TOPIC GROUPING BASED ON DESCRIPTION TEXT IN MICROSOFT RESEARCH VIDEO DESCRIPTION CORPUS DATA USING FASTTEXT, PCA AND K-MEANS CLUSTERING

Ahmad Hafidh Ayatullah[1], Nanik Suciati[2]

[1]Department of informatics, [2]Faculty of Intelligent Electrical and Informatics Technology, [3]Institut Teknologi Sepuluh Nopember
[1]6025211037@mhs.its.ac.id, [2]nanik@if.its.ac.id

Abstract

Video data retrieval can be done based on voice, image, or text data that represents video content. Searching for videos using text data can be done by calculating the similarity between the text descriptions provided by the user and the text descriptions of all the video data in the database. Only video data with a certain level of similarity will be provided to the user as a fetch result. Determining the similarity of the description text can be based on the clustering results of the feature representation of the description text with the word embedding used. This research groups topics of the Microsoft Research Video Description Corpus (MRVDC) based on text descriptions of Indonesian language dataset. The Microsoft Research Video Description Corpus (MRVDC) is a video dataset developed by Microsoft Research, which contains paraphrased event expressions in English and other languages. The results of grouping these topics show how the patterns of similarity and interrelationships between text descriptions from different video data, which will be useful for the topic-based video retrieval. The topic grouping process is based on text descriptions using fastText as word embedding, PCA as features reduction method and K-means as the clustering method. The experiment on 1959 videos with 43753 text descriptions to vary the number of *k* and with/without PCA result that the optimal clustering number is 180 with silhouette coefficient of 0.123115. The optimal clustering results in this study can be used for video data retrieval systems in the Indonesian language MRVDC dataset.

Keywords: **Microsoft Research Video Description Corpus, video data retrieval, silhouette coefficient**

1. Introduction

Information consists of several types, such as information in the form of images, text, video, and audio, and all this information requires an Information Retrieval (IR) process Feldi, et al (2021). One of the pieces of information that require a retrieval process (IR) is information in video form. Video is a technology used to capture, record, process, and display images with sound at the same time. In addition, videos can also present information and events in the form of stories, images, and sound, as well as text, while Information Retrieval is a procedure for recovering stored data and then providing information about the required subject Hidayati & Harjoko, (2012).

Information Retrieval in the video describes metadata searches for voice, text data, or images. The application of the Information Retrieval video concept can be carried out on an indexed video search system. The video search system works by receiving input from users in the form of queries or keywords. The video search system works by receiving input from users in the form of queries or keywords. The search does by calculating the similarity's videos in the index with the query.

Microsoft Research Video Description Corpus (MRVDC) is a dataset, which is developed by Microsoft Research. The dataset contains paraphrased expressions of an event both in one language and the other languages. The dataset is an important resource for developing and evaluating a semantic text similarity system. An intelligent system that can measure the similarity of texts by their meaning.

In the research, Rahutomo & Ayatullah, (2018) expand the Indonesian language dataset and the research collected 43,753 description texts from 1,959 videos and added more value of distance calculation such us Cosine Similarity, Jaccard, Euclidian Distance, and Manhattan Distance with the average results of 0,22, 0,33, 2,38, and 6,08.

In the research, Nurdin, et al (2020) compared the performance of word embedding from Word2vec, GloVe, and FastText on text classification. The results obtained show that Word Embedding FastText has the best performance compared to the other two word embeddings, namely Word2vec, and GloVe, although the difference is not too significant.

In the research, Hediyati & Suartana, (2021) evaluation of clustering data using the DB Index value shows the most optimal value in the dataset,

which is reduced to 1 PC and formed into 3 clusters, namely 0.4072. whereas with the same number of clusters, a dataset with 2 PCs produces a DB Index value of 0.6168, a dataset with 3 PCs produces a value of 0.6598, and a dataset without dimension reduction process produces a DB Index value of 0.4598.

In the research, Multi Fani & Santoso, (2021) the application of Text Mining to perform clustering using the K-means clustering method with the best number of clusters obtained from the Silhouette Coefficient method on the @bliblidotcom Twitter tweet data to determine the types of tweet content that are mostly retweeted by @bliblidotcom followers. Tweets with the most retweets and favorites are discount offers and flash sales, so Blibli Indonesia could use this kind of tweet to conduct advertising on social media Twitter because the prize quiz tweets are liked by the @bliblidotcom Twitter account followers.

Unfortunately, the existing MRVDC datasets are not grouped into the same cluster so it cannot be used in video data retrieval. To overcome this, it is necessary to group topics from the MRVDC dataset so videos with the same text description can have the same cluster. Grouping video data can be done based on description text from the Indonesian language Microsoft Research Video Description Corpus (MRVDC) dataset.

Principal Component Analysis (PCA) is a technique used to simplify data by converting it into a linear form so that a new coordinate system with maximum variance is formed. PCA can be used to reduce data dimensions without significantly reducing data characteristics Puspitaningrum, et al (2014).

In this study, PCA is used to reduce the number of features/columns in fastText Word Embedding, and K-means is used to group topics based on description text in video data Microsoft research video description corpus (MRVDC).

From the problems that arise, the purpose of this study is to group topics based on description text in video data Microsoft research video description corpus (MRVDC) using fast text word embedding, PCA reduction feature, and K-Means Clustering.

2. Literatur Review

2.1 Related Research

This study refers to previous research to help facilitate the research process carried out in determining systematic steps in terms of theory and concept.

In the research, Nurhadi, et al (2016) created dan application using CodeIgniter to display and describe videos according to the video events shown and to obtain test data. Test data in this study were analyzed using the Jaccard method with an overall average score of 0.159. This study uses 150 videos with 28 respondents who describe them.

In the research, Rahutomo & Ayatullah, (2018) expand the Indonesian language dataset and the research collected 43,753 description texts from 1,959 videos and added more value of distance calculation such us Cosine Similarity, Jaccard, Euclidian Distance, and Manhattan Distance with the average results of 0,22, 0,33, 2,38, and 6,08.

In the research, Kuyumcu, et al (2019) contributed to evaluating fastText on the Turkish TTC-3600 dataset without using pre-processing step and presented the algorithm's performance. The research used the TTC-3600 dataset obtained from https://github.com/denopas/TTC-3600. Experimental results showed fastText was significantly better than other algorithms in terms of accuracy and robustness from a preprocessing point of view. The research was the first study to use the fastText method in categorizing Turkish texts.

In the research, Rachmi, (2017) contributed to the selection of features to increase the accuracy of the Support Vector Machine and find the parameter values to get the highest accuracy in sentiment analysis review of the delivery of goods as well as produce a classification of negative and positive results of the review appropriately. The author used Principal Component Analysis and Genetic Algorithms as optimizations to increase the accuracy of the Support Vector Machine method. The accuracy of the Support Vector Machine algorithm is 86%, after being optimized using the Principal Component Analysis and Genetic Algorithm the accuracy has increased to 97%.

In the research, Putra, et al (2020) contributed to testing efficiency by adding process clustering to the information retrieval system and comparing test results with the clustering method that had been used previously. This study implements term weighting with TF-IDF and doc2vec and the K-means method, then the document-matching query is simplified into a vector-matching query with cluster centroid vectors. The results of the efficiency test show that the information retrieval system using the k-means method can find news more quickly. The evaluation test (precision, recall, and f-score) shows that the information retrieval system search process with TF-IDF weighting is best done on 4G LTE queries with a threshold of 0.05 in a total of 750 clusters with an f score of 0.818.

3. Research Method

3.1. Research Flow Diagram

The research flow diagram steps start from the Microsoft Research Video Description Corpus (MRVDC) with Indonesian language datasets. It is necessary to group topics based on description text with preprocessing (case folding, tokenizing, stopword), then vectorize the text by fastText word embedding, then PCA will be used to reduce the number of features/columns after the text is embedded in FastText. after that it is necessary to do

clustering using k-means, the results of k-means clustering will be evaluated using the Silhouette Coefficient, then the evaluation results can be known which cluster values are good for use, and also the evaluation results value can be compared with or without PCA. Figure 1 describes the flow diagram of this research.

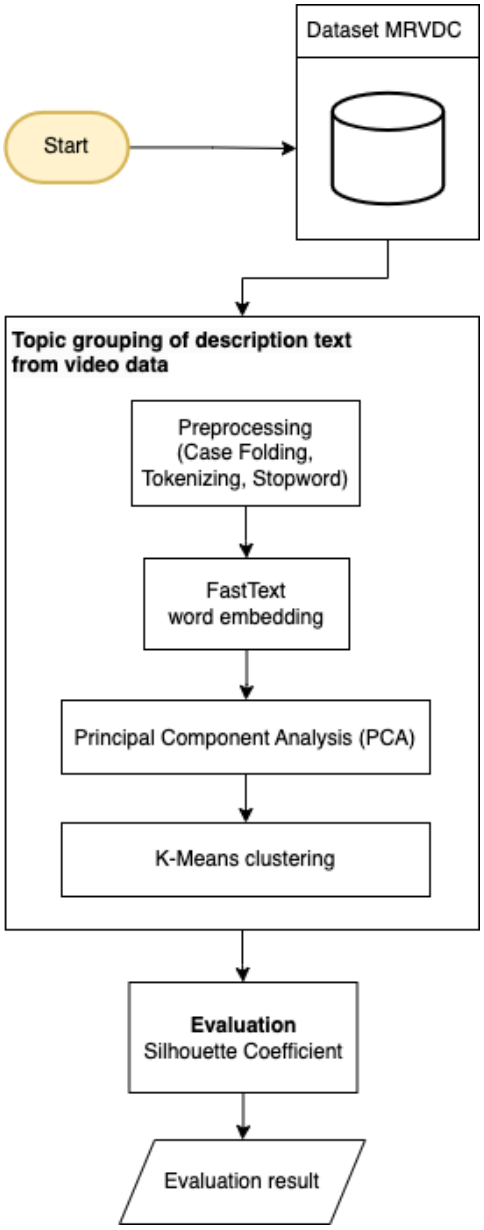

Figure 1. Research Flow Diagram

3.2. FastText

fastText is a toolkit developed by the Facebook Research Team to study effective word representation Joulin, et al (2016). One of the main contributions of the fastText embedding model is that the fastText algorithm takes internal word structure into account when learning word representations which are especially beneficial for morphologically rich languages like Turkish. Figure 2 describes the general architecture of the FastText.

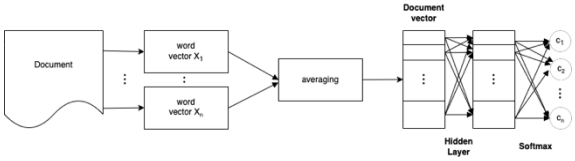

Figure 2. The general architecture of fastText

3.3. Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction method, it is also named the discrete Karhunen–Loève transform (KLT), and Hotelling transform singular value decomposition (SVD) and empirical orthogonal function (EOF). PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations (the principal components PCs) of the original variables with the largest variance. As per the number of the original variables there are as many PCs. The first several PCs explain most of the variance, so that disregarded the rest can be with minimal loss of information, for many datasets. Pang, et al (2008). To reduce the dimensionality of the huge data along with retaining as much information as possible in the original dataset, PCA is used.

To perform a Principal Components Analysis on data steps are. Smith, (2002).

a. Get image data: Suppose $X_1, X_2 \ldots X_M$ are represented as $N \times 1$ vectors
b. Compute the average of vector
$$\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i \quad (1)$$
c. Subtract the Mean: $\Phi i = x_i - \bar{x}$ (2)
d. Calculate the covariance matrix: Matrix A = [Φ], Φ2 …. ΦD] (N x M matrix) from this compute
$$C = \frac{1}{M}\sum_{n=1}^{m} \phi n \phi n^T = AA \quad (3)$$
e. Compute eigenvalues and eigenvectors of the covariance matrix.
$$c = \lambda_1 > \lambda_2 > \cdots > \lambda_N \quad \text{(eigenvalues) (4)}$$
$$c = u_1, u_{2,\ldots} u_n \quad \text{(eigenvectors)(5)}$$
f. Forming a feature vector: eigenvectors are order by eigenvalue, highest to lowest. This gives the components in order of significance. The eigenvector with the highest eigenvalue is the principle component of the data set. Feature vector is formed by choosing the highest eigenvalue. Smith, (2002).
g. Deriving new dataset: we have chosen the principal components (eigenvectors) to keep in our data and formed a feature vector; we simply take the transpose of the vector and multiply it on the left of the original data set, transposed. Smith, (2002).

Final data = row feature vector * row data adjust. PCA is a technique applied in many domains such as image compression, face recognition, patterns

recognition, eigenfaces, and text categorization and computer vision. Sachin, (2015).

3.4. K-means clustering

The K-Means algorithm starts with the formation of a cluster partition at the beginning and then iteratively repairs the cluster partition until there is no significant change in the cluster partition. Grouping that can be used is like non-hierarchical grouping which divides data into two or more groups. K-means is a group analysis method that leads to the division of N observation objects into K groups (clusters) and each observation object belongs to the group that has the closest mean (mean). The basic K-means algorithm is as follows:

a. Determine k as the number of clusters to be formed
b. Generates a random value for the initial cluster center (centroid)
c. Calculating the distance of each input data to each centroid using the Euclidean Distance formula until the shortest distance is found from each data to the centroid. Here's the equation (6):

$$d(xi, \mu j) = \sqrt{\sum(xi - \mu j)^2} \quad (6)$$

Where:
xi → criteria data,
μj → centroid in the *j* cluster

d. Group each data based on the proximity of each data based on its proximity to the centroid (smallest distance)
e. Update the centroid value. The new centroid value is obtained from the average cluster concerned using the equation (7):

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \epsilon Sj} xj \quad (7)$$

Where:
μj (t+1) → new centroid on iteration (t+1)
Nsj → lots of data on cluster Sj

3.5. Results Analysis

The topic grouping is based on description text in Microsoft Research Video Description Corpus (MRVDC) video data using fastText word embedding and compares the result with/without PCA reduction features, and clustering using K-means. The quality of topic grouping is evaluated using the silhouette coefficient. The software in this study is developed using Python programming language with Google Colabatory tools (Google Colab), and using some libraries such as NLTK (Natural Language Tool Kit), Scikit Learn, visualization library with matplotlib, and Gensim.

The Silhouette has three steps in its calculations. Following are the steps for calculating the silhouette coefficient according to Sholeh Hudin, et al (2018):

a. Calculating the average distance of objects with all documents in one cluster using the equation (8)

$$a(i) = \frac{1}{[A]-1} \sum j \epsilon_{A, J \neq i} \; d(i,j) \quad (8)$$

b. Then calculate the distance between objects with all documents between clusters using the equation (9)

$$d(i, C) = \frac{1}{[A]} \sum j \in C \; d(i,j) \quad (9)$$

c. Then calculate the silhouette value using the equation (10)

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (10)$$

4. Results and Discussion

This section will discuss the implementation results with the steps described in the previous chapter.

4.1. Preprocessing

The preprocessing used in the MRVDC dataset are case folding, tokenizing, stopword. Figure 3 on "*cleaned_text"* column describe the result of preprocessing used.

id_video	cleaned_text
11	lelaki berjenggot menggunting kertasnya
11	pria menggunting kertas
11	menggunting kertas
11	menggunting selembar kertas
13	anjing orang perempuan
...	...
1971	anak bermain
1972	seekor hewan makan
1973	anak menaburkan bunga
1974	sekumpulan orang makan
1975	anak bermain alat olahraga

Figure 3. The result of preprocessing used

4.2. FastText

The next step, "*cleaned_text"* from the preprocessing results above, needs to be vectorized using fastText word embedding. Figure 4.a and 4.b describes fastText word embedding results and Tabel 1 describes an example of embedded text with the "*makan"* keyword.

	0	1	2	3	4
0	0.258358	-0.808378	-2.785712	0.645353	0.394543
1	0.535369	0.069163	-2.509254	1.083359	0.601071
2	0.788125	0.804340	-2.840826	0.545587	-0.218872
3	2.035575	1.024577	-2.979814	0.958048	-0.076413
4	0.416541	-2.782364	0.337998	-0.412239	0.230827
...	...	...	...	...	...
43748	0.488412	-1.073198	1.095878	0.894342	-0.141339
43749	-0.801158	-5.835258	1.360255	2.183083	-0.231387
43750	2.249197	-1.525732	-1.285387	-0.244989	2.200906
43751	1.202480	-5.744614	1.525693	-1.896282	0.665683
43752	-2.419916	-2.396998	2.293241	3.066646	-0.433327

Figure 4.a. FastText word embedding results

95	96	97	98	99
-3.974369	1.408336	0.431874	-1.890903	-1.323124
-3.133759	1.100777	-1.047580	-2.475877	-1.747434
-3.164644	0.350076	-2.491153	-2.307849	-1.649091
-4.014539	1.094765	-3.361280	-1.308537	-1.352641
-0.439407	-1.176206	1.466181	2.588724	0.853627
...	...	...	...	...
-2.350640	-3.203275	0.753660	-1.055416	0.154231
-3.749290	0.141447	1.635334	1.038296	1.170709
-3.479937	-1.935302	-0.591071	-1.051412	-2.972410
-2.145702	-0.937151	-1.821381	1.908140	-1.018332
-0.353955	-3.762151	-1.174307	1.018953	0.946233

Figure 4.b. FastText word embedding results (continue)

Tabel 1. The example of embedded text

Text on MRVDC dataset	Embedded text
mamakan	0.967
makakan	0.914
dimakan	0.906
memakan	0.839
pemakan	0.817
makanlah	0.807
memakanlah	0.757
makanan	0.755
lupakan	0.755
iakan	0.745

4.3. Principal Components Analysis

Principal Components Analysis (PCA) is used to reduce the number of dimensions (features or columns). In this study, PCA is used to reduce the number of features/columns in word embedding fastText from 100 features to 50 features. Table 2 describes the PCA result.

Tabel 2. The PCA result

Features of word embedding fastText	Reduction features with PCA
100	50

4.4. K-Means clustering

Clustering based on the description text of the MRVDC dataset is performed in dimensionality reduction experiments with/without PCA. The experiment of k used are k = 10, k = 20, k = 40, k = 60, k = 80, k = 100, k = 120, k = 140, k = 160, k = 180, k = 200, then all of k is evaluated using the Silhouette coefficients. At this clustering stage, it takes a total of 32 minutes to process.

The results of clustering the MRVDC dataset have 180 clusters ranging from cluster 0 to 179. Table 3 shows the results of clustering the MRVDC dataset in 1 cluster which has a minimum of 26 description texts, a maximum of 1142 description texts, and an average of 243 description texts.

Tabel 3. The result of clustering MRVDC dataset.

Min	Max	Mean
26	1122	243

4.5. Evaluation

This evaluation test uses a silhouette score by including 11 times the cluster value (k) and also compares results with or without PCA. The silhouette score without PCA can be seen in Figure 6 and the silhouette score with PCA can be seen in Figure 7.

	K	Silhouette Score
0	10	0.087358
1	20	0.106640
2	40	0.103894
3	60	0.108685
4	80	0.111303
5	100	0.116219
6	120	0.118558
7	140	0.118529
8	160	0.122037
9	180	0.123115
10	200	0.120100

Figure 6. Result without PCA

	K	Silhouette Score
0	10	0.101947
1	20	0.100056
2	40	0.111714
3	60	0.110191
4	80	0.118740
5	100	0.121709
6	120	0.116226
7	140	0.121098
8	160	0.119345
9	180	0.122148
10	200	0.121542

Figure 7. Result with PCA

5. Conclusion

From the tests carried out in the previous chapter it can be concluded that clustering for grouping topics using k-means clustering can be done on the Indonesian dataset of Microsoft Research Video Description Corpus (MRVDC), the system can group topics using the k-means clustering algorithm. Topic grouping can be done based on the description text of each video data Microsoft research video description corpus (MRVDC).

The analysis result shows that in the several clusters used, the optimal value is obtained $k = 180$ with the resulting silhouette score without PCA of 0.123115. From these results, it can be concluded that K-means with PCA do not make the silhouette score evaluation better. So, it is recommended to use K-means clustering without PCA on the Indonesian dataset of Microsoft Research Video Description Corpus (MRVDC) because the results are better. And the optimal clustering results in this study can be used in future for video data retrieval systems on the Indonesian language MRVDC dataset.

References:

Feldi, M., Solahuddin, S., & Yuswanita, L. (2021). Implementasi Content Based Video Retrieval Menggunakan Speeded-up Robus Features (Surf). *Sintaksis*, *1*(1), 57–75.

Hediyati, D., & Suartana, I. M. (n.d.). *Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro*.

Hidayati, R., & Harjoko, A. (2012). Video Retrieval Berdasarkan Teks dan Gambar. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *7*(1), 77–88.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *ArXiv Preprint ArXiv:1612.03651*.

Kuyumcu, B., Aksakalli, C., & Delil, S. (2019). An automated new approach in fast text classification (fastText) A case study for Turkish text classification without pre-processing. *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, 1–4.

Multi Fani, S., & Santoso, R. (n.d.). *PENERAPAN TEXT MINING UNTUK MELAKUKAN CLUSTERING DATA TWEET AKUN BLIBLI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN K-MEANS CLUSTERING*. *10*, 583–593. https://ejournal3.undip.ac.id/index.php/gaussian/

Nurdin, A., Anggo, B., Aji, S., Bustamin, A., & Abidin, Z. (2020). PERBANDINGAN KINERJA WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT PADA KLASIFIKASI TEKS. *Jurnal TEKNOKOMPAK*, *14*(2), 74.

Nurhadi, R. A., Rahutomo, F., & Harijanto, B. (2016). Pengembangan Data Uji Sistem Komputasi Kemiripan Teks Secara Semantik Berbahasa Indonesia. *Seminar Informatika Aplikatif Polinema*.

Pang, Y., Yuan, Y., & Li, X. (2008). Effective feature extraction in high-dimensional space. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *38*(6), 1652–1656.

Puspitaningrum, D., Sari, D. K., & Susilo, B. (2014). Dampak Reduksi Sampel Menggunakan Principal Component Analysis (PCA) Pada Pelatihan Jaringan Saraf Tiruan Terawasi (Studi Kasus Pengenalan Angka Tulisan Tangan). *Pseudocode*, *1*(2), 83–89.

Putra, Y. P., Yunhasnawa, Y., & Rahutomo, F. (2020). Evaluasi Kmeans Clustering Pada Preprocessing Sistem Temu Kembali Informasi. *Seminar Informatika Aplikatif Polinema*.

Rachmi, H. (2017). *Penerapan principal component analysis dan genetic algorithm pada analisis sentimen review pengiriman barang menggunakan algoritma support vector machine*.

Rahutomo, F., & Ayatullah, A. H. (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*, 319–326.

Sachin, D. (2015). Dimensionality reduction and classification through PCA and LDA. *International Journal of Computer Applications*, *122*(17).

Sholeh Hudin, M., Fauzi, M. A., & Adinugroho, S. (2018). *Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus: Universitas Brawijaya)* (Vol. 2, Issue 11). http://j-ptiik.ub.ac.id

Smith, L. I. (2002). *A tutorial on principal components analysis*.