KINETIK, Vol. 3, No. 4, November 2018, Pp. 319-326
ISSN : 2503-2259
E-ISSN : 2503-2267

Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis

Faisal Rahutomo[*1], Ahmad Hafidh Ayatullah[2]
[1,2]Politeknik Negeri Malang/Information Technology Department
faisal@polinema.ac.id[*1], ahmad.hafidhayatullah@gmail.com[2]

Abstract

This paper describes the academic base of an openly Indonesian dataset in Mendeley Data with DOI: 10.17632/d7vx5cc92y.1*. The dataset is an Indonesian language expansion of Microsoft research video description corpus, an open dataset contains about 120 thousand sentences. The dataset is a useful resource because the sentences are a set of roughly parallel descriptions of more than 2,000 video snippets of 35 languages. Both paraphrase and bilingual relation are available but Indonesian description is not available in the dataset. Therefore, this paper describes the research effort to expand the dataset for the Indonesian language. The research collected 43,753 description texts of 1,959 short videos, parallel with Microsoft's dataset. Adding more value to the dataset, similarity metrics calculations of the texts were done. The metrics were Cosine, Jaccard, euclidian, and Manhattan with average results were 0.22, 0.33, 2.38, and 6.08 respectively.*

Keywords: *Microsoft Research Video Description Corpus, Indonesian descriptions, Cosine, Jaccard, Euclidean, Manhattan*

1. Introduction

Natural language processing (NLP) is an artificial intelligence branch that works on human language [1] [2]. The main purpose of the NLP study is making machines that are able to understand and grasp the meaning of human language and then give the appropriate response. Natural language processing is applied in many cases such as machine translation [3], information retrieval [4], and text mining [5]. In order to develop an appropriate system, scheme, and algorithm for a specific case, the researchers need a collection of data to work on it.

Microsoft Research Video Description Corpus (MRVDC) is a dataset, which is developed by Microsoft Research. The dataset contains paraphrase expressions of an event both in a language and the other languages. The dataset is an important resource for developing and evaluating a semantic text similarity system. An intelligent system that able to measure the similarity of texts by their meaning. The technical paper of the test set [6] is cited by 41 others, as recorded by ACM digital library. While Citeseer records the paper citation by the other works as 57. Several papers that use this dataset are [7], [8], and [9]. The citation reports show the dataset potential and usability.

Unfortunately, the Indonesian language is not available in the test set. Therefore, this paper exposes the effort to expand the test set in Indonesian. The research aim was to contribute to the development of the research environment in Indonesian information retrieval and natural language processing research area [9]. This research added more value to the test set by providing the similarity scores of the texts. The metrics were Cosine, Jaccard, Euclidean, and Manhattan.

Therefore, this paper is organized into five sections. The first section describes the introduction of this research. Section 2 describes the methodology being used. Section 3 describes the implementation, while Section 3 describes the results and discussion. Finally, Section 5 concludes this paper with several research suggestion based on the expansion dataset. The result of this research is available in Mendeley Data with DOI: 10.17632/d7vx5cc92y.1 [10]. The data is open with SQL dump format of MySQL database.

2. Indonesian Expansion of MRVDC

The MRVDC collection is approximately 120 thousand sentences of parallel texts of 35 languages. The texts are related to 2,089 short video clips. MRVDC is used to help researchers

Rahutomo, F., & Hafidh Ayatullah, A. (2018). Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 3(4). doi:http://dx.doi.org/10.22219/kinetik.v3i4.680
Receive August 14, 2018; Revise October 01, 2018; Accepted October 15, 2018

working on the computational linguistic research area, allowing them to compare systems with several types of automatic machine by the dataset [6]. The data was collected by showing a segment of a YouTube video to workers on Amazon's Mechanical Turk and asking them to give a one-sentence description of the main action/event in the video. Top 16 languages related to the dataset are English, Hindi, Romanian, Slovene, Serbian, Tamil, Dutch, and German with 85,550, 6,245, 3,998, 3,584, 3,420, 2,789, 2,735, and 2,326 texts respectively. Then, Macedonian, Spanish, Gujarati, Russian, French, Italian, Georgian, and Polish with 1,915, 1,883, 1,437, 1,243, 1,226, 953, 907, and 544 texts respectively.

The data file is provided in CSV format. The dataset contains several fields. “VideoID” field describes a YouTube video ID. The video can be viewed on a webpage by typing the link as http://www.youtube.com/watch?v=<video ID>. Where <video ID> is the ID of the particular video that wanted to be viewed. “Start” field specifies the starting time for the video segment in seconds. “End” field specifies the ending time for the video segment in seconds. “WorkerID” field describes an anonymized IDs to identify the worker who gave the description. “Source” field specifies only for English descriptions. Good annotators and their work are labeled as clean. “AnnotationTime” field describes the amount of time it took to annotate (including the time to watch the video) in seconds. “Language” field describes the language of the description is written in as identified by the annotators. “Description” field describes the description of the video segment.

In order to expand MRVDC, this research was started with searching the video clip source. The clips of MRVDC can be downloaded from [11], but not all of 2,089 clips are available. The link contains only 1,959 clips. After the clips were collected, then this research build the interface application [12]. Therefore, respondents were able to insert their description related to a video clip. The clips were divided into 39 sessions with about 50 clips for each session. This research developed an analysis application tool as well. The tool preprocessed the text and calculated the similarity of texts with several metrics. After all of the data were collected, the research application preprocessed the texts. Then similarity measurement of all possible pairs of texts of each video was calculated and stored in the database. Figure 1 describes the flow diagram of this research.

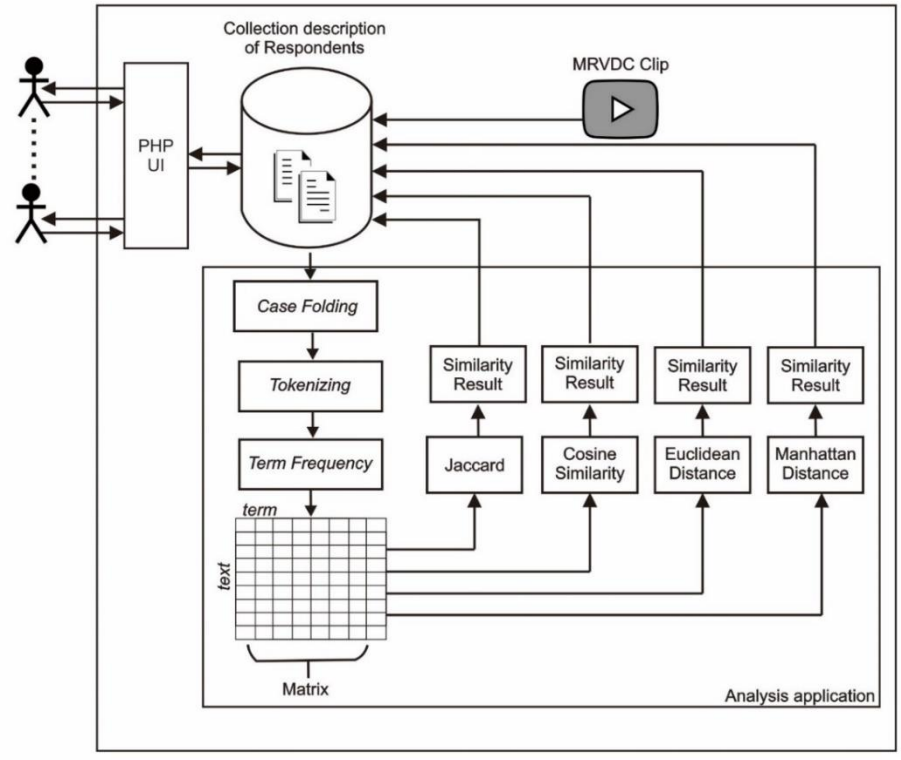


Figure 1. Research Flow Diagram

2.1 Text Collection

The videos of MRVDC are short videos with a duration of less than 1 minute. The videos display a clear event for everyone who watches it. The respondents were two classes of level 2

diploma programme in information technology department, State Polytechnic of Malang. Each class consists of 30 persons. The classes being participated in this research were TI-2G and TI-2H. The respondents were asked to express their natural language of an event in a video. The texts of expressions were collected in about 2 months, 8 weeks. With two sessions of each week, 1 hour of each session. Respondents were worked every Monday and Tuesday in two months. Each day the researcher ran 6 sessions with 3 sessions in the morning and 3 sessions in the afternoon. All of the respondents were native speaker of Indonesian. They were worked in one room with local network access, therefore each respondent was able to reach the local web server, a web application of this research.

Because of several reasons, such as respondent absent during class or unable to attend the class because of sickness, not all of the video clip has 30 descriptions in Indonesian. The descriptions were being collected by a minimum person of 12 and a maximum full class of 30 persons. After the data were collected, the data were validated manually. Correction of typography was done, as well as grammatical error. Further validation step was removing the texts with no meaning. Therefore, the Indonesian expansion texts of MRVDC were meaningful texts with free of grammar and typographical errors.

2.2 Text Preprocessing

Several preprocessing steps were involved. The first step was case folding, i.e. the texts form were transformed into lower case. Then, tokenizing was done based on white space. No stopword nor stemming processes were involved. Next step was calculating the term frequency. The result of those processes was a matrix of term vectors.

2.3 Cosine Similarity

Cosine Similarity is a metric to calculate the similarity between two vectors. Cosine similarity result is bounded between 0 and 1. If the result is 0 then can be said that no similarity between the two vectors. If the value is 1 then can be said the two vectors are perfectly the same. Equation 1 describes the calculation of Cosine similarity. Where A_k is term weighting k in document A and B_k is term weighting k in document B.

$$Cosine(A,B) = \frac{A.B}{|A||B|} = \frac{\sum_{k=1}^{n}(A_k.B_k)}{\sqrt{\sum_{k=1}^{n}(A_k)^2}.\sqrt{\sum_{k=1}^{n}(B_k)^2}} \quad (1)$$

2.4 Jaccard

Jaccard is the other metric to calculate the similarity of two vectors. The formula to calculate the similarity between vectors A and B is given in Equation 2. Similar to Cosine, the result is bounded between 0 to 1. In Equation 2, $|A \cap B|$ means how many tokens are apear both in texts A and B, while $|A \cup B|$ means the amount of unique tokens both in texts A and B.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (2)$$

2.5 Euclidean Distance

The Euclidean distance calculation of two vectors is defined by Equation 3. Where w_{Ak} is term weighting k in document A and w_{Bk} is term weighting k in document B. The equation describes that the similarity result is possible to be more than 1.

$$Euclidean(A,B) = \sqrt{\sum_{k=1}^{n}(w_{Ak} - w_{Bk})^2} \quad (3)$$

2.6 Manhattan Distance

The next distance metric being used in this research is Manhattan distance. Manhattan distance calculates the distance of two vectors by Equation 4. As same as Euclidean, w_{Ak} is term weighting k in document A and w_{Bk} is term weighting k in document B. The result value is possible to be more than 1.

$$Manhattan(A,B) = \sum_{k=1}^{n} |w_{Ak} - w_{Bk}| \quad (4)$$

2.7 Calculation Example

This sub-section describes the calculation example of Cosine, Jaccard, Euclidean, and Manhattan. For example, sentence *A* is "Lelaki berjenggot itu sedang menggunting kertasnya". And the sentence *B* is "pria sedang menggunting kertas". The processes are described as follows:

1. Case folding step
   Sentence *A* becomes "lelaki berjenggot itu sedang menggunting kertasnya".
   Sentence *B* becomes "pria sedang menggunting kertas".
2. Tokenizing step
   Sentence *A* becomes an array of words of {lelaki, berjenggot, itu, sedang, menggunting, kertasnya}.
   Sentence *B* become {pria, sedang, menggunting, kertas}.
3. Term frequency step
   Table 1 describes the results of this step.

Table 1. Preprocessing Results of Sentences A and B

	lelaki	berjenggot	itu	sedang	menggunting	kertasnya	pria	kertas	∑words
A	1	1	1	1	1	1			6
B				1	1		1	1	4

4. Cosine

$$Cosine(A,B) = \frac{1.0 + 1.0 + 1.0 + 1.1 + 1.1 + 1.0 + 0.1 + 0.1}{\sqrt{6}.\sqrt{4}} = \frac{2}{4{,}899} = 0{,}4082$$

5. Jaccard
   $A \cap B$= {sedang,menggunting}
   $|A \cap B|$ = 2
   $A \cup B$= {lelaki, berjenggot, itu, sedang, menggunting, kertasnya, pria, kertas}
   $|A \cup B|$ = 8

$$J(A,B) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}| - |\mathrm{A} \cap \mathrm{B}|} = \frac{2}{6+4-2} = \frac{2}{8} = 0{,}25$$

6. Euclidean

$$Euclidean(A,B) = \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (1-1)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2} = \sqrt{6} = 2.449$$

7. Manhattan

$$Manhattan(A,B) = |1-0| + |1-0| + |1-0| + |1-1| + |1-1| + |1-0| + |0-1| + |0-1| = 6$$

3. Analysis Application

The application was developed in a web-based environment. MySQL was chosen as the database management system. Then PHP framework CI was used to build the application. The respondents were given their special user interface. Therefore, they were able to watch and describe the moment by their own words. Each session contained 50 videos, that can be controlled by the researcher to be opened or closed to the respondent. Researchers had have their own user interface. Several functions were embedded: adding video, opening/ closing session, running metric calculation, and generating reports.

3.1 Database

Figure 2 describes the Physical Data Model (PDM) of "desvid" database. There were 9 tables in the database with different functions. Each table function is described as follows. Table "tb_admin" was used to store admin data. Table "tb_user" was used to store user data. Table

"tb_sesi" was used to store description collection session setting. Table "tb_video" was used to store the video data link path in local storage. Table "tb_deskripsi" was used to store the data descriptions from respondents. Table "tb_Cosine" was used to store the Cosine values. Table "tb_Jaccard" is used to store the Jaccard value of two pair texts. Table "tb_Euclidean" is used to store the Euclidean values. And table "tb_Manhattan" is used to store the Manhattan values.

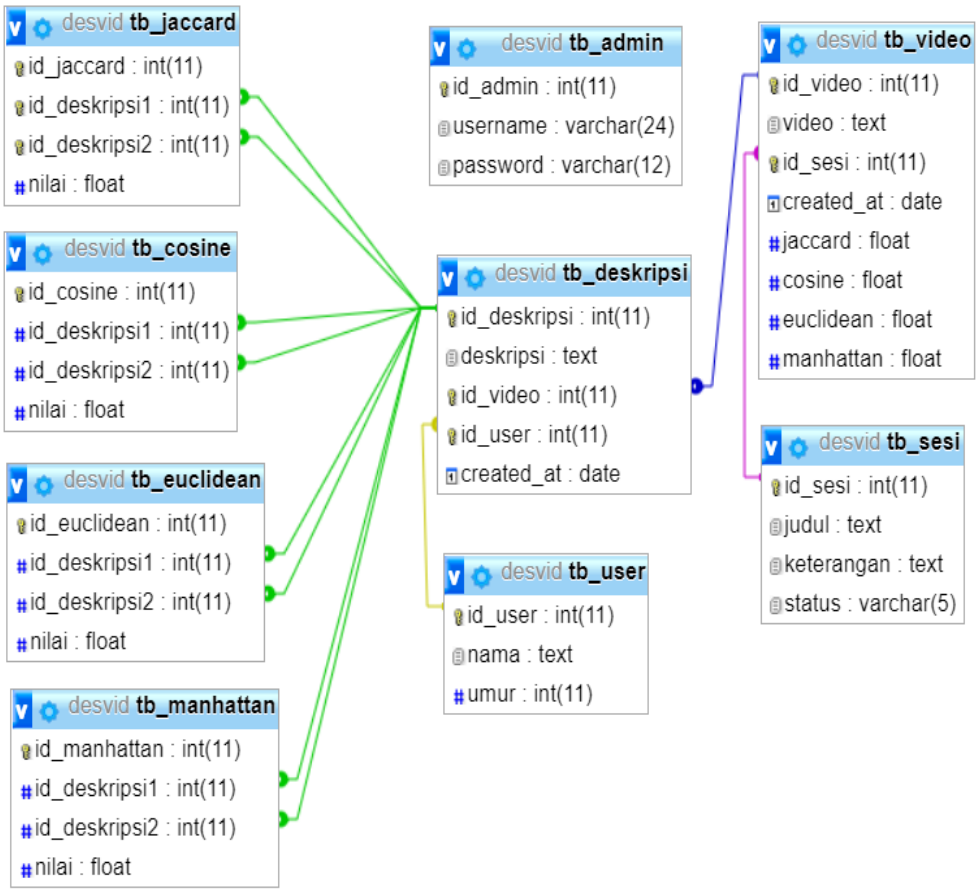


Figure 2. Physical Data Model

3.2 User Interface

Figure 3 describes the login page of the system. On the login page, the user should enter their login before being redirected into their appropriate dashboard.

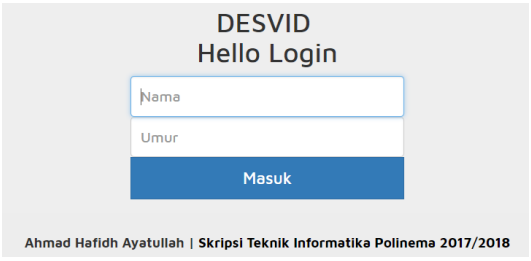


Figure 3. Main Page

Figure 4 illustrates the video description page. The user can watch the video clips one by one through the interface and filling out the description by their own words. Then, the session page can be accessed only by the researcher. The researcher was possible to add a new session and link the videos to the session. The researcher was also possible to open or close the session. If the session status was open then the session can be accessed by the respondents. The researchers were also able to add and modify data in the session. On a special page, researchers

were able also to add videos and remove videos. Video format in this system was only in MP4. The video files were uploaded in the folder “data_video”. While the database was only stores the name and path location of it.

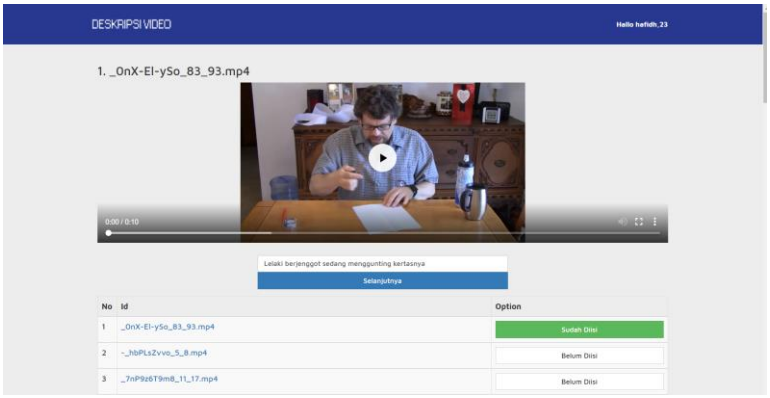


Figure 4. Description Video Page

4. Result and Discussion

This effort finally collected 43,753 description texts of 1,959 short videos. The videos’ name was the same as MRVDC’s name. Therefore the collection can be used for semantic text similarity research in Indonesian. The texts linkage to the other languages based on the same video ID makes it possible to be used for another area of researches, which are needed parallel language corpus. This dataset fills the gap both in MSRVDC and in Indonesian information retrieval system research. Table 2 shows the statistic facts of the dataset. The values were gathered from 1,959 videos. Because each video had has several values, the values were then being calculated its average. Therefore, row average actually shows the average value of several videos, where each video has one average value. Column “words” shows the maximum, minimum, and average value of words were being used by respondents. The values were 10.0714, 2.1538, and 4.3958 respectively. From the fact can be understood that the description forms were short texts. From a very short sentence, with only 2 words, to medium length sentence with about 10 words. The average value shows the descriptions in the dataset were mostly short sentences with about 4 words. Figures 5.a and 5.b show the dataset characteristic in scatter plot and its histogram respectively.

Figures 6, 7, 8 and 9 describe the dataset characteristics with metrics Cosine, Jaccard, Euclidean, and Manhattan respectively. While figure (a) describes the scatter plot, figure (b) describes the histogram. Average Cosine and Jaccard of the texts are low, only 0.33 and 0.22 respectively. The meaning of the fact was the similarity values of the texts in each video were low. This characteristic was very useful for semantic text similarity system, where the system calculates the similarity of two texts by based on their meaning. Euclidean and Manhattan values were low as well, close enough to the values of words being used by respondents. This characteristic values of the dataset can be used to challenge the researcher to develop their own system in Indonesian that able to catch the similarity of two different texts with the same meaning.

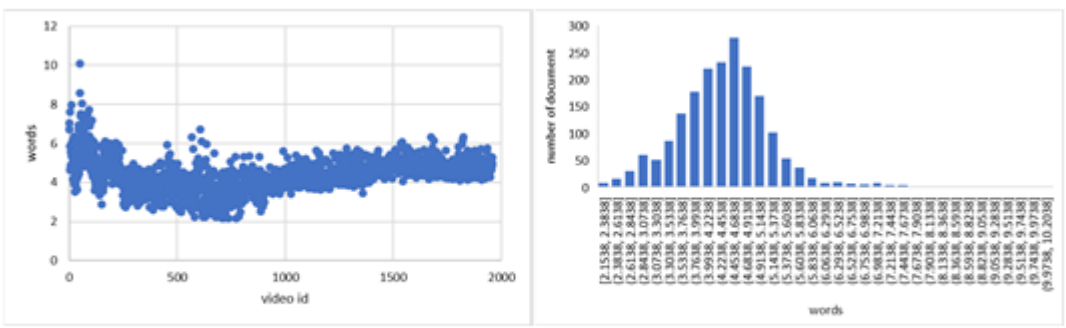
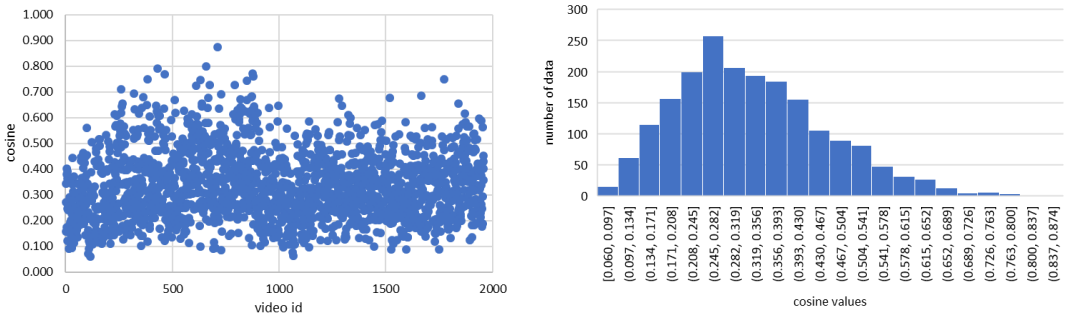
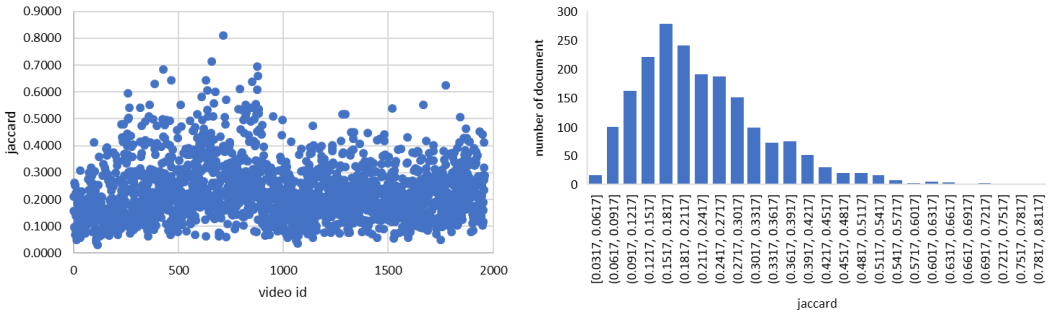


Figure 5. Description Texts Characteristic



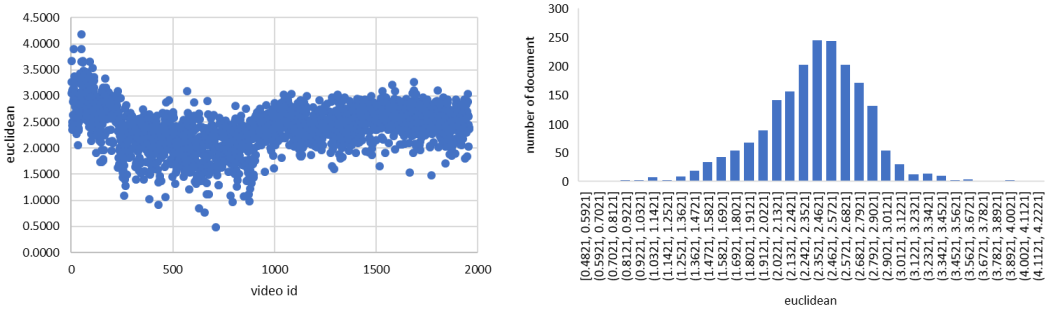
(a) (b)

Figure 6. Cosine Values Characteristic



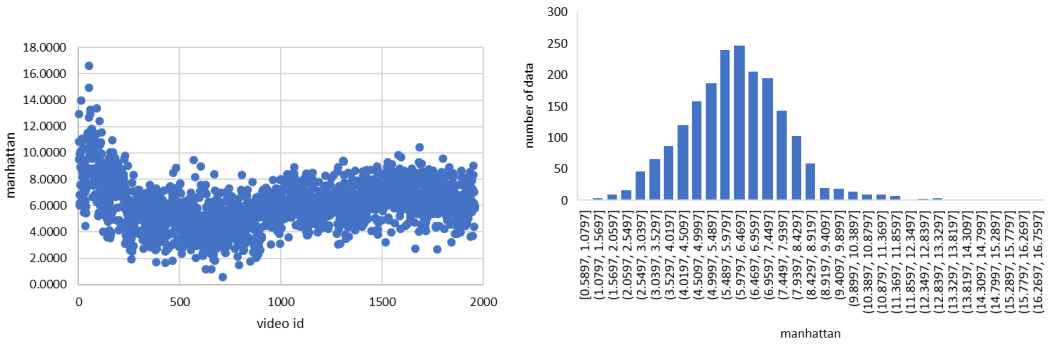
(a) (b)

Figure 7. Jaccard Values Characteristic



(a) (b)

Figure 8. Euclidean Values Characteristic



(a) (b)

Figure 9. Manhattan Values Characteristic

Table 2. Statistic Facts of the Indonesian Dataset

Statistic	Metrics				
	Cosine	Jaccard	Euclidean	Manhattan	Words
Max	0.8737	0.8098	4.1777	16.6323	10.0714
Min	0.0605	0.0317	0.4821	0.5897	2.1538
Average	0.3306	0.2278	2.3825	6.0886	4.3958

5. Conclusion

This research collected 43,753 description texts in Indonesian from 1,959 videos of MRVDC. The average Jaccard value was 0.22 while average Cosine was 0.33. Average Euclidean was 2.38 and average Manhattan was 6.08. Low values of Jaccard and Cosine indicate the data set was good enough to be used in Indonesian semantic similarity text research. The expansion nature of the dataset from MRVDC indicates the possibility to be used in machine translation research task. Automatic Indonesian paraphrasing system to avoid sentence plagiarism can also be developed based on the dataset.

References

[1] M. D. Harris, *"Introduction to Natural Language Processing,"* Reston, VA, USA: Reston Publishing Co., 1985.
[2] A. Kao and S. R. Poteet, *"Natural Language Processing and Text Mining,"* Springer Publishing Company, Incorporated, 2006.
[3] C. Goutte, N. Cancedda, M. Dymetman, and G. Foster, *"Learning Machine Translation,"* The MIT Press, 2009.
[4] R. Baeza-Yates and B. Ribeiro-Neto, *"Modern Information Retrieval: The Concepts and Technology Behind Search*," 2nd edition, USA: Addison-Wesley Publishing Company, 2008.
[5] S. Russell and P. Norvig, *"Artificial Intelligence: A Modern Approach*," 3rd edition, Upper Saddle River, NJ, USA: Prentice Hall Press, 2009.
[6] D. L. Chen and W. B. Dolan, *"Collecting Highly Parallel Data for Paraphrase Evaluation,"* in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Vol. 1, Pp. 190–200, 2011.
[7] F. Rahutomo, Y. Manabe, T. Kitasuka, and M. Aritsugi, *"Econo-ESA Reduction Scheme and the Impact of its Index Matrix Density,"* in Procedia Computer Science, Vol. 35, No. C, 2014.
[8] F. Rahutomo and M. Aritsugi, *"Econo-ESA in Semantic Text Similarity,"* Springerplus, Vol. 3, No. 1, 2014.
[9] F. Rahutomo and A. H. Ayatullah, *"Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis,"* 2018.
[10] F. Rahutomo and E. Rohadi, *"Pengembangan Piranti Penelitian Sistem Temu Kembali Informasi Bahasa Indonesia,"* in Seminar Nasional Sistem Informasi Indonesia *(SESINDO)*, Pp. 313–319, 2015.
[11] D. L. Chen and W. B. Dolan, *"Youtube Clips,"* 2011.
[12] N. Riza Akbar, R. Faisal, and H. Budi, *"Pengembangan Data Uji Sistem Komputasi Kemiripan Teks Secara Semantik Berbahasa Indonesia,"* in Seminar Informatika Aplikatif Polinema, 2016.