Available online at https://ejournal.unisayogya.ac.id/ejournal/index.php/ijhst

International Journal of Health Science and Technology, 3 (3), 2022, 109-117

Analysis and visualization of BPJS on twitter using K-means clustering

Andika Bayu Saputra[1*], Puji Winar Cahyo[2], Muhammad Habibi[3], Adri Priadana[4]

[1,2,3,4]Universitas Jenderal Achmad Yani Yogyakarta, Indonesia

Email: [1]dika.putra21@gmail.com*; [2]pwcahyo@gmail.com; [3]muhammadhabibi17@gmail.com; [4]adripriadana3202@gmail.com

*corresponding author

Submission date: March 24, 2021 Revision date: April 8, 2022 Receipt date: April 22, 2022

Abstract

Social security agency (BPJS) Health exists to provide national social security to meet the basic needs appropriate for all levels of society based on the principle of humanity. Originated from a change in the contribution premium policy, it is demanded by the organizers and health service providers to be able to provide safe, quality, affordable health facilities. But unfortunately, the government's efforts in realizing public welfare, especially in the field of health, are not fully supported by the community because of the ever- changing premium contribution policy and the health services they receive. The latest information developments related to BPJS on social media that can be easily accessed by the public. One of them is by using the Twitter platform as a place to exchange information using hashtags. The hashtag data can be processed and obtained information to be used as a tool for decision making. This study aims to analyze and visualize BPJS data on the Twitter platform using the K-Means clustering method. K-Means clustering method is a method of clustering data mining using the descriptive model concept. K- means method can use to explain the algorithm in determining an object into a specific cluster based on the nearest average.

Keywords: BPJS; health; K-means; clustering

INTRODUCTION

Social security agency (BPJS) Health exists to provide a national social security in order to meet basic life needs for the whole society based on the principles of humanity, benefits and fairness. BPJS Health is an institution that provides health insurance services using the insurance premium system. In the insurance premium system requires each participant to pay dues in accordance with the chosen class. The process of determining the policy on the amount of premium paid by the public has changed a lot (Rarasati 2017). Starting from the change in premium contribution policy, it is demanded by the organizers and health service providers to be able to provide safe, quality and affordable health facilities. But unfortunately, the government's efforts in realizing public welfare, especially in the field of health are not fully supported by the community because of the ever-changing premium contribution policy and the health services they receive. The latest information developments related to BPJS are spread on social media that can be easily accessed by the public. One of them is by using the Twitter platform as a place to exchange information using hashtags. The hashtag data can be processed and obtained information to be used as a tool for decision making.

Doi: 10.31101/ijhst.v3i3.2466

At present, data processing using data mining methods is very widely used to help speed up the decision making process appropriately, and manage the information contained in the data into a new knowledge (Sani 2018). The development of information technology in this case is that social media has become a necessity that cannot be abandoned in everyday life. The data of the largest Twitter social media user is Indonesia. Twitter users are becoming popular among Indonesian people because of the freedom to speak or voice (Abdulah et. al 2021).

Twitter is a social media that is widely used by the public and popular in Indonesia. This is based on the results of Hootsuite We are social research released in January 2020 social media users in Indonesia reached 175.4 million or 64% of the total population. The number is up 17% from the previous survey. While mobile social media (mobile) users reach 140 million or 50% of the total population. Meanwhile, based on data from inet.detik.com, Twitter's daily users in 2020 increased 24% to 166 million users (Jati 2020). Twitter users can make short messages called tweets, through which tweets, Twitter users can interact, share opinions, and find various information from around the world. To provide a link to the topic discussed specifically about BPJS Health, one of the things done is by using the hashtag #bpjs in uploads on the Twitter platform. The use of the hashtag makes it easier for the public to get information related to information about BPJS Health.

This study aims to analyses and visualize BPJS data on the Twitter platform using the K-Means clustering method. K-Means clustering method is a method of clustering data mining using the descriptive model concept. K-means method can be used to explain the algorithm in determining an object into a certain cluster based on the nearest average (Febianto and Palasara 2019). Cluster analysis is a multivariate technique whose main purpose is to group tweet data based on the number of tweet frequencies. The clusters formed have internal homogeneity and high external heterogeneity.

Research related to the use of the k-means clustering data method has been widely used, including the analysis of marketplace conversations trends on twitter platform using k-means (Nasron and Habibi 2020), the application of the k-means clustering analysis algorithm to human infectious diseases (case study in Kabupaten Majalengka) (Bastian, Sujadi, and Febrianto n.d.), the application of the k-means algorithm for clustering drug data at the Pekanbaru Public Hospital (Gustientiedina, Adiya, and Desnelita 2019).

RESEARCH METHODS

This research consists of several stages, including collecting data, pre-processing, feature extraction, and Clustering. The system architecture can see in Figure 1.

ISSN 2685-8673 (online)

A. Collecting Data

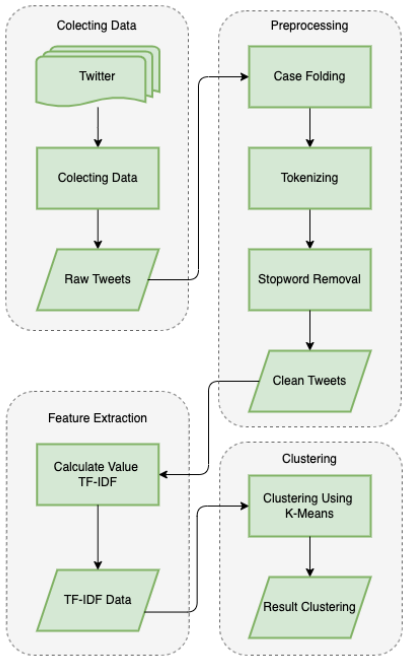

Figure 1. Architecture System

The first stage in this research is collecting data the data used in this study is the data of tweets and re-tweets related to BPJS contained in Indonesia. The data taken is the data of tweets and re-tweets related to the hashtag from BPJS health in the form of #bpjs health. The length of time used in collecting data is from November 19, 2020, to January 6, 2020, with the total number of tweets and re-tweets obtained as many as 59551. The data is processed to get the best Clustering.

B. Preprocessing

In the second stage, namely, text pre-processing is the initial stage of text mining, and this stage includes the process of preparing textual data that will use to process at the next scene. Data results are developing and cleaning up text data before the text is analysed (Yuan 2019). The steps in pre-processing used in this study include:

1. Case folding is the process of converting text in to one uppercase or lowercase letter.
2. Tokenizing is the process of dividing a text into tokens.
3. Stop-word removal is the process of removing words from documents that do not have an essential role in providing patterns or information.

C. Feature Extraction

The next step is Feature Extraction or feature extraction is an extraction process to identify the intended entities (Siqueira and Barros 2010). At this stage, the Term Frequency (TF) and Inverse Document Frequency (IDF) values were searched for the text that was obtained. This technique is useful for filtering out words that are too rare or too general to use TF-IDF (Term Frequency-Inverse Document Frequency),

ISSN 2685-8673 (online)

which gives low scores for words that are very rare or very frequent. Another option is to have a minimum frequency cut-off to filter out rare words and use a list of common words to filter out words that are too general. The formula for calculating the weight value of TF-IDF is obtained from the following equation (Manning, Raghavan, and Schutze 2009):

(1)
$$w_{t,d} = tf_{t,d} \times \log \frac{N}{df_t}$$

Information:

$tf_{t,d}$ = weighting term t on the document d
N = total number of documents
df_t = number of documents that contains term t.

D. Clustering

Clustering is the use of data mining techniques where groups of the same object are combined together to form clusters, this cluster is different from objects in other clusters (Studi et al. 2020). At the stage of clustering the data that has been collected is grouped into data that has similarity characteristics (similarity) between one data with another data. In the clustering process K-Means is used to cluster Twitter data related to BPJS. The K-Means calculation is obtained from the following equation (Elisawati, Wahyuni, and Arianto 2019):

(2)
$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n} (x_j - c_j)^2}$$

Information:

d = distance
n = many object
j = (starting from 1 until n)
X-j = feature object to j toward x
C-j = centroid feature to j

K-Means clustering is used to do the unsupervised learning modeling process, which is to make predictions or classifications, it does not need to be trained or the training process first. The purpose of the K-Means clustering algorithm is to group objects that are almost the same in a certain area (Nasron and Habibi 2020).

In K-Means clustering has the usual steps to group objects that are almost the same in certain areas as follows (Cahyo et al. 2018):

1. Determine the number of clusters (k)
2. Determine the centroid
3. Analysis of whether the centroid changes based on the means of each object feature (data)?

ISSN 2685-8673 (online)

a. [Yes], change centroid to new centroid
b. [No], done

4. Calculate distance space (proximity) of objects with centroids.
5. Grouping objects based on the object's proximity to centroids.

RESULTS AND DISCUSSION

This section discusses the results of the experiments that have been carried out. In the experiments that have been carried out, the TF-IDF results from corpus or dictionary have been collected. Next determine the best cluster using the Elbow method. Elbow method is a method used to produce information in determining the best number of clusters by looking at the presentation of the results of the comparison between the number of clusters that will form an elbow at a point. This method provides ideas / ideas by selecting a cluster value and then adding the value of the cluster to be used as a data model in determining the best cluster (Putu, Merliana, and Santoso n.d.). The following chart is the result of calculating the best number of clusters using the elbow method, can be seen in Figure 2.

Based on the calculation results, some K values have decreased the most and then the results of the K values will decline slowly until the results of the K values are stable. One of the values of K which experienced a large decrease is K = 14 and one that has decreased drastically to form an elbow at point K = 17. It can be concluded that the ideal K cluster value is K = 17, so that the K values that are suitable for use in the clustering process are 17 clusters.

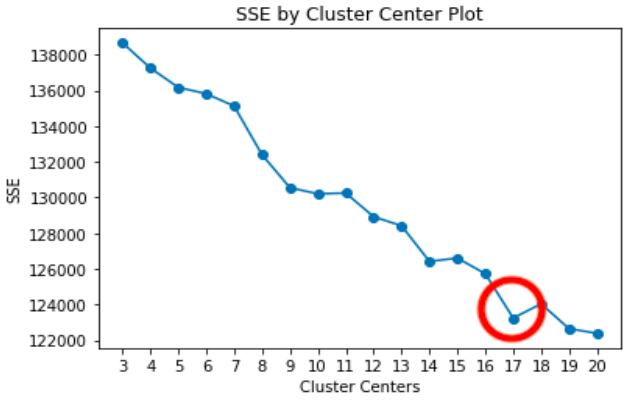

Figure 2. Elbow Graph Method

After the K value is found, which is 17 clusters in accordance with the results of the best number of clusters, the next step is to group tweet data into 17 tweet clustering data. In Twitter content analysis, we created a word cloud visualization model for all segments and there are 17 topic segments. words that are often used in conversation will be depicted in cloud visualization. That is, we can see what are the most significant words that make up a particular segment of the model being built.

Word cloud from all segments aims to summarize all keywords that appear in 17 topic segments. Based on the word cloud in Figure 17, the most dominant topics are the contribution of BPJS, BPJS deficits, health promotion, and BPJS participants. As shown in the word cloud, it is known that words like health, BPJS, destroyed, services, government, fees, deficits, jokowi dominate the most widely used key words in all segments, can be seen in Figure 3.

ISSN 2685-8673 (online)

Figure 3. Word Cloud all segment

Based on the results of the word cloud in Figure 3, a topic analysis was carried out for 17 segments. Topic analysis is made based on the relationship of words that often appear in 17 topic segments. The results of the topic analysis can be seen in Table 1.

Table 1. Content Analysis in each cluster

Cluster	Content Analysis
Cluster 0	Discussing about the government issuing official policies relating to regulations
Cluster 1	Discuss about health care facilities in hospitals.
Cluster 2	Discussing about the preparation of regional governments in the free policy for citizens
Cluster 3	Discusses Sharia Institutions in managing BPJS
Cluster 4	Discussing the technical services of the health office
Cluster 5	Discussing about stood for justice to cancel the fee hike BPJS
Cluster 6	Discussing about the rise in tuition rates BPJS
Cluster 7	Discussing the rise in health BPJS, increasing BPJS incentives, services and sick costs and the Jokowi government deficit
Cluster 8	Discussing about JOKOWI dizziness handle outbreaks corona
Cluster 9	Discussing about the deficit and the health BPJS JIWASRAYA case
Cluster 10	Discussing about the increase in the BPJS participant contributions to the independent class, it was decided by Commission III of the DPR
Cluster 11	Discusses mobile hospitals in assisting health services for healthcare professionals
Cluster 12	Discussing about defense and security, scholars, the military and the people and health services
Cluster 13	Discussing the policy of the Minister of Finance Sri Mulyani
Cluster 14	Discussing about Indonesia providing free facilities
Cluster 15	Discussing about the deficit JIWASRAYA

ISSN 2685-8673 (online)

Cluster	Content Analysis
Cluster 16	Discuss about private insurance

The most dominant topics from the BPJS tweet data are related to rising BPJS fees, intensive increases in BPJS, services and sick costs and the JOKOWI government deficit. This matter was warmly discussed through the twitter media by the people of Indonesia. Following a visual display of the results of the Clustering using the K-Means method using the bar diagram shown in Figure 4 below.

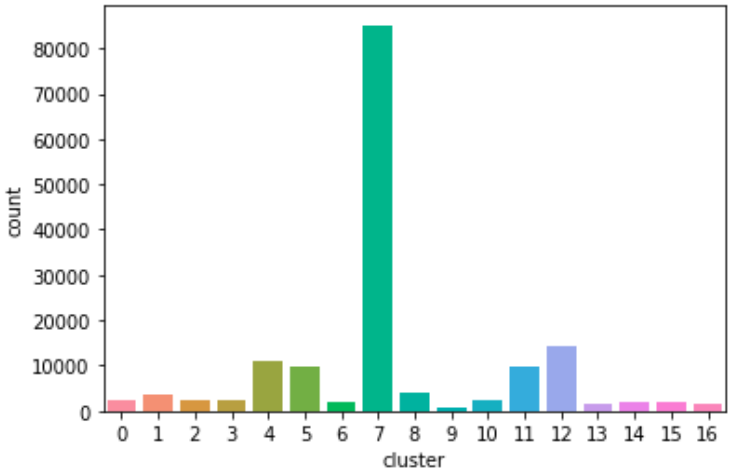

Figure 4. Cluster Diagram

The next step is to calculate the frequency with which words appear. For words with higher frequency of occurrence, the word will become a trend in the conversation on Twitter social media. Based on the results of frequency calculation, it can be concluded that contribution trends dominate the health BPJS data on twitter. The following 10 sequences of occurrences of the most words obtained in all clusters, can be seen in table 2.

Table 2. Word Frequency

Word	Frequency
Iuran	38825
Rakyat	18764
Peserta	17323
Kenaikan	15391
Dokter	13721
Pelayanan	12975
Kelas	12506
Jkn	11487
Defisit	11177
Bayar	10436

ISSN 2685-8673 (online)

CONCLUSION

This research succeeded in analyzing and visualizing BPJS data on the Twitter platform with k-means clustering. In this research, it is known that the trend of BPJS data on Twitter is about contributory tweets. The results of the calculation of tweets in each cluster obtained there are 5 clusters that have the most tweets, namely clusters 4, 5, 7, 11, and 12 with an increase in content bpjs fees and services. Cluster 7 has the highest number of tweets and dominates the data in this study. From the results of this study it was found that the application of the K-Means clustering method can produce information that can help in making decisions and as one of the considerations for making policies regarding health BPJS services.

REFERENCES

A. Bastian, H. Sujadi, and G. Febrianto, "Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka)," no. 1, pp. 26–32.

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R., & Hidayat, R. (2021). The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Quality & Quantity, 1-9.

Ahmar, A. S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., & Azmi, M. (2018, June). Using K-Means Clustering to Cluster Provinces in Indonesia. In Journal of Physics: Conference Series (Vol. 1028, No. 1, p. 012006). IOP Publishing.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

A. S. Jati, "Jumlah pengguna twitter meningkat" 04 Mei, 2020.

A. Sani, "Penerapan Metode K-Means Clustering Pada Perusahaan," J. Ilm. Teknol. Inf., no. 353, pp. 1–7, 2018.

D. H. Rarasati, "Dampak Kenaikan Tarif Bpjs Kesehatan terhadap pelayanan Kesehatan di Kota Malang," J. Polit. Muda, vol. 6, no. 1, pp. 34–40, 2017.

D. C. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval. 2009.

Elisawati, D. Wahyuni, and A. Arianto, "Analisa Clustering Pada Data Pelanggaran Lalulintas Di Pengadilan Negeri Dumai Dengan Menggunakan Metode K-Means," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 3, no. 3, pp. 50–61, 2019.

G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," J. Nas. Teknol. dan Sist. Inf., vol. 5, no. 1, pp. 17–24, 2019.

H. Siqueira and F. Barros, "A Feature Extraction Process for Sentiment Analysis of Opinions on Services," Proc. III Int. Work. Web Text Intell., 2010.

Kuswandi, D., Surahman, E., Thaariq, Z. Z. A., & Muthmainnah, M. (2018, October). K-Means clustering of student perceptions on project-based learning model application. In 2018 4th International Conference on Education and Technology (ICET) (pp. 9-12). IEEE.

ISSN 2685-8673 (online)

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE.

M. Habibi and P. W. Cahyo, "Clustering User Characteristics Based on the influence of Hashtags on the Instagram Platform," IJCCS (Indonesian J. Comput. Cybern. Syst., vol. 13, no. 4, p. 399, 2019.

N. I. Febianto and N. Palasara, "Analisa Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018," J. Sisfokom (Sistem Inf. dan Komputer), vol. 8, no. 2, p. 130, 2019.

N. Putu, E. Merliana, and A. J. Santoso, "Analisa Penentuan Jumlah Cluster Terbaik pada Metode K-Means," pp. 978–979.

P. Studi et al., "PENERAPAN ALGORITMA K-MEANS CLUSTER ING SEBAGAI STRATEGI PROMOSI PENERIMAAN MAHASISWA BARU PADA UNIVERSITAS HASYIM ASY ' ARI JOMBANG Aries Dwi Indriyanti Indana Lazulfa," vol. 04, pp. 20–27, 2020.

P. W. Cahyo, P. Studi, T. Informatika, U. Jenderal, and A. Yani, "KLASTERISASI TIPE PEMBELAJAR SEBAGAI PARAMETER Abstrak Data Collection," pp. 49–55, 2018.

U. A. Nasron and M. Habibi, "Analysis of Marketplace Conversation Trends on Twitter Platform Using K-Means," Compiler, vol. 9, no. 1, pp. 51–61, 2020.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, No. 1, p. 012017). IOP Publishing.

Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. J, 2(2), 226-235.

ISSN 2685-8673 (online)