BAREKENG: Journal of Mathematics and Its Application
September 2022 Volume 16 Issue 3 Page 939-948
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss3pp939-948

TEXT CLUSTERING ONLINE LEARNING OPINION DURING COVID-19 PANDEMIC IN INDONESIA USING TWEETS

Maulida Fajrining Tyas[1], Anang Kurnia[2*], Agus Mohamad Soleh[3]

[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University
Bogor, 16680, Indonesia

Corresponding author's e-mail: [2]anangk@apps.ipb.ac.id*

Abstract. *To prevent the spread of corona virus, restriction of social activities are implemented including school activities which reaps the pros and cons in community. Opinions about online learning are widely conveyed mainly on Twitter. Tweets obtained can be used to extract information using text clustering to group topics about online learning during pandemic in Indonesia. K-Means is often used and has good performance in text clustering area. However, the problem of high dimensionality in textual data can result in difficult computations so that a sampling method is proposed. This paper aims to examine whether a sampling method to cluster tweets can result to an efficient clustering than using the whole dataset. After pre-processing, five sample sizes are selected from 28300 tweets which are 250, 500, 2500, 10000 and 20000 to conduct K-Means clustering. Results showed that from 10 iterations, three main cluster topics appeared 90%-100% in sample size of 2500, 10000 and 20000. Meanwhile sample size of 250 and 500 tend to produced 20%-60% appearance of the three main cluster topics. This means that around 8% to 35% of tweets used can yield representative clusters and efficient computation which is four times faster than using entire dataset.*

Keywords: *Text Clustering, Sampling, K-Means, Twitter, Online Learning*

Article info:

Submitted: 20th July 2022 *Accepted: 25th August 2022*

How to cite this article:
M. F. Tyas, A. Kurnia and A. M. Soleh, "TEXT CLUSTERING ONLINE LEARNING OPINION DURING COVID-19 PANDEMIC IN INDONESIA USING TWEETS", *BAREKENG: J. Math. & App.*, vol. 16, iss. 3, pp. 939-948, September, 2022.

1. INTRODUCTION

COVID-19 pandemic has a direct impact on changes in community activities in Indonesia, including in the education sector [1], [2]. In the Circular Letter of the Minister of Education and Culture of the Republic of Indonesia Number 36962/MPK.A/HK/2020 on March 17, 2020, it is recommended that learning activities be carried out online and work from home in order to prevent the spread of the corona virus disease. In its implementation, online learning reaps the pros and cons of the community. Opinions about online learning are widely shared in several social media, including Twitter which is a social media that spreads a lot of information quickly written in a tweet.

Text mining is a research area that is included in the scope of text analytics, where the main focus is to find new and useful knowledge from a textual data source. Implementation of text mining used text from online news in predicting economic growth [3]. Text mining itself can be defined as a semi-automatic process using a computer that is used to extract patterns from a very large unstructured data set [4]. A collection of tweets about online learning can be utilized with the Text clustering method which is part of text mining that applies the Unsupervised Machine Learning algorithm to group textual data (tweets) into clusters that have the same characteristics [5], [6]. The results of grouping tweets are expected to be able to find out what topics are discussed regarding online learning in Indonesia.

However, problems often occur in the Text Clustering process where the amount of textual data available is usually very large (big data) and has high-dimensional variables which result in difficult computations so that the results are inefficient and complicated to interpret. Therefore, an exploration of how to overcome these problems from a statistical point of view is needed. An approach that usually used to improve time complexity in computation is to perform dimensionality reduction [7]. Sampling has a concept of estimating population parameters using information contained in a sample [8]. The random sampling method can be applied to solve the high dimensionality problem where in text data, reducing object means reducing the number of variables through sampling. The benefits of this reduction can result more meaningful clusters and have the same accuracy without having to use the entire data to conduct the clustering method.

The method that is commonly used and has good performance in the Text Clustering area for grouping tweets is K-Means. Tweets that have been converted into numeric form or vector space model, will form a document-term-matrix (DTM) which contains the weight of each word in each tweet. The DTM becomes the input for the K-Means algorithm which will group text by calculating the distance from the tweet vector to the centroid (center point) of the entire clusters and assigning the tweet vector into the closest cluster. Several studies that focus on analyzing Twitter data have carried out Text Clustering using the K-Means algorithm to obtain clusters or groups contained in a collection of tweets on various topics [7], [9]–[12].

Based on the previous explanation, the purpose of this study is to evaluate the sampling method in forming clusters compared to clustering using the entire text data with K-Means. This study focuses on exploring how the problem of high-dimensional data from a collection of tweets can be solved using a sampling method that can describe the topics towards the phenomenon of online learning during the COVID-19 pandemic in Indonesia.

2. RESEARCH METHODS

2.1 Data Preparation

The data used in this study is scrapped through Twitter using the *rtweet* package on Rstudio v4.0.2 software with being limited to Indonesian-language tweets. The data collection period started from July 18 to August 31, 2021 with the keywords "*pembelajaran daring*", "*pjj*", "*kuliah online*", "*kuliah daring*" and "*school from home*". Data collection is divided into 7 periods of time with details presented in Table 1.

Table 1. Number of Tweet

No	Period	Total	Cumulative
1	18 – 25 July	4551	4551
2	26 – 31 July	7505	12056
3	1 – 7 August	6459	18515
4	8 – 14 August	7650	26165
5	15 – 21 August	17946	44111
6	22 – 28 August	7188	51299
7	29 – 31 August	3341	54640

Twitter data that is taken in this research is in the form of tweets or text. Tweet pre-processing is a process to eliminate noise or clean tweet so that it can be used for further analysis without wasting important information that can be found in the data. The process is executed using the Python programming language by eliminating duplication and deleting URLs, @username, #hashtags, emoji, symbols and numbers, tokenization, case-folding, normalization, and removing stopwords [13].

Sampling is then applied on the pre-processed data by simple random sampling where if a sample of size *n* is taken from a population of size *N* such that every possible sample of size *n* has an equal chance of being selected [8]. The sample sizes used are 250, 500, 2500, 10000 and 20000 each taken 10 times. In its application to Twitter data, a study showed that simple random sampling can provide a more efficient performance in obtaining representative samples from Twitter data compared to the stratified random sampling [14].

In each sample size, the resulting text data will be transformed into numeric form by using TF.IDF (Term frequency-Inverse Document Frequency) which is the weight of the number of word frequencies that considers the presence of these words in the entire document. The number of t-word that appear in the i-th tweet (tf_{it}) and the number of tweets containing the t-word (df_t) in each sample size are calculated by the structure presented in Table 2.

Table 2. Term-frequency and Document Frequency

document	word$_1$	word$_2$	$\cdots$	word$_r$
tweet$_1$	tf_{11}	tf_{12}	$\cdots$	tf_{1r}
tweet$_2$	tf_{21}	tf_{22}	$\cdots$	tf_{2r}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
tweet$_n$	tf_{n1}	tf_{n2}	$\cdots$	tf_{nr}
df_t	df_1	df_2	…	df_r

Number of words and documents calculated are then used to calculate *t*-word in the *i*-th tweet weights (w_{it}) with TF.IDF using the following formula [15]:

$$w_{it} = tf_{it} \times idf_t \tag{1}$$

where :

w_{it} : TF.IDF weight of *t*-word in the *i*-th tweet
tf_{it} : number of words of *t*-word in the *i*-th tweet
idf_t : inverse document frequency of *t*-word, $idf_t = log(N/df_t)$
df_t : number of tweet containing the *t*-word
N : total tweet

The TF.IDF weights is then collected in the form of a matrix called document-term-matrix (DTM). Table 3 shows the data structure on DTM.

Table 3. Document-Term-Matrix

document	word$_1$	word$_2$	$\cdots$	word$_r$
tweet$_1$	w_{11}	w_{12}	$\cdots$	w_{1r}
tweet$_2$	w_{21}	w_{22}	$\cdots$	w_{2r}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
tweet$_n$	w_{n1}	w_{n2}	$\cdots$	w_{nr}

2.2 K-Means Clustering

Cluster analysis is applied to the data that had been formed into DTM with the K-Means method with the following algorithm [16]:

1. Randomly determine *k* seeds (tweet vectors) as the center of the clusters and each tweet is assigned to the nearest cluster using Euclidean distance:

$$D(\boldsymbol{d_i}, \boldsymbol{s_j}) = \sqrt{(\boldsymbol{d_i} - \boldsymbol{s_j})'(\boldsymbol{d_i} - \boldsymbol{s_j})} \tag{2}$$

2. Calculate the centroid M_i for each cluster:

$$M_i = |C_i|^{-1} \sum_{\boldsymbol{d} \in \boldsymbol{c_i}} \boldsymbol{d} \tag{3}$$

3. Update the cluster of each tweet vector that has the closest distance to the centroid and iterate until it converges.

In clustering with K-Means, optimization of the number of *k* clusters was carried out with the input value of *k* = 2 to *k* = 6 with 30 iterations. The output of this step is *k* optimal clusters that produce silhouette score with the minimum standard deviation. Silhouette score is calculated by the following formula [17]:

$$s(i) = \begin{Bmatrix} 1 - \frac{a_i}{b_i}, & if\ a_i < b_i \\ 0, & if\ \ a_i = b_i \\ \frac{b_i}{a_i} - 1, & if\ a_i > b_i \end{Bmatrix} = \frac{b_i - a_i}{max\ \{a_i, b_i\}} \tag{4}$$

where :

$s(i)$: silhouette score for *i*-th tweet of cluster C_i

a_i : average distance of *i*-th tweet with all tweet within cluster C_i

b_i : average distance of *i*-th tweet with tweet within other cluster

$$s(k) = \frac{1}{k} \sum_{r=1}^{k} s(C_r) \tag{5}$$

where :

$s(k)$: average silhouette score of all clusters

$s(C_r)$: average silhouette score of each object $s(i)$ within *k*-cluster

k : number of clusters

The results of clustering with optimal *k* clusters are visualized in the form of a wordcloud by identifying the topics visually by looking at the similarity of words in each cluster. Evaluation of the performance of text clustering on each sample size is done by comparing the number of main topics that appear in 10 iterations. An occurrence percentage of 90%-100% is considered a representative cluster.

3. RESULTS AND DISCUSSION

3.1. Tweet Exploration

Tweets can contain text, links, videos, photos and GIFs of up to 280 characters. One of the interesting elements in tweets used by Twitter users is the *#hashtag* which is used to index keywords or topics of interest so that other users can easily find them. Table 4 summarized the 10 most found #hashtags.

Table 4. Top 10 *#hashtag*

No	*Hashtag*	Total	No	*Hashtag*	Total
1	*#ShopeeMerdekaSale*	19649	6	*#ShopeeDiskonSupermarket*	2836
2	*#BeliSemuaDiShopee*	19636	7	*#BeliKebutuhandiShopee*	1711
3	*#MerdekaBersamaShopee*	19635	8	*#BelanjadiShopee88*	1709
4	*#RedmiNote105G*	6379	9	*#Belanjadishopee88*	933
5	*#RedmiBook15*	6372	10	*#BelikebutuhandiShopee*	933

For the raw tweet collection, it appears that most of the tweets contain #hashtags of promotional content from a marketplace such as *Shopee* (*#ShopeeMerdekaSale*) or products such as *Redmi* mobile phones *(#RedmiNote105G*) and *Redmi* laptops (*#RedmiBook15*). Illustration of tweet containing hashtag can be seen in Table 5.

Table 5. Tweet Containing #hashtag

Tweets
Redmi Note 10 5G memberikan kualitas tinggi dan tajam dalam video call akan memudahkan berbagai aspek kehidupan salah satunya proses belajar daring. "Dunia memberimu lebih dari yang pernah kamu berikan." Semoga dapet buat kuliah semester depan #RedmiNote105G #RedmiBook15 🎉1214

Based on the context of tweet in Table 5, the user made the tweet to participate in a promotional strategy in the form of a Giveaway which offered a product of a mobile phone which is an important item to support online learning activities. Therefore, in general, to get the giveaway, participants must follow certain requirements, such as sending as many tweets as in Table 5 which indicates that there are possibly thousands of duplicate tweets.

3.2. Pre-processing Tweet

From the exploration results, it was found that there were many indications of duplicate tweets that needs to be removed at the data cleaning stage. The data cleaning process starts by eliminating duplication, removing URLs, @usernames, #hashtags, emojis, symbols and numbers.

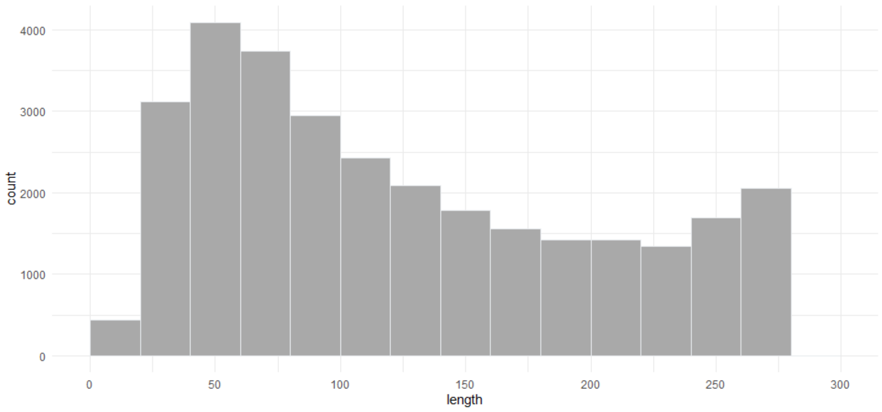


Figure 1. Distribution of Tweet’s Characters Length

In Figure 1, it can be seen that there are approximately 3000 tweets that only have less than 30 characters. This became a consideration to eliminate these tweets because most of them only contain keywords that are used in the scrapping process. From the process of eliminating duplication and tweets with less than 30 characters, the number of tweets was reduced by 48.2% from 54640 to 28300 tweets. An illustration of the tweet cleaning process is shown in Table 6.

Table 6. Ilusration of Tweet Cleaning Process

Before	After
Pandemi Covid-19 yang berkepanjangan di Indonesia ini membuat sebagian besar dosen dan mahasiswa mulai jenuh dengan kuliah daring. #Zonamahasiswa https://t.co/fgl0AidSXk	*Pandemi Covid yang berkepanjangan di Indonesia ini membuat sebagian besar dosen dan mahasiswa mulai jenuh dengan kuliah daring*

The next step is pre-processing which consists of tokenization, case-folding, non-standard word handling and stopword removal. The illustration of the pre-processing results is shown in Table 7.

Table 7. Ilustration of Pre-processing Tweet

Before	After
Pandemi Covid yang berkepanjangan di Indonesia ini membuat sebagian besar dosen dan mahasiswa mulai jenuh dengan kuliah daring	*pandemi, covid, indonesia, dosen, mahasiswa, jenuh*

To get an overview of the overall topics that are being discussed by the society about online learning phenomenon in Indonesia, a collection of tweets that have gone through the pre-processing step is formed into a wordcloud which is presented in Figure 2 with the most words being "offline".

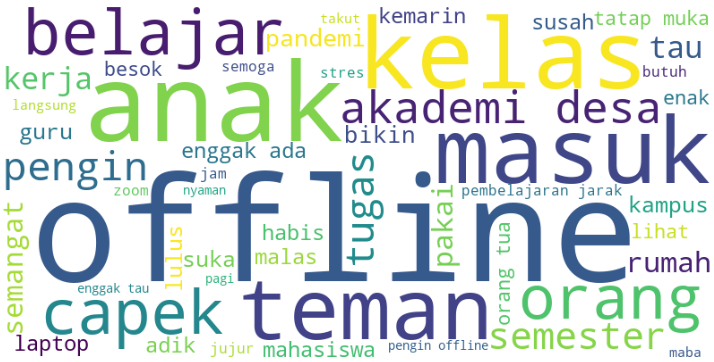


Figure 2. Wordcloud

A collection of pre-processed tweets is then converted into a vector space model that will represent each tweet as a vector containing the TF.IDF weight of each word. This collection of vectors will be a matrix called document-term-matrix. The DTM formation process was carried out on each iteration at each sample size. The number of words formed in a total of 28300 tweets is 13373 words, while the number of words (variables) of DTM in each sample size is summarized in Table 8.

Table 8. Number of Words on DTM

Sample Size	Iterations									
	1	2	3	4	5	6	7	8	9	10
250	324	325	326	340	350	321	313	347	354	312
500	632	633	637	688	660	668	631	685	670	647
2500	2487	2521	2451	2482	2549	2555	2532	2507	2494	2501
10000	6672	6847	6829	6761	6783	6784	6642	6840	6852	6846
20000	10707	10725	10738	10644	10613	10654	10642	10691	10732	10701

3.3. Clustering Evaluation

The results of clustering on text data are usually interpreted as a visualization to determine the "topic" of each clusters. Clustering was carried out on five sample sizes and was repeated 10 times to see the consistency of the resulting topics to explain the role of random sampling in producing representative clusters that were computationally and time efficient. Topic identification is done manually by looking at the similarity of words contained in the wordcloud in each iteration to determine the topic of the cluster.

Optimization in each iteration resulted a different value of *k* optimal clusters. Therefore, all wordclouds were identified to determine the total number of topics formed. There are nine topics formed in total. The distribution of topics is shown in Figure 3 to illustrate the frequency of occurrence of topics for each sample size. For all tweets, the main clusters formed have topics (3), (4), (5), (6) and (8).

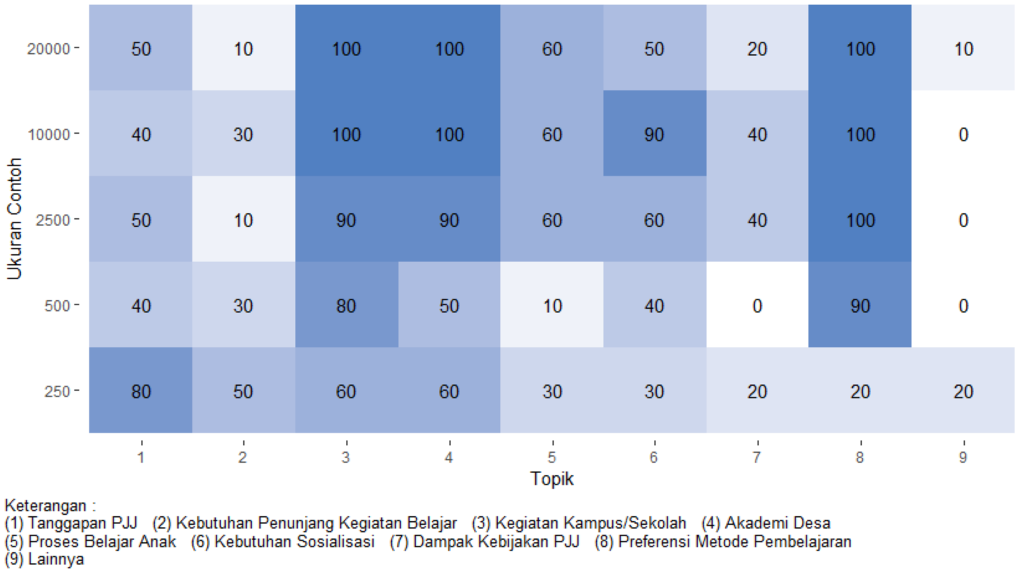


Figure 3. Heatmap of Cluster Topics Occurrence

The heatmap shows the color gradation of the points according to their values. Higher values will indicate darker colors than lower values. The percentage of certain topics appearing in 10 iterations in each sample size is interpreted as a value on the heatmap in Figure 3.

Small sample sizes, namely 250 and 500, look inconsistent, which is indicated by clusters with widely scattered topics with a low percentage of topic occurrences. In the sample size of 2500, or about 8.83% of all tweets, there are three main clusters appeared 90%-100% on each iteration, namely topics (3), (4) and (8). Consistency has started to show at the sample size of 2500 because the three topics also appear in the sample size of 10000 (35.34%), 20000 (70.67%) and clustering using all tweets. Compare to clustering with all tweets, the sample sizes of 2500, 10000 and 20000 are sufficient to cover the entire five main clusters with a 50% to 100% occurrence percentage.

Execution time is calculated to consider optimal clustering results with K-Means. Figure 4 showed the average execution time of 30 iterations of the K-Means algorithm for each sample size.

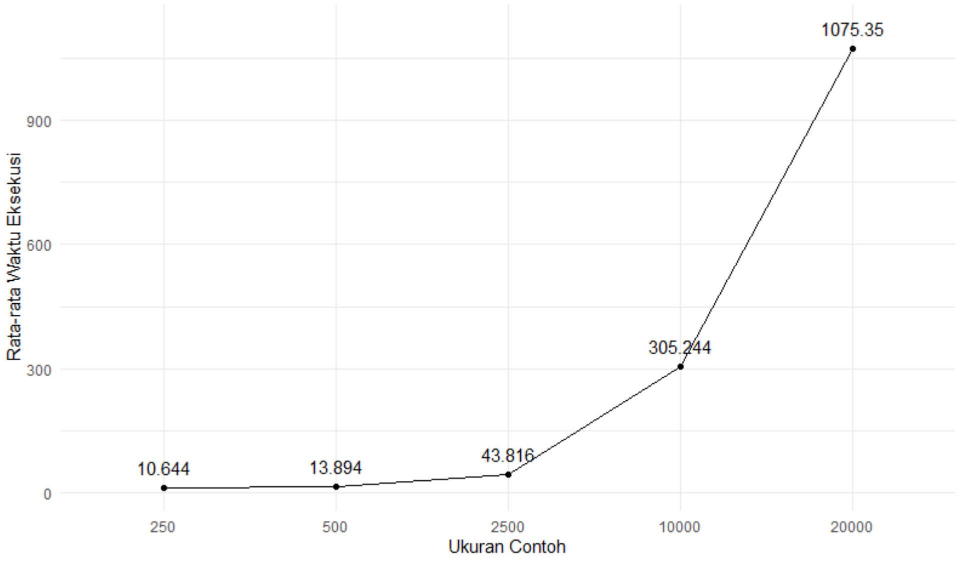


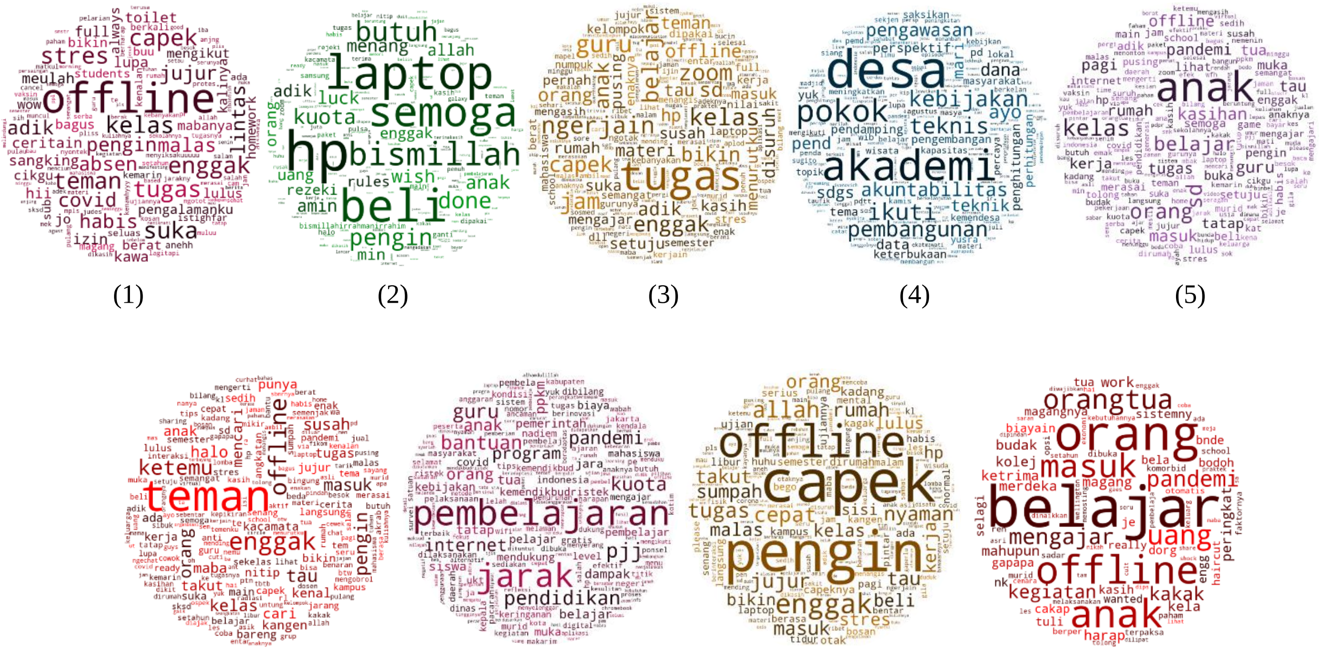
Figure 4. Average Execution Time

It can be seen that the sample sizes of 250, 500 and 2500 have not differ that much with execution time around 11 to 44 seconds compared to sample size of 10000 and 20000. The big difference is shown in the sample sizes of 10000 and 20000, where each takes about 5 and 18 minutes, while for a total of 28300 tweets it took 20 minutes.

The clustering performance at the sample size of 2500 still had a topic occurrence percentage of 90% on the two main clusters although it had a much faster performance compared to the sample size 10000 and 20000. However, the clustering at the sample size 10000 had covered the entire cluster by occurrences of 100% on the three main clusters with a much faster execution time than the sample size of 20000, which is around 5 minutes.

3.4. Cluster Visualization

From the clustering results on each sample size, wordclouds representing nine topics were taken for interpretation. The set of wordclouds is shown in Figure 5.



(1) (2) (3) (4) (5)

(6) (7) (8) (9)

Figure 1. Collection of Wordcloud of Each Topics

The topic is obtained by interpreting the collection of words contained in the cluster. Meanwhile, clusters that may form the same topic may not necessarily have the same exact words. The benchmark for determining the topic is to consider most frequent words which always appearing in clusters with the same topic. The topics formed were: (1) *Tanggapan PJJ*; (2) *Kebutuhan Penunjang Kegiatan Belajar*; (3) *Kegiatan Kampus/Sekolah*; (4) *Akademi Desa*; (5) *Proses Belajar Anak*; (6) *Kebutuhan Sosialisasi*; (7) *Dampak Kebijakan PJJ*; (8) *Preferensi Metode Pembelajaran*; (9) Others.

The visualization of the clusters formed is able to describe the distribution of topics discussed in the online learning phenomenon during the COVID-19 pandemic. In general, the tendency between the pros and cons of opinions about new learning methods that have been applied for approximately one year during the pandemic is not significantly appeared in the visualized clusters. Anxiety and difficulties experienced may often be discussed as an adaptation response of the offline method that was usually done before.

4. CONCLUSIONS

The sampling method can be a solution for textual data in reducing the dimensions of objects and variables to obtain optimal clustering results where 8% to 35% of the collection of tweets is able to cover clustering results that are quite representative and efficient in terms of execution time, which is four times faster rather than clustering using the entire data.

REFERENCES

[1] D. F. Murad, R. Hassan, Y. Heryadi, B. D. Wijanarko, dan Titan, "The Impact of the COVID-19 Pandemic in Indonesia (Face to face versus Online Learning)," 2020, doi: 10.1109/ICVEE50212.2020.9243202.

[2] A. B. Santosa, "Potret Pendidikan di Tahun Pandemi : Dampak COVID-19 Terhadap Disparitas Pendidikan di Indonesia," *CSIS Comment.*, hal. 1–5, 2020.

[3] F. Khairani, A. Kurnia, M. N. Aidi, dan S. Pramana, "Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest," *SinkrOn*, vol. 7, no. 2, hal. 532–540, Apr 2022, doi: 10.33395/sinkron.v7i2.11401.

[4] R. Sharda, D. Delen, dan E. Turban, *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, Fourth. Vivar, Malaysia: Pearson, 2017.

[5] M. W. Berry dan J. Kogan, *Text mining: Applications and Theory*. UK: John Wiley & Sons, 2010.

[6] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, dan R. A. Nisbet, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. USA: Academic Press, 2012.

[7] K. Nur'aini, I. Najahaty, L. Hidayati, H. Murfi, dan S. Nurrohmah, "Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter," in *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Okt 2015, hal. 123–128, doi: 10.1109/ICACSIS.2015.7415168.

[8] R. L. Scheaffer, W. M. III, R. L. Ott, dan K. G. Gerow, *Elementary Survey Sampling*, Seventh. USA: Cengage Learning, 2012.

[9] N. Garg dan R. Rani, "Analysis and visualization of Twitter data using k-means clustering," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Jun 2017, hal. 670–675, doi: 10.1109/ICCONS.2017.8250547.

[10] M. Sholehhudin, M. Fauzi Ali, dan S. Adinugroho, "Implementasi Metode Text Mining dan K-Means Clustering untuk Pengelompokan Dokumen Skripsi (Studi Kasus : Universitas Brawijaya)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 11, hal. 5518–5524, 2018.

[11] E. Yulian, "Text Mining dengan K-Means Clustering pada Tema LGBT dalam Arsip Tweet Masyarakat Kota Bandung," *J. Mat. "MANTIK,"* vol. 4, no. 1, hal. 53–58, Mei 2018, doi: 10.15642/mantik.2018.4.1.53-58.

[12] J. Rejito, A. Atthariq, dan A. S. Abdullah, "Application of text mining employing k-means algorithms for clustering tweets of Tokopedia," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, hal. 012019, Jan 2021, doi: 10.1088/1742-6596/1722/1/012019.

[13] A. F. Hidayatullah dan M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *J. Phys. Conf. Ser.*, vol. 801, no. 1, hal. 012072, Jan 2017, doi: 10.1088/1742-6596/801/1/012072.

[14] H. Kim, S. M. Jang, S.-H. Kim, dan A. Wan, "Evaluating Sampling Methods for Content Analysis of Twitter Data," *Soc. Media + Soc.*, vol. 4, no. 2, Apr 2018, doi: 10.1177/2056305118772836.

[15] C. D. Manning dan H. Schütze, *Foundations of Statistical Natural Language Processing*. London: The MIT Press, 1999.

[16] R. Feldman dan J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.

[17] J. Žižka, F. Dařena, dan A. Svoboda, *Text Mining with Machine Learning*. Florida: CRC Press, 2020.