khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Application of Mini-Batch K-Means Algorithm for Clustering Divorce Verdict Document Data Of Indramayu District Religious Court

Raudah Yasmin Ghozali[1], Nurissaidah Ulinnuha[2*], Wika Dianita Utami[3]

*correspondence: nuris.ulinnuha@uinsa.ac.id

[1,2,3]Mathematics Department

UIN Sunan Ampel

Surabaya

Abstract- Divorce occurs because married couples are no longer able to achieve the main goals of marriage. According to data from the West Java Central Bureau of Statistics, Indramayu Regency recorded the highest number of divorces in West Java during 2021-2023. This condition underscores the need for research to categorize the Plaintiff's or Applicant's arguments, as set out in the divorce decision issued by the Indramayu Regency Religious Court. The textual arguments of the Plaintiff or Petitioner will be pre-processed text and term weighting using Term Frequency-Inverse Document Frequency (TF-IDF). The term weighting results will be clustered using the Mini-Batch K-Means method. Mini-Batch K-Means speeds up computation by using a subset of data per iteration. In addition, the initial centroids are randomly initialized using K-Means++. The evaluation of Mini-Batch K-Means is measured based on the Silhouette coefficient, the number of iterations, and the speed of computation time. The results of this study show that Mini-Batch K-Means with random initialization is the best model, with a Silhouette coefficient of 0.5293, 4 iterations, and a running time of 0.0653 seconds. Based on the visualization results for each cluster, 2 topic groups were identified: quarrel and dispute factors, and work, children, and financial factors.

Keywords: Clustering, Divorce Verdict, Indramayu, Mini-Batch K-Means

Article info: *submitted May 28, 2025, revised March 13, 2026, accepted April 28, 2026*

1. Introduction

Marriage is a form of social being that involves a commitment to live together by forming a happy and lasting household [1]. However, in reality not all couples can maintain this commitment, resulting in divorce. Divorce is the legal termination and dissolution of the marriage bond due to the inability to fulfill the commitment to live together [2]. Divorce will be pursued if mediation efforts between the two parties have failed to maintain the integrity of the household. If the efforts made have failed, then divorce will be the best way to solve the problem.

Divorce is a form of family disintegration that has a negative impact on family members, namely the psychological disruption of children, which can affect growth and social interaction, the psychological condition of the husband and wife after divorce, the community's view of children who have divorced families, and the community's view of widows. In addition, divorce has an impact on the decline in marriage rates due to trauma from the perception of parents, relatives, and the surrounding environment, so that it puts aside the priority of getting married [3] [4]. The decline in marriage rates also indirectly affects the decline in population growth rates [5].

Several factors cause the adverse effects of divorce. Based on the Indonesian Central Bureau of Statistics, the factors of divorce are divided into several types, including adultery, drunkenness, madat (narcotics use), gambling, leaving one of the parties, being sentenced to prison, polygamy, domestic violence (KDRT), disability, constant disputes and quarrels, forced marriage, apostasy, and economy. Of all the factors, three factors dominate the occurrence of divorce, namely the factors of constant disputes and quarrels, economics, and leaving one of the parties [6]. Divorce trends in Indonesia during 2020-2023 reflect the influence of these factors. This can be seen from the significant increase in the number of divorce cases, where in 2020 there were 291,677 cases, in 2021 there were 447,743 cases, in 2022 there were 516,344 cases, and in 2023 there were 463,654 cases [7].

According to the West Java Central Bureau of Statistics, Indramayu Regency is the first district in West Java to report the highest number of divorces, with 100,000 cases in 2022-2023, making West Java the province with the most divorces in Indonesia. In three years Indramayu Regency has a divorce rate of 8,000 to 9,000 cases where in 2021 the number of divorces reached 8,026 cases then in 2022 a total of 9,152 and in 2023 reached 8,827 cases [8].

The high divorce rate in Indramayu Regency reflects complex social dynamics that require analysis to understand the underlying issues. One analysis that can be used is with a mathematical approach, namely clustering. Clustering is a data analysis technique that groups unlabeled data based on the similarity of certain patterns [9]. Clustering can be applied to both numerical and textual data. The first step in text clustering is preprocessing and word weighting to convert textual data into numerical data for analysis. Pre-processing text involves several stages: removing punctuation, case folding, stopword removal, stemming or lemmatization, and tokenization [10]. Furthermore, word weighting produces a matrix of words assigned weight values. One of the word weighting techniques is Term Frequency-Inverse Document Frequency (TF-IDF) [11].

Clustering has several methods and developments. One of them is K-Means, which has been developed and refined, and produces the K-Means++ and Mini-Batch K-Means algorithms. K-Means++ addresses the shortcomings of K-Means, namely its reliance on random centroid initialization. K-Means++ selects a centroid based on the probability distribution to produce stable and quality clusters [12]. Meanwhile, Mini-Batch K-Means is designed to handle large datasets by forming mini-batches or small samples randomly at each iteration. This approach not only speeds up the clustering process, but is also able to produce cluster results that are close to K-Means with a much shorter computation time [13].

Several previous studies have applied clustering methods. Research by Surianto used text clustering on 2019 presidential election tweet data with the K-Means algorithm and produced an optimal k of 9 clusters with a Silhouette coefficient value of 0.50189 [14]. Furthermore, text clustering on Arabic and English COVID-19 pandemic tweet data using the Mini-Batch K-Means and K-Means algorithms, and it was found that the computation duration of Mini-Batch K-Means was faster than K-Means [15]. Research by Kusmiran com paring K-Means, K-Medoids, and Mini-Batch K-Means on clustering earthquake datasets amounted to 1,039. The results stated that Mini-Batch K-Means tends to have a better Silhouette coefficient value in 4 cluster trials [16]. Further research compared K-Means and K-Means++ in clustering the zone of spread of COVID-19 cases in East Java where K-Means++ became the best algorithm by producing the highest Silhouette coefficient value of 0.882 [17]. However, so far, limited research have applied text clustering methods to divorce, especially to divorce decision documents. Divorce decision documents contain narrative information on the plaintiff's arguments for divorce, allowing a more detailed view of the pattern of divorce reasons. Thus, it can be concluded that there is a research gap that warrants further study. This study contributes by applying text clustering techniques to legal narrative documents to uncover patterns in divorce arguments. The novelty of this research lies in using divorce verdict documents as the primary data source and in integrating computational clustering results with qualitative interpretations of the underlying reasons for divorce.

Based on previous research, using mini-batches as an optimization strategy and K-Means++ as an initialization method can improve clustering results and reduce computation and convergence time. Therefore, this research will use Mini-Batch K-Means with two initializations, namely randomly and using K-Means++, and then compare with the basic algorithm. This research aims to apply text clustering to divorce verdict documents from the Indramayu Regency Religious Court, with the judge's ruling. The document will be analyzed in part, namely, the arguments of the plaintiff or applicant, to identify which words often appear in the arguments section. These results can provide a more systematic picture of the divorce problem and help identify its main factors, enabling the formulation of an appropriate policy to reduce the divorce rate in the future. In addition, this research can serve as a reference for using legal documents as research objects, as this dataset is relatively rarely used for text mining analysis.

2. Methods

This research will use text clustering with the Mini-Batch K-Means algorithm, implemented with two centroid initialization strategies: random and K-Means++ initialization. This study employs simple random sampling to select a representative sample of documents from the population. The total number of available divorce verdict documents is 5,115. Due to the manual extraction required to isolate the plaintiff's or petitioner's argument section from each document, a sampling approach was adopted to ensure data processing feasibility while maintaining representativeness. The sample size was determined using the Slovin formula with a margin of error of 0.03. Based on this calculation, 913 documents were selected as the research sample.

Figure 1 is the research flowchart.

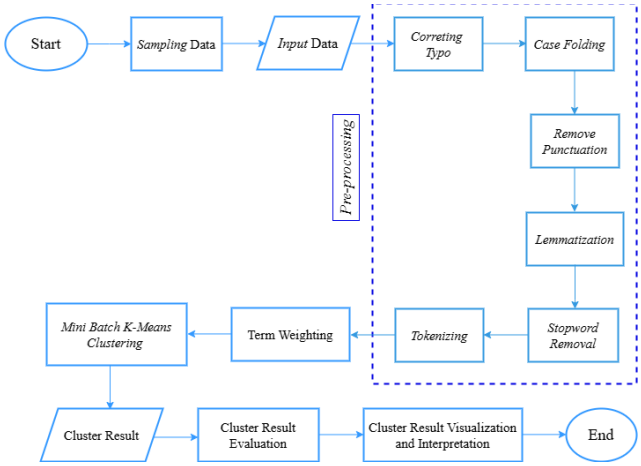

Figure 1. Flowchart

1. Sampling Data

Data sampling is a technique for selecting, processing, and analyzing a small portion of a population. There are many data sampling techniques, one of which is simple random sampling [18]. Simple random sampling is a sampling technique that selects sample members from the population at random, without regard to strata or criteria, so that all members of the population have the same opportunity to be included in the sample. Determining the number of samples using the Slovin equation, following equation (1) [19].

$$n = \frac{N}{1+N\times\varepsilon^2} \quad (1)$$

where,

n = sample size

N = population size

ε = margin of error or percentage of tolerance for inaccuracy due to errors in sampling

2. Pre-processing Text

Pre-processing aims to reconstruct the text by removing unnecessary characters, making it more structured and easier for the algorithm to process. Text pre-processing steps include correcting typos, case folding, removing punctuation, stemming or lemmatization, removing stop words, and tokenization [11][20].

a. Correcting typo: this process is to change an incorrect word due to a typing error to the correct order [21].

b. Case folding: this process converts all characters in the document that are capitalized into lowercase letters to avoid duplication of words[22].

c. Remove punctuation: this process removes meaningless characters such as numbers, punctuation marks, whitespace, and other symbols [23].

d. Lemmatization: this process normalizes vocabulary by mapping nouns to singular forms and verbs to infinitive forms [24].

e. Stopword removal: this process is used to remove words in the stopword or a common set of words that have no important meaning and relevance [25].

f. Tokenizing: this process separates all sentences in the text into words to simplify the word weighting process [26].

3. Term Weighting

Word weighting is used to assign values or weights to words contained in a document. Technically, word weighting is the process of converting textual data into numerical data. One of the methods in word weighting is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a feature extraction technique that assigns a weight to each word that appears to evaluate how relevant a word is to a document [11]. Term Frequency or TF focuses on measuring the value of the frequency of occurrence of a word in a text document. Meanwhile, Inverse Document Frequency (IDF) focuses on determining the importance of words across the entire document. Words that appear infrequently will be given a higher weight by IDF than words that appear frequently because. That is, the more the words are considered important, the more they rarely appear in the entire document. Meanwhile, the higher the word frequency (TF) value, the more occurrences of a word in the document. TF-IDF calculation using equations (2) and (3) [27].

$$W_{i,j} = TF_{i,j} \times IDF_i \quad (2)$$

$$IDF_i = \log\left(\frac{N}{df_i}\right) \quad (3)$$

where,

$W_{i,j}$ = weight value of word i in document j

$TF_{i,j}$ = the number of times word i appears in document j

IDF_i = the number of times word i appears in all documents

df_i= the number of documents containing word i

N = total number of data/documents

i = 1, 2, 3, …, p (number of variables/words)

j = 1, 2, 3, …, N (number of data/documents)

4. Mini-Batch K-Means Clustering

1) K-Means++

K-Means++ is a centroid initialization algorithm that addresses K-Means' weakness in its random selection of initial centroids. Randomly selecting data points in K-Means can lead to initialization sensitivity, affecting cluster formation. Therefore, K-Means++ improves centroid initialization by evenly distributing the initial centroids based on their probabilities. The following is the algorithm of K-Means++ [12][28].

a. Randomly select an initial centroid from all objects/data.

b. Calculate the squared distance between the previously selected nearest centroid and each data point using equation (4).

$$D(X_i, C_j)^2 = \min_{c\in C} \sum_{m=1}^{n} (x_{im} - c_{jm})^2 \quad (4)$$

di mana,

$D(X_i, C_j)^2$= squared distance between data point and centroid

x_{im} = i-th object of m-attribute

c_{jm} = j-th object of m-attribute

The minimum squared distance will be used to determine the third centroid initialization and subsequent stages.

c. Selecting the new centroid that has the maximum probability proportional to D(x)2 using equation (5).

$$P(X) = \frac{D(X)^2}{\sum_{x\in X} D(X)^2} \quad (5)$$

where,

$P(X)$= probability of selecting data point x as the new centroid

$D(X)^2$ = distance squared

$\sum_{x\in X} D(X)^2$ = sum of squared distances of all data points

d. Repeat steps 2 and 3 until k centroids are selected.

e. After determining the initial k centroids, run the K-Means clustering algorithm.

2) Mini-Batch K-Means

Mini-Batch K-Means is an algorithm that requires a mini-batch per iteration to update the centroids. The mini-batch is a random subset of the entire dataset. Using mini-batches reduces computation time, making it suitable for large datasets. Here are the steps of the Mini-Batch K-Means algorithm [13][29].

a. Randomly select k centroids as the initial cluster centroids from dataset X.

b. Randomly select a subset of data from the M \subset X data (batch size is smaller than the total dataset)

c. Calculate the Euclidean distance between each data point in the mini-batch and the existing centroid. Equation (6) is the Euclidean distance.

$$D(X_i, C_j) = \sqrt{\sum_{m=1}^{n}(x_{im} - c_{jm})^2} \quad (6)$$

where,
$D(X_i, C_j)$ = Euclidean distance
x_{im} = i-th object of m-attribute
c_{jm} = j-th object of m-attribute

d. Assign each data point to the nearest cluster based on the minimum distance with equation (7):

$$Cluster(X_i) = \min_{j \in \{1,2,\ldots,k\}} D(X_i, C_j) \quad (7)$$

where,
$Cluster(X_i)$ = cluster assigned to data point x_1
$D(X_i, C_j)$ = Euclidean distance

e. After the cluster sets a mini-batch, the cluster centroid is incrementally updated by using the mini-batch. The centroid update is done with equation (8).

$$c_j^{(t+1)} = (1 - \eta)c_j^{(t)} + \eta x_j \quad (8)$$

where,
$c_j^{(t+1)}$ = new centroid for cluster j
$c_j^{(t)}$ = old centroid for cluster j
x_j = data point in mini-batch assigned to cluster j
$\eta = \frac{1}{v[c]}$ or adaptive learning rate

f. Repeating steps 2 to 4 until convergence is achieved or until the maximum number of iterations is met so that no centroid changes significantly.

5. Cluster Result Evaluation

The silhouette coefficient is used to evaluate clustering models by assessing a model's ability to group data with similar characteristics and to place data points in close proximity to other data points. Equations (9), (10), and (11) are the silhouette coefficient equations. [30].

$$a(i) = \frac{1}{|C_i|-1}\sum_{j \in C_i, i \neq j} D(X_i, C_j) \quad (9)$$

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|}\sum_{j \in C_k, i \neq j} D(X_i, C_j) \quad (10)$$

$$S(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \quad (11)$$

where,

$S(i)$ = Silhouette coefficient value on the i-th data

$a(i)$ = average distance of the i-th object to other objects in the intra-cluster

$b(i)$ = average distance of the i-th object from other objects in the inter-cluster

C_i = number of cluster members i

C_k = number of members of cluster k

$D(X_i, C_j)$= Euclidean distance of the i-th data in cluster j

i, j = cluster member

The quality of clustering results, as measured by the silhouette coefficient, ranges from -1 to +1. When the value of S(i) is close to -1, it indicates that the object is placed in the wrong cluster because the inequality a(i) > b(i). When S(i) is close to +1, it indicates that the object is in the correct cluster because a(i) < b(i). Then, if S(i) = 0, it means that the object is not clear which cluster it should be placed in [31]. In Table 1, regarding the Silhouette coefficient criteria for evaluation [32].

Table 1. Silhouette coefficient criteria

Silhouette Coefficient Value	Criteria
0.7 < SC ≤ 1	Strong Structure
0.5 < SC ≤ 0.7	Standard Structure
0.25 < SC ≤ 0.5	Weak Structure
SC ≤ 0.25	No Structure

6. Cluster Result Visualization

Word cloud is a visualization technique for a collection of words that describe a piece of content. To measure the letter of a word in a word cloud, it is determined by the frequency of occurrence of the word. Equation (12) is used to determine the size of the word letter in the word cloud [33].

$$S_i = \begin{cases} \frac{f_{max}(t_i - t_{min})}{t_{max} - t_{min}}, & t_i > t_{min} \\ 1, & t_i = t_{min} \end{cases} \quad (12)$$

where,

S_i= font display size

f_{max}= maximum font size

t_i = number of occurrences of the i-th word

t_{max}= maximum number of words

t_{min}= minimum number of words

3. Result

1. Sampling Data

This research will use simple random sampling to take samples. Determining the number of samples from all data using the Slovin formula with a margin of error of 0.03. Using Slovin to determine

sample size is based on the limitations of manually extracting content from divorce decree documents. In addition, the method of data sampling uses simple random sampling, which can obtain a representative sample [34]. The following is the determination of the sample size.

$$n = \frac{5115}{1 + (5115 \times 0.03^2)} = 9128223432 \approx 913$$

The calculation results show that 913 documents are required from the entire data population of 5115 documents.

2. Pre-processing Text

To facilitate clustering analysis, the textual data will be preprocessed to extract the characters required for analysis. Table 2 is an example of the text of the divorce verdict document used in this research.

Table 2. Example text of divorce verdict document

Text Document of Divorce Verdict
Bahwa awalnya rumah tangga Penggugat dengan Tergugat berjalan dengan baik dan harmonis namun kurang lebih satu minggu setelah pernikahan antara Penggugat dan Tergugat mulai ada percekcokan yang diakibatkan dari masalah Diduga Tergugat memiliki wanita idaman lain, dimana Tergugat didapati sedang bersama yang diduga sebagai wanita idaman lain, namun Penggugat mencoba untuk sabar dan memberikan kesempatan kepada Tergugat serta dianggap sebagai bumbu dalam rumah tangga; 4. Bahwa dikarenakan Penggugat sudah mempunyai jadwal penerbangan yang sudah terdaftar sebagai pekerja buruh migran, maka dengan tekad yang kuat pada bulan September 2018 Penggugat keluar negeri dengan negara tujuan Taiwan, hingga sekarang; 5. Bahwa puncaknya kurang lebih satu bulan setelah Penggugat berada di luar negeri sekitar bulan Oktober 2019, antara Penggugat dengan Tergugat terjadi pertengkaran dan perselisihan yang besar yang diakibatkan dari adanya kabar dari perempuan yang yang diduga sebagai wanita idaman lain dari Tergugat yaitu melalui pesan inbox dan mengirimkan gambar yang tidak pantas antara Tergugat dengan yang diduga sebagai wanita idaman lain dari Tergugat, sehingga Penggugat Pun melakukan konfrimasi kepada Tergugat dan Tergugat Pun mengakuinya serta Tergugat seketika dengan entengnya menjatuhkan talak kepada Tergugat sehingga Penggugat sempat mengalami trauma yang mendalam dan akibatnya sesaat setelah hal tersebut terjadi sudah tidak komunikasi baik hingga ± 4,5 (empat tahun setengah) lamanya ; 6. Bahwa penggugat telah meminta nasihat kepada keluarga dan orang yang dituakan demi adanya kelangsungan Pernikahan yang sakinah, mawaddah, warahmah serta ketenangan batin namun tidak berhasil dan tiada jalan lain kecuali mengajukan gugatan ke Pengadilan Agama ini ; 7. Bahwa kehidupan Penggugat saat ini menjadi tidak menentu, sebagai seorang wanita tentunya dalam menghadapi rumah tangganya merupakan beban mental yang sangat berat serta Penggugat merasa tidak mampu dan tidak sanggup lagi meneruskan rumah tangganya dengan Tergugat.

a. Correcting Typo

The first stage is correcting typos. Table 3 is the result of the typo correcting process, where words that have been corrected are colored red.

Table 3. Text after correcting typo

Text after Correcting Typo
Bahwa awalnya rumah tangga Penggugat dengan Tergugat berjalan dengan baik dan harmonis namun kurang lebih satu minggu setelah pernikahan antara Penggugat dan Tergugat mulai ada percekcokan yang diakibatkan dari masalah Diduga Tergugat memiliki wanita idaman lain, dimana Tergugat didapati sedang bersama yang diduga sebagai wanita idaman lain, namun Penggugat mencoba untuk sabar dan memberikan kesempatan kepada Tergugat serta dianggap sebagai bumbu dalam rumah tangga; 4. Bahwa dikarenakan Penggugat sudah mempunyai jadwal penerbangan yang sudah terdaftar sebagai pekerja buruh migran, maka dengan tekad yang kuat pada bulan September 2018 Penggugat keluar negeri dengan negara tujuan Taiwan, hingga sekarang; 5. Bahwa puncaknya kurang lebih satu bulan setelah Penggugat berada di luar negeri sekitar bulan Oktober 2019, antara Penggugat dengan Tergugat terjadi pertengkaran dan perselisihan yang besar yang diakibatkan dari adanya kabar dari perempuan yang yang diduga sebagai wanita idaman lain dari Tergugat yaitu melalui pesan inbox dan mengirimkan gambar yang tidak pantas antara Tergugat dengan yang diduga sebagai wanita idaman lain dari Tergugat, sehingga Penggugat Pun melakukan konfirmasi kepada Tergugat dan Tergugat Pun mengakuinya serta Tergugat seketika dengan entengnya menjatuhkan talak kepada Tergugat sehingga Penggugat sempat mengalami trauma yang mendalam dan akibatnya sesaat setelah hal tersebut terjadi sudah tidak komunikasi baik hingga ± 4,5 (empat tahun setengah) lamanya ; 6. Bahwa penggugat telah meminta nasihat kepada keluarga dan orang yang dituakan demi adanya kelangsungan Pernikahan yang sakinah, mawaddah, warahmah serta ketenangan batin namun tidak berhasil dan tiada jalan lain kecuali mengajukan gugatan ke Pengadilan Agama ini ; 7. Bahwa kehidupan Penggugat saat ini menjadi tidak menentu, sebagai seorang wanita tentunya dalam menghadapi rumah tangganya merupakan beban mental yang sangat berat serta Penggugat merasa tidak mampu dan tidak sanggup lagi meneruskan rumah tangganya dengan Tergugat.

b. Case Folding

The second stage is to convert capital letters into lowercase letters. Table 4 below shows the results of the case-folded text.

Table 4. Text after case folding

Text after Case Folding
bahwa awalnya rumah tangga penggugat dengan tergugat berjalan dengan baik dan harmonis namun kurang lebih satu minggu setelah pernikahan antara penggugat dan tergugat mulai ada percekcokan yang diakibatkan dari masalah diduga tergugat memiliki wanita idaman lain, dimana tergugat didapati sedang bersama yang diduga sebagai wanita idaman lain, namun penggugat mencoba untuk sabar dan memberikan kesempatan kepada tergugat serta dianggap sebagai bumbu dalam rumah tangga; 4. bahwa dikarenakan penggugat sudah mempunyai jadwal penerbangan yang sudah terdaftar sebagai pekerja buruh migran, maka dengan tekad yang kuat pada bulan september 2018 penggugat keluar negeri dengan negara tujuan taiwan, hingga sekarang; 5. bahwa puncaknya kurang lebih satu bulan setelah penggugat berada di luar negeri sekitar bulan oktober 2019, antara penggugat dengan tergugat terjadi pertengkaran dan perselisihan yang besar yang diakibatkan dari adanya kabar dari perempuan yang yang diduga sebagai wanita idaman lain dari tergugat yaitu melalui pesan inbox dan mengirimkan gambar yang tidak pantas antara tergugat dengan yang diduga sebagai wanita idaman lain dari tergugat, sehingga penggugat pun melakukan konfrimasi kepada tergugat dan tergugat pun mengakuinya serta tergugat seketika dengan entengnya menjatuhkan talak kepada tergugat sehingga penggugat sempat mengalami trauma yang mendalam dan akibatnya sesaat setelah hal tersebut terjadi sudah tidak komunikasi baik hingga ± 4,5 (empat tahun setengah) lamanya ; 6. bahwa penggugat telah meminta nasihat kepada keluarga dan orang yang dituakan demi adanya kelangsungan pernikahan yang sakinah, mawaddah, warahmah serta ketenangan batin namun tidak berhasil dan tiada jalan lain kecuali mengajukan gugatan ke pengadilan agama ini ; 7. bahwa kehidupan penggugat saat ini menjadi tidak menentu, sebagai seorang wanita tentunya dalam menghadapi rumah tangganya merupakan beban mental yang sangat berat serta penggugat merasa tidak mampu dan tidak sanggup lagi meneruskan rumah tangganya dengan tergugat.

c. Remove Punctuation

Removing punctuation is done to remove unnecessary numbers, punctuation marks, double spaces, and characters/symbols. In Table 5, the result of the removal of punctuation process.

Table 5. Text after removing punctuation

Text after Remove Punctuation
bahwa awalnya rumah tangga penggugat dengan tergugat berjalan dengan baik dan harmonis namun kurang lebih satu minggu setelah pernikahan antara penggugat dan tergugat mulai ada percekcokan yang diakibatkan dari masalah diduga tergugat memiliki wanita idaman lain dimana tergugat didapati sedang bersama yang diduga sebagai wanita idaman lain namun penggugat mencoba untuk sabar dan memberikan kesempatan kepada tergugat serta dianggap sebagai bumbu dalam rumah tangga bahwa dikarenakan penggugat sudah mempunyai jadwal penerbangan yang sudah terdaftar sebagai pekerja buruh migran maka dengan tekad yang kuat pada bulan september penggugat keluar negeri dengan negara tujuan taiwan hingga sekarang bahwa puncaknya kurang lebih satu bulan setelah penggugat berada di luar negeri sekitar bulan oktober antara penggugat dengan tergugat terjadi pertengkaran dan perselisihan yang besar yang diakibatkan dari adanya kabar dari perempuan yang yang diduga sebagai wanita idaman lain dari tergugat yaitu melalui pesan inbox dan mengirimkan gambar yang tidak pantas antara tergugat dengan yang diduga sebagai wanita idaman lain dari tergugat sehingga penggugat pun melakukan konfirmasi kepada tergugat dan tergugat pun mengakuinya serta tergugat seketika dengan entengnya menjatuhkan talak kepada tergugat sehingga penggugat sempat mengalami trauma yang mendalam dan akibatnya sesaat setelah hal tersebut terjadi sudah tidak komunikasi baik hingga empat tahun setengah lamanya bahwa penggugat telah meminta nasihat kepada keluarga dan orang yang dituakan demi adanya kelangsungan pernikahan yang sakinah mawaddah warahmah serta ketenangan batin namun tidak berhasil dan tiada jalan lain kecuali mengajukan gugatan ke pengadilan agama ini bahwa kehidupan penggugat saat ini menjadi tidak menentu sebagai seorang wanita tentunya dalam menghadapi rumah tangganya merupakan beban mental yang sangat berat serta penggugat merasa tidak mampu dan tidak sanggup lagi meneruskan rumah tangganya dengan tergugat

d. Lemmatization

This stage converts a word to its base form. Table 6 is the result of lemmatization.

Table 6. Text after lemmatization

Text after Lemmatization
bahwa awal rumah tangga gugat dengan gugat jalan dengan baik dan harmonis namun kurang lebih satu minggu telah nikah antara gugat dan gugat mulai ada cekcok yang akibat dari masalah duga gugat milik wanita idam lain mana gugat dapat sedang sama yang duga bagai wanita idam lain namun gugat coba untuk sabar dan beri kesempatan kepada gugat serta anggap bagai bumbu dalam rumah tangga bahwa karena gugat sudah punya jadwal terbang yang sudah daftar bagai kerja buruh migran maka dengan tekad yang kuat pada bulan september gugat keluar negeri dengan negara tuju taiwan hingga sekarang bahwa puncak kurang lebih satu bulan telah gugat ada di luar negeri sekitar bulan oktober antara gugat dengan gugat jadi tengkar dan selisih yang besar yang akibat dari ada kabar dari perempuan yang yang duga bagai wanita idam lain dari gugat yaitu lalu pesan inbox dan kirim gambar yang tidak pantas antara gugat dengan yang duga bagai wanita idam lain dari gugat sehingga gugat pun laku konfirmasi kepada gugat dan gugat pun aku serta gugat ketika dengan enteng jatuh talak kepada gugat sehingga gugat sempat alam trauma yang dalam dan akibat saat telah hal sebut jadi sudah tidak komunikasi baik hingga empat tahun tengah lama bahwa gugat telah minta nasihat kepada keluarga dan orang yang tua demi ada langsung nikah yang sakinah mawaddah warahmah serta tenang batin namun tidak hasil dan tiada jalan lain kecuali aju gugat ke pengadilan agama ini bahwa hidup gugat saat ini jadi tidak tentu bagai orang wanita tentu dalam hadap
rumah tangga rupa beban mental yang sangat berat serta gugat rasa tidak mampu dan tidak sanggup lagi terus rumah tangga dengan gugat

e. Stopword Removal

Stopword removal removes unnecessary words, such as conjunctions, proper names, place names, and month names. Table 7 is the result of the stopword removal process.

Table 7. Text after stopword removal

Text after Stopword Removal
cekcok wanita idam wanita idam sabar anggap kerja tekad kuat tengkar selisih kabar perempuan wanita idam inbox kirim gambar wanita idam konfirmasi enteng jatuh trauma komunikasi nasihat orang langsung tenang orang wanita rupa beban mental berat sanggup

f. Tokenizing

This tokenization will split sentences that have undergone the previous extraction stage into words. With this partitioning, it will be easier to weigh the words. Table 8 is the result of tokenizing.

Table 8. Text after tokenizing

Text after Tokenizing
[cekcok, wanita, idam, wanita, idam, sabar, anggap, kerja, tekad, kuat, tengkar, selisih, kabar, perempuan, wanita, idam, inbox, kirim, gambar, wanita, idam, konfirmasi, enteng, jatuh, trauma, komunikasi, nasihat, orang, langsung, tenang, orang, wanita, rupa, beban, mental, berat, sanggup]

3. Term Weighting

After tokenization, 1,529-word features were obtained from 913 documents. These word features will be indexed and assigned weights by converting text data to numeric values to facilitate clustering analysis.

The method for assigning term weights is TF-IDF. TF-IDF is a combined term weighting method between Term Frequency and Inverse Document Frequency. The following is a table of word weighting with TF. Table 9 is the overall result of the TF process.

Table 9. Term weighting result (TF)

Doc.	Feature								
	1	2	3	…	795	….	1504	…	1529
1	0	0	0	….	0	…	3	…	0
2	0	0	0	….	1	…	0	…	0
3	0	0	0	…	0	…	0	…	0
4	0	0	0	…	0	…	0	…	0
5	0	0	0	…	1	…	0	…	0
:	:	:	:	…	:	…	:	…	:
912	0	0	0	…	0	…	0	…	0
913	0	0	0	…	1	…	1	…	0

The table shows that the word “malas” (lazy) in the 795th feature and “wanita” (woman) in the 1504th feature each have a weight of 1 in the 913th document. This means that both words occur once in the 913th document, while a value of 0 indicates that

the word does not occur in the document. Next, is to calculate TF-IDF. TF-IDF is determined by multiplying TF and IDF, where the IDF is the number of documents that contain a particular word. For example, the word "*wanita*" (woman) in the 1504th feature appears in 99 documents. Table 10 is the calculation result of TF-IDF.

Table 10. Term weighting result (TF-IDF)

Doc	Feature								
	1	2	3	…	795	…	1504	…	1529
1	0	0	0	…	0	…	2,8945	…	0
2	0	0	0	…	0.8366	…	0	…	0
3	0	0	0	…	0	…	0	…	0
4	0	0	0	…	0	…	0	…	0
5	0	0	0	…	0.8366	…	0	…	0
:	:	:	:	…	:	…	:	…	:
417	0	0	0	…	0.8366	…	0	…	0
:	:	:	:	…	:	…	:	…	:
912	0	0	0	…	0	…	0	…	0
913	0	0	0	…	0.8366	…	0.9648	…	0

4. Cluster Result Evaluation

Evaluation of cluster results will be based on the Silhouette coefficient (SC), number of iterations, and running time (RT). The Silhouette coefficient (SC) value will determine which model has good cluster structure quality, while the number of iterations and running time (RT) will measure which model has faster computation time and convergence speed. The algorithms to be evaluated are Mini-Batch K-Means with random initialization and K-Means++ initialization. First, the optimal batch size (b) and number of clusters (k) are identified, and then compared with the basic K-Means algorithm.

Table 11. Comparison of Mini-Batch K-Means evaluation results

k	b	Random Initialization			K-Means++ Initialization		
		SC	Iteration	RT (s)	SC	Iteration	RT (s)
2	50	0.2145	2	0.0409	-0.1226	1	0.0332
	100	**0.5293**	**4**	**0.0653**	**0.3410**	**4**	**0.0690**
	200	-0.0602	4	0.0707	0.0905	4	0.0782
	500	-0.1155	7	0.1068	-0.1610	8	0.1145
3	50	0.103	1	0.0316	-0.0491	1	0.0368
	100	0.0477	3	0.0581	0.0321	4	0.0713
	200	-0.1404	3	0.0538	-0.0231	4	0.0766
	500	0.2769	7	0.1103	0.0015	7	0.1192
4	50	-0.1342	4	0.0722	-0.0444	1	0.0359
	100	-0.1548	2	0.0436	-0.1542	4	0.0694
	200	-0.2015	4	0.0764	-0.0208	4	0.0730
	500	-0.1872	13	0.1573	-0.0439	7	0.1128

Based on Table 11, the best configuration is obtained when (k=2) and (b=100). In this configuration, Mini-Batch K-Means with random initialization achieves the highest Silhouette coefficient (0.5293). Meanwhile, Mini-Batch K-Means with K-Means++ initialization produces a lower Silhouette coefficient of 0.3410. In terms of convergence and computation time, both models show very similar performance, with 4 iterations and running times of 0.0653 and 0.0690 seconds, respectively.

Based on the evaluation of the number of iterations and RT, the two models show little difference in computation time, speed, and convergence. However, across all trials, Mini-Batch K-Means with K-Means++ initialization reaches a maximum of 8 iterations with a running time of 0.1192 seconds. Meanwhile, Mini-Batch K-Means with random initialization reaches a maximum of 13 iterations with a running time of 0.1573 seconds. The differences in the number of iterations and the running time of the two models are shown in Figure 2.

Based on Figure 2, random initialization has a higher average number of iterations, while K-Means++ initialization has a longer average RT. According to the theory, K-Means++ reduces the number of iterations by mitigating sensitivity to initial centroid initialization. On the other hand, the computation time with K-Means++ initialization is longer than with random initialization because it first runs several processes to determine the initial centroids, whereas random initialization picks them at once. It can be concluded that using K-Means++ as initialization can shorten convergence time but not necessarily reduce computation time, and it can also produce a lower Silhouette coefficient.

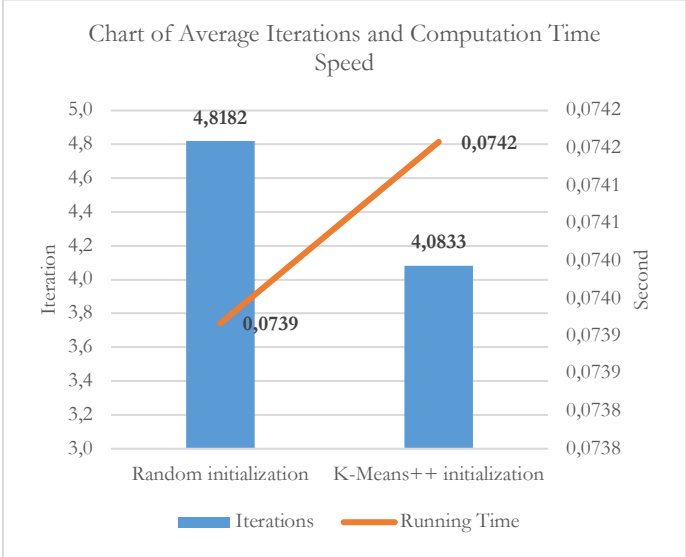

Figure 2. Chart of average iterations and computation time speed Mini-Batch K-Means

Table 12. Comparison of K-Means evaluation results

k	Random Initialization			K-Means++ Initialization		
	SC	Iteration	RT (s)	SC	Iteration	RT (s)
2	**0.5096**	**9**	**0.8447**	**0.4211**	**3**	**0.1003**

3	0.3026	22	0.5638	0.1086	2	0.0540
4	0.3249	20	0.3946	0.1159	2	0.0524

Furthermore, the results are compared with the basic K-Means algorithm using two initialization methods: random and K-Means++. Based on Table 12, the best configuration is obtained when k = 2, as indicated by the highest Silhouette coefficient (SC). In this configuration, random initialization yields a higher SC value (0.5096) than K-Means++ (0.4211), indicating better clustering quality.

In terms of computational performance, K-Means++ requires fewer iterations and shorter running time than random initialization. This occurs because K-Means++ selects initial centroids based on distance probabilities, which helps distribute the centroids more evenly across the dataset and reduces centroid movement during iteration. However, contrary to several previous studies, this research shows that K-Means++ initialization does not improve the Silhouette coefficient relative to random initialization. Although K-Means++ is designed to produce better-distributed initial centroids, its effectiveness can depend on the dataset's characteristics. In this case, random initialization produces higher clustering quality, while K-Means++ mainly improves computational efficiency.

Table 13. Comparison of Mini-Batch K-Means and K-Means (k=2)

Model		SC	Iteration	RT (s)
Mini-Batch K-Means	**Random Initialization**	**0.5293**	**4**	**0.0653**
	K-Means++ Initialization	0.3410	4	0.0690
K-Means	Random Initialization	0.5096	9	0.8447
	K-Means++ Initialization	0.4211	3	0.1003

Furthermore, comparing 4 models, namely Mini-Batch K-Means with random initialization, Mini-Batch K-Means with initialization using K-Means++, K-Means, and K-Means++. Based on the previous evaluation results, the best configuration among the four models is obtained when $k = 2$. In Table 13, Mini-Batch K-Means with random initialization and K-Means achieve the best SC value of 0.5. This means that these two models can group data well with a standard structure. Of these two models, Mini-Batch K-Means with random initialization is the best, and it can be proven that using mini-batches in each iteration can speed up computation time while producing a good Silhouette coefficient. In addition, the model can converge faster than other models.

5. Cluster Result Visualization and Interpretation

Based on the previous comparison results, Mini-Batch K-Means with random initialization is the best model and will be visualized in a word cloud. The following is the visualization form for both clusters.

Figure 3. Word cloud visualization for 1st cluster

The word cloud in Figure 3 contains the words "*tengkar*" (quarrel) and "*selisih*" (dispute) as the words with the highest frequency of occurrence. These two words illustrate that the dominant topics in the documents in cluster 1 are the reasons for divorce stemming from quarrels and disputes. These terms indicate that many divorce arguments in this cluster emphasize recurring interpersonal conflicts between spouses.

On the other hand, based on TF-IDF weights, 20 words were obtained that were most representative of cluster 1 with the highest weights, namely *kerja* (0.3138), *orang* (0.2802), *diam* (0.2579), *rumah* (0.2188), nafkah (0.2124), *anak* (0.1950), *retak* (0.1900), *pulang* (0.1883), *ekonomi* (0.1859), *pergi* (0.1796), *sulit* (0.1730), *bentuk* (0.1678), *wanita* (0.1668), *sikap* (0.1638), *nasihat* (0.1607), *hati* (0.1590), *pasang* (0.1562), *sanggup* (0.1553), *sabar* (0.1530), and *tahan* (0.1512). These findings indicate that the main issues in this cluster are not only quarrels and disputes but also work problems, livelihood issues, economic conditions, and dynamics of attitudes and communication within the household. Thus, the TF-IDF weighting results complement the findings in the word cloud, indicating that this cluster represents divorce triggered by gradually developing relationship conflicts that undermine the relationship's harmony. The following is an excerpt from one of the divorce verdict documents in cluster 1.

"*Bahwa kurang lebih sejak bulan Juni tahun 2023 rumah tangga Penggugat dengan Tergugat mulai retak, sering terjadi perselisihan dan pertengkaran yang penyebabnya penyebabnya karena rumah tangga penggugat dan Tergugat tidak harmonis dikarenakan suami (Tergugat) kedapatan atau ketahuan oleh Penggugat telah berselingkuh dengan wanita lain sehingga terjadi pertengkaran yang hebat.*"

Figure 4. Word cloud visualization for 2nd cluster

The word cloud in Figure 5 contains the words "kerja" (work), "anak" (child), and "uang" (money) as the words with the highest frequency of occurrence. These three words illustrate that the

dominant topics in the documents in cluster 2 are reasons for divorce related to work, children's interests, and household finances. The co-occurrence of these terms suggests that economic pressures and financial responsibilities toward children are important themes within this cluster of divorce arguments.

On the other hand, based on TF-IDF weight, 20 words were obtained that were most representative of cluster 2 with the highest weight, namely *uang* (2.9038), *kirim* (2.4954), *anak* (2.4887), *kerja* (1.1726), *biaya* (1.1589), *rupiah* (1.1340), *tabung* (1.1308), *beli* (1.0262), *engah* (0.9671), *sikap* (0.8843), *percaya* (0.8551), *marah* (0.8543), *pulang* (0.8193), *komunikasi* (0.8084), *kena* (0.7994), *utang* (0.7531), *lantai* (0.7253), *majikan* (0.7253), *picu* (0.7185), and *sakit* (0.7173). These findings indicate that the main issues in this cluster are household economic problems, particularly income, expenditure, remittances, and financial responsibility for children. Thus, the TF-IDF weighting results reinforce the findings from the word cloud. This cluster represents divorce triggered by economic pressures and household financial burdens, which contribute to relationship conflict. The following is an excerpt from one of the divorce verdict documents in cluster 2.

"*Bahwa seiring berjalannya waktu Penggugat mencoba bersabar, akan tetapi karena semakin terdesak ekonomi maka Penggugat memutuskan untuk bekerja di Luar Negeri/Singapura (pada bulan Desember 2020 sampai Desember 2022); Bahwa pada saat Penggugat di Singapura tersebut, komunikasi antara Penggugat dan Tergugat masih berjalan meskipun jarang, namun Penggugat sangat kecewa kepada Tergugat karena ternyata Tergugat tidak memberi nafkah kepada anaknya, apalagi kepada Penggugat sama sekali. Anak tinggal bersama orang tua Penggugat dan oleh Tergugat hanya dikasih uang jajan saja itupun sangat jarang;*"

4. Discussion

Based on the results, this study shows some findings that are not fully consistent with those of previous studies. In the context of centroid initialization, both in Mini-Batch K-Means and standard K-Means, using K-Means++ does not improve the SC value compared to random initialization. This finding contradicts previous studies [17] and [35], which report that K-Means++ can improve clustering quality by selecting initial centroids based on probability distributions. However, K-Means++ initialization is more efficient than random initialization in terms of convergence: it requires fewer iterations, resulting in shorter computational time. In addition, this study's results show that using mini-batches and K-Means++ can shorten both computation and convergence times compared to other models. Thus, the difference in evaluation results indicates a trade-off between cluster quality and process efficiency. In addition, the quality of clustering results depends on the characteristics of the data used. As in research [14][36][37], which reports Silhouette coefficients greater than 0.5 for text clustering with the K-Means and K-Means++ methods. However, based on this study, Mini-Batch K-Means with random initialization achieves better accuracy, faster convergence, and lower computation time than K-Means and K-Means++. Based on the results, Mini-Batch K-Means with random initialization shows better performance than the previous methods, namely K-Means and K-Means++, especially in terms of accuracy, convergence, and computational efficiency. Therefore, this method is worth considering as an alternative for document clustering.

In addition to the quantitative evaluation using the Silhouette coefficient, the clustering results also reveal meaningful qualitative patterns in the divorce arguments contained in the documents. Using frequency of occurrence and TF-IDF weights emphasized arguments that frequently appeared in specific documents, enabling each cluster to represent distinct narrative tendencies in divorce reasons. For instance, the dominance of terms such as *tengkar* and *selisih* in the first cluster reflects the centrality of interpersonal conflict in the argument. In contrast, the prominence of *kerja*, *anak*, and *uang* in the second cluster indicates socio-economic pressures related to employment, financial responsibility, and childcare. This shows that the clustering results are not merely computational groupings but are substantively aligned with real-world divorce dynamics.

5. Conclusion

Based on the Silhouette coefficient (SC), the best configuration is obtained when k = 2 and b = 100. In this configuration, Mini-Batch K-Means with random initialization achieves an SC value of 0.5293, while Mini-Batch K-Means with K-Means++ initialization produces a lower SC value of 0.3410. In terms of computational performance, both models show similar results. Mini-Batch K-Means with random initialization requires four iterations with a running time of 0.0653 seconds, whereas the K-Means++ initialization also requires four iterations with a running time of 0.0690 seconds. These results indicate that Mini-Batch K-Means with random initialization provides the best clustering performance in this study. Although K-Means++ is designed to improve centroid initialization, in this dataset, it does not significantly outperform random initialization in terms of clustering quality or computational efficiency.

Furthermore, Mini-Batch K-Means with the two centroid initialization strategies is compared with the standard K-Means algorithm. Using the same optimal configuration (k = 2), Mini-Batch K-Means with random initialization achieves the highest SC value of 0.5293 in only 4 iterations and takes 0.0653 seconds. Therefore, this model can be considered the optimal clustering approach in this study because it achieves better clustering quality while maintaining efficient computational performance. The results also indicate a trade-off between clustering quality and computational efficiency: random initialization tends to produce better clustering quality, while K-Means++ initialization may achieve slightly faster convergence in some cases.

The word cloud visualization shows that cluster 1 is dominated by the words "tengkar" (quarrel) and "selisih" (dispute), indicating that interpersonal conflicts are the dominant theme in divorce arguments in this cluster. Meanwhile, cluster 2 is characterized by the frequent occurrence of the words "kerja" (work), "anak" (children), and "uang" (money), reflecting themes of employment, financial responsibility, and childcare within the household. These findings suggest that divorce arguments in the documents tend to revolve around two main patterns: interpersonal conflict between spouses and socio-economic pressures within the household.

Reference

[1] J. Arifin dan I. Arifin, "Sociocultural Relations of Early Marriage

in Sidrap Regency," *AKSIOLOGI J. Pendidik. dan Ilmu Sos.*, vol. 5, no. 2, hal. 403–413, 2025.

[2] B. Ö. Cabilar dan A. E. Yilmaz, "Divorce and Post-divorce Adjustment : Definitions , Models and Assessment of Adjustment," *Psikiyatr. Guncel Yaklasimlar*, vol. 14, no. 1, hal. 1–11, 2022, doi: 10.18863/pgy.910766.

[3] S. Behtoei, F. Golshani, A. Baghdasarians, dan S. Emamipour, "Presenting a Model of Compromise and Withdrawal of Divorce in Couples Applying for Consensual Divorce : A Grounded Theory Study," *Iran. Evol. Educ. Psychol. J.*, vol. 6, no. 1, hal. 329–355, 2024.

[4] A. S. A. Alghawli, "Application of the Fuzzy Delphi Method to Identify and Prioritize the Social-Health Family Disintegration Indicators in Yemen," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, hal. 680–691, 2022.

[5] I. S. Ningtias, "Faktor yang Mempengaruhi Penurunan Angka Pernikahan di Indonesia," *J. Regist.*, vol. 4, no. 2, hal. 87–98, 2022.

[6] Badan Pusat Statistik Indonesia, "Jumlah Perceraian Menurut Provinsi dan Faktor, 2023." 2024. [Daring]. Tersedia pada: https://www.bps.go.id/id/statistics-table/3/YVdoU1IwVmlTM2h4YzFoV1psWkViRXhqTlZwRFVUMDkjMw==/jumlah-perceraian-menurut-provinsi-dan-faktor.html?year=2023

[7] Badan Pusat Statistik Indonesia, "Nikah dan Cerai Menurut Provinsi, 2023." 2024. [Daring]. Tersedia pada: https://www.bps.go.id/id/statistics-table/3/VkhwVUszTXJPVmQ2ZFRKamNIZG9RMVo2VEdsbVVUMDkjMw==/nikah-dan-cerai-menurut-provinsi.html?year=2023

[8] Badan Pusat Statistik Jawa Barat, "Jumlah Nikah dan Cerai, 2022-2023." 2024. [Daring]. Tersedia pada: https://jabar.bps.go.id/id/statistics-table/2/MzMyIzI=/jumlah-nikah-dan-cerai.html

[9] A. Chaerudin, D. T. Murdiansyah, dan M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indones. J. Comput.*, vol. 6, no. 1, hal. 53–60, 2021.

[10] W. A. Prabowo dan F. Azizah, "Sentiment Analysis for Detecting Cyberbullying Using TF-IDF and SVM," *RESTI J. (System Eng. Inf. Technol.*, vol. 4, no. 6, hal. 11–12, 2020.

[11] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *J. Phys. Conf. Ser.*, vol. 2171, no. 1, hal. 012021, 2022, doi: 10.1088/1742-6596/2171/1/012021.

[12] F. S. Mukti, A. Junikhah, P. M. A. Putra, A. Soetedjo, dan A. U. Krismanto, "A Clustering Optimization for Energy Consumption Problems in Wireless Sensor Networks using Modified K-Means ++ Algorithm," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 3, hal. 355–365, 2022, doi: 10.22266/ijies2022.0630.30.

[13] S. Igescu, V. Sanca, E. Zapridou, dan A. Ailamaki, "Improving K-Means Clustering Using Speculation," 2023.

[14] D. F. Surianto, "Clustering Tweets Data on Twitter Social Media Using K-Means Method," *J. Secur. Comput. Information, Embed. Network, Intell. Syst.*, hal. 44–51, 2023.

[15] M. A. Alanezi dan N. M. Hewahi, "Tweets Sentiment Analysis during COVID-19 Pandemic," in *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 2020, hal. 1–6.

[16] A. Kusmiran, "Clustering and Risk Analysis of The Earthquake in Sulawesi Using Mini-Batch K-Means, K-Medoids, and Maximum Likelihood Method," *Elkawnie J. Islam. Sci. Technol.*, vol. 9, no. 1, hal. 1–23, 2023.

[17] N. Nugroho dan F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means++ sebagai Clustering Data Covid-19 di Pulau Jawa," *Teknika*, vol. 11, no. 3, hal. 170–179, 2022.

[18] R. Iliyasu dan I. Etikan, "Comparison of Quota Sampling and Stratified Random Sampling," *Biometrics Biostat. Int. J.*, vol. 10, no. 1, hal. 24–27, 2021, doi: 10.15406/bbij.2021.10.00326.

[19] R. M. Ramdhan dan A. A. Rachman, "The Effect of The Awareness of Taxpayer and Tax Socialization on Taxpayer Compliance for Motor Vehicles," *Int. J. Financ. Accounting, Manag.*, vol. 5, no. 2, hal. 133–148, 2023.

[20] N. Ulinnuha dan J. G. Indriyani, "Ekstraksi Topik Pantun di Twitter Menggunakan K-Means Clustering," *KUBIK J. Publ. Ilm. Mat.*, vol. 8, no. 1, hal. 24–34, Mei 2023, doi: 10.15575/kubik.v8i1.29191.

[21] G. Eser dan C. Sahin, "Sentiment Analysis and Rating Prediction for App Reviews Using Transformer-based Models," 2024.

[22] M. Qorib, T. Oladunni, M. Denis, E. Ososanya, dan P. Cotae, "COVID-19 Vaccine Hesitancy: Text Mining, Sentiment Analysis and Machine Learning on COVID-19 Vaccination Twitter Dataset," *Expert Syst. Appl.*, vol. 212, hal. 118715, 2023.

[23] N. Y. Pradipta dan H. Soetanto, "Sentiment Classification of General Election 2024 News Titles on Detik. com Online Media Website Using Multinominal Naive Bayes Method," *J. Appl. Sci. Eng. Technol. Educ.*, vol. 6, no. 1, hal. 43–55, 2024.

[24] C. Dewi, F. A. Indriawan, dan H. J. Christanto, "Spam Classification Problems Using Support Vector Machine and Grid Search," *Int. J. Appl. Sci. Eng.*, vol. 20, no. 4, hal. 1–10, 2023.

[25] P. H. Prastyo, A. S. Sumi, A. W. Dian, dan A. E. Permanasari, "Tweets Responding to The Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *J. Inf. Syst. Eng. Bus. Intell*, vol. 6, no. 2, hal. 112, 2020.

[26] M. F. Hilman, R. Passarella, D. Kurniawan, dan S. Sutarno, "Automatic Text Summarization on Aviation Traffic Accident Report Synopsis Using Term Frequency-Inverse Document Frequency (TF-IDF) Algorithm," *Available SSRN 4758981*, 2024.

[27] M. Kamyab, G. Liu, dan M. Adjeisah, "Attention-based CNN

and Bi-LSTM Model Based on TF-IDF and Glove Word Embedding for Sentiment Analysis," *Appl. Sci.*, vol. 11, no. 23, hal. 11255, 2021.

[28] D. Arthur dan S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," 2006.

[29] D. Sculley, "Web-Scale K-Means Clustering," in *Proceedings of the 19th international conference on World wide web*, 2010, hal. 1177–1178.

[30] S. S. Momahhed, S. Emamgholipour Sefiddashti, B. Minaei, dan Z. Shahali, "K-Means Clustering of Outpatient Prescription Claims for Health Insureds in Iran," *BMC Public Health*, vol. 23, no. 1, hal. 788, 2023.

[31] K. R. Shahapure dan C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, 2020, hal. 747–748.

[32] I. M. K. Karo, S. Dewi, M. Mardiana, F. Ramadhani, dan P. Harliana, "K-Means and K-Medoids Algorithm Comparison for Clustering Forest Fire Location in Indonesia," *J. Ecotipe (Electronic, Control. Telecommun. Information, Power Eng.*, vol. 10, no. 1, hal. 86–94, 2023.

[33] O. Olabanjo *et al.*, "From Twitter to Aso-Rock: A Sentiment Analysis Framework for Understanding Nigeria 2023 Presidential Election," *Heliyon*, vol. 9, no. 5, 2023.

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

[35] R. F. Salsabila, D. D. Prasetya, T. Widiyaningtyas, dan T. Hirashima, "Comparison of Text Representation for Clustering Student Concept Maps," *Matrik J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 24, no. 2, hal. 259–272, 2025, doi: 10.30812/matrik.v24i2.4598.

[36] M. F. Fiqri, R. Muhammad, dan M. I. Ardimansyah, "Cluster Analysis of Emotions in Quranic Translations Using K-Means Clustering," *J. Softw. Eng. Inf. Commun. Technol. J.*, vol. 5, no. 2, hal. 123–134, 2024.

[37] D. S. Maylawati, R. F. Abdullah, Y. A. Gerhana, A. Wahana, I. Budiman, dan W. Uriawan, "Changes Analysis in Public Opinion Regarding Binary Option Trends using K-Means++," in *2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED)*, 2023, hal. 1–6.