ISSN 2355-0082

Analysis Sentiment in Bukalapak Comments with K-Means Clustering Method

Rena Nainggolan[1], Fenina Adline Twince Tobing[2], Eva J.G. Harianja[3]
[1,3] Komputer Akuntansi, Universitas Methodist Indonesia, Medan, Indonesia
filipiopat@gmail.com
[2] Prodi Informatika, Fakultas Teknik dan Informatika,
Universitas Multimedia Nusantara, Tangerang, Indonesia
fenina.tobing@umn.ac.id

Accepted 05 December 2022
Approved 04 January 2023

Abstract— Technological development is very fast this era of globalization, to facilitate the work of many aspect that can be utilized, as well as for the flow of information. By applying computer technology in various fields, such as education, entertainment, health, tourism and culinary. Clustering is one of the Data Mining techniques. Clustering works by combining many data or objects into one cluster, with the aim that each data in one cluster will have data that is as similar as possible and different from data or objects in other groups. K-Means Clustering can to perform computations that are relatively fast and efficient in combining large amounts of data. In this research, there are 1407 comments which will be training data and testing data.

Keywords-; Clustering; K-Means Clustering; Sentiment;

I. INTRODUCTION

With the convenience and advantages of using E-Commerce, companies carry out online promotions using electronics and internet marketing by designing their own websites and entering information about the facilities and advantages possessed by theis companies on well known websites, intending to make appeal to prospective customers in the hope that prsopective customers will be interested in choosing and deciding to use it. [1]

Especially in the current, it is very demanding for business people to use computer technology so they can complete. In the current era of globalization, which is called e-commerce, transactions in the internet world are defined as e-commerce [2]

For example Bukalapak.com, this site helps potential customers who want to order goods or services because it make it easier for potential customers to make goods purchase transactions

In addition to promoting products owned by each seller, the e-commerce department also handles reviews from E-Commerce users. It can be seen from the use of the internet that it has now become a communication tool in everyday life, both in seeking information uses.

Bukalapak is an Indonesian technology company whose mission is to create a fair economy for all. Through its online and offline platforms, Bukalapak provides opportunities and choices for everyone to achieve a better life.

With the increasing use of the internet encouraging e-commerce [3] managers to further increase the effectiveness of promotions, one of the services is a platform for ordering goods, purchasing, and payment systems such as credit card installments, virtual Accounts, Noncredit card installment, in-store payment, internet banking, payment gateways, direct debit, electronic money, chat, shipping and returning gods if the goods do not match the order. With online purchases, online customer reviews appear which help consumers to be able to interact directly with service providers or tourist attractions, consumers can bargain, provide suggestions, or give impressions or feedback in using the facilities provided, one of which is open stalls.

Over time, this site is used by consumers or buyers, to share experiences while making purchases of goods using the site's facilities. To minimize the negative impact, before purchasing goods. Prospective buyers must find as many information as possible about the items to be purchased. The easiest way is to see ratings from reviews or ratings on products that will be purchased by potential customers.

II. METHOD

A. Sentimen Analisis

Sentiment analysis is the process of extracting, process, and understanding data in the form of text that is not structured automatically to retrieve information and sentiment contained in a sentence of opinion.[4] Sentiment analysis is performed to assess opinions and the tendency of an opinion towards a good topic negative or positive. Sentiment analysis can be applied to opinions in all fields such as economics, political, society, and law. Twitter's social media opens a window for researchers to study emotions, moods hearts, and public opinion through sentiment analysis. Sentiment Analysis is a field of Natural Language Processing (NLP) that builds a system to recognize and extract opinions in text form. [5]

B. Data Mining

There are two terms in data mining namely; such as knowledge discovery and pattern recognition. Each of them has a different meaning and has a definition from each other. [6]According to [7] the purpose of data mining is to obtain knowledge that is still hidden in chunks of data, where as according to [8] it is pattern recognition in chunks. The data to be extracted is called pattern recognition

C. Clustering

The way clustering works is to form a group from a collection of various physical or abstract objects into the same group [9]. The one group consists of a set of data that is as similar as possible between one data and another and must be different from the data or objects in the other groups[10]

Clustering performs the grouping of data without being based on certain that have been defined from the start. This process is very different from the classification process which at the beginning of the process must provide data classes. So clustering is often referred to as unstructured data grouping

D. TF-IDF

At this stage, the Tfidf Vectorizer function from the library is used to carry out the word weighting process [11]. After the TFIDF process has been carried out and the weight value of each word in each tweet sentence has been successfully obtained, the sentiment data classification process will be carried out.[12]

E. K-Means Clustering

The way the K-means method works is by grouping data or objects that have very similar characteristics in one cluster/group and will group data/objects into other clusters that have different characteristics and will eventually produce a cluster or group that has a different level. Very high resemblance [13]. The steps for clustering with the K-Means method are as follows [14]

1. Determine K or the number of clusters 2.
2. Cluster centers are given an initial value with the random number. But in providing initialization of K cluster centers can be done in various ways, and the most common is randomly
3. In placing objects in which cluster they are placed, then it is calculated based on the distance between the two objects, as well as the distance between the cluster center and the object. To calculate the distance of all data to each cluster center point, you can use the Euclidean distance theory which is formulated as follows:
4. Recalculating the distance of the current cluster membership to the cluster center, the average of all data/objects in a particular cluster is the cluster center, using the mean (average value) or median
5. Do it again for each object with a new cluster, if the new cluster center does not change or change, then the clustering process stops, otherwise it will return to the third stage until the cluster center does not change.

$$D(i,j)= ((X1i-X1j)^1+(X2i-X2j)^1+..+(Xki-Xkj) \quad (1)$$

The following is an explanation of the flow of the research method

1. Identification of problems observing and finding problems that occur in the Bukalapak.com Application seen from the comments of the Bukalapak.com application user.
2. The determination of objectives serves to further clarify what framework is the target of this research. As already written in chapter I, the purpose of this research is to classify the comments of Bukalapak.com using the K-Means Clustering.
3. Library study aims to find out what theories will be used to solve the problems to be studied, as well as get strong reference bases for researchers.
4. Data collection the process of data collection and input in this system begins with web scraping data, the web scraping process is carried out using a data miner, a chrome extension that can be used for web data scraping. The output of this system is in the form of data testing with output values in the

form of positive and negative sentiments which are classified by the system based on training data learning.

5. Preprocessing

Preprocessing is one of the important stages for data in the mining process. The data used in the mining process is not always in ideal conditions for processing. Sometimes in the data, there are various problems that can interfere with the results of the mining process itself such as missing values, redundant data, outliers, or data formats that are not under the system. Therefore, to overcome these problems, the preprocessing stage is needed. Preprocessing is one of the stages of eliminating problems that can interfere with the result of data processing. This stage includes the following stages:

A. Cleaning Process

The cleaning Process that is carried out is changing text to lowercase (case folding, removing characters other than letters, removing user, usernames, or mention (@ 0, removing hashtags (#), and removing URLs or links from each comment.

B. Filtering

Filtering removes unnecessary words from token results. Apart from that, punctuation marks and stopwords were also removed. Stopwords are processed in a sentence if they contain words that often come out and are considered unimportant such as time, liaisons, and so on.

C. Stopword Removal

Stop words are common words that usually appear in large numbers but have no meaning. In indonesia like in, with, to, which, if, will and so on [15]. For that it is necessary to deletion process this word is required a data or list of words that you want to delete [16]

D. Stemming

At the stemming stage, namely changing affixed words into basic words. The stem process is removing the prefix and suffix words.[17]

Example:

Mempermainkan peranan 10 kuda di pementasan seni

Menjadi:

Main peran 1o kuda di pentas seni

After all the data is transformed into numbers, then the data has been obtained and grouped using the K-Means Clustering method. To be able to do to group the data into several clusters, several clusters, several steps need to be carried out, namely:

1. Determine in advance the number of clusters you want. In this study the data there into will be grouped into two clusters.

Determine the starting point of each cluster. In this study, the starting point was randomly generated.

III. Result and Discussion

A. Dataset

To carry out the clustering proposed in this and process ini this study. The authors tested the dataset based on the method proposed in this study, namely by using the K-Means Clustering method. The dataset used is crawling data from Bukalapak.com

Result of data collection the research was conducted using online customer reviews consisting of 1407 review

1. Barang bagus banget *_*
2. Alhamdulillah cocok 😊
3. Harga dan kualita ok
4. Makasih moga cocok @_@
5. Kualitas fungsi produk joss mantap
6. Respon cepat
7. JELEK!!
8. Barang sampe tujuan dengan cepat
9. Rekomendasi lapak ini mantap
10. Semoga berkhasiat dan cepat sehat
11. Lebih cepat dari perkiraan *8-)
12. Memuaskan
13. KECEWA…☹
14. Pasti order lagi
15. BARANG RUSAK %%

Dataset Processing method the dataset that has been obtained from the scrapping process still has a data structure of unsstructured, arbitrary and irregular data. Because of this, it is needed the data preprocessing stage is before the data set is tested with a model. This stage is done to clean data from noise and transform data into structured data, as for the stage in the processing of the data in this study namely:

1. Tokenize

Where in this process word splitting is done in the review sentence. This stage also removes certain character such as punctuation marks as well as filter by text length.

2. Stemming

Stemming in which the process find basic words by eliminating them all the affixes attached to the word.

3. Case Folding

Case Folding where the process utilizes the transform cases feature aims to homogenize the entire text into all lowercase

4. Stopwords Removal

Stopwords Removal where in this process the removal of the word is included in the stopword category. Stopword is words that appear frequently but is considered meaningless.

1. Paket hancur
2. Barang cepat sampai
3. Barang datang lama
4. Sesuai ukuran
5. Sangat kecewa
6. Barang bagus
7. Barang dikemas rapi
8. Penjual ramah
9. Respon cepat
10. Barang rusak

Define abbreviations and acronyms the first

C. String to word vector

The TF-IDF algorithm is applied to convert string data into word vectors, and generate a data matrix with the dimension of 25 attributes x 1407 data, from TF-IDF implementation. There are 25 terms in the data as shown in the following figure

Kece, persis, pas, cepat, profesional, elegan, mahal, murah, trendi, tukar, nyasar, parah, cocok, rapi, tebal, sobek, terimakasih, keras, enak, jelas, komplit, rusak, longgar, sama, salah

D. Converting Data to Numeric

The following table is an example of the final conversion result data.

18 0.004276, 23 0.005957, 27 0.005897, 34 0.025563, 37 0.01063, 43 0.084699, 72 0.504275, 91 0.017341, 94 0.006041, 95 0.005957, 99 0.0107 38,102 0.0188 66, 112 0.957791, 128 0.027989, 131 0.034241, 136 0.021461, 148 1.479877, 155 0.00586, 169 0.121277, 170 0.005802, 176 0.025946, 178 0.005328, 188 0.026011, 193 0.005868, 206 0.005904, 211 0.131904, 1 9 0.12829, 13 0.035806, 18 0.005942, 26 0.01387, 29 0.014772, 30 1.685916

E. Attribute Selection

The data above is still too large and ineffective, so the existing attributes must be filtered. By using the Cfs algorithm, there are 10 attributes

pas, cepat, parah, cocok, jelek, jelas, terimakasih, tebal, komplit, salah

F. K-Means Clustering Coding

Table 1 K-Means Clustering Coding

Data Ke-	Cluster 0	Cluster 1	Min Cluster
1	0.7148	0.8465	C0
2	1.3117	1.32	C0
3	2.1711	0.4824	C0
4	0.1005	0.2815	C0
5	1.1461	0.6353	C1
6	2.1636	2.2118	C0
7	3.1182	3.1818	C0
8	2.772	2.8258	C0
9	1.5869	1.4849	C1
10	2.3618	2.5538	C0
1407	1.2121	1.303	C0

Cluster Instance

Table 2 Cluster instances are shown in the table

	Cluster	Percentage
1	1278	91 %
2	129	9%

G. Graph Comparison

Graph comparison of the result of each cluster from 1407 test data with 2 clusters is shown in the picture below

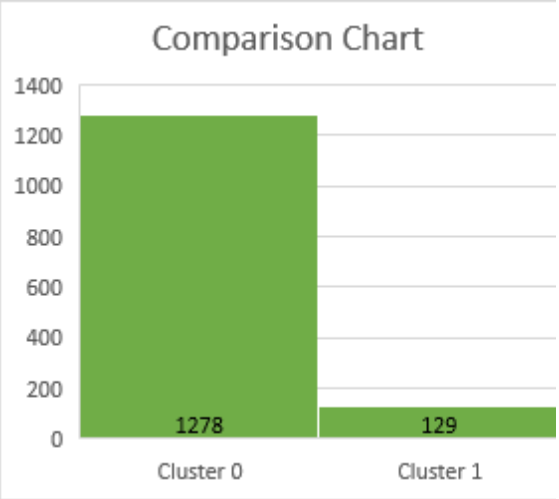

Fig 1. Graph Comparison Chart cluster 0 and Cluster 1

Comparison chart from the table above can be concluded that the result of testing on the1407 review yielded 2 clusters namely:

a. Cluster 0 produces 1278 (91%) reviews that was similar and very high which are grouped into 0 cluster
b. Cluster 1 resulted in 129 (9%) reviews which had very high similarity grouped into 1 cluster group

IV. CONCLUSION

Based on a test conducted on sentiment analysts at the store online with a total of 1407 data reviews, this study it based on the test. Conducted on sentiment analysts at the store online with a total of 1407 data reviews, this study offers a model clustering, and testing produces 2 clusters, and testing produces a tool to assist consumers in deciding to buy a product or service, because of the importance of sentiment analysis for increasing customer confidence. The information helps consumers to get the quality of the product they are looking for from reviews and experiences written by other consumers who have purchased the product from the previous online sellers.

The result obtained is 2 namely:

1. Cluster 0, which consists of 1728 comments or 91 % of customers who gives positive comments,
2. Cluster 1 consists of 129 comments or 9 % of customers who give negative comments.
3. By applying the K-means Clustering method to analyzing sentiment data products, it can help potential customers to determine whether the product is feasible or not to buy because sentiment data is used as a means for consumers to search and obtain information that will later influence decisions to purchase sentiment data also has a function as a tool decision making

REFERENCES

[1] E. Y. Nasution, P. Hariani, L. S. Hasibuan, and W. Pradita, "Perkembangan Transaksi Bisnis E-Commerce terhadap Pertumbuhan Ekonomi di Indonesia," *Jesya*, vol. 3, no. 2, pp. 506–519, 2020, doi: 10.36778/jesya.v3i2.227.

[2] K. Kasmi and A. N. Candra, "Penerapan E-Commerce Berbasis Business To Consumers Untuk Meningkatan Penjualan Produk Makanan Ringan Khas Pringsewu," *J. Aktual*, vol. 15, no. 2, p. 109, 2017, doi: 10.47232/aktual.v15i2.27.

[3] C. F. C. Dewantara, "ANALISIS DAMPAK PENGGUNAAN SITUS Bukalapak. Com, Perilaku, Terhadap Perilaku Pembelian Pada Komunitas Samarinda Photographer," *eJournal Ilmu Komun.*, vol. 3, no. 2, pp. 488–502, 2015.

[4] N. Ransi, L. Surimi, A. Tenriawaru, and L. O. Saidi, "Analisis Sentimen Masyarakat Terhadap Toko Online Aplikasi," pp. 1–8, 2020.

[5] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *J. SIMETRIS*, vol. 10, no. 2, pp. 681–686, 2019.

[6] Yuli Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5 Data mining merupakan bagian dari tahapan proses Knowledge Discovery in Database (KDD) . Jurnal Edik Informatika," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2019.

[7] D. Firdaus, "Penggunaan Data mining dalam kegiatan pembelajaran," vol. 6, no. 2, pp. 91–97, 2017.

[8] A. S. Aribowo, "Metode Data Mining Untuk Klasifikasi Kesetiaan," vol. 10, no. 1, pp. 2–4, 2013.

[9] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan," vol. 5, no. 1, pp. 62–66, 2019.

[10] T. Hardiani, "ANALISIS CLUSTERING KASUS COVID 19 DI INDONESIA MENGGUNAKAN ALGORITMA K-MEANS Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI | 157," vol. 11, pp. 156–165, 2022.

[11] M. Nurjannah and I. F. Astuti, "PENERAPAN ALGORITMA TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF) UNTUK TEXT MINING," vol. 8, no. 3, pp. 110–113, 2013.

[12] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF- IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," vol. 1, no. 2, pp. 58–64, 2019.

[13] I. Sumadikarta and E. Abeiza, "PENERAPAN ALGORITMA K-MEANS PADA DATA MINING UNTUK MEMILIH PRODUK DAN PELANGGAN POTENSIAL (Studi Kasus : PT Mega Arvia Utama)," pp. 12–23, 2014.

[14] R. K. Dinata, N. Hasdyna, and N. Azizah, "Analisis K-Means Clustering pada Data Sepeda Motor," vol. 5, no. 1, 2020.

[15] A. F. Hidayatullah, "Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 1, no. 1, pp. 1–4, 2016, doi: 10.14421/jiska.2016.11-01.

[16] A. Setiawan, E. Kurniawan, and W. Handiwidjojo, "Implementasi Stop Word Removal Untuk Pembangunan

Aplikasi Alkitab Berbasis Windows 8," *J. EKSIS*, vol. 6, no. 2, pp. 1–11, 2013.

[17] H. R. Pramudita, "Penerapan Algoritma Stemming Nazief & Adriani Dan Similarity Pada Penerimaan Judul Thesis," *J. Ilm. Data Manaj. dan Teknol. Inf.*, vol. 15, no. 4, pp. 15–19, 2014.

[18] Nainggolan, R., & Tobing, F. A. (2020). ANALISIS CLUSTER DENGAN MENGGUNAKAN K-MEANS UNTUK PENGELOMPOKKAN ONLINE CUSTOMER REVIEWS (OCR) PADA ONLINE MARKETPLACE. *METHODIKA: Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(1), 1-5.