DOI: 10.15625/vap.2023.0078

AN IMPROVEMENT IN MEASURING THE SIMILARITY OF VIETNAMESE DOCUMENTS

Pham Thi Thu Thuy

Information Technology Faculty, Nha Trang University

thuthuy@ntu.edu.vn

ABSTRACT: *Text similarity seems to be very common in the digital age. So how to detect the copying of documents, this topic have also been studied by many people. However, detecting the same text remains a challenge. The problem is more complicated when the text is written in another language or when the text is changed some words. Therefore, the detection of similarity and forgery is still being researched and developed to help protect intellectual property rights and prevent piracy in the digital world. This paper proposes the application of a vector model to measure text similarity based on the method of using the Cosine measure in combination with the word order measure to increase the accuracy and help improve the comparison results between two Vietnamese documents. The performance results show that our proposed model gives more accurate results than applying the traditional Cosine measure, and gives a faster execution time than current methods.*

Keywords: *Text similarity, Plagiarism, Pointwise, Cosine measure, Jaccard, Word order measure.*

I. INTRODUCTION

Text similarity measures the degree of similarity between texts. It can be applied to compare an entire document or part of it with another document, or to compare a test document with a set of other documents. That is, text similarity is the ratio of similarity between text units to be compared.

To measure text similarity, studies often use string similarity [1-2], by using distance measures [3, 11] or semantic similarity measures [16] to understand the meaning of text. The statement of the text similarity calculation problem is as follows: consider a document d consisting of n sentences: d = s1, s2,…, sn. The goal of the problem is to find a value of the function S(si, sj) with S(0,1), and i, j = 1, 2,… n. The function S(si, sj) is called the measure of similarity between two documents si and sj. The higher the value, the more similar the meaning of the two texts.

For example, considering the two sentences "It's raining" and "It's sunny", intuitively it can be seen that the two sentences above have a high similarity. Semantic similarity is a confidence value that reflects the semantic relationship between two sentences. In practice, it is difficult to get an exact value because the semantics are only fully understood in a particular context.

In summary, similar documents are documents with relatively similar word frequencies, so it is possible to measure the similarity between documents or between a document and other documents in the data warehouse to the word frequency table. In text mining, there are many different metrics to calculate the similarity of documents, of which the most commonly used measure is the Cosine measure.

In Vietnam, the issue of plagiarism has also received much attention, with seminars and conferences focusing on plagiarism in schools. But still very little research on plagiarism has been published. Using IT to detect and combat plagiarism is considered one of the most effective ways to detect plagiarism. However, Vietnamese semantics is very complicated, with many factors, so it is difficult to apply a semantic text comparison system with absolute accuracy. Therefore, studies need to exploit Vietnamese data stores to solve the most optimal semantic problem, although it is still more limited than the English language.

Therefore, it is difficult to apply a complete model from other languages to Vietnamese. This paper will focus on the method of pointwise word separation to increase semantics while comparing Vietnamese documents, using the Cosine measure combined with the word order measurement to show the similarity between the two documents with high accuracy, both the word and the semantics of the text. This research will serve for assessing the level of plagiarism to detecting plagiarism in the Journal of Naval Science and Training.

The paper consists of the following parts: (I) Introduction to the article and the purpose of the research as well as the obtained results. (II) Related work, (III) Method of implementation and results and (IV) Conclusion.

II. RELATED WORK

Various studies have proposed different methods to determine if a piece of document text appears in another document. These methods include calculating statistics and semantic relationships between words in the text.

With statistical methods: Using statistical methods to evaluate the semantics of sentences, for example, using the Cosine measure based on the frequency of occurrence of words. However, these methods have fast processing speed and low cost, but do not guarantee high accuracy in semantic evaluation.

Methods based on semantic relations between words: can be approached through analyzing grammatical structures or using semantic networks, such as Wordnet corpus or Brown corpus. However, these methods have slower processing speed and more cost than the previous method, but the semantic accuracy is higher.

Because of the characteristics of Vietnamese, the problem of semantic similarity in Vietnamese texts is more complicated than in English. Most of the existing solutions focus on statistics-based metrics without fully exploiting the potential of natural language processing. Among them, the method using Wordnet corpus has been highly appreciated for its accuracy. However, WordNet only supports English and does not have a Vietnamese version yet. Instead, several other methods have been proposed such as hidden topic analysis or using Wikipedia's semantic network.

In general, in word processing problems, the vector space model is the most popular. This model represents the text as a feature vector of terms/words that appear in the entire corpus. Feature weights are usually calculated using the TF-IDF measure. Some notable topics on comparing and assessing text similarity include from [1] to [7]. However, those methods proposed the calculating semantic similarity between two documents based on word-to-word similarity and word-specific frequency that does not consider the relationship of word/phrase structure and the position of the sentence, leading to still many cases where the test program gives incorrect results. Therefore, texts with high semantic similarity are not necessarily the same.

To evaluate the text similarity to determine the degree of plagiarism, the separation of sentences and words is very important. There are many tools to support the separation of sentences and words in Vietnamese documents, such as vnTokenizer, JVPTextPro. The most prominent is the study of author Luu Tuan Anh [8] with the Pointwise method with 98% accuracy and faster processing speed than vnTokenizer.

Our paper inherits this Pointwise method to separate words and sentences before including text similarity calculation.

Some other studies use distance measurement method to calculate text similarity. Specifically, the studies of Aggarwal [9], Khatibsyarbini [10] and Leonardo [11]. Those studies presented various approaches for measuring similarity and distance, including distance metrics, similarity coefficients, and kernel-based methods. Although those approaches may be effective for certain types of text, they may not be ideal for detecting more sophisticated forms of plagiarism, such as text that has been paraphrased or reworded.

Several recent studies from [12] to [14] related to our paper on plagiarism detection. Their methods relied on analyzing sentence patterns rather than word usage, which they claim was more effective in identifying plagiarism. The authors used a support vector machine (SVM) classifier with a set of features such as sentence structure, conjunctions and comparisons. Those studies focussed heavily on the structure, with a sentences rather than specific words, which could limit its effectiveness in detecting plagiarism in certain contexts. For example, if a student rephrases a sentence with different words and retains the same grammatical structure, the method may not classify it as plagiarism, and it could go undetected.

Therefore, in order to check the similarity in both structure and semantics of the text, we propose to use the Pointwise word separation method and combining the Cosine measure [15] to compute the semantic similarity with the word order measure to increase accuracy and improve the results of comparing Vietnamese documents.

III. APPROACH TO THE PROBLEM OF COMPARING VIETNAMESE TEXTS

A. Document preprocessing

In most word processing systems, documents are treated as sets of keywords and represent them in vector form, such as Di = (di1,di2, …, din) with dik representing the weight/ frequency of the word tk in Di. The comparison of similarity between two documents Di and Dj , denoted Sim(Di,Dj), is calculated according to similarity calculation formulas, eg Cosine similarity. If this similarity reaches a sufficiently large threshold, it is said that they are semantically related.

However, when dealing with Vietnamese documents, the grammar method is mainly based on the word order in the sentence. Therefore, applying a text representation model based on word occurrences may not give the expected results. This comes from the fact that the vector representation of the text does not guarantee the semantic relationship between words, the position of words, phrases and sentences in the text. Furthermore, two vectors with different word orders can still have exactly the same degree of similarity. In order to inherit the advantages of known methods, the approach to the Vietnamese text comparison problem is to determine the similarity of documents based on the similarity of sentences and sentence order, the similarity of sentences is based on word similarity and word order in a sentence. This approach inherits the advantages of known methods.

Before being included in the processing model, the text needs to be preprocessed to improve the efficiency of the model and reduce the complexity of the implemented algorithm. This process helps to reduce the number of words in the text representation. Usually the preprocessing steps include the following steps:

- Split the text into individual sentences and words for later use for calculation purposes.
- Remove stopwords, remove non-alphanumeric characters.
- Save sentences and words into a suitable data structure.

During analysis, special cases such as numbers, brackets, and punctuation are often ignored, because they do not contribute significant meaning to the text (except in a few special cases, for example in historical data collection). However, for compound words like "state-of-the-art", omitting the "-" can completely change the meaning of the word and is not allowed.

B. Text vectorization

To calculate the feature value for the text, the previous study was performed by the TF-IDF method through the following steps:

- Step 1: Calculate the values: N is the number of documents in the set D, n is the number of words in the text d, Nt,d is the number of occurrences of the word t in the text d.

- Step 2: Calculate the weight TF=tf(t,d).

- Step 3: Calculate weight IDF=idf(t,D).

- Step 4: Calculate the weight di of the word t: TF-IDF= TF IDF.

Thus, we have represented the text as a vector model with Word/Text weights. The next stage is to apply similarity measures to calculate the degree of similarity between documents.

C. Calculate text similarity

* Algorithm to calculate text similarity as follows:

- Input: The given text set and the text to be checked.

- Output: the similarity between the test text and the given text in terms of Cosine measure.

- Handle:

+ Step 1: Preprocessing (separating sentences, separating single words, creating vocabulary lists...).

+ Step 2: Build a word/text weight matrix by calculating TF-IDF weights.

+ Step 3: Calculate the weight of the sentences in the text vector q.

• Calculate the similarity between two sentences in the test text with wik being the weight of the kth word in the i-th sentence according to the formula (1):

$$\mathrm{Sim}(S_q, S_j) = \sum_{k=1}^{n} w_{qk} w_{jk} \quad (1)$$

• Calculate the weight of the sentence in the test text according to the formula (2):

$$\mathrm{Score}(S_q) = \sum_{j=1, j\neq q}^{m} \mathrm{Sim}(S_q, S_j) \quad (2)$$

+ Step 4: Build the sentence weight vector for the test text q.

+ Step 5: Calculate the weight of the sentence in the given text set D.

• Calculate the similarity between two sentences in a document with wik being the weight of the kth word in the i-th sentence according to the formula (3):

$$\mathrm{Sim}(S_i, S_j) = \sum_{k=1}^{n} w_{ik} w_{jk} \quad (3)$$

• Calculate the weight of the sentence in any text according to the formula (4):

$$\text{Score}(S_i) = \sum_{j=1, j\neq i}^{m} \text{Sim}(S_i, S_j) \quad (4)$$

+ Step 6: Build the weight matrix of Sentence/Text for the entire text.

+ Step 7: Calculate the similarity between two documents according to Cosine's formula (5):

$$Sim(D_{ij}) = \frac{\sum_{k=1}^{t} w^i{}_k w^j{}_k}{\sqrt{\sum_{k=1}^{t}(w^i{}_k)^2 * \sum_{k=1}^{t}(w^j{}_k)^2}} \quad (5)$$

* Measure word order similarity:

In order to increase the accuracy in the experimental training process, we have added a step to calculate the similarity in the order of words in the text to ensure higher accuracy. The algorithm that combines the Cosine measure and the word order measure is as follows:

- Input: The given text set and the text to be checked.

- Output: The similarity between the test text and the given text using the Cosine measure combined with the word order measure.

Processing: Follow these steps:

Step 1: Preprocessing (separate single words, delete stop words, create vocabulary lists, etc.).

Step 2: Calculate the weight of index words Wtd = TF-ID, specifically as follows:

• Calculate the TF weight according to the formula (6):

$$tf(t,d) = \frac{N_{t,d}}{N_d} \quad (6)$$

//From the index t, the text d, Nt,d is the number of occurrences of t in d, Nd the total number of words in the text d.

• Calculate IDF weight according to formula (7):

$$\text{idf}(t,D) = 1 + \ln\left(\frac{N}{N_t}\right) \quad (7)$$

where D is a set of documents, N: Number of documents in set D, Nt: Number of documents with the word t appearing.

• Calculate the TF-IDF weight of index word t in text d according to formula (8):

$$Wt,d = tf(t,d) \text{ x } idf(t,D) \quad (8)$$

+ Step 3: Build the Word/Text weight matrix representing the corpus D and the test text vector q.

+ Step 4: Calculate the similarity between the test text vector q with the text k in the set D using the Cosine measure according to formula (9):

$$\text{Sim}(\mathbf{q},\mathbf{k}) = \frac{\sum_{i=1}^{n} q_i k_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} k_i^2}} \quad (9)$$

+ Step 5: Create a word order measurement between two documents, rq and rk.

+ Step 6: Calculate the word similarity between 2 vectors q and k according to formula (10):

$$\text{Sim}_R(\mathbf{q},\mathbf{k}) = 1 - \frac{\sqrt{\sum_{i=1}^{m}(r_{iq} - r_{ik})^2}}{\sqrt{\sum_{i=1}^{m}(r_{iq} + r_{ik})^2}} \quad (10)$$

+ Step 7: Calculate the composite similarity between 2 vectors q and k according to formula (11):

$$S(q,k) = \delta Sim(q,k) + (1 - \delta)SimR(q,k) \quad (11)$$

Repeat steps 4 to 7 to calculate the similarity between the test text vector q and the rest of the documents in the corpus D.

D. Experiment and the results

To conduct simulation installation, we use Windows 10, 64bit operating system and C# programming language. The program uses Microsoft's .NET technology, so it fully supports Vietnamese Unicode.

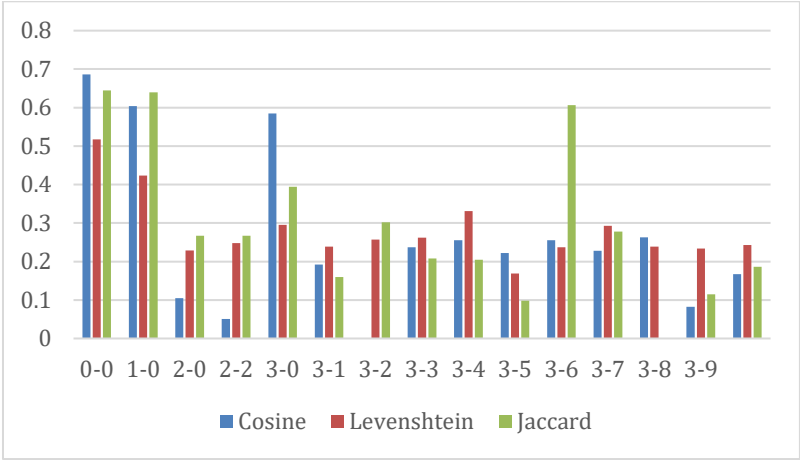

Figure 1. The similarity results among 3 algorithms Cosine, Levenshtein and Jaccard

The experimental installation program gets the database of journal articles of the Naval Science and Training Journal from 2010 to 2022, nearly 1500 articles with more than 3,600,000 words.

Using a text containing 600 words to check the similarity with 3 similarity measurement algorithms Cosine, Levenshtein and Jaccard gave the results as Figure 1.

Comparing the processing speed between 3 similarity measurement algorithms Cosine, Levenshtein and Jaccard gives the results as shown in Figure 2.

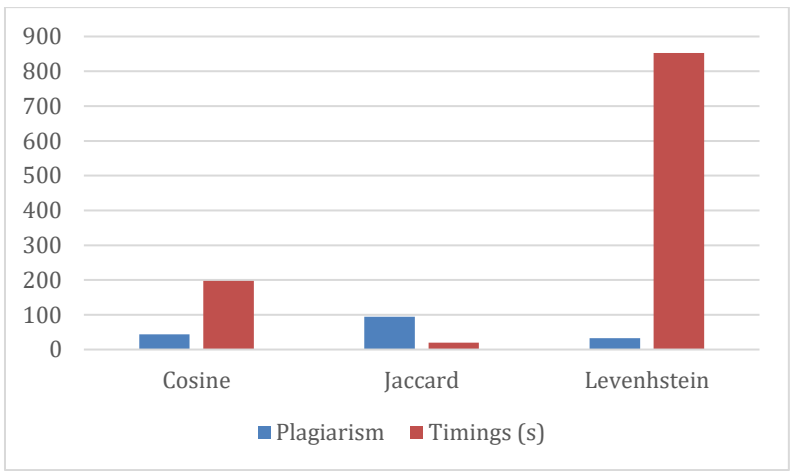

Figure 2. Processing speed and plagrarism between 3 algorithms Cosine, Levenshtein and Jaccard

The test results after improving the Cosine measure algorithm to calculate semantics combined with the word order measure are shown in Figure 3.

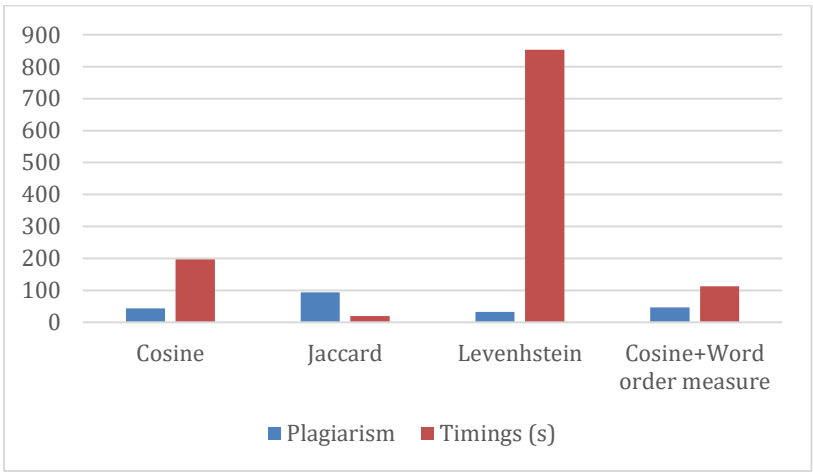

Figure 3. The results after improving the Cosine measure algorithm combined with the word order measure

General comment: Algorithms all compare results with a value of 100% when the two documents are completely similar and the result is 0% when the two documents are completely different. In the remaining cases, the result of the algorithm compared with the estimated value has a relative difference, namely:

- The method of improving the Cosine measurement speed combined with the word order measurement: the algorithm gives faster results than the Cosine method and the similarity between documents also increases the accuracy.

- Cosine method: The algorithm uses the word frequency method, because the test data are selected according to the set standards with absolute accuracy of the estimated value, the results are completely accurate. Therefore, the method of comparing text similarity based on Cosine measure is effective in detecting text duplication when comparing lexicographically.

- Jaccard method: based only on the number of common words on the total number of words. The advantage is giving results very quickly. The disadvantage is also because it is only based on the number of common words, so the results are not highly accurate.

- The Levenshtein method has the advantage of measuring the distance between two strings based on characters, so in case two documents are completely different in terms of words, they can also be similar in characters and spaces. The similarity measure of two completely different documents can also be greater than 0%. The processing speed of the Levenshtein method is very slow, and the accuracy is not high. So this method using text similarity measure will not work.

The improvement of this study is the application of the Cosine measure and the combined methods of measuring the word order to produce high accuracy results and fast processing speed. With quite high accuracy, I think this model can be applied to check plagiarism at agencies with available databases such as schools, newspapers and magazines.

- Improvements have greatly increased the processing speed compared to the simple cosine method and also increased accuracy when detecting plagiarism.

- Basically, the system has met the set requirements and the results of detecting the same text fragments have high accuracy.

- Regarding the data warehouse, up to the present time, in the data warehouse there are about 1500 articles of Naval Science and Training journals with about 3,600,000 words. In the future, it is possible to regularly update the database.

Thus, with the research results, proposed solutions and practical applications achieved in the research content, it has met the important requirements for assessing the similarity of Vietnamese texts to each others for text similarity detection system.

IV. CONCLUSION

This study used methods to improve the detection of plagiarism in Vietnamese documents. Using Pointwise method to separate Vietnamese words to increase semantics for Vietnamese comparative text. Using the Cosine measure of the semantics of the word combined with the measure of the word order to create the measure of accuracy closest to the semantics while still ensuring the speed of processing Vietnamese documents.

Cosine measure is one of the methods to measure the similarity between two documents. Compared with other methods such as Jaccard method or Levenshtein method, the Cosine measure has several advantages as follows:

- Cosine measure solves the problem of text indexing well.

- It can handle the wide and varied vocabulary of Vietnamese text.

- Cosine measure can also calculate the similarity between documents based on the occurrence of common words, helping to detect similar documents quickly and accurately.

- It is applicable to long documents and complex content.

- The Cosine measure is also very easy to use and understand, so it is suitable for beginners or inexperienced people.

- Cosine measure can also be applied to find similar documents in a large data set, helping to optimize search speed and reduce computer resources.

Compared with other methods, the Cosine measure has advantages such as simple calculation, high accuracy and good processing ability with text with complex content. In summary, the Cosine measure is a useful and powerful method to measure the similarity between documents, and it can be applied to many different fields such as plagiarism detection, similarity text search, etc.

Currently, one of the limitations of the Cosine measure is its ability to distinguish words with similar meanings. Therefore, the results of using the Cosine measure may not be accurate in some cases. In addition, the Cosine measure

can also be affected by the size of the data set and the input parameter values. Therefore, further research is needed to further improve the accuracy of the Cosine measure.

The author proposes several possibilities for further development and research to address these issues:

- Use machine learning algorithms to improve accuracy in plagiarism detection.

- Find ways to use Vietnamese text dataset to train and improve plagiarism prediction models.

Use natural language processing techniques to improve text analysis and plagiarism detection.

- Develop independent models to detect plagiarism in different types of texts, such as news texts, textbook texts, story texts, etc.

Research and development in this area can help optimize the power and accuracy of plagiarism detection models, making text analysis and processing easier and more efficient.

REFERENCES

[1] Do Thi Thanh Nga, "Calculation of text semantic similarity based on word-to-word similarity," Master thesis, Vietnam National University of Technology, Hanoi (2010).

[2] D. T. Long, "Research on the measure of text similarity in Vietnamese and its application to support the assessment of electronic copying," Institute-level scientific research project, Hanoi Open University (2014).

[3] Ho Phan Hieu, Vo Trung Hung, Nguyen Thi Ngoc Anh, "Some methods for calculating text similarity based on vector models," *UD Journal of Science and Technology*, no. 11(120), K.M. 112-117 (2017).

[4] Le Quy Tai, "Research on Vietnamese language processing methods, applications for text summarization," Master thesis, Vietnam National University, Hanoi (2011).

[5] Nguyen Trung Kien, "Vietnamese segmentation using CRFs model," Master thesis, Vietnam National University, Hanoi (2006).

[6] Nguyen Dinh Manh, "Research on methods to calculate similarity of Vietnamese legal documents," University of Technology (2020).

[7] Pham Thi Hai Van, Pham Huu Loi, "Research on methods to compare text similarity by Cosine measure," *Vietnam Science and Technology Journal*, no. 12 (1), pp. 6-11 (2017).

[8] Luu Tuan Anh, "Applying Pointwise method to the word separation problem for Vietnamese," Nagaoka University, Japan (2012).

[9] C.C. Aggarwal, *Similarity and Distances, in Data Mining*, Ed: Springer, Cham, pp. 63-91 (2015).

[10] M. Khatibsyarbini, et al., "A hybrid weight-based and string distances using particle swarm optimization for prioritizing test cases," *Journal of Theoretical and Applied Information Technology*, vol. 95, pp. 2723-2732 (2017).

[11] B. Leonardo, and Hansun, S., "Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, pp. 462-471 (2017).

[12] Daisuke Sakamoto, Kazuhiko Tsuda, "A Detection Method for Plagiarism Reports of Students," *Procedia Computer Science*, volume 159, Pages 1329-1338, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.09.303 (2019).

[13] K.M. Jambi, Khan, I.H.; Siddiqui, M.A., "Evaluation of Different Plagiarism Detection Methods: A Fuzzy MCDM Perspective," *Appl. Sci.*, 12, 4580. https://doi.org/10.3390/app12094580 (2022).

[14] Hamed Arabi, Mehdi Akbari, "Improving plagiarism detection in text document using hybrid weighted similarity," *Expert Systems with Applications*, volume 207, 118034, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.118034 (2022).

[15] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining: Concepts and Techniques* (section 2.4.7 Cosine Similarity), Elsevier, 2012.

[16] Fei Lan, "Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method," *Advances in Multimedia*, vol. 2022, Article ID 7923262, 11 pages, 2022. https://doi.org/10.1155/2022/7923262

CẢI TIẾN TRONG ĐO ĐỘ TƯƠNG ĐỒNG CỦA VĂN BẢN TIẾNG VIỆT

Phạm Thị Thu Thúy

TÓM TẮT: *Văn bản trông giống nhau và vi phạm bản quyền dường như rất phổ biến trong thời đại kỹ thuật số. Vậy làm thế nào để phát hiện việc sao chép tài liệu, chủ đề này cũng đã được nhiều người nghiên cứu. Tuy nhiên, việc phát hiện văn bản giống nhau vẫn là một thách thức. Vấn đề càng phức tạp hơn khi văn bản được viết bằng ngôn ngữ khác hoặc khi văn bản bị thay đổi một số từ. Vì vậy, việc phát hiện sự giống nhau vẫn đang được nghiên cứu và phát triển nhằm giúp bảo vệ quyền sở hữu trí tuệ và ngăn chặn vi phạm bản quyền trong thế giới kỹ thuật số. Bài báo này đề xuất ứng dụng mô hình vector để đo độ tương tự của văn bản dựa trên phương pháp sử dụng độ đo Cosine kết hợp với độ đo trật tự từ để tăng độ chính xác và giúp cải thiện kết quả so sánh giữa hai văn bản tiếng Việt. Kết quả thực hiện cho thấy mô hình đề xuất của chúng tôi cho kết quả chính xác hơn so với việc áp dụng độ đo Cosine truyền thống và cho thời gian thực hiện nhanh hơn các phương pháp hiện tại.*

Từ khóa: *Text similarity, Plagiarism, Pointwise, Cosine measure, Jaccard, Word order measure.*