e-ISSN: 2614-8404

A Grouping of Song-Lyric Themes Using K-Means Clustering

Dionisia Bhisetya Rarasati[1*)]
[1]Program Studi Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia
email: [1]dionisia_rara@yahoo.com

***Abstract* – One of the automatic way of theme grouping that can be used is K-Means Clustering. In this research, the song theme is taken from the text of song lyrics. The aim of this study is developing a system that can automatically group the song lyric theme and know the accuracy level of the grouping. The process stage is started with the data processing or text processing called as text mining. In text mining, there are some processes. First, the text operation. The text operation consists of tokenizing, stopword, steeming, and word weighting then can be processed using K-Means clustering. In clustering process, it consists of initial centroid initialization uses Variance Initialization, next counts the centroid distance on the data using Euclidean distance until get the proper grouping accurately. The accuracy counting uses confusion matrix. The next step to see the suitability system that has been made, new data is added which then is processed by a system. After that, it can decide the new data is classified into one specific theme. From the research that has been conducted as case study in Masdha Radio Yogyakarta, total data available 400 and divided into four clusters. The clusters consist of love cluster, friendship cluster, religion cluster, and fighting cluster. The result of research song lyric grouping based on the theme works well with 93.25% accuracy for the unique word frequency numbers 121 maximum and unique word 0 minimum.**
Keywords – K-Means clustering, Text Operation, Variance Initialization, Confusion Matrix.

I. INTRODUCTION

Themes are needed to explain feelings or emotions that can be arranged through song lyrics. Many song themes are known to the public, such as love, friendship, struggle, religion, and other types of themes. So to avoid placing a song theme that is not following the song lyrics, a method is needed to determine the song theme according to the song lyrics. The words in the song lyrics are processed so that later they can be classified according to the suitability of the theme. Word processing or text processing is called text mining. After the word processing is complete, it is necessary to categorize or cluster the theme of the song lyrics, one of which is using the K-Means method. K-Means clustering is one method that is often used because it has high accuracy and easy-to-understand processing so that it is considered quite efficient, as shown by its complexity O (tkn), provided that n is the number of data objects, k is the number of clusters formed and t the number of iterations. Usually, the values for k and t are much smaller than the values for n. Also, the iteration of this method will stop under local optimum conditions (Williams, 2006)[1]. So the main problem to be answered in this study is to use the K-Means clustering method, can the system automatically classify song lyric themes properly, so that we can find out which song lyrics are appropriate to the theme grouping.

II. LITERATURE REVIEW

The data used for research in this paper is data obtained from Masdha Radio Yogyakarta, divided into four themes, namely love, struggle, religion, and friendship. Below is a literature review, among others:

2.1 Information Retrieval

Information Retrieval is a set of algorithms and technologies for processing, storing, and retrieving information (structured) in a large data collection (Manning, 2008)[2]. The data used can be in the form of text, tables, images, and videos. A good IR system allows the user to determine quickly and accurately whether the contents of the document received to meet their needs.

2.1.1 Text Operation

Text mining has the definition of mining data in the form of text where data sources are usually obtained from documents, and the goal is to find words that can represent the contents of the document so that analysis of the relationship between documents can be carried out. (Harlian, 2006)[3]. The work process of text mining has several stages, namely:

In-text mining goes through a text preprocessing process, which is a process that is applied to text data that aims to produce numerical data. The stage in this process, namely, the first stage of tokenizing, namely the stage of cutting the input string based on each word that composes it. So that it changes all letters in the document to lowercase and characters other than letters are removed. The stopword stage is the stage of filtering important words from the tokenizing results, where irrelevant words are discarded. The stemming stage is the stage of finding the root word of each stopword result. Furthermore, the stage of combining words (synonyms) is a form of language that has the same meaning or the same as other language forms. The synonym process will be carried out when there are different words but have the same meaning, to minimize the number of words in the system, without eliminating the number of frequencies. Weighting words by calculating the term frequency multiplied by the inverse document

 e-ISSN: 2614-8404

frequency to find the weight (Wij), with the following formula:

$$w_{ij} = tf_{ij} \times idf_j$$

$$w_{ij} = tf_{ij} \times log\ (D/df_j)$$

The next stage is normalization using the Z-Score. Z-score is a normalization method based on the mean (average value) and standard deviation of the data. This method is very useful if we do not know the actual minimum and maximum values of the data. (Martiana. 2013)[4]. Below is the Z-Score normalization formula.

newdata = (data-mean)/std

The next stage enters the Variance Initialization process. Variance Initialization is a multivariate analysis technique that functions to distinguish the mean of more than two groups of data by comparing the variance. Analysis of variance is included in the category of parametric statistics. (Ghozali, 2009)[5]. Below is the variance initialization formula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Explanation:
xi = The x with the value i
$\bar{x}$ = Average
n = sample size
s^2 = Variant

2.2 *K-Means Clustering*

K Means clustering is a popular method used to obtain a description of a set of data by revealing the tendency of each individual data group to group with other data individuals. The tendency of grouping is based on the similarity of the characteristics of each individual data available. The basic idea of this method is to find the center of each possible data group and then group each individual data into one of these groups based on the distance. (Turban, dkk, 2005)[6]. The closer the individual data is, say X1 to one of the centers of the existing group, call it A, the more clear it is that X1 is a member of the group centered on A, and the clearer it is that X1 is not a member of the other groups. (illustration can be seen in the image below). Quantitatively, this is shown by the fact that d1A, which is the distance from X1 to A, has the smallest value when compared with d1B and d1C.

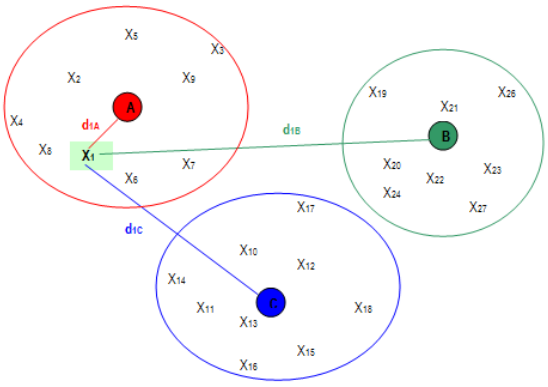

Figure 1. Illustration of Determining Group Membership Based on Distance (Turban, dkk, 2005)

2.3 *Confusion Matrix*

At the testing stage for the accuracy of a clustering, what is processed is the labeling of each cluster. Labeling is obtained from case studies, where label one is love, label two struggles, label three is religion, and label four is friendship. After obtaining four labels in each cluster, it is processed using the Confusion Matrix test.

The Confusion Matrix contains actual and predictable information (Kohavi dan Provost, 1998), where system performance can be evaluated using the data in the matrix.

The table below shows the confusion matrix for the two classess (Kohavi dan Provost, 1998)[7] :

Table I. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Explanation:

- a is the number of correct predictions that the example is negative.
- b is the number of incorrect predictions that an example is positive.
- c is the number of incorrect predictions that an example is positive
- d is the number of correct predictions that the example is positive.

Accuracy is the number of correct predictions, with the following formula:

$$AD = \frac{a+d}{a+b+c+d}$$

III. RESEARCH METHODOLOGY

At the Research Methodology stage, several processes were carried out, among others:

3.1 *Preprocessing*

Before the data mining process is carried out using clustering, the data used first goes through the preprocessing stage. The processing stages are carried out:

1.*Tokenizing*

In the tokenizing process, the process that occurs is the breaking of sentences into individual words, the words are changed to lowercase and eliminating characters that are not included in the word.

2. *Stopword*

After experiencing the tokenizing process, the next step is the stopword process. Stopword is an important step for filtering words, so that irrelevant words can be discarded. Irrelevant words have their word dictionary, so the system checks the words that appear in the song lyric document against the stopword word dictionary. If a word in the song

 e-ISSN: 2614-8404

lyric data exists with a stopword word dictionary, the word is discarded.

3. *Stemming*
The next process is the stemming process, which looks for the basic words from the song lyric data obtained.

4. Word Weighting
In the word weighting process, the steps taken are to give the frequency value of a word as a weight, which can later be processed in K-means clustering.

5. Word Combination
The process of combining words is carried out when there are different words but have the same meaning, then they can be combined into one word, without changing the frequency value.

6.Normalisasi *Z-Score*
After finding the weighting, the next step is the normalization process using the z-score, which functions so that the weighted words can be compared with each other.

3.2 *Clustering* and Accuracy
After carrying out the text operation process, the next grouping step is using K-Means Clustering. Initial centroid = 4 centroid, four centroids were chosen because they were limited by the grouping of topics that were assumed to be four groups/clusters, namely love, struggle, friendship, and religion, then for centroid determination using variance initialization, the largest variance was sought, then the lyrics were sorted using the variance results. the largest and the sorted lyrics are divided into four parts, each part of the group/cluster is looked for the average/mean, then that is the initial centroid. After finding the initial centroid, the next step is to find the closeness between the centroid and each of the lyrics using the Euclidean distance.
In testing accuracy using a confusion matrix. Confusion Matrix is used to find out how much success the system is. The step is to create class groups, namely the actual class and the prediction class. The actual class is the class that will be checked against the prediction class.

3.3 *User Interface*
Matlab version 8.0.0.783 (R2012b) is a tool for making song lyrics grouping systems using the K-Means clustering method. A system was formed as a means of preprocessing to determine the accuracy of data grouping with ¬K-Means clustering. The system formed can immediately display the results of the entire process. The following is an example of how the entire system has been formed.

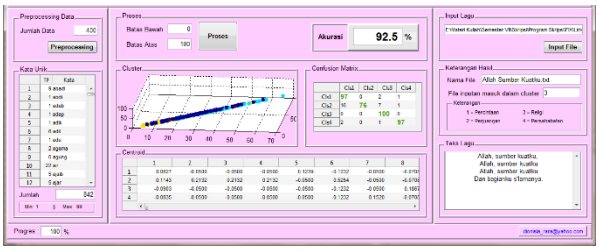

Figure 2. User Interface Implementation

IV. RESULTS AND DISCUSSION
After implementing it, it can help to analyze the grouping of themes in the song lyrics. The analysis is carried out on all song lyric data so that it can be seen the song lyric group in all the data. The stages that have been passed starting from the information retrieval stage which consists of tokenizing to separate the lyric data into individual words and removing punctuation marks, stop words to eliminate words that have no meaning (hyphens), stemming to eliminate affix words into basic words then match back to the stopword dictionary. The next stage is the word weighting process, which functions to calculate the frequency of appearance of words in each lyric data so that words that appear more frequently in lyric data are considered more important. Then enter the k-means clustering process, the stage for this process is the initialization of the centroid based on the largest variant, then the lyric data is sorted according to the largest variant, then the lyric data is grouped into four parts, and each part is searched for the average/mean, The next step is normalization using the Z-score to get closer to the values that span too far, after that comparing the centroid with the lyric data using the Euclidean distance proximity, the cluster group is formed. The last stage is the stage of calculating accuracy, calculating accuracy using confusion matrix where the prediction data is compared with the actual data, then the total of the prediction data results and the corresponding actual data is divided by all the many words contained. To measure the level of success of this writing is to experiment. The following are the steps for the experiment:

1. Determine the number of clusters = 4, according to a predetermined theme.
2. Song lyric data = 400.
3. Enter the lower and upper limits that ultimately determine the level of accuracy.
4. Performed each experiment 36 times.

Table II. Experiment with an upper limit of 121-70 and a lower limit of 0-5

Lower limit		**0**	**1**	**2**	**3**	**4**	**5**
Upper limit	**121**	93.25%	88.75%	80.25%	78.25%	77.50%	76.50%
	110	92.00%	85.25%	82.50%	80.50%	74.50%	73.25%
	100	92.00%	85.25%	82.50%	80.50%	74.50%	73.25%
	90	84.50%	83.50%	73.50%	69.25%	69.25%	64.50%
	80	84.50%	83.50%	73.50%	69.25%	69.25%	64.50%
	70	84.50%	83.50%	73.50%	69.25%	69.25%	64.50%

The table above is the result of accuracy with 36 experiments, the upper or lower limits are the limits used to limit the total term frequency results contained in unique words. If the user enters an Upper Limit of 121, the maximum unique word is 121, then with an upper limit of 121, the program will process from the Upper Limit of 121, vice versa with the Lower Limit. The lower limit is the minimum limit. So that if the user enters a Lower Limit of 0, the program starts processing from a Lower Limit of 0.
Based on the results of the experiment above, the graph below is a graph with an upper limit = 121 and an upper limit = 70 with each lower limit = 0-5:

e-ISSN: 2614-8404

a. Upper Limit = 121 with Lower Limit= 0-5

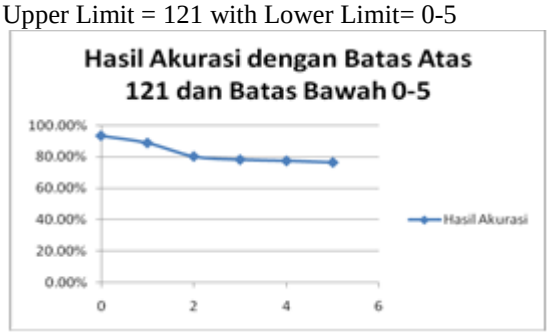

Figure 3. Upper Limit Accuracy Results 121

In the Figure above it is known that the user inputting the Upper Limit is 121 and the Lower Limit is 0-5, with the highest accuracy at a value of 93.25% when the Lower Limit = 0, while for the lowest accuracy the value 76.50% is when the Lower Limit = 5.

b. Upper Limit = 70 with Lower Limit = 0-5

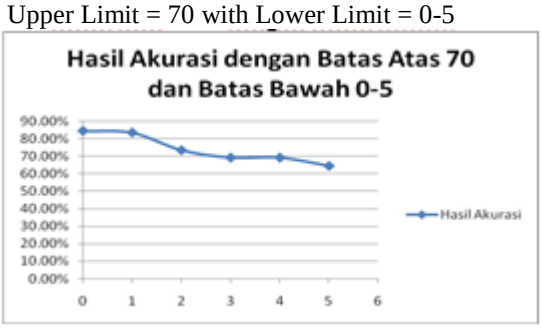

Figure 4. Upper Limit Acurcary Results 80

In the Figure above it is known that the user inputting the Upper Limit is 70 and the Lower Limit is 0-5, with the highest accuracy at 84.50% when the Lower Limit = 0, while the lowest accuracy value is 64.50% when the Lower Limit = 5. The discussion of the results of the research and testing obtained is presented in the form of theoretical descriptions, both qualitatively and quantitatively. The results of the experiment should be displayed in graphs or tables. The discussion must be following the research methodology being carried out.

V. SUMMARY

Based on the results of the analysis of theme grouping based on song lyrics with song lyric data amounting to 400 and cluster = 4, the following conclusions are obtained:

1. The K-Means clustering method can be used to group data in the form of text.
2. The results of grouping the entire data depend on the number of frequencies.
3. Based on Experiment Table 4.8, it can be concluded, which has the highest level of accuracy is at, the upper limit = 121 the lower limit = 0, namely 93.25%. Meanwhile, those with the lowest level of accuracy in the experiment with an upper limit of 70 and a lower limit of 0 to 5, namely the upper limit = 70 and the lower limit = 5 with an accuracy rate of 64.5%. The upper limit or lower limit is the limit used to limit the total term frequency results contained in unique words. If the user enters an Upper Limit of 121, the maximum unique word is 121, then with an upper limit of 121, the program will process from the Upper Limit of 121, vice versa with the Lower Limit. The lower limit is the minimum limit. So that if the user enters a Lower Limit of 0, the program starts processing from a Lower Limit of 0.

BIBLIOGRAPHY

[1] William,Graham., 2005, *Data Mining Algorithms Cluster Analysis*, Journal of Introduction Requirements Measuring Similarity.
[2] Manning, C. D., Ragvana, P., Schütze, H., 2008, *Introduction to Information Retrieval*, Cambridge University Press.
[3] Harlian, Milkha., 2006, Jurnal, *Melakukan Analisa Keterhubungan Antar Dokumen.*
[4] Martiana, E., 2013, *Data Preprocessing*, Institut Teknologi Surabaya.
[5] Ghozali, Imam., 2006, Statistik Nonparametrik, Semarang: Badan Penerbit UNDIP.
[6] Turban, E., Aronson, J. E., Liang, T. P., 2005, *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset
[7] Kohavi dan Provost., 1998, *Confusion Matrix.*

.