Identifying Twitter Topics Using K-Means Clustering and Association Rule Mining for Improved Insights

[1]Cristiany Gunu Lengari, [2]Ira Puspitasari
[1]Master of Human Resources Development, Postgraduate School, Universitas Airlangga, Indonesia
[2]Information System Study Program, Faculty of Science and Technology, Universitas Airlangga, Indonesia
Email: [1]cristianylengari@gmail.com, [2]ira-p@fst.unair.ac.id

Article Info

Article history:
Received Jul 19th, 2024
Revised Sep 3rd, 2024
Accepted Nov 4th, 2024

Keyword:
Association Rules
Clustering
Improved Insights
K-Means
Twitter

ABSTRACT

The annual growth in social media users has led businesses to increasingly leverage these platforms for marketing, promotion, and addressing public complaints. Twitter, now known as X, stands out as one of the most widely used social media platforms. It serves as a forum for various opinions and complaints regarding services provided by businesses. This study focuses on analyzing public opinions related to Indihome services, as expressed on the @indihomecare Twitter account. These opinions range from expressions of support to complaints about internet services and Indihome's responses to these issues. This study employs a text clustering approach using the K-means algorithm on Twitter data, complemented by association rules to identify topics related to Indihome customer complaints. The optimal number of clusters is determined using the Elbow method, while Word Cloud visualizations are utilized to illustrate frequently occurring words within each cluster. The application of association rules revealed that the most frequently appearing words, with a support value of 0.057, were "indihome," "account," "whatsapp," and "channel." These findings provide insights into the primary concerns and communication channels used by Indihome customers on Twitter.

Corresponding Author:
Cristiany Gunu Lengari
Master of Human Resource Development, Postgraduate School,
Universitas Airlangga,
4-6 Airlangga, Gubeng, Surabaya City 60115, East Java.
Email: cristianylengari@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v8i1.31720

1. INTRODUCTION

These days, a lot of people have been using social media to communicate with each other. According to Dataindonesia.id, the number of social media users in Indonesia reached 167 million people in 2023. Twitter is one of the most widely used social media in Indonesia and is usually used by several companies to market their products. In 2023, the number of Twitter users in Indonesia reached 14.75 million. It ranked sixth in the world [1]. Social media platforms such as Facebook, Instagram, Twitter, LinkedIn, TikTok, YouTube and Facebook provide online spaces for socializing and learning. They also contain a large amount of user-generated information, which can help academics and organizations in their understanding of human behaviour [2].

Social media is used in society as a place for sharing or exchanging ideas, content and other types of information. Texts, photographs, videos and other types of information are shared on social media platforms on a massive scale [3]. Analyzing large amounts of social media data can provide in-depth knowledge about how people like, care about, and behave. There are various applications of the data, such as sentiment analysis, content recommendation, and market research [3]. Furthermore, social media text data is a key area for big data studies due to its ability to identify current trends and topics and extract current trends and topics [4].

The increase in the number of users on social media some companies are using social media as an extension of their business, where users can find other businesspeople to become friends or followers, as well as a place to advertise and receive criticism or complaints from the community that can be followed up on by businesspeople. One company that uses social media as a place to receive public complaints is indihome. The @indihomecare account allows customers to submit problems related to the grid, periodic payments, and other issues. Given the many complaints made through Twitter, this research explores clustering data in the form of text in the social media Twitter, using the K-means algorithm. The elbow method is used to determine the number of clusters and the use of word clouds to visualize the words in each cluster. In this research using clustering techniques and association rules. Data retrieval or Twitter's data crawling process using Twitter's API key with the help of Rapid miner.

People can now generate, share or exchange ideas, content and other types of information thanks to the emergence of social media as one of the most important pillars of society. Texts, photos, videos and other types of information shared or traded on social media platforms can all be found among the vast amounts of data generated by users of these networks. Analyzing large amounts of social media data can lead to in-depth knowledge of user patterns, interests, and behaviors. There are several uses for the data, such as sentiment analysis, content recommendation, and market research [3]. In addition, social media text data is an important field for big data research because it can identify current trends and subjects and extract current trends and topics [4].

Text mining is applied to extract data from social media. Text mining is a multidisciplinary discipline that refers to information retrieval, data mining, machine learning, statistics, and computational linguistics. Information is in the form of text content such as news, technical papers, books, digital libraries, emails, blogs, and web pages. Examples of text mining generally include text categorization, text clustering, concept/entity extraction, granular taxonomy production, sentiment analysis, document summarization, and entity relationship modeling (i.e., studying relationships between named entities) [5].

Clustering is one of the well-known techniques in data mining that is used when no class is predicted and grouped based on data similarity. One of the clustering methods is k-means, which is often used because it is easy to use and efficient. It is easy to show that the total squared distance between each cluster point and its center is minimized when the cluster center is assigned as the centroid. The main objective is to reduce the total squared distance between each point and its cluster center. This is achieved by assigning each point to the nearest cluster center after a stable iteration [6].

Clustering techniques have been widely used in previous research, such as research conducted by [7] related to creating a program that can automatically create clusters from geotagged text data based not only on content but time, location and plotting cluster locations on a map. This was because for 10 years crisis and disaster management organizations were unable to effectively utilize the information provided by twitter data. This research shows that the performance of the tool is highly scalable and the technology can be used to identify where a disaster is, when it occurred and what resources should be used to stabilize the situation.

The use of big data analysis for social media is growing rapidly, along with the increasing use of big data in various sectors. Research conducted by [4] analyzed the text taken using big data text mining. Researchers used R to collect, analyze data and twitter to collect data. Text grouping is analyzed through a cluster dendrogram and produces a corpus and then groups similar entities from the term-document matrix, and eliminates rare terms. With this study, researchers can confirm the opinions and ideas of various participants and analyze the complex relationships between phenomena and predicted problems.

With false and misleading content on social media increasingly harmful, [8] conducted research on disinformation, fake news and propaganda in Finnish tweets using word cloud clustering, topic modeling, and word count clustering. From this study, it was found that topics on twitter related to disinformation, fake news, and propaganda often mention politics, both at home and abroad.

2. MATERIAL AND METHOD

2.1 Social Media

People can now generate, share or exchange ideas, content and other types of information thanks to the emergence of social media as one of the most important pillars of society. Texts, photos, videos and other types of information shared or traded on social media platforms can all be found among the vast amounts of data generated by users of these networks. Analyzing large amounts of social media data can lead to in-depth knowledge of user patterns, interests, and behaviors. There are several uses for the data, such as sentiment analysis, content recommendation, and market research [3]. In addition, social media text data is an important field for big data research because it can identify current trends and subjects and extract current trends and topics [4].

2.2 Text Mining

Text mining is applied to extract data from social media. Text mining is a multidisciplinary discipline that refers to information retrieval, data mining, machine learning, statistics, and computational linguistics. Information is in the form of text content such as news, technical papers, books, digital libraries, emails, blogs, and web pages. Examples of text mining generally include text categorization, text clustering, concept/entity extraction, granular taxonomy production, sentiment analysis, document summarization, and entity relationship modeling (i.e., studying relationships between named entities) [5].

2.3 Clustering

Clustering is one of the well-known techniques in data mining that is used when no class is predicted and grouped based on data similarity. One of the clustering methods is k-means, which is often used because it is easy to use and efficient. It is easy to show that the total squared distance between each cluster point and its center is minimized when the cluster center is assigned as the centroid. The main objective is to reduce the total squared distance between each point and its cluster center. This is achieved by assigning each point to the nearest cluster center after a stable iteration [6].

Clustering techniques have been widely used in previous research, such as research conducted by [7] related to creating a program that can automatically create clusters from geotagged text data based not only on content but time, location and plotting cluster locations on a map. This was because for 10 years crisis and disaster management organizations were unable to effectively utilize the information provided by twitter data. This research shows that the performance of the tool is highly scalable and the technology can be used to identify where a disaster is, when it occurred and what resources should be used to stabilize the situation.

2.4 Big Data

The use of big data analysis for social media is growing rapidly, along with the increasing use of big data in various sectors. Research conducted by [4] analyzed the text taken using big data text mining. Researchers used R to collect, analyze data and twitter to collect data. Text grouping is analyzed through a cluster dendrogram and produces a corpus and then groups similar entities from the term-document matrix, and eliminates rare terms. With this study, researchers can confirm the opinions and ideas of various participants and analyze the complex relationships between phenomena and predicted problems.

With false and misleading content on social media increasingly harmful, [8] conducted research on disinformation, fake news and propaganda in Finnish tweets using word cloud clustering, topic modeling, and word count clustering. From this study, it was found that topics on twitter related to disinformation, fake news, and propaganda often mention politics, both at home and abroad.

2.5 K-Means Clustering

The clustering process uses the k-mean operator. Clustering does not require any label attributes, as it falls into the category of unsupervised learning. K-means is a grouping of "n" observations which are combined using a clustering method to produce "k" clusters, which are then combined on the basis of certain similarities [12], This can be seen in Figure 2.

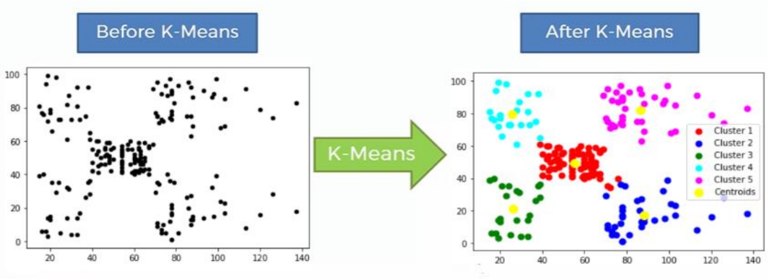

Figure 2. K-Means Clustering

The goal of clustering is to find similarities, regardless of the type of data being clustered. Therefore, before a clustering method can start to work, it must be able to calculate the similarity. In this research, the Elbow method is used to calculate the number of clusters. The objective function of k-means is given in Equation [13].

$$J = \min \sum_{j=1}^{k} \sum_{x_i \in C_i} \parallel x_i - \mu_j \parallel^2 \tag{1}$$

Where k is the number of clusters, μ_j is the centroid of the j^{th} cluster, x is the a data point, $\parallel x_i - \mu_j \parallel^2$ is the distance from the data point x_i to the cluster center, which is μ_j of the j^{th} cluster.

2.6 Association Rules

Association rule mining, or ARM One of the most important methods for finding and extracting valuable information from a big data set. Association rules show the relationships and interdependencies between a large set of data item. Certain criteria known as "Support" and "Confidence" are employed to identify these significant associations [14]. An association rule is shown by an if-then rule: X → Y, where X and Y are sets of attributes, and X is an antecedent part and Y is a consequent part [15]. Support: This measurement is computed using the parameters found in Formula (2). and indicates the proportion or total amount of transactions that contain both X and Y items.

$$\text{Support } ((X \rightarrow Y) = (X \cup Y) \tag{2}$$

Confidence: This measure, which is similarly computed using formula (3), indicates the extent to which a specific item depends on another; that is, it computes the extent to which the two sets X and Y are dependent on one another. It is used as a gauge to assess the strength of a rules.

$$\text{Confidence } (X \rightarrow Y) = \frac{P\ (X \cup Y)}{P\ (X)} \tag{3}$$

High-quality connections between the data can be found by association rule mining [16].

2.7 Flowchart Methodology

The research data used is Twitter text data totaling 3390 lines. The data taken is related to an internet service provider company in Indonesia. The company also uses twitter social media as one of the platforms to receive community complaints as input or suggestions. The research will apply two different algorithms, namely k means clustering and association rules. The fundamental principle of the K-means algorithm, which is one of the most commonly used clustering techniques, is to divide data into K clusters by minimizing the sum of distances between each data point and the center of each respective cluster [9]. Association rule mining is a technique that is widely used in many areas of data mining that allows the identification of trends, frequent patterns, and relationships among the data. The method is designed to show relationships among various data elements within a dataset [10]. Figure 1 shows the stages of the research conducted.

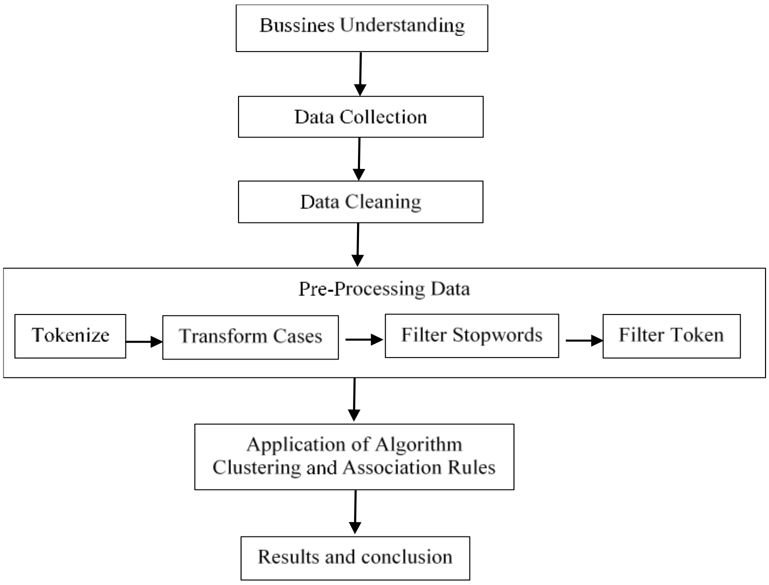

Figure 1. Research Methodology

1. Collecting Data
   The data that will be used will be retrieved from the twitter social media with the help of the twitter search operator provided by Rapidminer. Access tokens are used for accessing and retrieving data from the tweet database. Once the parameters for the twitter search operator have been set, a collection of examples of twitter tweets will be generated [11]. In this experiment, the query @indihomecare was applied. The results of the parameterized search are either the most recent or the most popular tweets. These results contain different types of attributes and labels, such as the language, the date and time of creation, the original text, the number of retweets, the user name, and the tweet ID. Data was retrieved incrementally, yielding 3390 data rows. Microsoft Excel is used to summarize the resulting data. In June 2023 the data collection was done.

2. Data Cleaning
   This stage is used to clean text data from characters like HTML links, mentions, usernames, hashtags, retweets, numbers, punctuation, and spaces. Filter examples are used to return data according to the required conditions and remove duplicates to remove the same data in the data set.

3. Pre-Processing Data
   The next step is to process the textual information to ensure the accuracy and consistency of the record and to correct or remove unnecessary information. Perform the following steps:
   a. Tokenization is the separation of text into words, which is the first step in the process.
   b. Transform Cases is used to convert the word characters in the text to lowercase letters
   c. The stopwords filter is used to eliminate terms that are not used in the document. The stopwords filter is Indonesia. The stopwords filter used is Indonesia.
   d. Token filter is used to remove tokens on the basis of the specified length. The minimum and maximum length is defined as 4 and 25 characters respectively.

3. RESULTS AND ANALYSIS

The study applies the k-means clustering technique and association rules. Applying clustering technique, it produces 4 clusters on Twitter topics on indihomecare Twitter accounts related to complaints. Applying the Association Rules technique produces word relationships on Twitter topics related to complaints by determining the support value. Support values dictate how many associations are produced, with lower support values resulting in more word use [17]. Results are visualized using word cloud and conclusions can be drawn.

3.1. Data Cleaning

This stage is used to clean text data from characters such as HTML links, mentions, usernames, hastags, retweets, numbers, punctuation marks and excess spaces. Setelah melalui. The result shown in table 1.

Table 1. Data Cleaning

Before	After
Aduh bajingan, ini kok loading terus twt gue. Eror lagi kah Indihome?	Aduh bajingan ini kok loading terus twt gue Eror lagi kah Indihome
malam, cek tagihan indihome saya berapa ya?	malam cek tagihan indihome saya berapa ya
@notturbsns Hi, Kak Neng. DM-nya akan segera kami balas, silakan lakukan pengecekan secara berkala DM dari kami.	notturbsns Hi Kak Neng DMnya akan segera kami balas silakan lakukan pengecekan secara berkala DM dari kami

3.2. Pre-Processing Data

There are several sub-process operators applied. The following is the set of operators used:

3.2.1. Tokenize

Tokenize is used to split the text into a series of words, and the result is shown in table 2.

Table 2. Document Tokenize Process

Before	After
Aduh bajingan ini kok loading terus twt gue Eror lagi kah Indihome	[‘Aduh’. ‘bajingan’, ‘ini’, ‘kok’, ‘loading’, ‘terus’, ‘twt’, ‘gue’, ‘Eror’, ‘lagi’, ‘kah’, ‘Indihome’]
malam cek tagihan indihome saya berapa ya	[‘malam’, ‘cek’, ‘tagihan’, ‘indihome’, ‘saya’, ‘berapa’, ‘ya,]
notturbsns Hi Kak Neng DMnya akan segera kami balas silakan lakukan pengecekan secara berkala DM dari kami	[‘notturbsns’, ‘Hi’, ‘Kak’, ‘Neng’, ‘DMnya’, ‘akan’, ‘segera’, ‘kami’, ‘balas’, ‘silakan’, ‘lakukan’, ‘pengecekan’, ‘secara’, ‘berkala’, ‘DM’, ‘dari’, ‘kami’]

3.2.2. Transform Cases

The stage of transforming cases, it is done to homogenize the letters by changing the text from uppercase to lowercase. The result is shown in table 3.

Table 3. Document Transform Cases Process

Before	After
Aduh bajingan ini kok loading terus twt gue Eror lagi kah Indihome	aduh bajingan ini kok loading terus twt gue eror lagi kah indihome
malam cek tagihan indihome saya berapa ya	malam cek tagihan indihome saya berapa ya
notturbsns Hi Kak Neng DMnya akan segera kami balas silakan lakukan pengecekan secara berkala DM dari kami	notturbsns hi kak neng dmnya akan segera kami balas silakan lakukan pengecekan secara berkala dm dari kami

3.2.3. Filter stopwords

This operator is used to remove terms that are not used in the document. The result is shown in table 4.

Table 4. Document Filter Stopwords Process

Before	After
Aduh bajingan ini kok loading terus twt gue Eror lagi kah Indihome	Aduh bajingan loading twt gue eror kah indihome
malam cek tagihan indihome saya berapa ya	malam cek tagihan indihome ya
notturbsns Hi Kak Neng DMnya akan segera kami balas silakan lakukan pengecekan secara berkala DM dari kami	notturbsns hi kak neng dmnya balas silakan lakukan pengecekan secara berkala dm

3.3. Filter Token

This operator is used to remove tokens based on the specified length. The result in shown in table 5.

Table 5. Document Filter Token Process

Before	After
Aduh bajingan ini kok loading terus twt gue eror lagi kah Indihome	Aduh bajingan loading eror indihome
malam cek tagihan indihome saya berapa ya	malam cek tagihan indihome
notturbsns Hi Kak Neng DMnya akan segera kami balas silakan lakukan pengecekan secara berkala DM dari kami	notturbsn dmnya balas silakan lakukan pengecekan berkala

TFIDF vectorizer tokenizes text documents by converting them into vectors based on word relevance. Each vocabulary will be checked and the frequency weight of the data will be inverted. Transformed into a feature vector, the text can be used as input by the estimator. The vocabulary has a feature index for each token, which serves as a dictionary to convert each word or token into a feature index in a matrix. The weight of each vector, which is represented by an integer, corresponds to its feature [18].

3.4. K-Means

The data will produce 4 cluster models after modeling the data set using the k-means algorithm with parameter k = 4. The clusters start with 0 because in programming languages 0 is the first number of the numbering sequence. The number of clusters is obtained by applying the elbow method, which shows that additional clusters do not improve the overall clustering when added, and the total number of squares only decreases to a minimum [19]. Figure 3 shows a relationship between average centroid distance and cluster value k, where the plot shows the number of elbows at cluster 4.

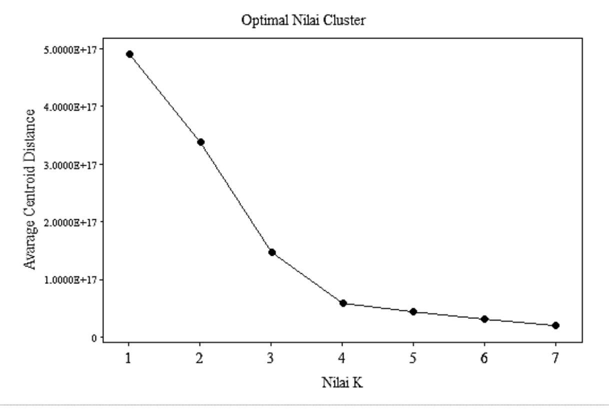

Figure 3. Elbow Method

Getting the value k=4 produces a relationship related to the topic on indihome's Twitter account when there is a complaint. The word cloud is the description of the words that are most frequently used in each cluster. A word cloud is a graphical display of textual data that can be used to view free form text or to display the metadata (tags) of keywords on a web page [20].

In Figure 4, cluster 0 is related to public complaints on the indihome care account and is responded to by indihome. Indihome care will answer these problems by checking and it is hoped that people can check regularly on social media accounts. Based on cluster 1 the word that often appears is the word “berkala” followed by the words “terima”, “balas” and “pengecekan”. Based on cluster 0, this relates to indihome's treatment of customers who make complaints.

In Figure 5, cluster 1, in case of a complaint, IndiHomecare users can report through whatsapp or through official social media accounts. When connected to the indihomecare Twitter account, the word that appears most often is “Whatsapp”. There are words that have the most occurrences in comparison to other words, namely “channel” and “akun”. Based on cluster 1, this is related to the service reports provided by IndiHomecare for questions about products and complaints related to IndiHomecare services.

In Figure 6, cluster 2 customers complained about the unstable network, so through social media X customers complain or report when it is connected to the Twitter account of indihomecare, the word that appears most often is “Jaringan”. There are some words that have the highest number of occurrences when compared to the other words, namely “Widhi”, “Bantu”. Based on cluster 2, this is related to customers complaining about unstable Internet networks.

In Figure 7, cluster 3 When a complaint occurs, IndiHome asks the customer to facilitate verification. When it is connected to the Twitter account of indihomecare, the word that appears most often is “Data”. There are some words that appear the most compared to other words, namely “Nama”, “Pemilik”, which is related to checking customer data. Based on cluster 3, this is related to the verification of customer data.

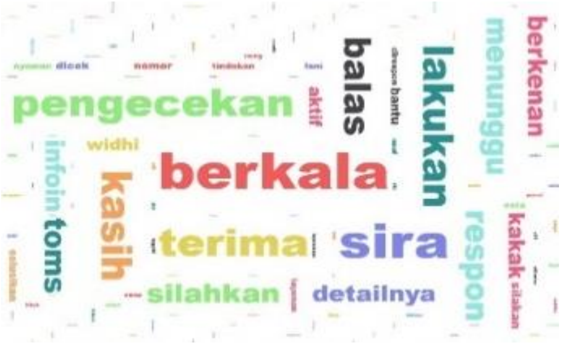

Figure 4. Word Cloud Cluster 0

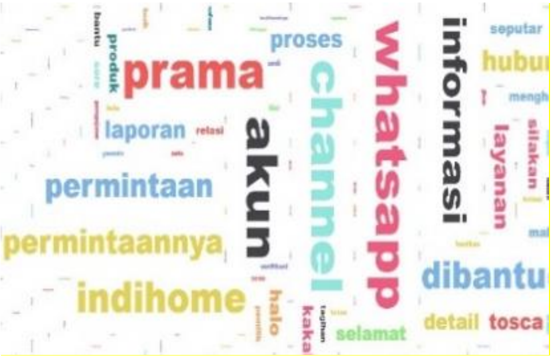

Figure 5. Word Cloud Cluster 1

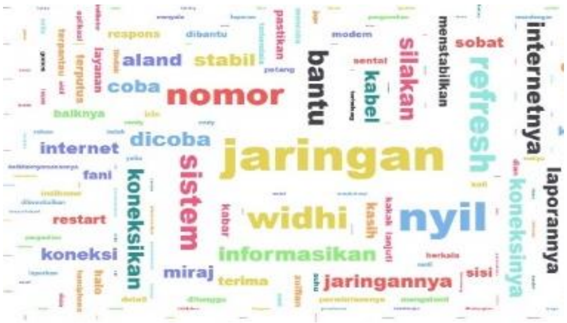

Figure 6. Word Cloud Cluster 2

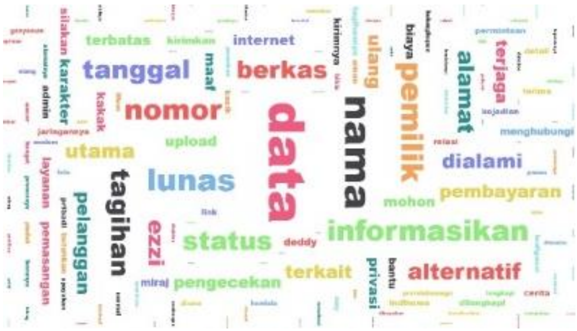

Figure 7. Word Cloud Cluster 3

3.5. Assosiation Rules

Several criteria were chosen from the obtained data, including the support and confidence values that were higher than the rest. The 5 support values with the highest values are taken, so they are shown in the table 6.

Table 6. Support Value and Confidence Value

No	Premises	Conclusion	Support	Confidence
1	Indihome, channel, akun, whatsapp	Dibantu	0,057	0,829
2	Akun	Whatsapp	0,071	0,831
3	Indihome, akun	Channel, Whatsap	0,069	0,833
4	Whatsapp	Indihome, Channel, Akun	0,069	0,833
5	Akun	Indihome, Channel	0,072	0,838

In table 6 the word indihome is most often associated with assisted, account and whatsapp. the 2850 twitter data, the words indihome, channel, account, whatsapp have a support of 0.057 followed by the word account. Discussion on the @indihomecare twitter account, if there are complaints that occur, the words above are the most widely used. Customers make complaints when problems occur and most problems are related to internet connections that are dead or slow, payments that are not appropriate and related to service. From the word relationship above, it can be seen that the @indihomecare twitter account always responds to service complaints by asking users to contact the whatsapp channel or via the indihomecare DM account so that it can explain the handling procedures that will be carried out clearly to customers. It is hoped that indihome will be more alert and fast, especially in handling complaints and continue to improve services by listening to every customer complaint.

4. CONCLUSION

This study analyzed 2,850 data points after processing an initial dataset of 3,390. Applying the elbow method resulted in an optimal K-value of 4, yielding four distinct clusters: Cluster 0 with 578 items, Cluster 1 with 1,498 items, Cluster 2 with 461 items, and Cluster 3 with 313 items. Each cluster revealed unique themes based on frequently occurring words. Cluster 0 focused on community complaints addressed by Indihome, with "Berkala" (periodic) as a key term. Cluster 1 centered on Indihome's services and reporting channels, prominently featuring "WhatsApp." Cluster 2 highlighted customer complaints about network instability, frequently mentioning "Jaringan" (network). Cluster 3 emphasized Indihome's requests for user data to facilitate issue resolution, with "Data" as a recurring term.

The application of association rules revealed that the words "indihome," "channel," "akun" (account), and "WhatsApp" appeared most frequently, with a support value of 0.057. These terms were often associated with the word "Dibantu" (assisted), suggesting a connection to customer support interactions. In conclusion, this analysis indicates that customers experiencing network disruptions can report issues through Indihome's social media accounts or WhatsApp, reflecting a multi-channel approach to customer service. These findings provide valuable insights for Indihome to enhance their customer service strategies and potentially improve their network stability based on the prevalent issues identified in the clusters. The results underscore the importance of efficient communication channels and prompt issue resolution in maintaining customer satisfaction in the telecommunications industry.

REFERENCES

[1] P. Iswara, “Jumlah Pengguna Twitter di Indonesia Capai 14,75 Juta per April 2023, Peringkat Keenam Dunia,” Databoks.

[2] M. Saari, L. Haapanen, and P. Hurmelinna-laukkanen, “Social media and international business : views and conceptual framing,” vol. 39, no. 7, pp. 25–45, 2022, doi: 10.1108/IMR-06-2021-0191.

[3] I. Ardhanayudha, F. Nurrohman, I. Haryani, and A. Alamsyah, “Understanding service quality concerns from public discourse in Indonesia state electric company,” *Heliyon*, vol. 9, no. 8, p. e18768, 2023, doi: 10.1016/j.heliyon.2023.e18768.

[4] Y. Jeong and Jin-Heeku, “A study on social big data analysis using text clustering,” vol. 7, pp. 1–4, 2018.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edit. 2012.

[6] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, Third Edit. United States of America: Elsevier, 2011.

[7] S. A. Barnard, S. M. Chung, and V. A. Schmidt, “Content-based Clustering and Visualization of Social Media Text Messages,” 2017.

[8] J. Jussila, A. H. Suominen, and A. Partanen, “Text Analysis Methods for Misinformation – Related Research on Finnish Language Twitter,” pp. 1–16, 2021.

[9] X. Chen, C. Liu, B. Lin, J. Lai, and D. Miao, “AHA-3WKM: The optimization of K-means with three-way clustering and artificial hummingbird algorithm,” *Inf. Sci. (Ny).*, vol. 672, no. November 2023, p. 120661, 2024, doi: 10.1016/j.ins.2024.120661.

[10] B. Mudumba and M. F. Kabir, “Mine-first association rule mining: An integration of independent frequent patterns in distributed environments,” *Decis. Anal. J.*, vol. 10, no. February, p. 100434, 2024, doi: 10.1016/j.dajour.2024.100434.

[11] A. S. Halibas, “Application of Text Classification and Clustering of Twitter Data for Business Analytics,” pp. 1–7, 2018.

[12] O. Iparraguirre-villanueva *et al.*, “Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm,” vol. 13, no. 6, pp. 571–578, 2022.

[13] A. Şenol, “ImpKmeans: An Improved Version of the K-Means Algorithm, by Determining Optimum Initial Centroids, based on Multivariate Kernel Density Estimation and Kd-Tree,” *Acta Polytech. Hungarica*, vol. 21, no. 2, pp. 111–131, 2024, doi: 10.12700/APH.21.2.2024.2.6.

[14] M. F. Kabir, S. A. Ludwig, and A. S. Abdullah, “Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining,” *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data 2018*, pp. 2433–2441, 2018, doi: 10.1109/BigData.2018.8622028.

[15] S. Mabu, T. Higuchi, and T. Kuremoto, “SemiSupervised Learning for Class Association Rule Mining Using Genetic Network Programming,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 15, no. 5, pp. 733–740, 2020, doi: 10.1002/tee.23109.

[16] Z. F. Sokhangoee and A. Rezapour, “A novel approach for spam detection based on association rule mining and genetic algorithm,” *Comput. Electr. Eng.*, vol. 97, no. January 2022, 2022, doi: 10.1016/j.compeleceng.2021.107655.

[17] J. Tamaela, E. Sediyono, and A. Setiawan, “Implementasi Metode Association Rule untuk Menganalisis Data Twitter tentang Badan Penyelenggara Jaminan Sosial dengan Algoritma Frequent Pattern-Growth,” *J. Sist. Inf. Bisnis*, vol. 8, no. 1, p. 25, 2018, doi: 10.21456/vol8iss1pp25-33.

[18] P. S. Reddy, D. Renu Sri, C. S. Reddy, and S. Shaik, “Sentimental Analysis using Logistic Regression,” *Int. J. Eng. Res. Appl. www.ijera.com*, vol. 11, no. 7, pp. 36–40, 2021, doi: 10.9790/9622-1107023640.

[19] R. Sitaram, “An employee segmentation model,” no. June, 2021.
[20] N. Garg and R. Rani, “Analysis and Visualization of Twitter Data using k-means Clustering,” *Int. Conf. Intell. Comput. Control Syst.*, pp. 670–675, 2017.

BIBLIOGRAPHY OF AUTHORS

Cristiany Gunu Lengari, Graduated with a bachelor's degree in chemical engineering and continued her master's degree in human resource development specializing in data analytics at Surabaya Airlangga University.

Ira Puspitasari, Lecturer at Faculty of Science and Technology Universitas Airlangga with interests in IT-Business alignment, data analytics, and e-health. she pursued her bachelor's and master's degrees in engineering at the Bandung Institute of Technology, then continued her doctoral education at Osaka University, Japan. During his career as a lecturer, there are currently 19 researches that have been successfully published since 2006 - 2020. Her latest publication in 2020 is titled “Making the Information Technology (IT) business alignment works: a framework of IT-based competitive strategy”. She is known as a lecturer in the S1 Information Systems study program who teaches subjects such as Information Systems Innovation and Technology, Enterprise Architecture Planning, E-health Interaction System Design.