JURNAL PENDIDIKAN TEKNOLOGI KEJURUAN

Vol. 3, No. 3 August 2020

P-ISSN: 2621-3273
E-ISSN: 2621-1548

Talk Show Segmentation System Based on Twitter Using K-Medoids Clustering Algorithm

Kharisma Jevi Shafira Sepyanto, Yulison Herry Chrisnanto and Fajri Rakhmat Umbara
Universitas Jenderal Achmad Yani
*Corresponding author: kharismashafiraa@gmail.com

Abstract— Innovations on a talk show on television can be a threat. Audience will be divided into groups so that it can make a downgrade rating program. Program ratings affect companies that will use advertising services. Television companies will go bankrupt. The biggest source of income is sales of advertising services. One way to overcome them can be analyzed in public opinion. The results of the analysis can provide information about the attractiveness of the community towards the program. But the analysis process takes a long time and can be done only by a competent person so another process is needed to get the results of the analysis that is fast and can be done by anyone. In this study using K-Medoids Clustering in the process of identifying public opinion. The clustering process known as unsupervised learning will be combined with the labeling process. The previous episode's tweet data will be labeled and then used to obtain the predicted labels from other cluster members. Before going through the clustering stage, the tweet data will go through the text preprocessing stage then transformed into a numeric form based on the appearance of the word. Transformation data will be clustered by calculating proximity using Cosine Similarity. Labels from the Medoids cluster will be used on unlabeled tweet data. The cluster results were tested using the Silhouette Coefficient method to get 0.19 results. However, this method successfully predicted public opinion and achieved an accuracy of 80%.

Keywords: **twitter segmentation, k-medoids clustering, cosine similarity, data transformation, silhouette coefficient**

I. INTRODUCTION

Data that is always increasing every second can be used to obtain information. The process of processing data or data mining can be done in various ways, one of which is clustering. Clusterization is a method of grouping data based on the degree of similarity of characteristics to one another. The clustering method is divided into two, namely hierarchical clustering and partitional clustering. On partitional clustering there are K-Means Clustering, K-Medoids Clustering (PAM) and CLARA algorithms. In the field of energy efficiency in finding and analyzing energy periodicity using the K-Medoids Clustering and CLARA algorithms, the K-Medoids Clustering algorithm is the best algorithm in the case of matrix calculations using Euclidean Distance (Ruiz, Pegalajar, Arcucci, & Molina-Solana, 2020). In the application of 10,000 KEEL transactions with the K-Means and K-Medoids algorithms, the results show that K-Medoids is superior in execution time and is not noise sensitive (Arora, Deepali, & Varshney, 2016). Many other fields use clustering in processing data.

Clusterization can be applied to the broadcasting field. Analysis of public opinion succeeded in producing the necessary information. A study processes data to produce information on people's lifestyles (Li, 2020). Clustering with the K-Means algorithm gets information on program production decisions to determine broadcast schedules so as to maintain the rating of the program (Kui et al., 2020). This makes advertisers interested in using advertising services on TV. They think that television stations with high ratings are more competent (Pribadi, Yoedtadi, & Siswoko, 2017). In this study using unsupervised learning types.

In this study, built a system for segmenting public opinion on social media twitter using K-Medoids Clustering. The difference with other segmentation research is that in this research, the clustering process of public opinion will use a labeling process. In addition, tweet data that has been labeled will be used at the label prediction stage in the tweet episode reruns data. Analysis of public opinion succeeded in producing the necessary information (Devika, Sunitha, & Ganesh, 2016). The "free" nature of social media makes everyone tend to express their views in the form of comments (Hutto

& Gilbert, 2014)(Ahuja, Chug, Kohli, Gupta, & Ahuja, 2019). A literature review identified 13 studies that applied different clustering methods. The results show that the use of unsupervised types of learning in mining social media data has several weaknesses (Guftar, Ali, Raja, & Qamar, 2015).

A study to try to reduce costs in the clustering process tried to implement a labeling process and succeeded in reducing costs by 50-60% (Shuyang, Heittola, & Virtanen, 2017)(Darnstadt, Meutzner, & Kolossa, 2014). This process was also successfully carried out on the K-Medoids algorithm in the active learning method (Shuyang, Heittola, & Virtanen, 2018)(Ji, Wang, & Ma, 2019). Seeing from this research, this study will combine supervised learning and unsupervised learning, which is expected to help determine the label on the tweet data that will be processed. The tweet data will then go through the preprocessing stage and be transformed using term frequency calculations. Tweet data in numeric form will be clustered using K-Medoids Clustering with Cosine Similarity calculation. K-Medoids is a partitional clustering algorithm that has the aim of breaking the dataset into groups. The KMedoids algorithm can reduce new data outside the segment to enter the cluster center (Tan, 2018). The label for each cluster will be predicted based on the majority of existing labels.

II. METHOD

A. Data Mining

Data mining is the process of finding meaningful information through new correlation patterns and trends by sorting through large amounts of data stored in repositories using pattern recognition techniques as well as statistical and mathematical techniques. The stages in data mining are data selection, data cleaning, transformation, data mining and interpretation.

B. Twitter

Twitter social media is a service for friends, family, and co-workers to communicate and stay connected through exchange of messages. Besides being used as a means of communication, Tweets can also be used in the analysis stage (Hutto & Gilbert, 2014). Twitter social media is quite good when it becomes an object in analyzing a text.

Analytical techniques can be proven tools for extracting information. Every information is obtained from various reviews or Tweets uploaded by Twitter users. Twitter social media is quite good if it becomes an object in analyzing a text (Hutto & Gilbert, 2014)(Ahuja et al., 2019). The text is limited to 140 characters, making the information conveyed by the public more meaningful (Dos Santos & Gatti, 2014).

C. Preprocessing

Preprocessing is done so that the data is ready to be processed. The preprocessing stage is carried out such as the text mining stage for information retrieval, text classification and text grouping (Vijayarani, Ilamathi, & Nithya, 2016). In the process, using 4 stages, namely case folding, tokenizing, filtering and stemming.

The case folding stage is the stage of changing it to the same type, which can be upper and lowercase letters and eliminating notations other than letters. The tokenizing stage is the stage of cutting sentences into words. The filtering stage is the stage of data collection and deletion of unused words. The last is the stemming stage, which is a grouping of other words that have a similar root.

D. Transformation Data

The transformation stage is carried out by term frequency. Term Frequency (TF) is the frequency of appearance of a term in the document concerned. The greater the number of occurrences of a word in the document, the greater its weight or will provide a greater suitability value Tahap transformasi dilakukan dengan term frequency (Chrisnanto & Abdillah, 2015).

E. Clustering

Clustering is a method of grouping data into different groups, so that the data in each group has the same trends and patterns. K-Medoids is an algorithm that represents clusters formed using a central point originating from cluster members. The stages are:

a. Initialize early medoids. To determine the optimal k value, you can use the Elbow technique (Guftar et al., 2015).
b. Initial distance.
c. Apply each data (object) to the nearest cluster using the Cosine Distance equation in calculating the distance.
d. Randomly selecting non-medoids as candidates for new medoids.
e. Calculation of the distance between old medoids and new medoids. If the difference is greater than 0, repeat to step c.

$$Cosine\ Distance = 1 – Similarity \quad (1)$$

$$Similarity\ (x,y) = \cos\theta \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \qquad (2)$$

The equation used in calculating distance is Cosine Distance which can be seen in (1) and Cosine Similarity in (2). where x & y is the document being compared, $\sum_{i=1}^{n} x_i y_i$ is the total word weight of document x to document y, $\sqrt{\sum_{i=1}^{n} x_i^2}$ is the root of the total document weight x and $\sqrt{\sum_{i=1}^{n} y_i^2}$ is the root of the total document weight y. The cluster results will be tested using the Silhouette Coefficient method. The Silhouette Coefficient is a stage where the accuracy of the calculations performed by K – Medoids is tested to see how accurate the classification is performed by the K – Medoid method (Dos Santos & Gatti, 2014). Calculation using equation 3.

$$S = \frac{b-a}{\max\,(a,b)} \qquad (3)$$

[redacted] average distance between the n [redacted] e average distance between the [redacted] jects outside the cluster. The calculation of the distance in the cluster and the calculation of the distance outside the cluster using Euclidean Distance with equation 4.

$$d\ (x,y) = \sum \sqrt{(x_i - y_i)^2} \quad i = 1\ ;\ 1, 2, 3, \ldots n \qquad (4)$$

d is distance, x and y are centroid variable values.

III. RESULT AND DISCUSSION

The system is built through three stages, namely input, process and output. In the input, there is twitter data which is divided into two, namely training data and test data. This test data appears because the method used is a combination of supervised and unsupervised as has been presented in the background.

At the input stage, there are two types of processes described in the experimental data process. At the process and output stages, there are several stages starting from preprocessing to getting labels. The illustration of the twitter social media segmentation system can be seen in Figure 1.

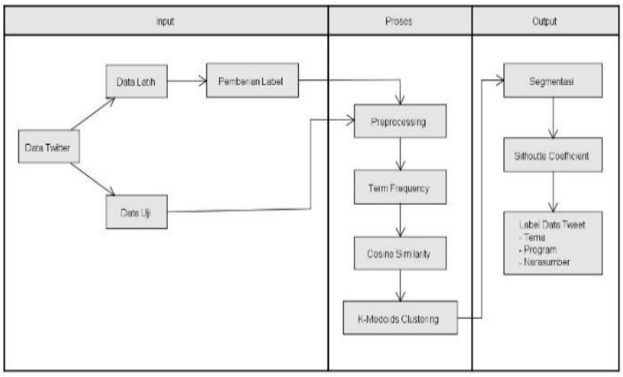
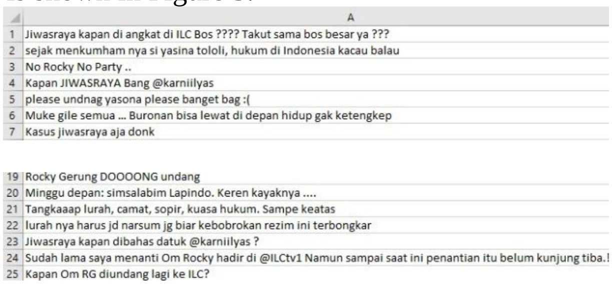


Figure 1 Illustration

A. Data

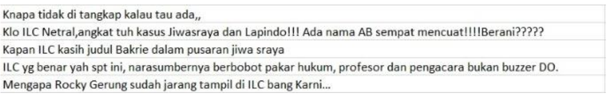
The input process uses data from social media twitter from each talkshow episode. The Tweet used is the 2020 posting year. The Tweet data used will be divided into two types, namely Tweet data during the first impression and Tweet data during the replay. The training data is written using the CSV format as shown in Figure 2 and the test data is shown in Figure 3.



	A
1	Jiwasraya kapan di angkat di ILC Bos ???? Takut sama bos besar ya ???
2	sejak menkumham nya si yasina tololi, hukum di Indonesia kacau balau
3	No Rocky No Party ..
4	Kapan JIWASRAYA Bang @karniilyas
5	please undnag yasona please banget bag :(
6	Muke gile semua ... Buronan bisa lewat di depan hidup gak ketengkep
7	Kasus jiwasraya aja donk

19	Rocky Gerung DOOOONG undang
20	Minggu depan: simsalabim Lapindo. Keren kayaknya ....
21	Tangkaaap lurah, camat, sopir, kuasa hukum. Sampe keatas
22	lurah nya harus jd narsum jg biar kebobrokan rezim ini terbongkar
23	Jiwasraya kapan dibahas datuk @karniilyas ?
24	Sudah lama saya menanti Om Rocky hadir di @ILCtv1 Namun sampai saat ini penantian itu belum kunjung tiba.!
25	Kapan Om RG diundang lagi ke ILC?

Figure 2 Tweet Data

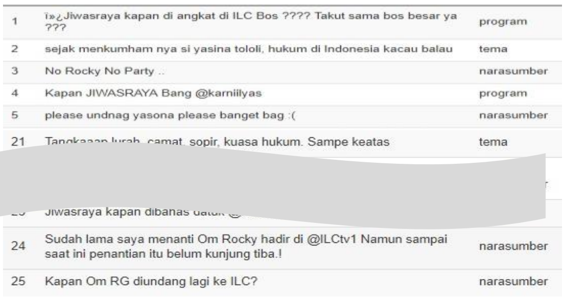


Knapa tidak di tangkap kalau tau ada,,
Klo ILC Netral,angkat tuh kasus Jiwasraya dan Lapindo!!! Ada nama AB sempat mencuat!!!!Berani?????
Kapan ILC kasih judul Bakrie dalam pusaran jiwa sraya
ILC yg benar yah spt ini, narasumbernya berbobot pakar hukum, profesor dan pengacara bukan buzzer DO.
Mengapa Rocky Gerung sudah jarang tampil di ILC bang Karni...

Figure 3 Data Testing

B. Labeling Process

At the input stage, there is a process for tagging Tweets. The clustering process is one type of unsupervised learning. However, at this stage, labeling is used to make it more optimal in segmentation results. There are three types of labels used. The determination of the label is carried out by the authorized party in the company. Data that is labeled is Tweet data at the time of the first broadcast (not replay). The data can be seen in Figure 4.



1	ï»¿Jiwasraya kapan di angkat di ILC Bos ???? Takut sama bos besar ya ???	program
2	sejak menkumham nya si yasina tololi, hukum di Indonesia kacau balau	tema
3	No Rocky No Party ..	narasumber
4	Kapan JIWASRAYA Bang @karniilyas	program
5	please undnag yasona please banget bag :(	narasumber
21	Tangkaaan lurah, camat, sopir, kuasa hukum. Sampe keatas	tema
23	Jiwasraya kapan dibahas datuk @	
24	Sudah lama saya menanti Om Rocky hadir di @ILCtv1 Namun sampai saat ini penantian itu belum kunjung tiba.!	narasumber
25	Kapan Om RG diundang lagi ke ILC?	narasumber

Figure 4 Data with Label

C. *Preprocessing*

Preprocessing is done so that the data is ready to be processed. The preprocessing stage is carried out such as the text mining stage for information retrieval, text classification and text grouping. In the process, using 4 stages, namely case folding, tokenizing, filtering and stemming.

a. Case Folding

This first stage is called case folding. In the process all data is converted to lowercase. In addition, characters other than letters 'a' through 'z' will be removed as delimiters. Delimiters covering all characters other than ASCII notation will be deleted.

@KarniIlyas jgn lupa diundang @AdhieMassardi	⇨	karniilyas jgn lupa diundang adhiemassardi

b. Tokenizing

The second stage is the stage of cutting sentences. This stage broadly breaks down a set of characters in a text. Characters that are discussed, such as spaces, punctuation or others that have a function as a pause.

karniilyas jgn lupa diundang adhiemassardi	⇨	Karniilyas jgn lupa diundang adhiemassardi

c. Filtering

The third stage is taking words from the results of the previous stage. One of the applications is by removing words that are considered stopwords. Types of stopwords such as conjunctions, affixes and other words that have the same function. Additional stopwords are made to suit the case. Words that are mentioned too often will be deleted.

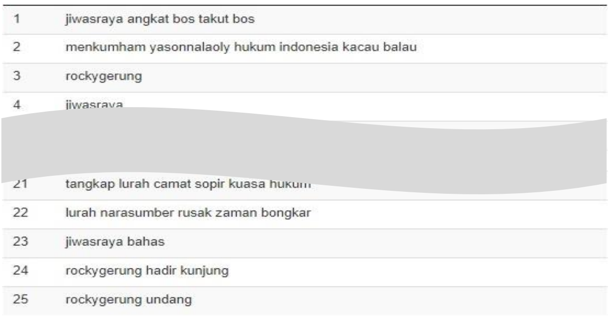
Karniilyas jgn lupa diundang adhiemassardi	⇨	Karniilyas jgn lupa undang adhiemassardi

d. Stemming

The last step is needed to reduce the number of different indexes of a document. This stage performs a grouping of other words which have the same root but have different forms. The stemming stage also processes non-standard written words, typos and languages other than Indonesian. Words that have the same meaning or synonym will be counted as different words.

Karniilyas jgn lupa undang adhiemassardi	⇨	Karniilyas jangan lupa undang adhiemassardi

The results of Tweet preprocessing can be seen in Figure 5.



1	jiwasraya angkat bos takut bos
2	menkumham yasonnalaoly hukum indonesia kacau balau
3	rockygerung
4	jiwasraya
21	tangkap lurah camat sopir kuasa hukum
22	lurah narasumber rusak zaman bongkar
23	jiwasraya bahas
24	rockygerung hadir kunjung
25	rockygerung undang

Figure 5 Preprocessing Result

D. *Transformation using TF*

Data transformation is the process of calculating the weight of the processed Tweet text. In determining the cluster, the data needs to be calculated using a formula. So Tweets need to be transformed into numeric data. One of the processes using data will be calculated based on term frequency (TF).

E. *Cluster using K-Medoids*

In the fifth stage, the process of determining clusters using the K-Medoids Algorithm. The transformed data set was calculated using the k-medoids formula to obtain a fixed medoid. The condition for getting a fixed medoid is if the difference between the total distance and the total distance on the new medoid is less than 0. The ideal k value is determined by the Elbow method. Elbow method is a method to find out information in determining the best number of clusters. The results can be seen in Figure 6.

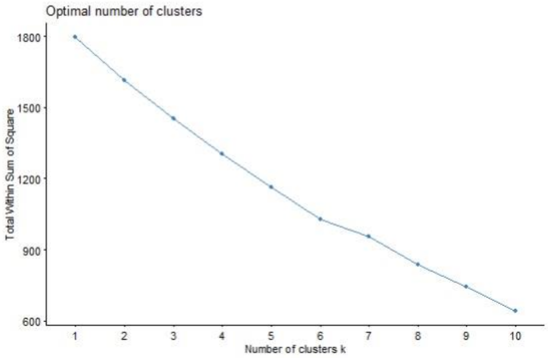


Figure 6 K using Elbow

The graphic above shows the elbow in figure 6.The ideal k value used when calculating the cluster with K-Medoids is 6. The results of the cluster can be seen in Figure 7.

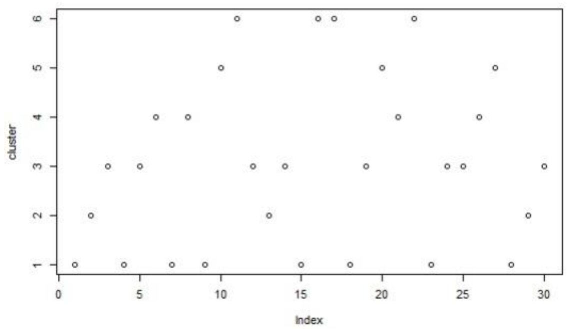


Figure 7 Grafik Cluster

In Figure 7 above the x-axis shows the Tweet number and the y-axis shows the location of the cluster.

F. Segment Identification

The sixth stage is segment identification. This stage will show the cluster results and Tweet data labels. Conclusions can be drawn from the majority of labels that appear in each cluster formed. Tweet data on the reruns episode will have labels based on the majority of labels formed in the cluster. Cluster results can be seen in Table 1.

Table 1 Cluster Result

Cluster	Data	Majority of Labels
C1	8 tweet {1, 4, 7, 9, 15, 18, 23, 28}	Program
C2	3 tweet {2, 13, 29}	Tema, Program, Narasumber
C3	8 tweet {3, 5, 12, 14, 19, 24, 25, 30}	Narasumber
C4	4 tweet {6, 8, 21, 26}	Tema
C5	3 tweet {10, 20, 27}	Program
C6	4 tweet {11, 16, 17, 22}	Tema, Narasumber

Tweets that will be searched for labels are Tweets 26, 27, 28, 29 and 30. Based on Table 1, it can be concluded that the labels of the Tweets you are looking for correspond to the majority of labels. The results of label classification can be seen in table 2.

Table 2 Prediction Result

Tweet	Cluster	Label
26	C4	Tema
27	C5	Program
28	C1	Program
29	C2	Tema, Program, Narasumber
30	C3	Narasumber

G. Testing

The last stage is testing the system that has been built. Tests were carried out on the data for the cluster quality test and the label results accuracy test. Testing the quality of the cluster formed using the Silhouette Coefficient method. The quality test results can be seen in Figure 8.

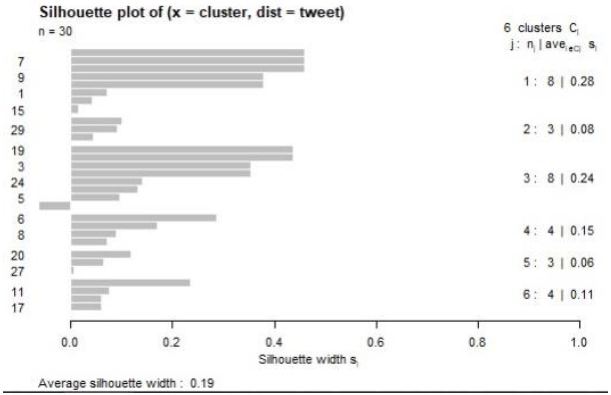


Figure 8 Result Silhouette Coefficient

On average, the 6 clusters formed get 0.19 quality, which means that the cluster formed has a weak structure. Label accuracy testing can be seen in Table 3.

Table 3 Testing Result

Testing	Result
1	79%
2	84%
3	79%
4	60%
5	98%

The results of the five tests obtained an average of 80%.

IV. CONCLUSION

K-Medoids Clustering as one of the clustering algorithms has successfully classified Tweet data. Test data in the form of Tweet episodes reruns of 5 Tweets and training data in the form of Tweet data on the first broadcast of 25 Tweets resulting in 6 clusters. The results of the cluster quality test are considered to have a weak structure because they only have a value of 0.19. This happens because the data set is transformed using the term frequency method. Words that have the same meaning will be considered different by this method so that the transformation results are less representative. Unlike the label accuracy testing, the system built from the five tests managed to get 80% points and was declared successful in showing public opinion on Twitter about the program.

REFERENCES

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science*, *152*, 341–348. https://doi.org/10.1016/j.procs.2019.05.008

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Procedia Computer Science*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Chrisnanto, Y. H., & Abdillah, G. (2015). Gambaran Umum Kemampuan Akademik Mahasiswa Unjani Dengan Algoritma Partitioning Around Medoids (PAM) Clustering. *Seminar Nasional Ilmu Pengetahuan Dan Teknologi*, 285–290.

Darnstadt, M., Meutzner, H., & Kolossa, D. (2014). Reducing the Cost of Breaking Audio CAPTCHAs by Active and Semi-supervised Learning. *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, 67–73. https://doi.org/10.1109/ICMLA.2014.16

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*, *87*, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

Dos Santos, C. N., & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. *International Conference on Computational Linguistics*, 69–78. Ireland.

Guftar, M., Ali, S. H., Raja, A. A., & Qamar, U. (2015). A Novel Framework for Classification of Syncope Disease using K-Means Clustering Algorithm. *SAI Intelligent Systems Conference*, 127–132. https://doi.org/10.1109/IntelliSys.2015.7361135

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *International AAAI Conference on Weblogs and Social Media*, 216–225. https://doi.org/10.1210/en.2011-1066

Ji, W., Wang, R., & Ma, J. (2019). Dictionary-Based Active Learning for Sound Event Classification. *Multimedia Tools and Applications*, *78*(3), 3831–3842. https://doi.org/10.1007/s11042-018-6380-z

Kui, X., Lv, H., Tang, Z., Zhou, H., Yang, W., Li, J., … Xia, J. (2020). TVseer: A Visual Analytics System for Television Ratings. *Visual Informatics*, *4*(3), 1–11. https://doi.org/10.1016/j.visinf.2020.06.001

Li, S. S. (2020). Lifestyles, Technology Clustering, and the Adoption of Over-the-top Television and Internet Protocol Television in Taiwan. *International Journal of Communication*, *14*, 2017–2035.

Pribadi, M. A., Yoedtadi, M. G., & Siswoko, K. H. (2017). Perspektif Praktisi Televisi Indonesia terhadap Konvergensi Televisi dan Internet dalam Persaingan Penyajian Informasi di Internet. *Jurnal Muara Ilmu Sosial, Humaniora, Dan Seni*, *1*(1), 319. https://doi.org/10.24912/jmishumsen.v1i1.372

Ruiz, L. G. B., Pegalajar, M. C., Arcucci, R., & Molina-Solana, M. (2020). A Time-Series Clustering Methodology for Knowledge Extraction in Energy Consumption Data. *Expert Systems with Applications*, *160*, 113731. https://doi.org/10.1016/j.eswa.2020.113731

Shuyang, Z., Heittola, T., & Virtanen, T. (2017). Active Learning for Sound Event Classification by Clustering Unlabeled Data. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 751–755. https://doi.org/10.1109/ICASSP.2017.7952256

Shuyang, Z., Heittola, T., & Virtanen, T. (2018). An Active Learning Method Using Clustering and Committee-Based Sample Selection for Sound Event Classification. *16th International Workshop on Acoustic Signal Enhancement, IWAENC 2018 - Proceedings*, 116–120. https://doi.org/10.1109/IWAENC.2018.8521336

Tan, Y. (2018). An Improved KNN Text Classification Algorithm Based on K-Medoids and Rough Set. *International Conference on Intelligent Human-Machine Systems and Cybernetics*, *1*, 109–113. https://doi.org/10.1109/IHMSC.2018.00032

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2016). Preprocessing Techniques for Text Mining -An Overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.