

Mapping Public Conversation on Indonesia's Free Nutritious Meal Program on Twitter (X) Using TF-IDF and K-Means Clustering

Cahyasari Kartika Murni¹, Faisal Basbeth², Silviana Widya Lestari³

Department of Information Technology, Faculty of Engineering and Informatics, Institut Teknologi dan Bisnis Widya Gama Lumajang, Indonesia^{1,2,3}

Corresponding Author: Cahyasari Kartika Murni (cahyasarikartikamurni@gmail.com)

ARTICLE INFO

Date of entry:
17 February 2026
Revision Date:
13 August 2026
Date Received:
01 October 2026

ABSTRACT

Public programs now generate their own online commentary, and that commentary is worth reading systematically rather than anecdotally. This paper maps the topic structure of Indonesian-language conversation about the Free Nutritious Meal program (Makan Bergizi Gratis, MBG) on Twitter (X). We scraped 3,459 tweets posted in January and February 2025, and cleaned them through case folding, filtering, tokenization, stopword removal, and stemming, which left 2,926 documents, or 84.59% of the raw set. These were weighted with Term Frequency-Inverse Document Frequency (TF-IDF) into a 2,926 x 5,000 document-term matrix and partitioned with K-Means. The Elbow Method placed the optimal solution at $K = 3$, with an SSE of 2,760.3617 and a Silhouette Score of 0.0092. Three topics emerged: institutional endorsement tied to food security and self-sufficiency (94 tweets, 3.21%), everyday talk about how the program is being delivered and paid for (2,012 tweets, 68.76%), and the benefits of the program for children and their health (820 tweets, 28.02%). The Silhouette Score sits close to zero. We read this as a property of the corpus rather than a failure of the procedure: every document discusses the same program, so the clusters share a large common vocabulary and their boundaries are soft. The partition should therefore be treated as a first thematic map rather than a clean separation, and contextual representations such as Word2Vec, FastText, or IndoBERT are the obvious next step.

Keywords: Text Mining; TF-IDF; K-Means; Clustering; Free Nutritious Meal



Cite this as: Murni, C. K., Basbeth, F., & Lestari, S. W. (2026). Mapping Public Conversation on Indonesia's Free Nutritious Meal Program on Twitter (X) Using TF-IDF and K-Means Clustering. *Journal of Informatics Development*, 5(1), 18–27. <https://doi.org/10.30741/jid.v5i1.2081>

INTRODUCTION

Social media is where Indonesians first react to government policy. Opinions, complaints, and jokes appear on Twitter (X) within hours of an announcement, and they accumulate into a record of public reaction that no survey can match for speed or for volume [1]. The properties that make this record valuable also make it awkward to use. Tweets are short, unstructured, full of abbreviations, and far too numerous to read by hand, so any serious attempt to understand what people are saying has to

be computational. The Free Nutritious Meal program, known in Indonesia as Makan Bergizi Gratis and abbreviated as MBG throughout this paper, is a useful case study. It touches child nutrition, schooling, household budgets, and a very large line in the state budget, and it has been argued about loudly online since it was announced. Academic work on MBG so far has been normative, reading the program through the lens of social justice and political dynamics [6]. That work is valuable, but it does not tell us what the public conversation about MBG actually consists of, nor how public attention is distributed across its themes.

Text mining offers a way to answer that question. The familiar sequence of collection, preprocessing, feature construction, pattern analysis, and interpretation turns unstructured text into something an algorithm can work with [2]. For Indonesian tweets, the preprocessing stage carries unusual weight because the raw text mixes URLs, mentions, hashtags, numbers, and a great deal of non-standard spelling [7]. Once the text is clean, TF-IDF remains the workhorse for turning documents into vectors, rewarding terms that are frequent in one document and rare across the corpus [3]; it has also performed well on Indonesian material [8]. K-Means is the usual partner for TF-IDF when no labels are available, largely because it is cheap and scales [4]. The number of clusters is typically determined using the Elbow Method [5], and the result is assessed using the Silhouette Score [9]. Published work using these tools tends to fall into one of two groups. The first is supervised: hybrid deep learning for Indonesian sentiment analysis [1], or convolutional and bidirectional LSTM architectures for text classification [2]. Such models are accurate, but they need labeled data, and hand-labeling several thousand tweets is exactly the cost a policy analyst wants to avoid. The second group is methodological and concerned with the tools themselves: refinements to TF-IDF [3], surveys of K-Means [4], elbow-based selection of K [5], and a soft variant of the Silhouette Score for deep clustering [9]. Studies in this group usually stop at a metric. They rarely carry a clustering through to a substantive reading of a particular policy debate. Where the pipeline is applied, the domain is typically e-commerce, marketing, or general opinion rather than a government program.

What is missing is an unsupervised map of the MBG conversation itself, built from recent data and honest about how well the clusters actually separate. That last point matters more than it may seem. Applications of K-Means to social media data often present cluster interpretations without reporting the metrics that would allow a reader to judge them, leaving the interpretations unverifiable. This paper does three things differently. It applies the TF-IDF and K-Means pipeline to an Indonesian-language MBG corpus collected directly from Twitter (X). It uses the Elbow Method and the Silhouette Score together, treating the second as a diagnostic of the corpus rather than a score to be maximized. And it reads the resulting clusters against a practical question: how is this program being talked about, and by whom? Four objectives follow. The first is to build a TF-IDF representation of MBG-related tweets. The second is to partition that representation with K-Means. The third is to identify the terms carrying the most weight in each cluster. The fourth is to evaluate the partition and interpret it as a picture of public attention to the program.

THEORETICAL FRAMEWORK AND HYPOTHESES

Text Mining and the Preprocessing of Social Media Text

Text mining extracts usable information from unstructured text and, in practice, is a sequence of decisions rather than a single method [2]. For Indonesian Twitter data, the sequence is fairly settled [7]. Case folding normalizes capitalization. Filtering strips URLs, mentions, hashtags, digits, punctuation, and symbols. Tokenization splits the remaining string into terms. Stopword removal drops high-frequency function words, and stemming reduces the rest to their roots. Every one of these steps discards information. That is the point of them, but it also means the choices made at this stage constrain everything downstream, which is why we report them in full in the next section.

TF-IDF

TF-IDF weights a term by how often it occurs in a document and how rare it is across the corpus [3]. The weight of term t in document d is $w(t,d) = tf(t,d) \times idf(t)$, where $idf(t) = \log(N / df(t))$, $tf(t,d)$ is the frequency of t in d , N is the number of documents, and $df(t)$ is the number of documents containing t . A term that appears almost everywhere receives a low weight; a term confined to a subset of documents receives a high one. The output is a document-term matrix that can be fed directly to a clustering algorithm [8].

K-Means, the Elbow Method and the Silhouette Score

K-Means partitions data into K clusters based on proximity to a centroid [4]. It fixes the number of clusters, initializes the centroids, assigns each document to the nearest centroid, recomputes the centroids, and repeats until the assignment no longer changes. Compactness is summarised by the Sum of Squared Errors, the total squared distance between each document and its own centroid. The Elbow Method sets K at the point where adding another cluster no longer buys a meaningful reduction in SSE [5]. The Silhouette Score asks a different question, namely how well each document fits its own cluster compared with the nearest rival, through $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$, where $a(i)$ is the mean distance from document i to the other members of its cluster and $b(i)$ is the mean distance to the members of the nearest neighboring cluster. The score runs from -1 to 1. Values near zero indicate that the clusters are close together and their boundaries are soft [9].

Position of This Study and Propositions

Table 1 sets the present study against the work reviewed above. The comparison makes the contribution clear enough: none of the methods used here is new, but their application to the MBG debate, with cluster-quality metrics reported alongside the interpretation, has not been done.

Table 1. Position of This Study Relative to Prior Work

No	Author and Focus	Main Finding	Position of This Study
1	Lin and Nuha [1] - sentiment analysis of Indonesian datasets	A hybrid deep-learning strategy captures public opinion in Indonesian text effectively.	Uses an unsupervised design to map topics rather than to classify sentiment.
2	Huan et al. [2] - CNN-BiLSTM text classification	Deep-learning architectures handle unstructured text with high accuracy.	Uses a classical, inexpensive pipeline that requires no manual labelling.
3	Liu et al. [3] - improved TF-IDF algorithm	TF-IDF weights terms effectively according to their importance within a document.	Applies TF-IDF as the input representation for clustering an MBG corpus.
4	Ahmed et al. [4] - survey of the K-Means algorithm	K-Means is simple, efficient, and well suited to large datasets.	Applies K-Means to a TF-IDF matrix of Indonesian-language MBG conversation.
5	Cui [5] - the Elbow Method	The Elbow Method supports the choice of an optimal cluster count through the SSE curve.	Uses the Elbow Method as the primary basis for setting K , and states why.
6	Kiftiyah et al. [6] - MBG and social justice	MBG is examined normatively as a question of social justice and political dynamics.	Adds empirical evidence drawn from public conversation on social media.
7	Hidayatullah and Ma'arif [7] - preprocessing Indonesian tweets	Preprocessing determines the quality of any analysis of Indonesian Twitter messages.	Applies five preprocessing stages and reports the resulting data loss.

No	Author and Focus	Main Finding	Position of This Study
8	Sihombing et al. [8] - TextRank with TF-IDF	TF-IDF is reliable for processing Indonesian text in web-based applications.	Uses TF-IDF for clustering features rather than for summarisation.
9	Vardakas et al. [9] - soft silhouette for deep clustering	The Silhouette Score measures the compactness and separation of clusters.	Reports the Silhouette Score as a statement of limitation, not only as validation.

Because the design is exploratory and unsupervised, we do not test inferential hypotheses. We state three propositions instead, each of which the clustering and its metrics can support or undermine. P1: TF-IDF weighting differentiates terms in the MBG corpus well enough for an interpretable group structure to emerge. P2: K-Means with K set by the Elbow Method produces groups that differ substantively, not merely numerically. P3: Because the corpus is confined to a single government program, vocabulary overlap will hold the Silhouette Score low even where the groups remain readable.

METHODS

This is a quantitative study based on a text-mining pipeline. The approach is quantitative in that tweets are converted into numbers using TF-IDF and then analyzed algorithmically, with the outcome measured by explicit metrics rather than by a human coder's reading. The pipeline runs in seven stages, shown in Figure 1.

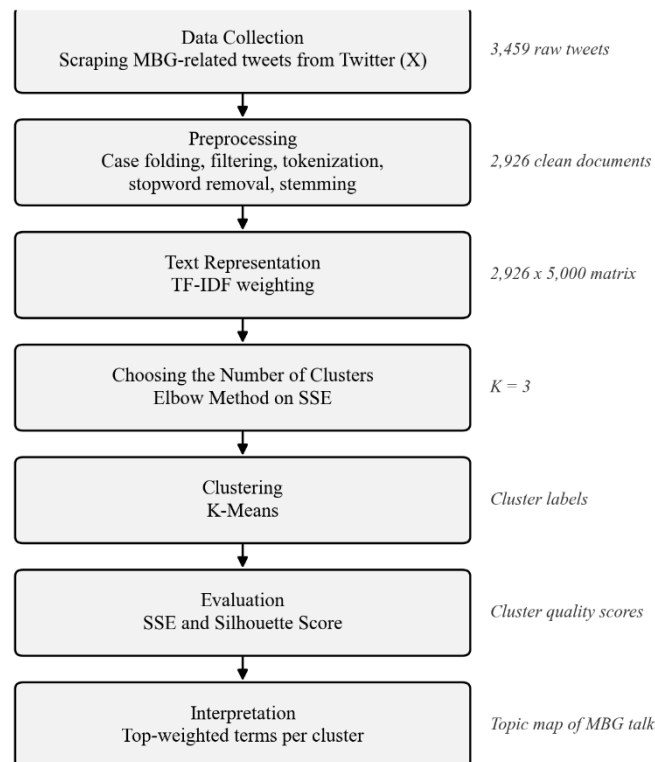


Figure 1. Research Flow

Data Collection

Tweets were collected from Twitter (X) via web scraping using the keywords "makan bergizi gratis" and its abbreviation, "MBG," during January and February 2025. Those two months cover the first weeks of the program and the argument that followed, which is when the conversation was at its busiest. The text field of each tweet is the unit of analysis; the username and timestamp were retained only to trace the origin of a record. Scraping returned 3,459 raw tweets. Table 2 lists the keywords, collection window, parameter settings, and software, enabling the procedure to be reproduced.

Table 2. Study Configuration and Parameters

Component	Value or Setting
Data source	Twitter (X)
Collection technique	Web scraping
Search keywords	"makan bergizi gratis", "MBG"
Collection window	January to February 2025
Document language	Indonesian
Field analysed	Tweet text
Raw records	3,459 documents
Clean records	2,926 documents
Preprocessing stages	Case folding, filtering, tokenization, stopword removal, stemming, removal of empty and duplicate records
Text representation	TF-IDF, capped at 5,000 features
Feature matrix	2,926 x 5,000
Selection of K	Elbow Method over K = 2 to K = 6
Clustering algorithm	K-Means, Euclidean distance, K = 3
Evaluation metrics	Sum of Squared Error (SSE) and Silhouette Score
Software	Python; Scikit-Learn (TfidfVectorizer, KMeans); Sastrawi (Indonesian stemmer)

Preprocessing

Preprocessing removes noise and standardizes the text. It runs through case folding; filtering of URLs, mentions, hashtags, digits, punctuation, and symbols; tokenization; stopword removal; and stemming with Sastrawi, the standard stemmer for Indonesian, to recover root forms. Empty records and duplicates were then dropped so that every remaining document is unique. Table 3 traces a few real examples through the stages; the tweets are shown in the original Indonesian, since that is what the pipeline operates on.

Table 3. Worked Examples of the Preprocessing Stages

Stage	Before	After
Case folding	Makan Siang Bergizi Gratis	makan siang bergizi gratis
Filtering	Makan Siang Bergizi Gratis https://t.co/f27aIthAdX	makan siang bergizi gratis
Stopword removal	Semoga program makan bergizi gratis tetap berlanjut	Semoga program makan bergizi gratis berlanjut
Stemming	Momen Prabowo Tanda Tangani Sepatu Siswa di Bogor saat Tinjau Makan Bergizi Gratis	Momen Prabowo tanda tangan sepatu siswa Bogor tinjau makan gizi gratis

Representation, Clustering and Evaluation

The cleaned documents were weighted using the Scikit-Learn implementation of TF-IDF, capped at 5,000 features, resulting in a 2,926 x 5,000 matrix. K was searched over the range 2 to 6 using the

Elbow Method on SSE, and the Silhouette Score was computed over the same range for comparison. K-Means was then run on the TF-IDF matrix under Euclidean distance. Two metrics report the outcome: SSE for internal compactness, the Silhouette Score for separation between clusters. Interpretation was carried out by selecting the ten highest-weighted terms for each centroid and comparing them against what is known about the program and its coverage.

RESULTS AND DISCUSSION

Collection and Preprocessing

Scraping produced 3,459 tweets about the MBG program, covering delivery in schools, the experience of pupils receiving meals, the role of government figures, and general reaction to how the program is being run. Preprocessing left 2,926 documents. Table 4 gives the breakdown.

Table 4. Data Reduction During Preprocessing

Description	Documents	Share
Raw records from scraping	3,459	100.00%
Discarded (empty, duplicate, irrelevant)	533	15.41%
Clean records used in the analysis	2,926	84.59%

The loss of 533 documents, 15.41% of the raw set, is accounted for mostly by identical reposts of news items and by tweets carrying nothing but a link. Removing them matters more than the headline number suggests, because repeated identical sentences would otherwise inflate the weight of whatever phrasing a widely shared news item happened to use. This is consistent with the emphasis earlier work places on preprocessing as a determinant of validity for Indonesian Twitter data [7].

TF-IDF Representation

Weighting the 2,926 clean documents produced a 5,000-term matrix. The weight distribution is itself informative. The core vocabulary of the program, makan (eat), gizi (nutrition), gratis (free), and program, appears in almost every document and therefore carries a low inverse document frequency. Terms that mark a specific context score higher even though they are rare include swasembada (self-sufficiency), kapolri (the national police chief), anggar (the stem of anggaran, budget), and tinjau (to inspect). This supports P1. It is also an early warning: a corpus organized around one program carries a large shared core, and that shared core places a ceiling on how sharply any lexical method can separate it [3].

Choosing the Number of Clusters

SSE was computed for K from 2 to 6, with the results in Figure 2. The largest single drop comes from K = 2 to K = 3, at 28.04 units. After that, the returns shrink and no longer behave regularly: 19.50 units from K = 3 to K = 4, 22.62 units from K = 4 to K = 5, and 11.57 units from K = 5 to K = 6. The bend in the curve, therefore, occurs at K = 3.

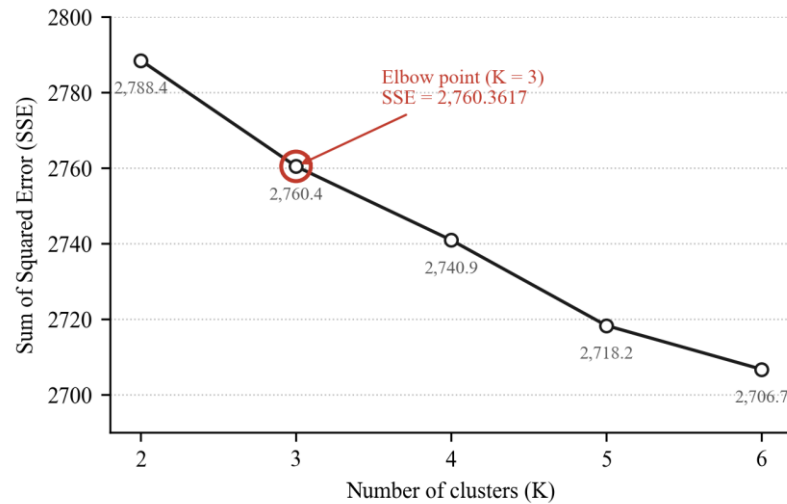


Figure 2. Elbow Curve of SSE Across the Number of Clusters

The Silhouette Score was computed over the same range as a cross-check, and is shown in Figure 3. Its numerical maximum occurs at $K = 5$, with a value of 0.0117, but every value in the range lies between 0.0086 and 0.0117. The gap between the maximum and the value at $K = 3$ is 0.0025, which is far too small to establish that one partition separates better than another. Put plainly, the Silhouette Score cannot discriminate here, and selecting K on the strength of a difference in the third decimal place would be reading noise. We therefore set K to 3 using the Elbow Method [5], which has the additional merit of yielding the most readable groups. Running the pipeline at four and five clusters only split the general-conversation group further; no new theme appeared.

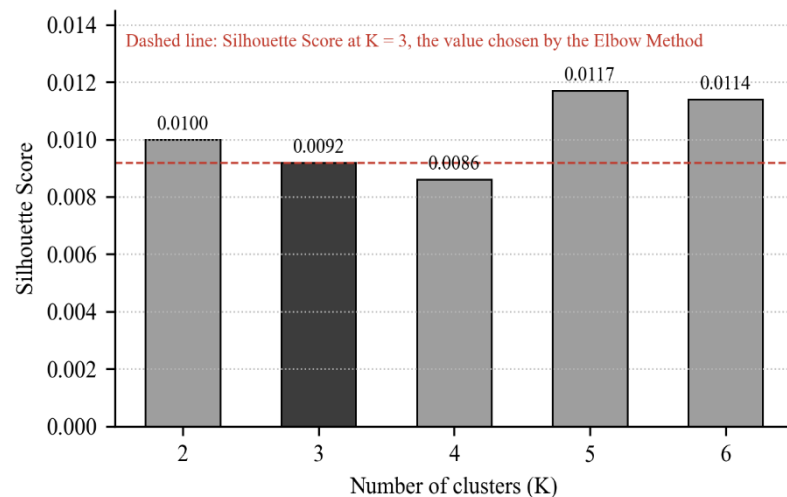


Figure 3. Silhouette Score Across the Number of Clusters

Clusters and Their Interpretation

K-Means was run on the TF-IDF matrix at $K = 3$. Table 5 reports the ten highest-weighted terms for each centroid, along with cluster size, share, and interpretation, and Figure 4 shows the size distribution.

Table 5. Top Terms, Size and Interpretation of Each Cluster

Cluster	Ten Highest-Weighted Terms	Size (Share)	Interpretation
0	Polri, pangan, jenderal, tni, dukung, wujud, program, Kapolri, tahan, swasembada	94 (3.21%)	Institutional endorsement of MBG, framed through national food security and self-sufficiency.
1	siang, makan, gratis, yg, program, gizi, prabowo, anggar, gak, gibran	2,012 (68.76%)	Everyday public talks about how the program is delivered, what it costs, and whether it works.
2	gizi, Prabowo, program, anak, makan, gratis, presiden, sehat, tinjau, mbg	820 (28.02%)	Benefits for children and pupils, health outcomes, and the link to government policy.

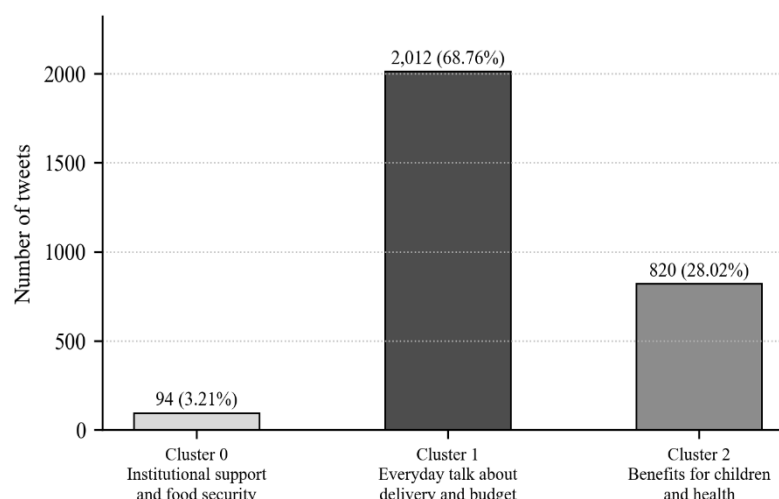


Figure 4. Distribution of Tweets Across Clusters

Cluster 0 holds 94 documents, or 3.21%, and is dominated by polri and kapolri (the national police and its chief), tni (the armed forces), pangan (food), tahan and swasembada (resilience and self-sufficiency), dukung (support) and wujud (to realize). These terms mark a framing in which MBG belongs to a national food-security agenda, and the accounts producing them are largely institutional or news accounts reporting official activity. The size of the cluster is the interesting part: this framing is broadcast frequently but picked up by almost nobody. Cluster 1 is by far the largest at 2,012 documents, or 68.76%, and turns on siang (midday), makan, gratis, program, gizi, the political names Prabowo and Gibran, anggar (budget), and the colloquial forms yg and gak, standing for yang and tidak. The presence of the colloquial forms is diagnostic.

This is ordinary conversation rather than reporting, and it runs on personal opinion, complaints about how the program is being delivered, and arguments about what it costs. Cluster 2 contains 820 documents, or 28.02%, with gizi, anak (child), sehat (healthy), presiden, tinjau (to inspect), prabowo and mbg among its leading terms. Talk here centers on what the program does for children and pupils and on its link to government policy, including coverage of official visits to schools. The tone is noticeably warmer than in Cluster 1. The three groups differ substantively, which supports P2. They also share a core, since makan, gizi, gratis, and program appear near the top of multiple centroids. Separation is therefore doing its work through the accompanying vocabulary rather than through the vocabulary of the program itself.

Cluster Quality

The final configuration at $K = 3$ returns an SSE of 2,760.3617 and a Silhouette Score of 0.0092. A Silhouette Score this close to zero means that most documents sit at roughly equal distance from their own centroid and from the nearest rival, so the boundaries between clusters are soft [9]. This is what P3 anticipated, and there are two reasons for that. The first is the corpus: every document discusses the same program, so lexical overlap is high by construction. The second is representation, since TF-IDF operates on surface forms and lacks access to meaning, which can leave two tweets making the same point in different words far apart in the feature space [3]. The high dimensionality and sparsity of that space compound the problem, because Euclidean distances between sparse vectors lose much of their discriminating power, and K-Means depends on exactly those distances [4]. The partition should be read accordingly: as a first map of where attention falls, not as a clean separation of discourse into three sealed compartments. We think reporting this openly is worth the loss of a tidier result. Studies that present cluster interpretations without the accompanying metrics leave their readers with no way of telling a real structure from an arbitrary cut.

Against the supervised approaches reviewed earlier, whether hybrid deep learning for Indonesian sentiment [1] or convolutional and bidirectional LSTM classification [2], the pipeline used here is far cheaper to run and needs no labeled data at all. The price is contextual sensitivity, and the Silhouette Score is where that price shows up. The finding fits the established picture of TF-IDF as a strong lexical device for surfacing salient terms [3], [8], which weakens when it is asked to separate documents that are thematically close to begin with. Set beside the normative literature on MBG [6], the contribution here is a set of proportions. Almost seven of ten tweets fall into the group concerned with delivery and cost. In contrast, fewer than one in thirty belong to the institutional framing of food security and self-sufficiency. The gap between those two figures is the practical finding of this paper. The narrative that institutions are producing is not the narrative circulating among the public, and program communication keeps restating the first.

In contrast, the second, going unanswered, is unlikely to land. Questions about daily delivery and about where the money goes are what the public is actually asking. There is a methodological point as well: in a corpus organized around a single program, the Silhouette Score is better treated as a description of the corpus than as a criterion for choosing K . Several limitations bound these conclusions. The data come from a single platform over a single collection window, so nothing here should be read as national opinion. The corpus is modest at 2,926 documents. The representation ignores context, and no special handling was given to non-standard spelling, slang, or sarcasm, all of which are common in the material. The authors performed cluster interpretation without an inter-rater reliability check, so a degree of subjectivity remains. Each of these points marks a direction for the work that follows.

CONCLUSION

This study mapped the topic structure of public conversation about the Free Nutritious Meal program on Twitter (X) using TF-IDF and K-Means. From 3,459 raw tweets, preprocessing retained 2,926 documents, which were represented as a 2,926 x 5,000 TF-IDF matrix. The Elbow Method fixed the number of clusters at three, yielding an SSE of 2,760.3617 and a Silhouette Score of 0.0092. The three groups are institutional support tied to food security, with 94 documents or 3.21%; everyday conversation about delivery and budget, with 2,012 documents or 68.76%; and the benefits of the program for children and health, with 820 documents or 28.02%. That distribution is the central result: public attention focuses on how the program is run day-to-day and on its cost, while the food-security framing favored by institutions occupies a small corner of the conversation. The near-zero Silhouette Score shows that the three groups share a large common vocabulary and that their boundaries are soft, so the partition is best used as an initial thematic map rather than as a definitive segmentation of the debate. Work that follows should widen the collection window and the range of

sources, test context-aware representations such as Word2Vec, FastText, or IndoBERT against the same corpus, and add dimensionality reduction so that the structure of the feature space can be inspected directly rather than inferred from a single score.

REFERENCES

- C. H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *Journal of Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00782-9.
- H. Huan, Z. Guo, T. Cai, and Z. He, "A text classification method based on a convolutional and bidirectional long short-term memory model," *Connection Science*, vol. 34, no. 1, pp. 2108–2124, 2022, doi: 10.1080/09540091.2022.2098926.
- C.-Z. Liu, Y.-X. Sheng, Z.-Q. Wei, and Y.-Q. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018.
- M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, Aug. 2020, doi: 10.3390/electronics9081295.
- M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," *Clausius Scientific Press*, 2020, doi: 10.23977/accaf.2020.010102.
- A. Kiftiyah, F. A. Palestina, F. U. Abshar, and K. Rofiah, "Program Makan Bergizi Gratis (MBG) dalam Perspektif Keadilan Sosial dan Dinamika Sosial-Politik," *Pancasila: Jurnal Keindonesiaan*, vol. 5, no. 1, pp. 101–112, Apr. 2025, doi: 10.52738/pjk.v5i1.726.
- A. F. Hidayatullah and M. R. Ma'Arif, "Pre-processing Tasks in Indonesian Twitter Messages," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Mar. 2017, doi: 10.1088/1742-6596/801/1/012072.
- J. J. Sihombing, A. Arnita, S. I. Al Idrus, and D. Y. Niska, "Implementation of text summarization on Indonesian scientific articles using text rank algorithm with TF-IDF web-based," *Journal of Soft Computing Exploration*, vol. 5, no. 3, pp. 310–319, Dec. 2024, doi: 10.52465/joscex.v5i3.475.
- [9] G. Vardakas, I. Papakostas, and A. Likas, "Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters," Feb. 2024, doi: 10.1007/s10994-026-07026-w.