TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 5, October 2020, pp. 2480~2487
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i5.14027

Detecting Indonesian ambiguous sentences using Boyer-Moore algorithm

Risky Aswi Ramadhani, I Ketut Gede Darma Putra, Made Sudarma, I. A. D. Giriantari
Udayana University, Indonesia

Article Info

Article history:

Received Sep 2, 2019
Revised Apr 2, 2020
Accepted May 1, 2020

Keywords:

Ambiguous
Boyer-Moore
Grammatical
Indonesian sentences
String
Text

ABSTRACT

Ambiguous sentences are divided into 3 types namely phonetic, lexical, and grammatical. This study focuses on grammatical ambiguous sentences, grammatical ambiguous sentences are ambiguities that occur due to incorrect grammar, but this ambiguity will disappear once it is used within a sentence. Ambiguous sentences become a big problem when they are processed by a computer. In order for the computer to interpret ambiguous words correctly, this study seeks to develop detection of Indonesian ammbiguous sentences using Boyer Moore algorithm. This algorithm matches ambiguous sentences that are inserted as input with the data set. Then the sentence is being detected whether it contains ambiguous sentences, by calculating the percentage of similarity using cosine similarity method. Cosine similarity system is able to find out the meaning of the sentence. In the data set, the number of ambiguous sentences that can be collected is 50 words. The 50 words consist of ambiguous words data, ambiguous sentences, and ambiguous sentence meanings. This system trial was carried out for 200 times and the accuracy level was 0.935, precision was 0.9320, and Recall was 0.8. While the F-Measure was 0.8061. While the speed for word search 0.003275 seconds

Corresponding Author:

I Ketut Gede Darma Putra,
Udayana University,
P. B. Sudirman St., Denpasar, Bali, Indonesia.
Email: ikgdarmaputra@unud.ac.id

1. INTRODUCTION

Ambiguous sentences are sentences that have more than one meaning. Ambiguous sentences are divided into 3 types, namely phonetic, lexical, and grammatical. This research will focus on grammatical ambiguity. Grammatical ambiguity occurs due to incorrect grammar usage. However, this ambiguity would disappear once it is used within a sentence [1-4]. In Indonesian, the unability to understand ambiguous sentences often occurs due to different levels of language use, different levels of education, and culture [5]. Ambiguous word is a word hat has a vague (unclear) nature, in Indonesian, there are a number of grammatical ambiguous words such as "*bulan* (moon/ month)". "*Bulan*" has two meanings, the first meaning is "an astronomical object orbitting the earth", and the second meaning means "a period of time" [6, 7]. Grammatical ambiguous sentences would not pose a big problem when used in direct conversation, direct dialogue between humans, and sentences read by humans [8]. Because humans have intelligence that can process, and absorb ambiguous words in accordance with the topic of conversation, and words related to the ambiguous sentence. This is very different from computers, computers do not have the intelligence to detect ambiguous sentences. By using the grammatical ambiguous sentence detection system, the system is able to find out the meaning of

an ambiguous sentence, and translate it according to the meaning [9]. This system is aimed to enable computer to understand ambiguous sentences in Indonesian properly.

Research on grammatical ambiguous sentences has not been widely developed, especially regarding the detection of Indonesian grammatical ambiguous sentences. Currently, researches related to ambiguous sentences were only able to find ambiguous sentences but were not able to understand the meaning of ambiguous sentences [10]. So far, the data sets covering Indonesian ambiguous grammatical sentences are still not available yet. While the availability of grammatical ambiguous sentences detector is highly needed. Fro instance, in order to improve the accuracy of a translator system, and to make it easier for computers to understand a text. So, in this research, the expected novelty that will be achieved is to create a grammatical ambiguous sentence detection system in Indonesian, using the Boyer-Moore algorithm

2. RESEARCH METHOD

Figure 1 explains the processes involved in the ambiguous sentence detection system using Boyer-Moore algorithm. This flowchart explains the sentence being entered, then the sentence is checked using Boyer-Moore algorithm, so that it can be selected wether the sentence contains any ambiguous words. If the sentence is stated to contain ambiguous words, then the meaning of the sentencewould be searched using Cosine Similarity method. Several steps are needed to build this research; the following is the research method used.

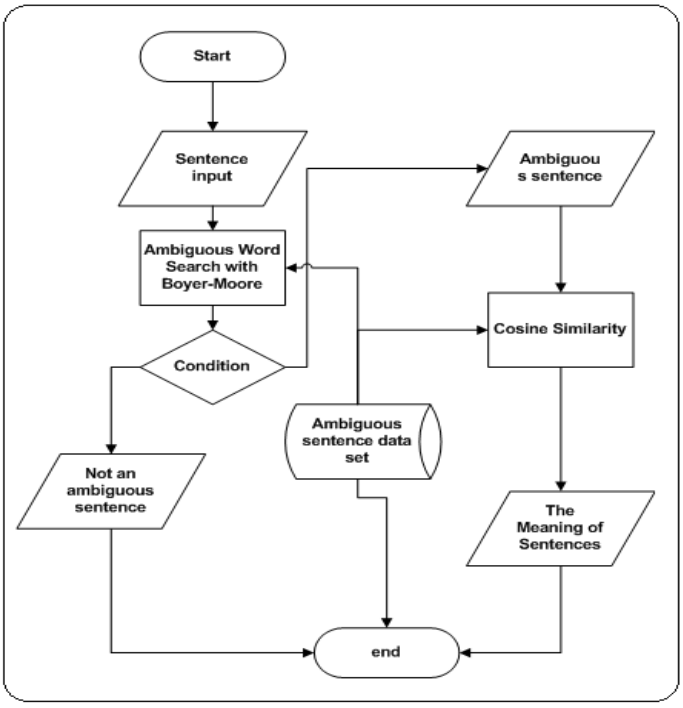


Figure 1. Flowchart detection of ambiguous Indonesian sentences with Boyer-Moore algorithm

2.1. Sentence input

Sentence input consists of sentences which are still unknown whether it contains ambiguity. The sentences are conversational sentences in Indonesian. In Indonesian, there are several types of ambiguous sentences, namely grammatical, lexical, and phonetic [11]. This research will focus on grammatical ambiguity. Grammatical ambiguous sentences are ambiguous sentences that occur due to incorrect grammar use, but this ambiguity will disappear once it is used in a sentence. The following are examples of ambiguous sentences

"Setiap awal bulan kami gajian (We are paid at the beginning o feach month)*"*

The sentence above contains an ambiguous word that is "*bulan* (month)", the word "*bulan*" has two meanings, which are;
- *Bulan* (month) = a period of time
- *Bulan* (moon) = sky object

In Figure 2, it is explained that the word "*bulan*" has two meanings, in which one refers to a particular unit of time (month), and *the "bulan*" which shows the earth's satellites (moon--an object in the sky). At this stage the meaning of the word is unknown. The following is a simple description of an ambiguous word.

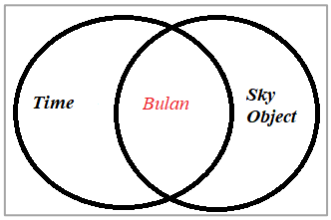


Figure 2. Grammatical ambiguous word description

2.2. Ambiguous word search using Boyer-Moore

Boyer-Moore algorithm is an algorithm used for string searching [12-20]. In conducting string searching, the Boyer-Moore algorithm is highly accurate. Following are the steps conducted by Boyer-Moore algorithm to find ambiguous sentences.

2.2.1. 1st step

Figure 3 explains the process of searching for the ambiguous word "*bulan*" in the sentence "*setiap awal bulan kami gajian* (at the beginning of each month we are paid)." This search is carried out from the first string, the search is carried out from the left side to the right side. If the word has not been found, the search would be repeated again, starting with the second string.

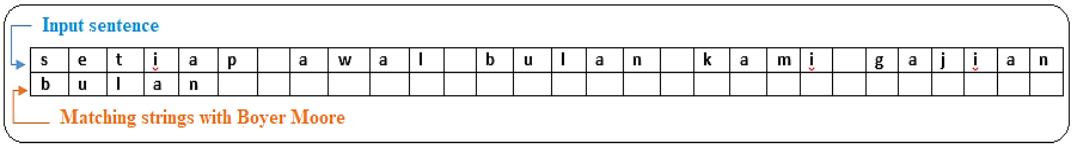


Figure 3. Ambiguous word search step 1

2.2.2. 2nd step

Figure 4 explains the process of searching for the ambiguous word "*bulan*" in the sentence "*setiap awal bulan kami gajian* (at the beginning of each month we are paid)". This process is a continuation of the first process, the search string starts from the second string.

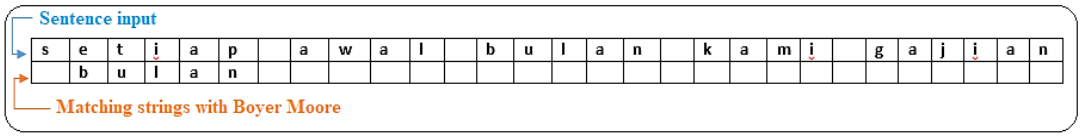


Figure 4. Ambiguous word search step 2

2.2.3. 13th step

Figure 5 shows that the process of searching for the ambiguous word *"bulan"* in the sentence "*setiap awal bulan kami gajian* (at the beginning of each month we are paid)" has been successful. The word "moon" is found on the 13th process, the word *"bulan"* was found in the 13th string. On the 13th step, a grammatical ambiguous word was found; the word is the word *"bulan"*. In this study, the Boyer-Moore algorithm is used to check strings. Inputs (sentences that are not yet known to be grammatically ambiguous) are being matched with data sets of words that have been identified as grammatical ambiguous. At present, the number of data sets that can be stored is only 50; this happens because there are no researchers who have developed applications related to grammatical ambiguous sentences in Indonesian.

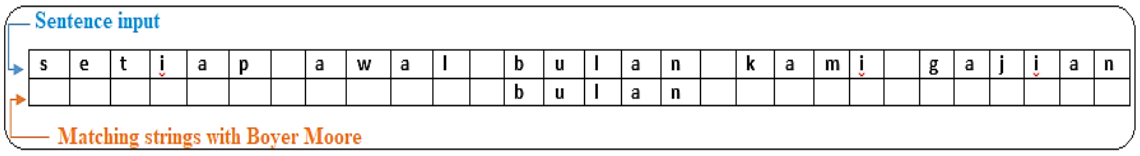


Figure 5. Ambiguous word search step 13

Figure 6 explains the Flowchart where the ambiguous word in a sentence is searched, on the Flowchart it is shown that if the ambiguous word has not been found, then a search is carried out on the next string, until the word is found or declared to be missing. If the word is found from the beginning, the system will immediately be terminated and it can be decided that the ambiguous word exists. The following is a flowchart describing the Boyer-Moore string searching process.

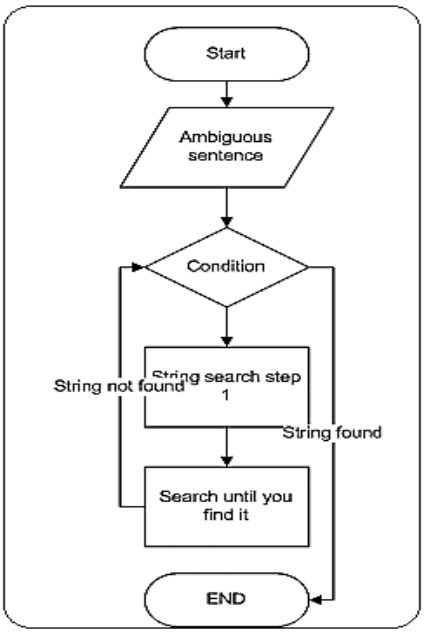


Figure 6. String searching flowchart using boyer-moore algorithm

2.3. Ambiguous sentences data SET

Grammatical ambiguous sentence dataset is a collection of ambiguous words and sentences used as a benchmark [21-23]. Since up to this stage, there was no ambiguous sentences found, this research has collected data on ambiguous words and sentences from Indonesian linguists. In this research, the resource person is an Indonesian language lecturer, Encil Puspitoningrum, M. Pd. The following is a table of ambiguous words and sentences obtained from her. Table 1 consists of 3 rows, line 1 is "Ambiguous Words" which contains the list of ambiguous words. Line 2 "Sentences" contains sentences that usually use ambiguous words. Line 3 is the “Meaning” ean which contains the meaning of the ambiguous sentences.

Table 1. Ambiguous words and sentences

Ambiguous words	Sentence	Meaning
Budi (Mind)	*Aku mengenang budi baikmu* (I remember your kindness)	*Kebaikan* (Kindness)
*Salam (*Regards*)*	*Gus kamu kemarin mendapatkan salam dari anggi* (Gus, Anggi sent you regards yesterday)	*Sapaan* (Greetings)
*Tahu (*Tofu*)*	*Agus kesini tadi memberi tahu* (Agus came here to give us tofu)	*Makanan* (Food)
Bunga (Interest)	*Bunga deposito di bank jatim lumayan tinggi* (The deposit interest rate in Bank Jatim is quite high)	*Keuntungan* (Profit)
Bangku (Bench)	*Dia tidak pernah makan bangku sekolah* (He never went to school)	*Pendidikan* (Education)
Kemas (Organized)	*Acara ini dikemas dengan sangat baik* (This event is very well organized)	*Melakukan pekerjaan* (Doing work)
Bulan (Month)	*Awal Bulan Kamu gajian* (You are paid at the beinning of the month)	*Waktu* (Time)

2.4. Cosine similarity

The sentences inputted are calculated in terms of their resemblance to the sentences in the data set. In order to calculate similarities between sentences, cosine similarity is used [24-28]. In this research, the sentence to be used as input is "*setiap awal bulan kami gajian*". The sentence has been identified to contain a grammatical ambiguous word, "*bulan* (month)". Sentences related to "*bulan* (month)" are:

- *Awal bulan kami gajian* (We are paid at the beginning of the month).
- *Bumi dan bulan merupakan benda langit* (The Earth and The Moon are sky objects).

By using cosine similarity, we can find the similarity of a string. After the string similarity is known, the system is able to show the meaning of the sentence. The following is the cosine similarity algorithm processing. In a more detailed explanation, the example used is the closeness between "*awal bulan kami gajian"* (We are paid at the beginning of each month) and *"awal bulan kamu gajian"* (You are paid at the beginning of each month) ".

- S1 = *awal bulan kami gajian* (We are apid at the beginning of the month).
- S2 = *awal bulan kamu gajian* (You are paid at the beginning of the month).

Table 2 explains the existence of each word in a sentence. If the word is contained in the sentence, code 1 will be given in line A. Conversely, code 0 will be given when the word is not found in the sentence.

Table 2. Ambiguous words and sentences

Word	Count				
	A	B	A. B	A^2	B^2
*Awa*l (beginning)	1	1	1	1	1
Bulan (month)	1	1	1	1	1
Kami (we)	1	0	0	1	0
Kamu (you)	0	1	0	0	1
Gajian (paid)	1	0	0	1	0
			2	4	3

Cosine Similarity is a method used to calculate the degree of similarity between two objects. For the purpose of data clustering, a good function is the Cosine Similarity function. For the set notation the formula as shown in (1):

$$Similarity = \cos(\theta) = \frac{A.B}{||A||B||} = \frac{2}{(4x3)} = 0.166 \quad (1)$$

After being calculated using the Cosine Similarity method, the highest closeness is 0.166. More detailed explanation is shown in Table 3. In Table 3 two values appear, which are 0.16 and 0.05. Given the high similarity value of the sentences *"setiap awal bulan kami gajian"* and *"setiap awal bulan kamu gajian"*. It can be condluded that the word *"bulan"* in the sentence means "a period of time".

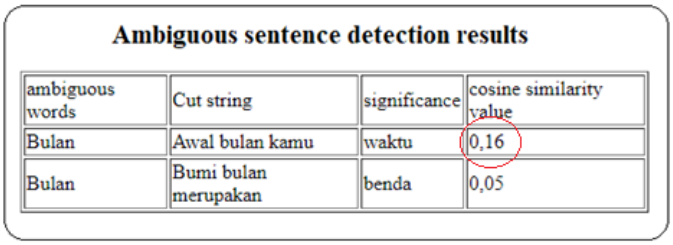
Table 3. The results of anayzing the meaning of sentences using confusion matrix method

Id	Input Sentences	Data Set	Value of Similarity
1	*Setiap awal bulan kami gajian* (We are paid at the beginning of each month)	*Setiap awal bulan kamu gajian* (You are paid at the beginning of each month)	0.16
2	*Setiap awal bulan kami gajian* (We are paid at the beginning of each month)	*Bumi dan bulan merupakan benda langit* (Earth and moon are sky objects)	0.05

2.5. Determining the meaning of sentences when beingprocessed in the program

From the Boyer-Moore Algorithm and Cosine Similarity processes some results are obtained [29, 30]. These results stated that "*bulan*" in the input sentence means a period of time (month). Following are the results obtained, the results are also implemented on the web. Figure 7 discusses the results of calculations performed by usingCosine Similiarity method. At this stage the sentence "*Setiap awal bulan kami gajian* (we are paid at the beginning of each week) " had been tested for the smiliarity with the sentence "*Setiap awal bulan kamu gajian* (you are paid at the beginning of each month) and "*Bumi dan Bulan merupakan benda langit* (The Earth and The Moon are sky objects)". After being calculated using Cosine Similarity method, the sentence has been

proven to have closeness in meaning with the sentence "*Setiap awal bulan kami gajian* (We are paid at the beginning of each month) " this sentence contais *word "bulan"* which means a period of time (month) with the value of similarity +of 0.16.



Ambiguous sentence detection results

ambiguous words	Cut string	significance	cosine similarity value
Bulan	Awal bulan kamu	waktu	0,16
Bulan	Bumi bulan merupakan	benda	0,05

Figure 7. The results of detecting grammatical ambiguous sentences using cosine similarity

3. RESULTS AND ANALYSIS

3.1. Accuracy, precision, recall and F-measure test

At this stage, the system is tested using a confusion matrix, which is often used to find out precision, accuracy, and recall [31-33]. With the confusion matrix, it can be seen how well the system is able to understand grammatically ambiguous sentences. This system experiment has been carried out 200 times. While there are 50 words in the database which are ambiguous words, this word is called True Positive (TP). When separating ambiguous words there is also an error, which is an unambiguous word but an ambiguous word is captured, this word is called false positive (FP). In some cases, there are ambiguous words but cannot be recognized by the system, this word is called false negative (FN). Whereas words that are not ambiguous are called true negatives (TN). The calculations can be seen in Table 4.

Table 4. Confusion matrix value

TP=40	FP=3
FN=10	TN=147

$$Accuracy = \frac{40+187}{40+147+3+10} = 0.935 \quad (2)$$

$$Precision = \frac{40}{40+3} = 0.9302 \quad (3)$$

$$Recall \frac{40}{40+3} = 0.8 \quad (4)$$

The value scale of matrix confusion ranges from 0-1. From the above calculation it is obtained the value of accuracy which is 0.935, Precision is 0.9320, and Recall is 0.8. Judging from the recall value, the system is able to recognize ambiguous words as much as 80%. Meanwhile, the lack of data sets has made the system unable to recognize ambiguous words. F-Measure is one of the evaluation calculations method in retrieving information that combines recall and precision. The values of recall and precision in a situation might bear different weights. The measurement that displays the reciprocity between recall and precision is the F-Measure, which is the weight of the harmonic mean of the Recall and Precision. The f-measure range is between 0-1. From the above calculation, the F-measure value is 0.86.

$$F1 = 2 * \frac{0.9302*0,8}{0.9302+0,8} = 2 * \frac{0.744}{1.73} = 0.8601 \quad (5)$$

3.2. The speed in detecting ambiguous words

In understanding grammatical ambiguous sentences, the system requires different time to process each sentence; the processing of this sentence depends on the number of characters understood [34]. The average sentence search value is 0.003275. There is a need for speed calculations to analyze system performance. Ambiguous sentence detection speed is presented in Table 5. The highest speed in this speed detector is 0.0024 to detect the sentence "*Dia bagai kuda hitam* (He is like a dark horse)". While the lowest speed is 0.0042 in the sentence "*Acara ini dikemas dengan sangat baik* (This event is very well organized)". The following table shows the rate of speed in detecting ambiguous words:

Table 5. Speed in detecting ambiguous words

Ambiguous words	Sentences	Speed
Kemas (organized)	*Acara ini dikemas dengan sangat baik* (This event is very well organized)	0.0042
Budi (Mind)	*Aku mengenang budi baikmu Aku mengenang budi baikmu* (I rememberi your kindness)	0.0036
*Salam (*Regards*)*	*Gus kamu kemarin mendapatkan salam dari anggi* (Gus, Anggi sent you regards yesterday)	0.0039
*Tahu (*Tofu*)*	*Agus kesini tadi memberi tahu* (Agus came here to give us tofu)	0.0039
Bunga (Interest)	*Bunga deposito di bank jatim lumayan tinggi* (The deposit interest rate in Bank Jatim is quite high)	0.0037
Bangku (Bench)	*Dia tidak pernah makan bangku sekolah* (he never went to school)	0.0043
Kuda (Horse)	*Dia bagai kuda hitam* (He is like a dark horse)	0.0024

4. CONCLUSION

Grammatical ambiguous sentences in Indonesian are sentences that have two meanings. To recognize ambiguous sentences, we need a Boyer-Moore algorithm and cosine similarity algorithm. Boyer-Moore algorithm is used to find strings (ambiguous sentences). While the cosine similarity algorithm is used to calculate the degree of similarity between two objects. Cosine similarity can be used to find out the meaning of a sentence, by calculating the similarity of the test data to the data set. The Boyer-Moore algorithm and the Cosine similarity algorithm are very effective for detecting ambiguous words. This can be proven by the success rate of the system in retrieving information (recall) of 80%. While the average speed of the Boyer-Moore algorithm when detecting ambiguous sentences takes 0.003275 seconds.

REFERENCES

[1] Charina I. N., "Lexical and Syntactic Ambiguity in Humor," *International Journal of Humanity Studies*, vol. 1, no. 1, pp. 120-131, September 2017.

[2] Soyusiawaty D., Ariwibowo E., "Designing and Implementing Parsing for Ambiguous Sentences in Indonesian Language," *Journal of Theoretical and Applied Information Technology*, vol. 84, no. 3, pp. 339-347, 2016.

[3] Andarini S. R. P., Anugerahwati M. G., "Structural Ambiguity in The Jakarta Post Newspaper's Headline News," *English Language Education Universitas Negeri Malang*, vol. 4, no. 2, pp. 1-15, 2013.

[4] Moukrim C., et al., "An innovative approach to autocorrecting grammatical errors in Arabic texts," *Journal of King Saud University - Computer and Information Sciences*, February 2019.

[5] Dennis D., Dewi I. I., "Students understanding of ambiguous sentences in websites," *Humaniora*, vol. 2, no.1, pp. 381-394, April 2011.

[6] Uliniansyah M. T., et al., "Solving Ambiguities in Indonesian words by morphological analysis using minimum connectivity cost," *Journal of natural language Processing*, vol. 2, no. 3, July 1995.

[7] Clare Q. C., "Language ambiguity: A curse and a blessing," *Literary Translation*, vol. 7, no. 1, pp. 1-4, 2003.

[8] Ritan Y. C. G., "Ambiguty and tree structure of setences in home movie," Thesis Universitas Sanata Darma Yogyakarta, 2018.

[9] Gatt A., Kramher E., "Survey of the state of the art in natural language generation: core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, no. 1, pp. 65-70, 2018.

[10] Rizki T. D., Yusliani N., "Design and build an ambiguity checking system for Indonesian sentences using harmony search algorithm (in Bahasa: Rancang bangun sistem pengecekan ambiguits kalimat berbahasa Indonesia mengunakan harmony search algorithm)," *Annual Research Seminar*, vol. 2, no. 1, 2016.

[11] Anwari Y., et al., "Analysis of ambiguous sentences in the novel of a semantic punch (in Bahasa: Analisis kalimat ambigu dalam novel suatu tinjuan semantic)," *Bung Hatta*, vol. 2, no. 3, 2013

[12] Lin Y, et al., "Sphere classification for ambigous data," *2006 International Confrence on Machine Learning and Cybernetics*, pp. 2571-2574, 2016.

[13] Xiong Y., "A composite Boyer-Moore algorithm for the string-matching problem," *International Conference on Pararlel and Distributed Computing*, *Application and Technologies*, Dec 2010.

[14] Wang Y., "A new method to obtain the shift-table in Boyer-Moore's algorithm," *19th International Confrence on Pattern Recognation*, pp. 1-4, December 2008.

[15] Goldschlag D. M., "Mechanically veriying concurret programs with the Boyer-Moore prover," *IEEE Transaction on Software Engineering*, vol. 16, no. 9, pp. 1005-1023, September 1990.

[16] Domingues A., et al., "Prommable Soc platform for deep packet inspection using enhanced Boyer- Moore algorithm," *Internasional Symposium on Reconfigurable Commication-Centric System-on- Chip*, pp. 1-8, July 2017.

[17] Danvy O., Rohde H. K., "On obtaining the Boyer-Moore String-Matching algorithm by partial evaluation," *Information Processing Letters*, vol. 99, no. 4, pp. 158-162, August 2006.

[18] Saleh A. Z. M., et al., "A method for web application vulnerabilities by using Boyer-Moore string matching algorithm," *Procedia Computer Science*, vol. 72, pp. 112-121, 2015.

[19] Watson W. B., "A new regular gramar pattern matching algorithm," *Theoretical Computer Science*, vol. 299, no. 1-3, pp. 509-521, 2003.
[20] Sabriye A. O. J., Zainon W. M. N. W., "An aprroch for detection syntax and syntatic ambiguity in sofware requitment specification," *Journal of Theretical and Applied Infoemation Tecnology*, vol. 96, no. 8, April 2018.
[21] Ramadhani R. A., et al., "Database of Indonesian Sign Systems," *ICSGTEIS*, pp. 225-228, Oct 2018.
[22] Jiang H., "Study on information retrieval model based on rough set theory," *Internasional Symposium on Intelligent Ubiquitous Computing and Education*, pp. 440-444, May 2009.
[23] Lynn T., et al., "Data set for automatic of online misogynistic speech, *Data in Brief*, vol. 26, Oct 2019.
[24] Gokul P. P, et al., "Setence Similarity detection in Malayalam languange using cosine similarity," *International Confrence on Recent Trend in Electronics*, *Informationm & Communication Techonology*, pp. 221-225, May 2017.
[25] Akbas C. E., et al., "L1 norm-based multiplication-free cosine similarity measures for big data analysis, "*International Workshop on Computational Intellegence For Multimedia Understanding*, pp. 1-5, Nov 2014.
[26] Alodad M., Janeja, "Similarity in patient support forums using tf-idf and cosine similarity metrics," *International Conference on Healthcare Informatics*, pp. 521-522, Oct 2015.
[27] Zhu S., et al., "Top-K cosine similarity intersting pairs search," 2*010 Sevent International Confrence Fuzzy System and Knowledge Discovery*, pp. 1479-1483, Aug 2010.
[28] Hermandes A. F. R., Garcia N. Y. G., "Distributed processing using cosine similarity for mapping big data in hadoop," *IEEE Latin America Tracsaction*, vol. 14, no. 6, pp. 2857-2861, June 2016.
[29] Karim M. S., et al., "Implemtation and performance evaluation of semantic features analysis system for bangla assertive, impperative and interrogative senteces," *ICBSLP*, pp. 1-5, September 2018.
[30] Zitnick C. L., et al., "Learning the visual interpretation of setence," *2013 IEEE Internasional Confrence on Computer Vision*, pp. 1681-1688, Dec 2013.
[31] Fergyanto E., et al., "A Simple classifier for detecting online child grooming coversation," *TELKOMNIKA Telecommunication Computing Electronics and Control,* vol. 16, no. 3, pp. 1239-1248, June 2018.
[32] Garcia-Balbola J. L., et al., "Homogeity test for confusion matrices: A method and example," *Internasional Geoscience and Remote Sensing Syposium*, pp. 1203-1205, July 2018.
[33] Bhalla R., Bagga A., "Oppinion mining framework using proposed rb-bayes model for text classification," *Internasional Journal of Electrical and Computer Engineering (IJECE),* vol. 9, no. 1, pp. 477-484, January 2019.
[34] Gokay R., Yalcin H., "Improving low Resource Turkish speech recognition with Data Augmentation and TTS," *International Multi-Confrence on Sytems, Signals & Device (SSD)*, pp. 357-360, March 2019.