Clustering box office movie with Partition Around Medoids (PAM) Algorithm based on Text Mining of Indonesian subtitle

A D Alfarizy\textsuperscript{1}, Indahwati\textsuperscript{1} and B Sartono\textsuperscript{1}

\textsuperscript{1}Department of Statistics, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, Indonesia

Email: athikah.dalfarizy@gmail.com, indah.stk@gmail.com, bagusco@apps.ipb.ac.id

Abstract. Indonesia is the largest Hollywood movie industry target market in Southeast Asia in 2015. Hollywood movies distributed in Indonesia targeted people in all range of ages including children. Low awareness of guiding children while watching movies make them could watch any rated films even the unsuitable ones for their ages. Even after being translated into Bahasa and passed the censorship phase, words that uncomfortable for children to watch still exist. The purpose of this research is to cluster box office Hollywood movies based on Indonesian subtitle, revenue, IMDb user rating and genres as one of the reference for adults to choose right movies for their children to watch. Text mining is used to extract words from the subtitles and count the frequency for three group of words (bad words, sexual words and terror words), while Partition Around Medoids (PAM) Algorithm with Gower similarity coefficient as proximity matrix is used as clustering method. We clustered 624 movies from 2006 until first half of 2016 from IMDB. Cluster with highest silhouette coefficient value (0.36) is the one with 5 clusters. Animation, Adventure and Comedy movies with high revenue like in cluster 5 is recommended for children to watch, while Comedy movies with high revenue like in cluster 4 should be avoided to watch.

Keywords: gower similarity, partition around medoids, text mining, silhouette coefficient

1. Introduction

Hollywood is the biggest movie industry in the world with the fastest growth and the highest income from distribution in all countries. Motion Pictures Association of America [1] recorded 2015’s global income from all Hollywood box office movies released in each country over the world combined, reached the number of US$38 billion. The highest portion is contributed by Asia Pacific region, including Indonesia which reached US$ 250 million at 2015, topping all South-East Asia countries. It is supported by the fact that Indonesia has huge number of population spread in wide areas which almost all of the areas has cinemas.

Movies distribution in Indonesia targeted all people from any range of age and gender, including children. They can be accessed through many devices and any places like in cinemas, local television channel, DVD, and even smartphones. All these devices are on daily life and can be accessed easily by children. The movies has information that can reach out and influence the people who watches them, so it’s necessary for children to have adults as a company while watching movies in order that they would not have wrong perception of the information from the movies.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
However, adults’ awareness to the importance of guiding children while watching movies is yet very low though it is very crucial. As a matter of fact, children who are roaming alone or with group of same age friends can be spotted easily in cinemas. Even parents who should guide their children, bring them along to watch unsuitable movies for their children’s age. This resulted in children who can watch any kind of movies which are not appropriate for them. We know that movies that are distributed in Indonesia should have passed the censorship phase and translated to Indonesian. However there are still words or pictures that troubling for children to watch. So, adults need references to choose suitable movies for the children to watch.

Nowadays, most of the natural information retrieval is texts, like social media, movie subtitles, messages and any other. Texts are unstructured data that can’t be processed by machine, because they can only be read by human (human learning). Text need to be treated so that it can change into structured one and read by the machine to find a lot of information from it. This method can be done by a method called text mining, which can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools to extract useful information from data sources through the identification and exploration of interesting patterns [2].

There are already many researches in any areas applying text mining method, [3] used text mining to movie rating classifications depended on the keywords in English, that are used in the movies and other attributes such as genres. They focused on the movies which have rating PG, PG-13 and R. They used text mining as preprocessing step to create prototype model based on decision tree (J48).

In this research, text mining will be applied to extract information from Indonesian language subtitle of 624 box office movies from 2006 until first half of 2016, to cluster the movies. Some words from the outcome of the method, will be grouped into three groups of words which are bad words, sexual words and terror words. These attributes and other attributes like IMDb user rating, genres and revenue, will be used to cluster movies using Partition Around Medoids (PAM) or K-Medoids algorithm. This algorithm is used since it is robust with outlier [4], while assumed that data consists of mixed type data, [5] gower similarity is suitable to use in this research. Then after some numbers of cluster is created, the best number of clustered will be chosen based on silhouette coefficient value [6].

2. Backgrounds
In this section, we present the background of the works as following

2.1. Text Mining
Text mining is a method that can be defined as a process to extract useful information from groups of documents which consist of unstructured data, which is text. This method processes those unstructured data which can only be read by human into structured data, through some identification and exploration of some patterns.

According to [2] the most important process from text mining is the pre-process, which focused on identification and extraction representative feature for natural language documents to make the structured data. The text mining process usually needs input text making based on grammar of the language used in the documents [4]. This would be then followed by extraction of already structured data, evaluation and interpretation. The text mining usually used are text categorization, text clustering, text classification and any other analysis process.

2.2. Gower Similarity
Almost all of matrix proximity measurement can only be used for one type of data, but condition of data used in many researches is mostly consisting of mixed type data. Gower similarity is proximity matrix measurement that suitable for mixed type data [5].

Gower similarity coefficient $s_{ij}$ compares two observations $i$ and $j$, and is defined as follows:
\[ S_{ij} = \sum_{k=1}^{n} s_{ijk} \omega_{ijk} / \sum_{k=1}^{n} \omega_{ijk} \]  

(1)

where \( s_{ijk} \) is the contribution provided by \( k \)-th attribute and \( \omega_{ijk} \) is the weight for \( k \)-th attribute between two observations \( i \) and \( j \). If \( s_{ijk} \) can be defined or there’s no missing data, we can use \( w_{ijk} = w_k \), where one value of weight is for one attribute. Usual combination of the weights that is used is 1, for all attributes. If there’s missing data on dataset, then \( w_k = 0 \) for \( k \)-th attribute. Weights can be specified beforehand and the value can be assigned according to user.

Gower similarity concept is for each attribute with each type of data will use suitable measurement depends on the data type of attribute. Then linearly combined using weight for each attribute to calculate final distance. If attribute’s type of data is continuous, following equation is used:

\[ s_{ijk} = 1 - \left| x_{ik} - x_{jk} \right| / R_k \]  

(2)

where \( x_{ik} \) is \( i \)-th observation of \( k \)-th attribute, \( x_{jk} \) is \( j \)-th observation of \( k \)-th attribute and \( R_k \) is range of \( k \)-th attribute.

If attribute’s type of data is nominal, the value of \( s_{ijk} \) is 1 if \( x_{ik} = x_{jk} \) and valued 0 if \( x_{ik} \neq x_{jk} \). Or in other words, \( s_{ijk} = 1 \) if \( i \) and \( j \) have same state for \( k \)-th attribute and \( s_{ijk} = 0 \) if \( i \) and \( j \) have different state. If attribute’s type of data is binary, Gower defines as the table below:

| Value in \( k \)-th attribute | \( i \)-th observation | \( j \)-th observation |
|-----------------------------|------------------------|------------------------|
| \( s_{ijk} \)             | +                       | -                       |
| \( \omega_{ijk} \)       | 1                       | 0                       |

where “+” denotes that present and “-” denotes that absent, for the observation in \( k \)-th attribute.

2.3. Partition Around Medoids (PAM) Algorithm

Clustering is the process of dividing a set of data into some subsets, which called cluster, so that objects in a cluster are similar to each other and dissimilar with other objects in other clusters [4]. One of the major fundamental categories is partitional methods, which creates an initial partitioning then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Partition Around Medoids (PAM) or usually called as K-Medoids is one of the partitional clustering methods that choose one of the object (medoids) as initial. Medoids has similar concept as mean or centroid, but Medoids is always an object from dataset. This method is more robust to outliers not like K-Means, because it minimize the absolute error.

Below is the algorithm of Partition Around Medoids (PAM) [4]:

1. Arbitrarily choose \( k \) objects as initial representative objects of each cluster
2. Assign remaining objects to the cluster with nearest representative objects
3. Randomly select nonrepresentative objects as medoids candidate, \( O_{\text{random}} \)
4. Calculate total cost, \( S_{\text{random}} \), of the distance between remaining objects with initial object, \( O_j \), and total cost, \( S_j \), of the distance between remaining objects with \( O_{\text{random}} \)
5. If \( S_{\text{random}} < S_j \), then swap \( O_j \) with \( O_{\text{random}} \) as new medoids of \( k \) representative object
6. Repeat step 2-5 until there is no change

2.4. Silhouette Coefficients

Silhouette coefficient is a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its
cluster [6]. Silhouette value used to measure how similar an object to its own cluster compared to other cluster. The value range from -1 to 1, where high value indicates that the object is well matched to its own cluster and poorly matched to other cluster.

After the data have been clustered by any methods into $k$ cluster, we can evaluate with silhouette coefficient defined by the following equation:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$ (3)

where $a(i)$ is average of dissimilarity between $i$-th object with other object in its own cluster, $b(i)$ is the lowest average of dissimilarity between $i$-th object with other object in its other clusters. Cluster with the lowest average of dissimilarity is the neighbor of cluster where $i$-th object is and $s(i)$ is silhouette coefficient for $i$-th object. Coefficient for each object will be calculated and the average of silhouette value of entire dataset will be used to compare, to choose the best number of cluster.

3. Methodology

The most important and the key to do text mining is the data preprocessing. Indonesian language subtitle is used as it contains the words used in the movie. Since even after get through censorship process, there are still words that are uncomfortable for children to watch.

First, we collect the box office movies data from IMDb and subtitles data for each movie from various places like Subscene, Open Subtitle and other site that allow the user to free download the content. There are many records in the IMDb and we choose some of the records to help us clustering the movie. They are IMDb user rating, which the value given by the user to score movie after they watched it, Revenue, which is the income of a movie and genres of movies. Each movie can have more than one genre.

We manually extract these records of the movies from the website and inspect them. We strain the movies that released from 2006 until first half of 2016. The movies that later be used are box office movies with number of revenue more than US$ 50 million. Then, because each movie has possibly more than one genre, we divided each kind of the genre into each attribute consist of binary data.

Second is the preprocessing data of the Indonesian subtitles using text mining method, to prepare text data into structured data. The step of preprocessing, can be vary according to the needs of the user. In this research, follow the preprocessing step from [7]. Figure 1 shows the step of preprocessing.

![Figure 1. Step of preprocessing Indonesian subtitle texts](image)

From group of documents sources, tokenizing step is to remove any objects beside letter, like number and punctuation, and transform all the letter into lowercase. For each sequence of letters will be called as token. Then we do the filtering, to remove any tokens that are included in stopwords. Stopwords are the words usually refered to the most common words in a language. Any group of words can be chosen as stopwords for a given purpose. In this research, we are using group of stopwords listed by [8].

The next step is stemming, to transform any tokens into their root words. In Indonesian language there are many conditions to transform a word into its root word according to the dictionary, like remove suffix or prefix. So we are using algorithm of stemming Indonesian language developed by
After all tokens in its root word formed, we count the frequency of each word. Then it will be divided into three attribute which each attribute consist of group of chosen words. The three groups are group of words related to bad words, sexual words and terror words, the list of words is given in the Appendices. Then these attributes will be integrated with previous data containing information of the movies and can be analysed.

4. Results and Discussion

Based on the data obtained from previous step, a set containing 624 movies data is analysed. They included nine genres of popular movies which are Sci-Fi, Family, Fantasy, Adventure, Animation, Comedy, Drama, Horror and Action. The total frequency of appearance of three groups of words are calculated. Some most frequent words are shown in Figure 2.

![Figure 2. Most frequent words from group of words](image)

Figure 2 shows the frequency of some words from the group words that appear at least 300 times in total from all movies. Two words of the highest frequency are words from terror words group, which are “bunuh” and “mati”, that appear almost 6000 times. The third most used word is “sial” from bad words group that appear almost 5000 times. It can be seen that beside the censorship process, there are still many uncomfortable words making appearance in movies that can influence the filmgoers, especially children, if they watch them without any guidance from adults.

We need to divide these dataset into some clusters that has different characteristics, Partition Around Medoids (PAM) algorithm is used. This method is used to cluster the dataset into some variation number of cluster. Variations of number of clusters obtained from 2 clusters until 10 clusters. Then, silhouette coefficient is used to compare the effectiveness of the method in each variations. It can be seen in figure 3, that number of cluster which has the highest score of silhouette coefficient is the data with 5 clusters with the value of 0.29. So, this 5 clustered data is the one that will be used for the next analysis.

Since the dataset are high dimensional data, to help visualize the distribution of the clusters into two dimensional data using t-distributed stochastic neighbor embedding (t-SNE) [10]. It can help reduce the high dimensional data into two or three dimensions, which can then be visualized in a scatter plot. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points.
From figure 4, it can be seen that objects is grouped into 5 different clusters which represented in different colors. Each cluster has its own area consist of data objects included in their cluster. Though there are some objects that positioned quite far from the area of its cluster. It is reflected in silhouette value as evaluation for the making of cluster with the algorithm, 0.36, that is still quite far from value of 1. It can be happened because movies data have huge variation of information that need more advanced method to cluster, such as two-steps clustering or any other, or the number of clusters is too few. But just by seeing the plot, the clusters are quite distanced from another cluster and quite similar with other objects in its own cluster.

Here are the list of characteristics for each clusters shown in table 2.
Table 2. Characteristics for each clusters

| Characteristics               | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|-------------------------------|-----------|-----------|-----------|-----------|-----------|
| Number of movie               | 162       | 40        | 193       | 143       | 86        |
| Average Revenue (US$ Million) | 113.617   | 70.019    | 172.476   | 93.719    | 153.743   |
| IMDb Rating                   | 7.03      | 6.09      | 6.71      | 6.23      | 6.61      |
| Movie with genre Sci-Fi       | 10.49%    | 15.00%    | 30.57%    | 0.00%     | 0.00%     |
| Movie with genre Family       | 10.49%    | 0.00%     | 11.40%    | 13.29%    | 30.23%    |
| Movie with genre Fantasy      | 9.26%     | 25.00%    | 18.65%    | 8.39%     | 2.33%     |
| Movie with genre Adventure    | 14.20%    | 2.50%     | 81.87%    | 2.10%     | 76.74%    |
| Movie with genre Animation    | 0.62%     | 0.00%     | 8.29%     | 0.70%     | 89.53%    |
| Movie with genre Comedy       | 11.73%    | 7.50%     | 7.25%     | 100.00%   | 93.02%    |
| Movie with genre Action       | 97.53%    | 2.50%     | 6.74%     | 18.18%    | 2.33%     |
| Movie with genre Horror       | 12.35%    | 12.50%    | 89.64%    | 16.78%    | 3.49%     |
| Movie with genre Drama        | 3.09%     | 95.00%    | 1.55%     | 0.00%     | 0.00%     |
| Average Bad words             | 14.57     | 19.88     | 12.37     | 33.69     | 7.47      |
| Average Sexual words          | 1.86      | 0.83      | 0.65      | 17.50     | 0.27      |
| Average Terror words          | 30.04     | 30.60     | 36.56     | 26.81     | 16.19     |

From the characteristics above, shown that all cluster has similar average of IMDb user rating, so this attribute is not too significant to characterize this dataset. Cluster 4 has the highest average of bad words and sexual words appearance. It consist of movies with average revenue US$93 million and 100% of movies with Comedy genre. Movies with comedy genre itself actually vary from adult comedy until global comedy. Since the appearance of sexual words are the highest from other cluster combined, it can be assumed that the comedy movie in cluster 4 is mostly adult comedy.

In cluster 2, the characteristics of mostly the movie in this cluster, about 95%, have genre Drama and have income average about US$70 million or not quite popular. There are still many appearances of bad words and terror words in the movies in these cluster. So the movies in these cluster are still quite unsuitable for children.

In cluster 3, the characteristics of mostly the movie in this cluster, about 89%, have Horror genre and 82% have Adventure genre and have income average about US$172 million or popular movies. There are still many appearances of bad words and terror words in the movies in these cluster. So the movies in these cluster are still quite unsuitable for children.

Cluster 1 mostly consisted by movies with huge income with average US$113 million, it means that many people including children is interested in watching these movies. It mostly consisted by movies with Action genre, about 97%, which mostly require many violence scene followed with many terror words. So it’s still best for adults to accompany their children while watching these kind of movies.

Cluster 5 has the least appearance of those three groups of words. Since this cluster mostly consist of high income movie which popular among children with genre Adventure with 77%, Animation with 89% and Comedy 93%. Animation movies are mostly targeted children, though there are still these bad words appearances, these movie are quite suitable for children. One of the most appeared genre movies are ones with Comedy movies, while quite similar with cluster 4 that have the most bad words and sexual words appearance, in this cluster can be assumed that the comedy contained is kind of comedy that appeal to children.

5. Conclusion
There are 5 cluster formed with PAM in this dataset, in which cluster 4 has the most appearances of bad words and sexual words, that movies in this cluster should not be watched by children without guidance or better avoiding them. The cluster with the least appearances of the words is cluster 5, which consisted of popular movies with children targeted genre like animation. So it is best to choose
movie from cluster 5 for the children to watch. Even though movies in cluster 5 are quite comfortable to watch, there are still bad words that appear. The best way is to guide children to watch every movies they watch, even though the movie look harmless for them. So they can process the information correctly and will not influenced by bad things shown in movies.

References
[1] Motion Picture Assosiation of America (MPAA) 2015 Theatrical Market Statistics. http://www.mpaa.org/wp-content/uploads/2016/04/MPAA-Theatrical-Market-Statistics-2015_Final.pdf
[2] Feldman R and Sanger J 2007 The text mining handbook: advanced approaches in analyzing unstructured data (New York: Cambridge University Press)
[3] Kabinsingha S, Chindasorn S and Chantrapornchai C 2012 International Journal of Engineering and Innovative Technology (IJEBT) 2 I
[4] Han J, Kamber M and Pei J 2012 Data mining concepts and techniques third edition. (Massachusetts: Morgan Kaufmann Publishers.)
[5] Gower J C 1971 Biometrics 27 4 857-871
[6] Rousseuw P J 1987 Computational and Applied Mathematics 20 53–65.
[7] Palupiningsih P 2013 Pemodelan klasifikasi SMS berindikasi tindak penipuan menggunakan text mining [Thesis] (Bogor: Bogor Agricultural University)
[8] Tala F Z 2003 A study of stemming effects on information retrieval in Bahasa Indonesia M.S. M.Sc. [Thesis] (Amsterdam: Universiteit van Amsterdam The Netherlands)
[9] Adriani M Asian J Nazief B Tahaghoghi SM. And Wiliams HE 2007 Stemming Indonesian: a confix-stripping approach. ACM Transactions on Asian Language Information Processing. 6 4 1-33
[10] Van der Maaten L J P and Hinton G E 2008 Journal of Machine Learning Research 9 2579–2605

Appendices

| Table A1. List of bad words |
|-----------------------------|
| bajingan | banci | bangsat | bedebah | bencong | bego | berandal |
| bodoh | brengsek | cebol | cecunguk | cundang | goblok | idiot |
| jalang | keparat | lacur | persetan | sial | sialan | tolol |

| Table A2. List of sexual words |
|-------------------------------|
| anal | aseksual | asusila | berahi | biseksual | cabul | cumbu |
| ejakulasi | ereksi | erotika | erotis | erotisme | heteroseksual | homoseksual |
| homoseksualitas | kondom | lesbian | masturbasi | payudara | pemerkosa | penis |
| prostitutes | seks | sekstan | seksual | seksualitas | selingkuh | sensual |
| sensualitas | sodom | testis | vagina | zina |

| Table A3. List of terror words |
|-------------------------------|
| ancam | bantai | begal | benci | bengal | bengis | brutal | teror |
| bunuh | cambuk | cekik | culik | eksekusi | jagal | korupsi | teroris |
| koruptor | kriminal | mati | mutilasi | narapidana | overdosis | penggal | terorisme |
| psikopat | pukul | mesum | siksa | takut | tampar | tembak |