FIGURE 1. Architecture of the Siamese MaLSTM

Unlike other language modeling of Recurrent Neural Network (RNN) architectures which predict next words, the network computes the similarity between pairs of sequences. A main feature of the Siamese architecture is the shared weights across the subnetworks, which reduces not only the number of parameters but also the tendency of overfitting.

MaLSTM uses the Siamese structure along with the Manhattan distance; hence, it is named MaLSTM model. The similarity of two vectors that capture the underlying meaning of each question is calculated using Manhattan similarity function. In Equation (7), the x_1 and y_1 refer to the output of the first model $LSTM_a$ and x_2 and y_2 to the second model $LSTM_b$. The absolute difference between them shows the similarity measure between the two inputs given to the model.

3.4. **Evaluation model.** To evaluate whether a model that has been trained produces good performance, we use measurement of accuracy and loss value. Model evaluation was performed at each epoch both during learning and testing. As seen in the line chart in Figure 2, there are two different colored lines indicating the results of the model evaluation carried out during the training and testing processes. From the results of training, the best accuracy was obtained at a value of 97.2% and a loss value of 0.075, while the best accuracy at the time of model validation was at 91.6%. This number is considered good enough to ensure the use of the pre-trained model for the originality check system which will be discussed in the next section.

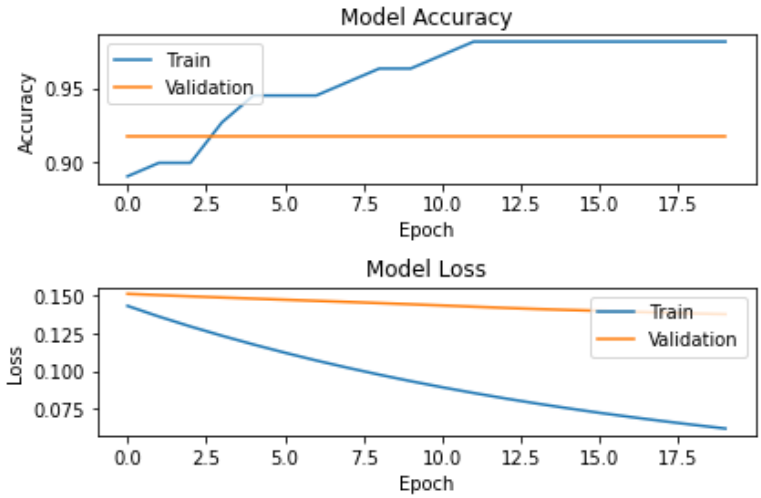

FIGURE 2. Evaluation of metric detection of duplicate poems using Siamese MaLSTM

4. Implementation.

4.1. **Poetry community service.** The system is a web-based application that can be accessed by anyone on the Internet. This application is intended for members or visitors who want to search the list of poems available in the poetry collection. Not only looking for poetry, visitors can register to be members of the poetry community, express appreciation for poetry, save and register poetry by their own creation, and participate in activities that are being organized both as poetry readers and as participants. In general, this application consists of three features, 1) account authentication, to identify authorizations from visitors, members or administrators of the application, 2) registration of new communities or following existing communities for the media to discuss, create, store and share poetry as well as to appreciate poetry, and 3) organization of poetry reading activities.

In poetry reading activities, the activity registrar is a member who already has an account in the application. Members are required to include data related to the poetry to be registered, such as the theme, poetry description, title and content of the poetry itself. The system will perform an originality check on the poetry that is registered; if it successfully passes the check, the poem will be included in the list of poems to be read. Poetry registration can be done as long as the poem list does not meet the predetermined

quota. The organization of this poetry reading activity is carried out regularly in a certain period of time. Poems that have been read in the activity will be printed in book format and can be downloaded by anyone who visits the application.

4.2. **Siamese MaLSTM similarity based detection service.** The second system is an application based on API service which is used to check the originality of poetry by comparing the poetry work registered with the collection of poetry in the database. The submitted data is in the form of an Indonesian poem text, while the server response is in the form of a Boolean value which indicates whether the poem submitted is original or not. The complete procedure of the process of checking the originality of poetry is described in Algorithm 1.

Algorithm 1. Poetry similarity detection

Require: New Poetry p_1
Poetry Collection $P_2 = \{p_{21}, p_{22}, \ldots, p_{2i}\}$
Ensure: Similarity score

```
similar ← false
for each p_2i ∈ P_2 do
    sentences ← 0
    for each s_j ∈ p_1 do
        for each s_k ∈ p_2i do
            DataSet ← (s_j, s_k)
            DataSet ← preprocessing (DataSet)
            XTest ← embeddings (DataSet)
            score ← predict (XTest)
            if score < threshold do
                sentences ← sentences + 1
    if sentences > 4 do
        similar ← true
    break
end for
```

This application employs several iteration processes. The first iteration is the comparison process between poems registered p_1 with the collection of poems in the database P_2. The second iteration is the comparison of each line of poetry s_j with another line of poetry s_k, resulting in two lines of sentences that will be the value entered in pre-trained model to check the value of similarity. The data entered will then be processed with several techniques text preprocessing, namely case folding, regular expression filter, stemming, stop words removal and tokenization. The text entered was broken into word lists and then transformed into numerical values using the FastText library [21]. The features of the two lines of poetry that have been prepared are then included in the pre-trained model based on Siamese MaLSTM. The prediction value resulting from the model will be compared with the value of threshold; if the value is higher than the value of threshold, then the poem line is sure to be similar to the other poem lines. If there are at least four lines of poetry that are semantically similar, then the system will stop the iteration process and send a response in the form of a Boolean value that indicates the poem is not original. Here is some example of the already running poetry community program in Figure 3.

5. **Discussion and Conclusions.** This paper proposes a method of examining originality in the creation of new poetry using the machine learning algorithm. The method used in this paper is by comparing the semantic meaning on each line between the poetry creations and the collection of poems from previous poets' works. Based on previous studies, the semantic-based method is a fairly good method used to overcome some types of

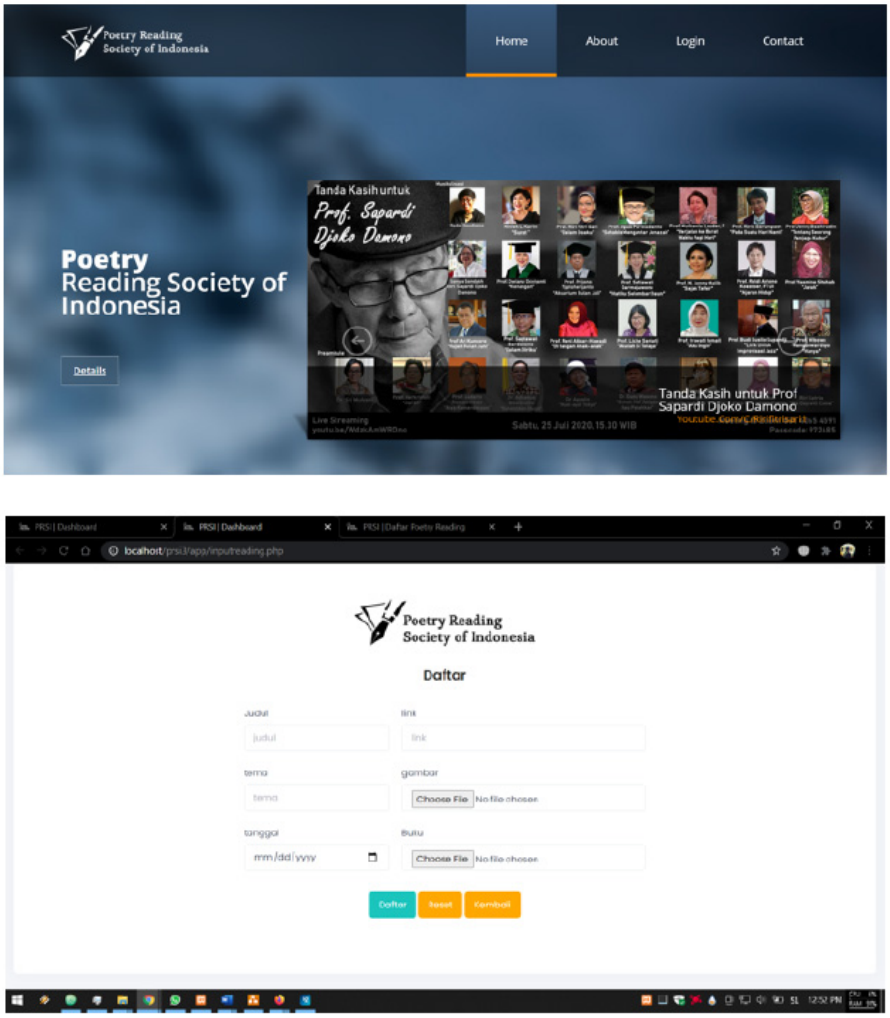

FIGURE 3. The interface of poetry community application

plagiarism such as copy, near copy, restructuring and paraphrasing. This study presents the stages of building the machine learning model to detect the similarity of the semantic context of two lines of poetry. The stages carried out to build the model are the collection of Indonesian poems, dataset preparation, text preprocessing, word embedding with FastText and the learning process with the Siamese MaLSTM algorithm. The results of the evaluation of the model that has been built provide a good accuracy value of 91.6%. The model that has been built is then used in the application of the poetry community which is intended to produce original Indonesian poems.

This research certainly cannot be separated from the various disadvantages, first the process to check the originality of poetry takes a long time, this happens because the process of examination is carried out on every line of poetry and carried out on all collections of poetry in the database. The process of sorting poems in the database must be carried out to reduce the number of poems that must be compared. Sorting can be done by first checking the similarity of poems such as detection of themes, genres, sentiments, language styles, and backgrounds. Then, the machine learning model built in this study also has a limited ability to only understand the semantic meaning of the line of poetry, the combination of language styles in poetry often gives other meanings, and the model should be able to understand the pragmatic context of the line of poetry. The model in this study also did not pay attention to the structure of lines and paragraphs of poetry. Some poets pay attention to this to bring out certain meanings. This research is also limited to comparing Indonesian poetries. In fact, the comparative literary process must be

carried out not only in an area. Plagiarism may be done on poetry with other languages, for example with regional languages or languages from other countries.

Acknowledgment. This work is supported by the Ministry of Research, Technology, and Higher Education (Kemristekdikti) of Indonesia under PDD Grant with contract number NKB-337/UN2.RST/HKP.05.00/2021.

REFERENCES

[1] H. J. Waluyo, *Theory and Appreciation of Poetry*, Erlangga, Jakarta, 1995.
[2] Z. K. Mabruri, Typographical study of Indonesian poetry, *Journal of Indonesian Language Research, Literature and Teaching*, vol.3, no.1, 2020.
[3] R. D. Pradopo, *Poetry Studies*, Gajah Mada University Press, Yogyakarta, 2010.
[4] A. R. Fadhilah, *Genius and Originality in Creation of Art*, Faculty of Cultural Sciences, Universitas Indonesia, Depok, 2007.
[5] S. Endraswara, *Comparative Literature Research Methodology*, Bukupop, Jakarta, 2011.
[6] G. Shao, Y. Kobayashi and J. Kishigami, Traditional Japanese Haiku generator using RNN language model, *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pp.263-264, 2018.
[7] K. Wang, J. Tian, R. Gao and C. Yao, The machine poetry generator imitating Du Fu's styles, *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp.261-265, 2018.
[8] S. Ahmad, M. Z. Asghar, F. M. Alotaibi and S. Khan, Classification of poetry text into the emotional states using deep learning technique, *IEEE Access*, vol.8, pp.73865-73878, 2020.
[9] J. Wang and Y. Dong, Measurement of text similarity: A survey, *Information*, vol.11, no.9, 421, 2020.
[10] D. Verma and S. Muralikrishna, Semantic similarity between short paragraphs using deep learning, *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp.1-5, 2020.
[11] T. Foltýnek, N. Meuschke and B. Gipp, Academic plagiarism detection: A systematic literature review, *ACM Computing Surveys (CSUR)*, vol.52, no.6, pp.1-42, 2019.
[12] S. P. Gunawan, L. D. Krisnawati and A. R. Chrismanto, Analysis of stylometric features and segmentation strategies in intrinsic plagiarism detection system, *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, vol.4, no.5, pp.988-997, 2020.
[13] M. Timofeeva, Comparative analysis of reasoning in Russian classic poetry, *Applied Sciences*, vol.11, no.18, 8665, 2021.
[14] B. Nagy, Metre as a stylometric feature in Latin hexameter poetry, *Digital Scholarship in the Humanities*, vol.36, no.4, pp.999-1012, 2021.
[15] O. Calin, Statistics and machine learning experiments in English and Romanian poetry, *Sci*, vol.2, no.4, 92, 2020.
[16] P. Plecháč, K. Bobenhausen and B. Hammerich, Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry, *Studia Metrica et Poetica*, vol.5, no.2, pp.29-54, 2018.
[17] M. R. Balbuena, Dibaxu: A comparative analysis of Clarisse Nicoïdski's and Juan Gelman's bilingual poetry, *Romance Studies*, vol.27, no.4, pp.283-297, 2009.
[18] C. Geofany and D. Triananda, Comparison of the poetry of Amir Hamzah's Doa and Sanusi Pane's Doa, *Asas: Literary Journal*, vol.7, no.3, 2018.
[19] L. M. Parsa, A comparative study of Wordsworth and Sepehris poetry in the light of Ibn Arabis Philosophy, *International Journal of Comparative Literature and Translation Studies*, vol.6, no.1, pp.10-17, 2018.
[20] F. F. Rahman and F. Rahman, Translation or intertextuality: A literature comparative analysis of "The Young Dead Soldiers Do Not Speak" by Archibald MacLeish and "Krawang Bekasi" by Chairil Anwar, *ELSYA: Journal of English Language Studies*, vol.1, no.3, pp.110-117, 2019.
[21] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, vol.5, pp.135-146, 2017.
[22] Z. Imtiaz, M. Umer, M. Ahmad, S. Ullah, G. S. Choi and A. Mehmood, Duplicate questions pair detection using Siamese MaLSTM, *IEEE Access*, vol.8, pp.21932-21942, 2020.