Application of text mining employing \(k\)-means algorithms for clustering tweets of Tokopedia

J. Rejito*, A Atthariq, and A S Abdullah
Department of Computer Science, Universitas Padjadjaran, Jl.Raya Bandung
Sumedang km. 21 Jatinangor, Sumedang 45363, Indonesia

*juli.rejito@unpad.ac.id

Abstract. In this current digital era, people tend to shop online. Because of that, there are currently many e-commerce companies that can satisfy the various needs of society in shopping. Each company certainly has a strategy to attract consumers to shop at their shopping place. One of the media commonly used to attract consumers is social media. Tokopedia is one of the biggest marketplaces in Indonesia and is also active in utilizing Twitter as their social media mean. Therefore, it is essential for Tokopedia to pay attention to tweet contents interesting enough to be publicized. By applying text mining using K-Means Clustering algorithm, it can be seen which types of tweet contents that are attractive for Tokopedia consumers. Out of 885 Tokopedia tweets that have been collected, a clustering is then done using K-means algorithm, resulting 48 cluster tweets. Then, from the 48 clusters, they are further grouped into 5 major groups. Based on the results of the grouping, it can be seen that the most interesting content deals with quiz prizes and the least attractive content is on lifestyle.

1. Introduction

Technology has become an important aspect for the global society currently, starting from carrying daily activities to work, entertainment, etc. Technological development that has been applied numerously in several sectors has given positive impacts to the society itself. One of the important sectors in life supported by technological development is the economic sector which is specifically related to buying and selling activities.

The current impact of technological development has been able to convert the shopping behaviour of Indonesian society from conventional, which is shopping directly to any shopping center, to modern, which is shopping through marketplace or known better as e-commerce. The number of e-commerce users in Indonesia is 31.65 million people and is expected to increase by 22.24 million people shopping online in 2022. So it is estimated that in 2022, 43.9 million Indonesians will shop online [1].

Currently in Indonesia, there are many established marketplaces that can satisfy consumer needs to shop with various conveniences and products that are not limited by time and place. The large number of marketplaces in Indonesia means they have to be able to compete against each other to attract consumers so that they would willingly shop in their marketplaces. Numerous strategies are carried out by each marketplace such as campaigning massive promotions and advertising their products and services in several media.

One of the media often used by business people to advertise their marketplace massively is social media. Lots of social media can be used to help advertise such marketplace. According to Statista (2018), it is explained that 44% of Indonesia’s population actively use social media with Twitter being one of the social media often used by Indonesian people.

Tokopedia is one of the marketplaces actively using Twitter as its social media. Now Tokopedia has 177 thousand followers with tweets as many as 29 thousand. Since its establishment 9 years ago, Tokopedia has currently developed into one of the biggest marketplaces in Indonesia. According to the data from the iprice.co.id website, Tokopedia sits comfortably on the first place based on the number of website visitors, application rank, social media followers, and the number of employees.
With the employment of Twitter as an effective media for Tokopedia to advertise their products, they must reassess their tweets that are going to be broadcast to their followers, because the number of retweets obtained reflects how interesting the content is to be broadcast. Therefore, the higher the retweets obtained on one tweet, it can be concluded that the tweet is more interesting for the followers.

Based on the explanation above, it is necessary to apply text mining in classifying Tokopedia’s tweet contents to figure out the most appropriate tweet content type in order to attract customer interest by using K-Means Clustering method. With this study, it is expected that it may help Tokopedia promote with Twitter.

2. Method

2.1. Text Mining

Text Mining is one of the special fields of data mining or alternatively known as data mining with text as its data input [2]. Text Mining is a process of data mining in the form of text where the data source is usually obtained from documents and its purpose is to find words that can represent the content of such documents so that it is possible to analyse the relation among documents. Text mining aims to extract useful information from data source. So, the data source used in text mining is a collection of documents that has unstructured format through identification and exploration of interesting patterns. Several special functions of text mining among others include text categorisation and text clustering [3].

Basically the working process of text mining adopts many of the mining research, but the difference is that the patterns used by text mining are taken from a set of unstructured natural languages while in data mining the patterns are taken from a structured database [4]. The stages of text mining in general are text preprocessing and feature selection [2] [5].

2.2. Text Preprocessing

Text Preprocessing is one of the components in text mining which functions to convert unstructured textual data into structured data and stored into a database [6]. The text preprocessing stage is the initial stage of text mining. This stage includes all routine and process to prepare the data which will be used in the knowledge discovery operation of text mining system [2]. The purpose of preprocessing is to produce an index term set that can represent documents [7]. The process performed in text processing among others are [8]

1. Case folding, which is the conversion of all letter characters in a sentence into lowercase and eliminating characters deemed as invalid such as numbers, punctuation marks. And Uniform Resources Locator (URL)
2. Tokenizing, which is cutting a sentence according to each word that compiles it.
3. Stemming, which is the conversion of several affixed words into their basic forms. This stage is generally performed for English texts, because English texts have fixed affixed structure.
4. Tagging, which is converting several words in past tense into their basic forms. This stage is generally performed for English texts or any other language which has past forms.

2.3. Weighting Term Frequency Inverse Document Frequency (TF-IDF)

The data that has gone through the preprocessing stage must be numeric. To convert the data into numeric, the TF-IDF weighting method is used. The Term Frequency Invers Document Frequency (TF-IDF) method is a method used to determine how far is the relation between a term and a document by weighting each word. The TF-IDF method combines two concepts, namely the frequency of occurrence of a word in a document and the inverse frequency of documents containing that word [9].
2.4. Clustering
The process of categorising one set of physical objects or abstract into similar object classes is called clustering. Cluster is a collection of data objects that is similar to one another in the same cluster and is different than objects in another cluster [10]. Clustering categorizes data into categories that have objects having the same characteristics. The criteria to determine the similarity depend on their implementations. Clustering algorithm can be grouped into two major classes which are hierarchical and partitioning [11].

2.5. K-Means
K-Means is one of the partition clustering techniques that is often used. This method partitions data into clusters/groups so that the data having similar characteristics are grouped into one same cluster and the data with different characteristics are grouped into other clusters. K-means begins by initializing the cluster centre. Every bit of data will be put into any available cluster according to its proximity to the cluster centre. The next step is to calculate the mean of every cluster to update the cluster centre. Update occurs as a result of the alteration in the cluster categories. The process repeats itself until the cluster is unchanged [12].

The K-Means Algorithm Stages
The following is the statement of the K-Means algorithm, which are [13]

a. Choosing the k value.
b. Choosing the k object randomly. Use it as the initial k centroid.
c. Assigning every object to the cluster closest to the centroid centre.
d. Recalculating centroid from k cluster.
e. Repeating steps 3 and 4 until the centroid no longer moves.

![K-Means algorithm flowchart](image.png)

2.6. Silhouette Coefficient
Silhouette Coefficient is one of the methods used to assess the quality and the strength of a cluster. The silhouette coefficient method is a combination of the cohesion and the separation methods. The cohesion method itself is a method used to measure how close the relation is between objects in the same cluster. While the separation method is used to measure how far a cluster is separated from other clusters. Silhouette has three stages in its calculation [7]. The following is the silhouette coefficient calculation according to [6]

a. Calculating the average object distance with all documents contained in one cluster using the equation:
(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j). \quad (1)

b. Then followed by calculating the object distance with all documents among clusters using the equation:

d(i, c) = \frac{1}{|A|} \sum_{j \in C} d(i, j). \quad (2)

c. Followed by calculating the silhouette value using the equation:

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

2.7. Tokopedia
Tokopedia is a company running in the e-commerce sector and carries a marketplace business model that allows every businessman in Indonesia to manage their business online easily and for free [14]. Tokopedia is considered as the biggest market in Indonesia and is also a platform for individuals and small-scale businesses to create their very own online shop. Tokopedia was established by two ambitious young people, William Tanuwijaya and Leontinus Alpha Edison on 6 February 2009 but then was launched publicly exactly on Indonesia’s Independence Day, 17th of August 2009. Its tagline is: “Marketplace is the Most beautiful Business Model in The World”. As seen from its tagline, Tokopedia aims to achieve success in the business market which is only achievable by making individuals and their businesses successful. The Tokopedia service claims that its service allows thousands of small and medium enterprises in Indonesia to deliver millions of products per month to its customers across Indonesia which also creates more jobs across the country (http://www.demystifyasia.com/tokopedia/),

2.8. Preprocessing
Preprocessing is carried out on the tweet contents, with the aim of making the clustering more effective and efficient and obtaining more maximum results. The following is an example of a document which will be preprocessed.

```
Mainkan map terbaru dari #PUBG Mobile ! ,  
2 kali lebih luas dari sanhook ;)  
https://pubg.mobile.com
```

In the preprocessing, 3 stages of the process are carried out, namely casefolding, stemming and eliminating the stopword and from the example document above produces the following results.

```
main map baru mobile kali luas sanhook
```

After the preprocessing is run, it can be seen that the whole words are converted into lowercase, then numbers, special characters, and url are also removed. The words “dari” and “lebih” are eliminated because they are considered as stopwords. The words “mainkan” and “terbaru” are converted into their basic forms which are “main” and “baru”.

2.9. The TF-IDF Weighting Method
Before the clustering process is carried out, data in the form of text must first be converted into numeric form. The method used is the TF-IDF weighting method.

- Calculating the Term Frequency (TF).
Term Frequency (TF) is the frequency of appearance of term \((t)\) in the document \((d)\). The following is an example of calculating the TF value:

**Document 1 (D1):** This is Budi

**Document 2 (D2):** Budi is playing football

**Document 3 (D3):** Mr. Amir is Budi’s father

**Document 4 (D4):** Asta is the kid of Mr. Nino

**Document 5 (D5):** Asta and Budi are friends

| Term(t) | D1 | D2 | D3 | D4 | D5 |
|---------|----|----|----|----|----|
| amir    | 0  | 0  | 1  | 0  | 0  |
| Kid     | 0  | 0  | 0  | 1  | 0  |
| Asta    | 0  | 0  | 0  | 1  | 1  |
| Father  | 0  | 0  | 1  | 0  | 0  |
| Play    | 0  | 1  | 0  | 0  | 0  |
| Befriend| 0  | 0  | 0  | 0  | 1  |
| Ball    | 0  | 1  | 0  | 0  | 0  |
| budi    | 1  | 1  | 1  | 0  | 0  |
| And     | 0  | 0  | 0  | 0  | 1  |
| This    | 1  | 0  | 0  | 0  | 0  |
| nino    | 0  | 0  | 0  | 1  | 0  |
| Mr      | 0  | 0  | 1  | 1  | 0  |

Calculating Inverse Document Frequency (IDF). To calculate the IDF value, the following equation is used:

\[
IDF = \log \left( \frac{n}{df} \right) + 1. \tag{4}
\]

| Term(t) | df | idf         |
|---------|----|-------------|
| amir    | 1  | 2.609439    |
| Kid     | 1  | 2.609439    |
| Asta    | 2  | 1.916291    |
| Father  | 1  | 2.609439    |
| play    | 1  | 2.609439    |
| befriend| 1  | 2.609439    |
| ball    | 1  | 2.609439    |
| budi    | 4  | 1.223144    |
| and     | 1  | 2.609439    |
| this    | 1  | 2.609439    |
| nino    | 1  | 2.609439    |
| Mister  | 2  | 1.916291    |
• Calculating TF-IDF.
   After obtaining the TF and IDF values, the TF-IDF value can be calculated by multiplying the two values.

| Table 3. TF-IDF value calculation |
|----------------------------------|
|       | D2          | D3        | D4  | D5  |
| D1    | 0           | 2,609439  | 0   | 0   |
| 0     | 0           | 0         | 2,609439 | 0  |
| 0     | 0           | 0         | 1,916291 | 1,916291 |
| 0     | 2,609439    | 0         | 0   | 0   |
| 0     | 0           | 0         | 0   | 2,609439 |
| 0     | 2,609439    | 0         | 0   | 0   |
| 1,223144 | 1,223144   | 1,223144 | 0   | 1,223144 |
| 0     | 0           | 0         | 0   | 2,609439 |
| 2,609439 | 0           | 0         | 0   | 0   |
| 0     | 0           | 0         | 2,609439 | 0  |
| 0     | 0           | 1,916291  | 1,916291 | 0  |

• Normalization.
   The results of TF-IDF calculations are normalized using the euclidean normalization method with the following formula

\[
\bar{\mathbf{v}}_{\text{norm}} = \frac{\mathbf{v}}{||\mathbf{v}||_2} = \frac{\mathbf{v}}{\sqrt{v_1^2+v_2^2+v_3^2+\cdots+v_n^2}}
\]

(5)

2.10. Determining the number of clusters
To determine the number of clusters, the Silhouette Coefficient method is used. The calculation of the silhouette coefficient can be done on data that has been clustered using the formula \((b - a) / \max(a, b)\), where \(b\) is the average distance of an object with all members of its nearest cluster, and \(a\) is the average distance of an object with the whole members of its cluster.

2.11. K-Means Clustering
The stages of K-Means algorithm begin with reading the data which will be clustered where in this study the tweets that are already in their numeric forms. The initial centroid value which will be used is determined randomly. Then the object distance to the centroid is calculated. The distance calculation is done using the Euclidean Distance, which is:

\[
D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}
\]

(6)

where:
- \(D_e\) is Euclidean Distance
- \(i\) is the object index
- \((x, y)\) is the object coordinate and
- \((s, t)\) is the centroid coordinate
The object grouping is done by calculating the minimum distance of an object with the centroid. After all objects are grouped based on their closest distance with their centroids, then a new centroid value calculation will be carried out. The process is carried out until no noticeable change of the centroid or there is no change on the cluster members. To determine the centroid value which is the stage of iteration (new centroid), the following equation is used:

\[
\bar{v}_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj}
\]

where:
\(\bar{v}_{ij}\): is the centroid / average \(i^{th}\) cluster for the \(j^{th}\) variable
\(N_i\): is the amount of data becoming the \(i^{th}\) cluster member
\(i\): is the cluster’s index
\(j\): is the index of the variable
\(X_{kj}\): is the \(k^{th}\) data value belonging inside the cluster for the \(j^{th}\) variable

### 3. Result and Discussion

#### 3.1. Implementation

Tweet search under the username “Tokopedia” and the time of publication between the 1\(^{st}\) of August 2018 until the 30\(^{th}\) of November 2018. The tweet data obtained are tweet contents, the number of retweets and publication date. The number of data obtained in the crawling process are 855 lines of data consisting of 3 columns which are tweets, the number of retweets and the time of publication. The data are later saved as csv files. After the tweet data are obtained, preprocessing is carried out which consists of casefolding, stemming, and eliminating stopword.

Furthermore, the determination of the number of clusters is carried out using the silhouette coefficient method, namely by calculating the silhouette value for each number of clusters from \(k = 2\) to \(k = 50\). The best silhouette value is 1 (one) so the greatest value will be taken or close to 1. The best clustering results were obtained, namely at \(k = 48\) with a silhouette value of 0.025079.

Tweet data that have been converted into numerical form using the TF-IDF method are then grouped into 48 clusters using the \textit{K-Means} Clustering algorithm. Then, the calculation of the average number of retweets and the determination of words that appear most frequently in each cluster is performed.

![Figure 2. Main page](image_url)

On this page the user inputs the time span of tweets that will be clustered. The user also inputs the file name which will be used to save the clustering result.
On this page as shown in figure 3, the clustering results are presented in the form of tables and graphs. In the table there is a cluster column that shows the cluster, the word column that appears most often which shows the most words that appear in the cluster, the sum tweet column which shows the number of tweets in the cluster, the RT sum column which shows the number of retweets in the cluster and the Avg Retweet column which shows the average retweet on cluster. Meanwhile presented in the graphs is the comparison of the number of retweets of each cluster. From the graphs it can be concluded that cluster 17 is the cluster which earns the most retweets and cluster 14 is the cluster which earns the least retweets.

### 3.2. Discussion

From the results of the clustering, the content type naming or labelling was carried out in each cluster based on the words that appeared most frequently. Furthermore, the 48 clusters were grouped into 10 content groups with the following results.

| Big Group            | Sum Tweet | Sum RT | Avg RT    |
|----------------------|-----------|--------|-----------|
| info and tips        | 303       | 334    | 1,10231   |
| promo                | 303       | 246    | 0,811881  |
| quizz with gifts     | 59        | 153    | 2,59322   |
| lifestyle            | 116       | 93     | 0,801724  |
| program and event    | 74        | 154    | 2,081081  |
Figure 4. Graph on the average of the number of retweets for each content group

Figure 4 is the comparison graph of average number of retweets of each content group. It can be seen that of five content groups that the content group of prized quiz group is the group that has the most retweets, followed by the program and event groups. Meanwhile the group with the lowest retweets is the lifestyle group, followed by the promo group.

4. Conclusion
The analysis of tweet content type from Tokopedia’s Twitter account can be done using the K-Means Clustering Method. The data obtained are preprocessed to lighten the clustering process and to obtain the maximum result. Before the clustering is carried out, the tweet data which are in text forms are converted into numerical forms using the TF-IDF method.

The application is built using the python programming language as the backend and html as the frontend. There are 2 interface pages which are the main page to input the tweet date which will be clustered and the result page to show the clustering result in the forms of table and graph.

The results of the analysis in this study found that the type of content that earned a lot of retweets has something to do with prized quizzes, while the type of content that earned the least retweet is on lifestyle.

Acknowledgments
The authors thank to Rector Universitas Padjadjaran who provided financial assistance for the dissemination of research through Academic Leadership Grant 2020 with contract number: 1427/UN6.3.1/LT/2020. Also thanks to the reviewers for the valuable review of this paper.

References
[1] Wadolowska A Eshopworld 2018 Indonesia e-commerce insights 2018 [Online] Available: https://www.eshopworld.com/blog/indonesia-e-commerce-insights-2018/
[2] Feldman R and Sanger J 2007 The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data (New York: Cambridge University Press)
[3] Putri E K and Setiadi T 2014 Implementation of Text Mining In Email Spam Classification System Using Naive Bayes Jurnal Sarjana Teknik Informatika 2(3) 73-83
[4] Han J and Kamber K 2006 Data Mining: Concepts and Techniques, Second Edition San Francisco: Morgan Kaufmann Publisher
[5] Berry M W and Kogan J 2010 Text Mining: Applications and Theory (United Kingdom: WILEY)
[6] Handoyo R, Rumani R and Michrandi S 2014 Comparison of Clustering Methods Using Single Linkage and K-Means Methods for Document Grouping

[7] Sholehuddin M, Fauzi M A and Adinugroho S 2018 Implementation of Text Mining and K-Means Clustering Methods for Grouping Thesis Documents (Study Case: Universitas Brawijaya) *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 2(11) 5518-5524

[8] Indraloka D S and Santosa B 2017 Application of Text Mining to Cluster Indonesian Shopee Tweet Data *Jurnal Sains dan Seni ITS* 6(2) A51-A56

[9] Fitri M 2013 The Design of Information Retrieval Systems Using The TF-IDF Combination Weighting Method for Searching Indonesia Language Documents *Jurnal Sistem dan Teknologi Informasi* 1(1) 80-85

[10] Han J, Kamber M and Pei J 2012 *Data Mining Concept and Techniques, 3rd ed* (Amsterdam: Morgan Kaufmann-Elsevier)

[11] Tan P N, Steinbach M and Kumar V 2006 *Introduction to Data Mining* (Boston: Pearson Education)

[12] Nugroho G A P 2016 *Sentiment Analysis of Twitter Data Using K-Means Clustering* Yogyakarta

[13] Bramer M 2007 *Principles of Data Mining* (London: Springer)

[14] Rizki A 2016 *Analysis Quality of The Website in The 2016 Tokopedia Buying and Selling Forum Tugas Akhir* (Bandung: Telkom University)