ICIC Express Letters ICIC International ©2022 ISSN 1881-803X
Volume **16**, Number **6**, June **2022** pp. **639–646**

AUTOMATICALLY IDENTIFYING OF PLAGIARIZED SUBJECTIVE ANSWERS FOR THAI USING TEXT-BASED SIMILARITY ANALYSIS METHOD

Chumsak Sibunraung[1], Jantima Polpinij[1,*], Thananchai Khamket[1]
Jatuphum Juanchaiyaphum[1] and Bancha Luaphol[2]

[1]Intellect Laboratory
Faculty of Informatics
Mahasarakham University
Kantaravichai District, Mahasarakham 44150, Thailand
{ chumsak.s; thananchai.k; jatuphum.j }@msu.ac.th
*Corresponding author: jantima.p@msu.ac.th

[2]Department of Digital Technology
Faculty of Administrative Science
Kalasin University
Namon District, Kalasin 46230, Thailand
bancha.lu@ksu.ac.th

Received October 2021; accepted December 2021

Abstract. *In the context of education, many researchers design and develop methods or tools to identify plagiarism and maintain study quality. Text-based plagiarism often occurs in the academic domain, including online subjective examinations. Each one of the numerous proposed techniques has limitations in plagiarism detection. Here, a method is presented to identify plagiarized subjective answers in Thai when the subjective examination is performed online using natural language processing techniques (e.g., POS tagging) and cosine similarity analysis. The proposed method is called "similarity analysis of linguistic syntax and words used". Results gave scores of true positive rate (TPR) as 0.81. Furthermore, the proposed method was compared with the baseline and when compared to the baseline, our proposed method improved the average TPR by 7.69%. This may demonstrate the success of our proposed method in identifying plagiarized subjective answers.*

Keywords: Subjective examination, Plagiarized subjective answers, Thai, Syntax analysis, POS tagging, Similarity analysis

1. **Introduction.** The COVID-19 pandemic has greatly impacted the education system. With schools shut, students have had to adapt to online learning as a teaching format. Examinations performed through the online system [1-3]. In Thailand, a subjective examination is the usual method for an educational evaluation to assess student knowledge, skills, attitudes and concepts [4]. Online subjective examination techniques increase the chances of corruption, whereby students submit plagiarized answers and attain good grades without achieving the desired learning outcomes [5,6]. Online plagiarism corrupts results and represents a severe threat to the educational process.

In general, plagiarism occurs when someone uses words, ideas or work products attributable to another identifiable person or source, without referencing the source from which it was obtained, in a situation in which there is a legitimate expectation of original authorship to obtain some benefit, credit or gain which need not be monetary [7,8]. plagiarism is one of significant problems in educational activity and educational challenges

DOI: 10.24507/icicel.16.06.639

[7,8]. The task of plagiarism detection is to identify content similarity matches and this task is an ongoing and long-standing study [9-11].

Plagiarism is common in Thailand and reported in 80% of graduate theses [2,3]. However, no definitive study has concentrated on investigating plagiarized subjective answers in the Thai language to identify cheating and uphold academic integrity. Plagiarism for subjective answers has become a very real threat for Thai Universities. First and second year students have to enroll for general education (GE), with numbers exceeding 1,000 for many subjects. These large student numbers render identifying plagiarized subjective answers by evaluators difficult as the task is very time-consuming. To achieve success, the evaluators require a tool or system for automatic detection of subjective answer similarity that can identify students who present answers or ideas attributable to others when completing subjective online examinations. The system must be able to process and analyze data quickly before grade submission deadlines, especially for popular subjects chosen by many students. Unfortunately, a system for automatic detection of subjective answer similarity for Thai students has not yet been adopted.

Previous studies related to this problem addressed automated scoring techniques for short answers to subjective tests in the Thai language [12-14], using prominent keyword-based analysis together with similarity techniques [2,3,8]. These methods returned good results but were still insufficient [9-11] because in the Thai language sentence syntax is not flexible for plagiarized subjective answer identification [13]. To improve the accuracy of identifying the correct answer and students' answers for automatic scoring of short subjective answers, the position or role of words must also be considered to reduce false negatives and positives [9]. Furthermore, linguistic structural analysis to identify plagiarized content may also help to increase the accuracy of text similarity identification [15-18]. Combining these methods can improve similarity identification accuracy for automatic scoring for short answers of subjective tests in Thai, with satisfactory accuracy of identifying plagiarized subjective answers. Therefore, a method for identifying plagiarized subjective answers was presented based on keyword analysis together with similarity and syntactic structural analysis. This proposed method can be used to reduce false negatives and positives by automatically identifying plagiarized subjective answers. Syntactic sentence construction considered in this study included verb phrases and noun phrases. We also compared our proposed method with the baseline method proposed by Saipech and Seresangtakul [4]. Their method automatically analyzed Thai text using word segmentation and stop word elimination by cosine similarity. Their research objective was similar to our study.

2. Related Work.

The most commonly found piracy issues in educational activities are textual-based plagiarism and citation-based plagiarism [9,19,20]. Textual-based plagiarism involves copying and pasting text from other documents [9,19]. However, if the text is paraphrased or translated by a human, it becomes difficult to automatically detect plagiarism. Research into the development of automated tools to detect textual-based plagiarism is ongoing, with some examples as follows.

Jadalla and Elnagar [21] presented a novel plagiarism detection system for Arabic text-based documents utilizing a structure based on a search engine to reduce the cost of pairwise similarity. They used the winnowing n-gram with fingerprinting algorithm to reduce the index size. Results showed an improvement in search time. The detection process was accurate and robust, achieving recall at 0.94 and precision at 0.99. In 2014, Jiffriya et al. [22] designed an effective plagiarism detection tool to evaluate text-based electronic assignments. Their proposed tool identified suitable intra-corpal plagiarism detection for text-based assignments by comparing unigram, bigram and trigram vector space models with cosine and Jaccard similarity measures. Results suggested that a trigram vector space model with cosine similarity was better than a trigram vector space model with

Jaccard similarity measurements. Later, Elkhidir et al. [23] presented a tool called free text plagiarism detection software (FTPDS). The main mechanism of this tool concerned a document's fingerprint algorithm, which detected the likelihood that documents were plagiarized from each other. Their proposed system detected plagiarism between two given documents, a given document and a group of local documents, and between a given document and online available documents.

In 2016, Bagai et al. [24] introduced a computer-based plagiarism detection technique by combining the functionality of substring matching and keyword similarity to give more accurate results. They also improved the efficiency of the clustering algorithm as the main mechanism of the system by ranking the documents using longest common subsequence (LCS) matching. Their proposed system returned satisfactory results. In 2017, Eisa et al. [25] studied how to detect plagiarism related to scientific figures using textual-reference representation for figure plagiarism detection techniques. Their method improved textual feature extraction and similarity computation methods. Improved features allowed extraction of textual references such as captions and descriptive texts, while the similarity detection method classified a given figure into either plagiarized or non-plagiarized classes using a certain threshold value. After testing by recall and precision, their proposed method returned recall and precision at 0.67 and 0.78, respectively. However, this method was still unable to detect some figures that were plagiarized because the related textual references of these figures had been changed or modified by paraphrasing or summarization techniques. In 2018, Sinaga and Hansun [26] implemented Rabin-Karp and Confix-Stripping algorithms to assess the occurrence of plagiarism in scientific papers written in Bahasa Indonesia. The existing software was designed for text written in English and not suitable for text written in Bahasa Indonesia.

As detailed above, many text-based plagiarism detection systems or tools have been proposed and research is ongoing. In Thailand, a high rate of 80% plagiarism has been reported in graduate theses [2,3], and there are no studies that directly address plagiarism detection for subjective tests. The most relevant studies concerning this problem investigated automated scoring for short answers of subjective tests in Thai [12-14]. These studies used keyword-based analyses, along with similarity for automatic scoring of answers to subjective tests in Thai. This method lacked accuracy because the position or role of words was not considered, thereby increasing the rates of false negatives and positives. Our method used linguistic structural analysis in conjunction with keyword-based analysis and similarity techniques to reduce the rates of false negatives and positives.

3. **Method of Identifying Plagiarized Subjective Answers in Thai.** This section explains how to identify plagiarized subjective answers in Thai using the similarity analysis method of linguistic syntax and words used. The overall picture of the proposed method can be shown as Figure 1. It consists of five processing stages, with each stage explained as follows.

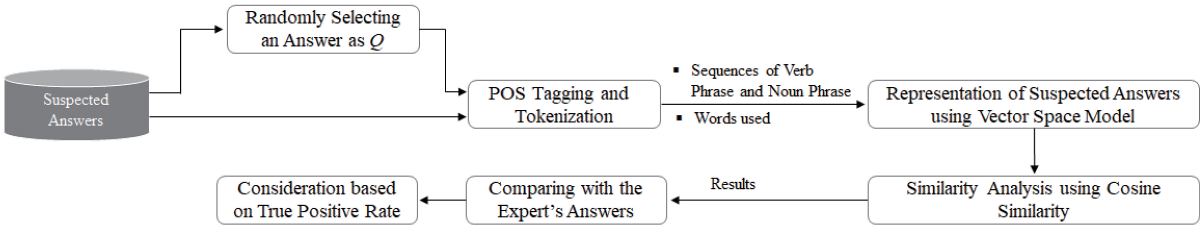

Randomly Selecting an Answer as Q
Suspected Answers
POS Tagging and Tokenization
▪ Sequences of Verb Phrase and Noun Phrase
▪ Words used
Representation of Suspected Answers using Vector Space Model
Similarity Analysis using Cosine Similarity
Results
Comparing with the Expert's Answers
Consideration based on True Positive Rate

Figure 1. An overview of the proposed method

Stage 1: Generating Linguistic Syntax of Thai Noun-phrase and Verb-phrase Patterns

This stage subjects the suspected subjective answers to part of speech (POS) tagging to obtain a sequence of POS tags for each sentence [27]. The ORCHID POS tag set was used for the Thai language [27]. For the question, "*Who was the first king of the Ayutthaya Kingdom*?", there are two likely suspected answers as "*กษัตริย์พระองค์แรกแห่งอาณาจักรอยุธยาคือพระเจ้าอู่ทอง* (The first king of the Ayutthaya Kingdom was King Uthong)" and "*กษัตริย์พระองค์แรกของอาณาจักรอยุธยาคือพระเจ้าอู่ทอง* (The first king of the Ayutthaya Kingdom was King Uthong)". Examples of generating and leveraging the POS tag sequence of these Thai subjective answers are illustrated in Figure 2. These answers seem to be similar but in reality they are different. There is one word in each suspected answer that makes them different. The first answer uses "*แห่ง* (*of*)" in the sentence, while the second uses "*ของ* (*of*)". In English, these words are similar but they have a different meaning in Thai. Using the examples shown in Figure 2, we leveraged three structures of the noun group and one structure of the verb group. The three structures of the noun group were leveraged as "NCMN|NCMN|DONM|RPRE|NCMN", "NCMN|NCMN|DONM|CNIT|NCMN" and "NCMN", while the structure of the verb group was detected as "JSBR". Then, these structures were represented in the vector space model (VSM) format (Figure 3). In this format, a structure present in the suspected answer is represented by 1; otherwise it is represented by 0, meaning absent.

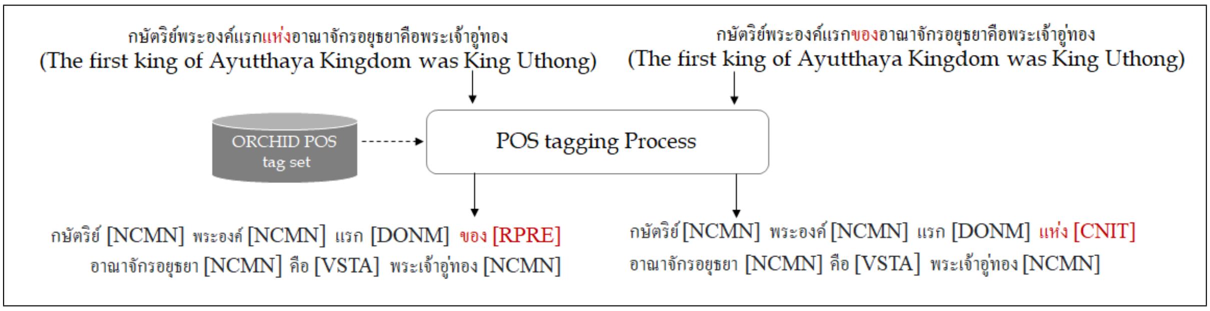
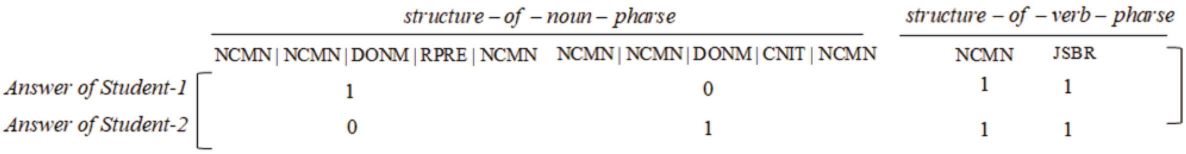

กษัตริย์พระองค์แรกแห่งอาณาจักรอยุธยาคือพระเจ้าอู่ทอง
(The first king of Ayutthaya Kingdom was King Uthong)
กษัตริย์พระองค์แรกของอาณาจักรอยุธยาคือพระเจ้าอู่ทอง
(The first king of Ayutthaya Kingdom was King Uthong)
ORCHID POS tag set
POS tagging Process
กษัตริย์ [NCMN] พระองค์ [NCMN] แรก [DONM] ของ [RPRE] อาณาจักรอยุธยา [NCMN] คือ [VSTA] พระเจ้าอู่ทอง [NCMN]
กษัตริย์ [NCMN] พระองค์ [NCMN] แรก [DONM] แห่ง [CNIT] อาณาจักรอยุธยา [NCMN] คือ [VSTA] พระเจ้าอู่ทอง [NCMN]

FIGURE 2. An example of POS tag sequence of a Thai subjective answer



	structure – of – noun – pharse		structure – of – verb – pharse	
	NCMN | NCMN | DONM | RPRE | NCMN	NCMN | NCMN | DONM | CNIT | NCMN	NCMN	JSBR
Answer of Student-1	1	0	1	1
Answer of Student-2	0	1	1	1

FIGURE 3. An example of POS tag sequence found in the suspected answers and represented them in the VSM format

Stage 2: Tokenizing the Suspected Subjective Answers

This stage splits the subjective answers into smaller units such as individual words. For Thai text tokenization, a dictionary-based word segmentation with longest matching algorithm [28] is applied to separating the text into Thai words. The Thai dictionary consists of 120,800 words. An example of Thai tokenization is presented in Figure 4.

Stage 3: Representation of Suspected Subjective Answers

After performing stages 1 and 2, the POS tag sequences of VP and NP and words used are represented in the format of a vector space model (VSM). In general, VSMs are used to represent documents and queries as vectors of weights, where each weight is a measure of the importance of an index term in a document or a query, respectively. The term weighting scheme used in this study is *term frequency* (*tf*) because its computational

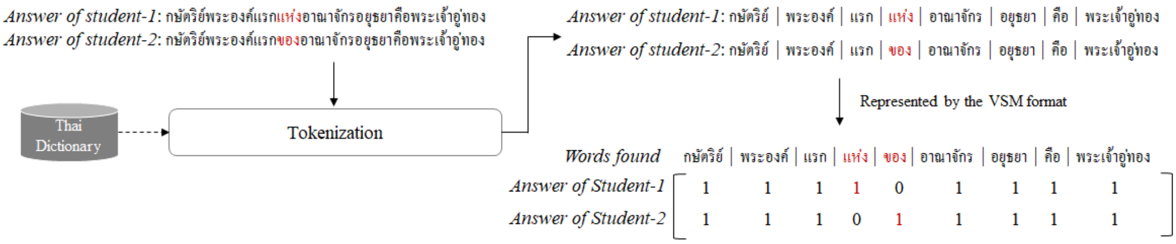

Answer of student-1: กษัตริย์พระองค์แรกแห่งอาณาจักรอยุธยาคือพระเจ้าอู่ทอง
Answer of student-2: กษัตริย์พระองค์แรกของอาณาจักรอยุธยาคือพระเจ้าอู่ทอง
Thai Dictionary
Tokenization
Answer of student-1: กษัตริย์ | พระองค์ | แรก | แห่ง | อาณาจักร | อยุธยา | คือ | พระเจ้าอู่ทอง
Answer of student-2: กษัตริย์ | พระองค์ | แรก | ของ | อาณาจักร | อยุธยา | คือ | พระเจ้าอู่ทอง
Represented by the VSM format
Words found กษัตริย์ | พระองค์ | แรก | แห่ง | ของ | อาณาจักร | อยุธยา | คือ | พระเจ้าอู่ทอง
Answer of Student-1 1 1 1 1 0 1 1 1 1
Answer of Student-2 1 1 1 0 1 1 1 1 1

FIGURE 4. An example of Thai text tokenization and representation

time is very low [29]. The equation of *tf* is

$$tf_{w,d} = \log\left(1 + tf_{w,d}\right) \tag{1}$$

The value of $tf_{w,d}$ is the raw count of w in a suspected subjective answer d. In this study, w can be a POS tag sequence of a VP, a POS tag sequence of an NP, or a word in the suspected answers.

In this study, the structure used to represent the suspected subjective answers is called *vector space model* (*VSM*) [29]. Then, a suspected answer is randomly chosen as a query, denoted as Q. Other subjective answers are considered as documents, denoted as $D = \{d_1, d_2, \ldots, d_n\}$. A representative example of the POS tag sequences of VP, the POS tag sequences of NP and words is illustrated as Figure 5.

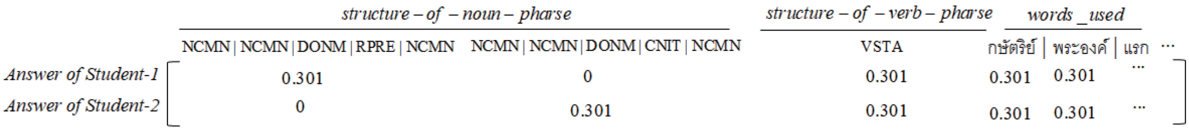

structure – of – noun – pharse
structure – of – verb – pharse
words _used
NCMN | NCMN | DONM | RPRE | NCMN
NCMN | NCMN | DONM | CNIT | NCMN
VSTA
กษัตริย์ | พระองค์ | แรก ···
Answer of Student-1 0.301 0 0.301 0.301 0.301 ···
Answer of Student-2 0 0.301 0.301 0.301 0.301 ···

FIGURE 5. An example representation of suspected answers

Stage 4: Analyzing the Similarity of Suspected Subjective Answers

After representing the verb phrase structure, noun phrase structure and words used as Q and D in the format of VSM, *cosine similarity* (*CS*) was applied to estimating the similarity of Q and D, where *CS* measures the similarity between two non-zero vectors using a Euclidean dot product space to measure the cosine of the angle between them [29]. The *CS* equation is presented as Equation (2).

$$sim(Q, D) = \frac{\sum_{i=1}^{N} Q_i \times D_i}{\sqrt{\sum_{i=1}^{n} Q_i^2}\sqrt{\sum_{i=1}^{n} D_i^2}} \tag{2}$$

The outcome of cosine similarity is bounded by $[0, 1]$. If the similarity score is close to 1, this means close similarity between Q and D. However, if the similarity score is 1, this means that Q and D are the same. It is noted that the CS similarity was chosen for this study because the equation of this similarity measure is not complicated. Therefore, its processing time is low.

Stage 5: Analyzing the Plagiarism Rate

A proposed solution for similarity rate consideration is presented as follows. Let N be the number of similarity scores. In this study, these are similarity scores of verb phrase structure noun phrase structure and words used in the suspected answers. S is a set of the *similarity scores* (*sim*) of verb phrase structure, noun phrase structure and words used, denoted as $\{sim_1, sim_2, sim_3\}$. The similarity rating between suspected subjective answers can be calculated by the following equation.

$$Score = \frac{\sum_{i=1}^{N} sim_i}{N} \times 100 \tag{3}$$

Suppose the similarity rates of verb phrase structure, noun phrase structure and words used found in the suspected answers are 1, 0.33 and 0.66, respectively. Then, the similarity rate of this example is

$$Score = \left(\frac{1.00 + 0.33 + 0.66}{3}\right) \times 100 \approx 66.33\%$$

Thus, the suspected subjective answer has a similarity rate of 66.33%. However, if the plagiarism rate returns as 100%, those suspected answers might be considered as plagiarized answer.

4. **Results and Discussion.** The dataset utilized in our experiment comprised information collected from 100 students enrolled in the Fundamentals of Computer Science, Computer Organization and Architecture, and Thai History courses, with each course comprising five questions in the exam. The students were asked to provide subjective answers. Results for automatically identifying of plagiarized subjective answers were compared with the expert's answers. Experimental results were evaluated for true positive rate (TPR), false negative (FN) and false positive (FP). The TPR measures the proportion of positives that are correctly identified. FN may appear to be negative when it is not, while FP determines something to be true, when in readily it is false [29]. However, we also compared our method to the baseline method proposed by Saipech and Seresangtakul [4]. Their method analyzed automatic Thai subjective examination using cosine similarity, and commenced by segmenting subjective answers into words using the longest matching algorithm. These words were then represented into vectors using *tf-idf*. The highlight of their study considered word synonyms, while they used cosine similarity to measure the plagiarism of the answers. The results can be shown in Table 1.

Table 1. The experimental results by considering TPR, FN, FP and computational time

Courses of subjective exam	**Proposed method**				**Baseline method**			
	TPR	**FN**	**FP**	**Time (ms.)**	**TPR**	**FN**	**FP**	**Time (ms.)**
Fundamental of Computer Science	0.84	0.14	0.15	0.051	0.80	0.17	0.17	0.050
Computer Organization and Architecture	0.88	0.11	0.12	0.050	0.88	0.16	0.14	0.050
Thai History	0.81	0.16	0.15	0.060	0.73	0.18	0.18	0.063
Average	**0.84**	**0.14**	**0.14**	**0.054**	**0.80**	**0.18**	**0.16**	**0.054**

Results in Table 1 showed that our proposed method was satisfactory in terms of average TPR, FN and FP scores. Consideration of the linguistic structures of nouns and verbs used in sentences improved the accuracy of identifying plagiarized subjective answers and also increased the accuracy of similarity analysis compared to the baseline. The baseline returned lower results than our proposed method but still performed satisfactorily by considering synonyms that reduced language ambiguity during linguistic processing. However, two possible reasons for the error rate of our proposed method were discussed. Firstly, Thai language is written continuously without punctuation. This makes it difficult to segment the correct words and align them to their tags, and hampers the extraction sequences of POS tags. If the system returns wrong words, erroneous tags are also applied to these words. One solution to address this issue is to add more Thai words to the dictionary because dictionary-based word segmentation based on the longest matching algorithm is the main mechanism of tokenization. Therefore, if the Thai dictionary contained insufficient words, poor results might be returned, and this increased the chances of mistakes when analyzing the VP and NP similarity of suspected subjective answers. Secondly, named entity detection may be required to identify proper names in the suspected subjective answers. Without using named entity detection, some proper names could be separated into general words. If this issue is handled, errors in the similarity analysis can also be induced in the suspected subjective answers.

5. **Conclusion.** This study presents a method of text-based plagiarism detection that can be applied to identifying plagiarized subjective answers in Thai for examinations performed online. The main mechanisms of the proposed method utilize natural language processing techniques (e.g., POS tagging) and cosine similarity analysis. After comparing the results of the plagiarized subjective answers with the expert's answers, the results were satisfactory both in terms of TPR and computational time. However, the efficiency of the proposed system needs to be further improved in terms of TPR. Firstly, the Thai dictionary used in this study can be improved by adding more Thai words. Secondly, we plan to integrate named entity techniques to identify proper names in future studies. Name entities should be identified, otherwise mistakes will occur in the similarity analysis of suspected subjective answers. Without using the named entity identification method, those name entities found in suspected subjective answers can be separated into many words. Consequently, this can make the meaning of those words incorrect. In addition, when compared to the baseline, our proposed method improved the average TPR by 7.69%.

Acknowledgment. This work is financially supported by Mahasarakham University Grant Year 2018.

REFERENCES

[1] A. Aristovnik, D. Keržič, D. Ravšelj, N. Tomaževič and L. Umek, Impacts of the COVID-19 pandemic on life of higher education students: A global perspective, *Sustainability*, vol.12, no.20, 2020.

[2] K. Nagi and V. K. John, Plagiarism among Thai students: A study of attitudes and subjective norms, *2020 6th International Conference on e-Learning (econf)*, pp.45-50, 2020.

[3] K. Nagi and V. K. John, A study of attitude towards plagiarism among Thai university students, *European Journal of Foreign Language Teaching*, vol.5, no.4, pp.21-35, 2021.

[4] P. Saipech and P. Seresangtakul, Automatic Thai subjective examination using cosine similarity, *The 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICA ICTA)*, pp.214-218, 2018.

[5] S. K. Dey and M. A. Sobhan, Impact of unethical practices of plagiarism on learning, teaching and research in higher education: Some combating strategies, *The 7th International Conference on Information Technology Based Higher Education and Training (ITHET)*, pp.1-6, 2006.

[6] P. Šprajc, M. Urh, J. Jerebic, D. Trivan and E. Jereb, Reasons for plagiarism in higher education, *Organizacija*, vol.50, no.1, pp.33-45, 2017.

[7] A. S. Bin-Habtoor and M. A. Zaher, A survey on plagiarism detection systems, *International Journal of Computer Theory and Engineering*, vol.4, no.2, pp.185-188, 2012.

[8] M. Pauland and S. Jamal, An improved SRL based plagiarism detection technique using sentence ranking, *Procedia Computer Science*, vol.46, pp.223-230, 2015.

[9] B. Gipp, N. Meuschke and J. Beel, Comparative evaluation of text- and citation-based plagiarism detection approaches using GuttenPlag, *Proc. of the 2011 ACM/IEEE Joint Conference on Digital Libraries*, pp.255-258, 2011.

[10] N. Charya, K. Doshi, S. awkar and R. Shankarmani, Intrinsic plagiarism detection in digital data, *International Journal of Innovative and Emerging Research in Engineering*, vol.2, no.3, pp.23-30, 2015.

[11] H. A. Chowdhury and D. K. Bhattacharyya, Plagiarism: Taxonomy, tools and detection techniques, *The 19th National Convention on Knowledge*, pp.1-18, 2016.

[12] M. Sudjai and S. Mungsing, Development of an automated subjective answer scoring system with full-text search technique and PHP text comparison function, *Journal of Information Science and Technology*, vol.11, no.1, pp.8-17, 2021.

[13] K. Anekboon, Automated scoring for short answering subjective test in Thai's language, *2018 International Conference on Image and Video Processing, and Artificial Intelligence*, pp.324-329, 2018.

[14] S. Jaihuk and S. Mungsing, Scoring Thai language subjective answer automaic system by semantic, *Information Technology Journal*, vol.16, no.1, pp.15-23, 2020.

[15] S. Tachaphetpiboon, N. Facundes and T. Amornraksa, Plagiarism indication by syntactic-semantic analysis, *2007 Asia-Pacific Conference on Communications*, pp.237-240, 2007.

[16] E. Stamatatos, Plagiarism detection based on structural information, *Proc. of the 20th ACM Conference on Information and Knowledge Management (CIKM2011)*, pp.1221-1230, 2011.

[17] Ö. Uzuner, B. Katz and T. Nahnsen, Using syntactic information to identify plagiarism, *Proc. of the 2nd Workshop on Building Educational Applications Using NLP*, pp.37-44, 2005.

[18] W. Massagram, S. Prapanitisatian and K. Kesorn, A novel technique for Thai document plagiarism detection using syntactic parse trees, *Engineering and Applied Science Research*, vol.45, no.4, pp.290-300, 2018.

[19] L. Sindhu and S. M. Idicula, A plagiarism detection system for Malayalam text based documents with full and partial copy, *Procedia Technology*, vol.25, pp.372-377, 2016.

[20] B. Gipp and J. Beel, Citation based plagiarism detection: A new approach to identify plagiarized work language independently, *Proc. of the 21st ACM Conference on Hypertext and Hypermedia*, pp.273-274, 2010.

[21] A. Jadalla and A. Elnagar, A plagiarism detection system for arabic text-based documents, *Pacific-Asia Workshop on Intelligence and Security Informatics*, pp.145-153, 2012.

[22] M. Jiffriya, M. A. Jahan and R. G. Ragel, Plagiarism detection on electronic text based assignments using vector space model, *The 7th International Conference on Information and Automation for Sustainability*, pp.1-5, 2014.

[23] M. Elkhidir, M. M. Ibrahim, T. A. Khalid, S. Ibrahim and M. Awadalla, Plagiarism detection using free-text fingerprint analysis, *2015 World Symposium on Computer Networks and Information Security (WSCNIS)*, pp.1-4, 2015.

[24] M. Bagai, S. Gupta and R. Ali, Text based plagiarism detection, *International Journal for Technological Research in Engineering*, vol.3, no.8, pp.1710-1714, 2016.

[25] A. E. Eisa, S. Salim and S. Alzahrani, Figure plagiarism detection based on textual features representation, *The 6th ICT International Student Project Conference (ICT-ISPC)*, pp.1-4, 2017.

[26] D. D. Sinaga and S. Hansun, Indonesian text document similarity detection system using Rabin-Karp and Confix-Stripping algorithms, *International Journal of Innovative Computing, Information and Control*, vol.14, no.5, pp.1893-1903, 2018.

[27] V. Sornlertlamvanich, N. Takahashi and H. Isahara, Building a Thai part-of-speech tagged corpus (ORCHID), *Journal of the Acoustical Society of Japan*, vol.20, no.3, pp.189-198, 1999.

[28] C. Haruechaiyasak, S. Kongyoung and M. Dailey, A comparative study on Thai word segmentation approaches, *The 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp.125-128, 2008.

[29] B. Luaphol, J. Polpinij and M. Kaenampornpan, Mining bug report repositories to identify significant information for software bug fixing, *Applied Science and Engineering Progress*, vol.15, no.3, pp.1-14, 2021.