Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Reliability-Aware Public Issue Priority Mapping for Indonesia's Free Nutritious Meal Program Using Sentiment Calibration and BERTopic

Rizaldi[1], Dewi Anggraeni[2], Abdul Kholiq[3]

[1,2]Information Systems, Faculty of Computer Science, Universitas Royal, North Sumatra, Indonesia

[3]Information Systems, Faculty of Computer Science, Universitas Satya Negara Indonesia, DKI Jakarta, Indonesia

Received:
February 11, 2026
Revised:
June 15, 2026
Accepted:
July 26 2026
Published:
August 22, 2026

Corresponding Author:
Author Name*: Rizaldi
Email*:
rizaldipiliang@gmail.com

DOI:
10.63158/journalisi.v8i4.1753

Abstract. The Free Nutritious Meal Program (MBG) has generated extensive public discussion on platform X. This study develops a reliability-aware framework for mapping public issue priorities by integrating sentiment analysis, probability calibration, and topic modeling. A final dataset of 1,641 public X posts was de-identified, preprocessed, relevance-filtered, manually labeled, validated, and proportionally split. TF-IDF + SVM was used as a classical baseline, while IndoBERTweet was fine-tuned and calibrated using temperature scaling. BERTopic was applied to generate topics, followed by manual interpretation into substantive issue groups. TF-IDF + SVM outperformed IndoBERTweet, achieving 0.8138 accuracy and 0.7999 Macro-F1, while IndoBERTweet achieved 0.7611 accuracy and 0.7142 Macro-F1. IndoBERTweet was retained because it provides probability-based confidence scores for calibration and priority mapping. Calibration modestly reduced NLL from 0.5239 to 0.5199 and ECE from 0.0700 to 0.0682. BERTopic produced 19 non-noise topics with a coherence score of 0.4552. The highest-priority public discussion theme concerned food safety, poisoning-related discourse, and consumption quality. This framework provides initial public-opinion monitoring input, not definitive policy evaluation, and requires external validation before formal policy use.

Keywords: Probability Calibration, BERTopic, Public Issue Priority Mapping, Policy Monitoring, Free Nutritious Meal Program

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The Free Nutritious Meal Program (MBG) is one of Indonesia's national strategic programs designed to improve human resource quality through the provision of nutritious meals for targeted beneficiaries, particularly students and vulnerable groups. In national budget policy documents, the MBG Program is planned on a large implementation scale in terms of budget allocation and beneficiary coverage [1]. This scale makes MBG not only a nutrition-related intervention, but also a public policy issue that attracts broad public attention. As a large-scale government program, MBG is suitable as a case for public issue priority mapping because it involves multiple dimensions, including nutrition, education, public health, budget allocation, governance, implementation readiness, and public trust.

The development of social media has expanded the space in which citizens express responses to public policies. Platform X, formerly Twitter, provides a digital arena where users can rapidly express support, criticism, concerns, complaints, and information related to public issues. From the perspective of data-driven governance, social media data may serve as a bottom-up source of public feedback that complements formal policy evaluation mechanisms [2], [3]. However, public opinion on platform X should not be interpreted as representing all Indonesian citizens. Social media data are shaped by platform-specific user demographics, user activity patterns, hashtag or keyword selection, political amplification, and coordinated posting behavior. Therefore, this study treats platform X data as digital public discourse rather than a representative measure of national public opinion.

The use of platform X data for policy-related analysis also requires conceptual and methodological caution. Posts on social media may reflect perceptions, concerns, expectations, political narratives, media reactions, or personal experiences, but they do not directly verify actual policy implementation conditions. Thus, discussions about issues such as food safety, poisoning-related discourse, budget, governance, or public trust in this study should be understood as public discussion themes emerging from digital discourse, not as verified factual evidence of program incidents. For this reason, the framework proposed in this study is intended to support preliminary public-opinion monitoring and academic NLP framework development, rather than definitive policy evaluation.

Analyzing public opinion on platform X presents several methodological challenges. Posts are usually short, unstructured, informal, and often contain abbreviations, slang, code-mixing, irony, sarcasm, and implicit expressions. These characteristics make manual analysis inefficient and may cause simple computational approaches to lose contextual meaning. In the Indonesian language context, this challenge is more complex because social media language often differs from formal language used in official documents, news, or surveys [4], [5]. Therefore, Natural Language Processing (NLP) is needed to systematically extract sentiment patterns and issue structures from digital public opinion data.

Previous studies have widely used sentiment analysis to understand public responses to government issues, public services, and digital policies. Classical machine learning approaches such as Support Vector Machine (SVM), Naïve Bayes, and TF-IDF representations have been used to classify public opinion into positive, negative, and neutral categories [6], [7]. Transformer-based language models have also been increasingly used to improve contextual understanding in text classification tasks [8], [9]. In the Indonesian social media context, IndoBERTweet is relevant because it was specifically developed using an Indonesian Twitter corpus [10]. Nevertheless, the use of more complex Transformer-based models does not automatically guarantee better classification performance or more reliable probability estimates, particularly when the dataset is limited, imbalanced, and linguistically diverse [11], [12].

In addition to sentiment polarity, policy-related public opinion analysis requires an understanding of the issues underlying the sentiment. Sentiment analysis alone can identify whether a post is positive, negative, or neutral, but it cannot explain what substantive issue is being discussed. Topic modeling is therefore needed to group posts into interpretable issue themes. Traditional topic modeling methods such as Latent Dirichlet Allocation (LDA) often face limitations when applied to short and informal texts because they rely heavily on word co-occurrence and may not fully capture semantic proximity [13], [14]. BERTopic offers an alternative by combining embeddings, dimensionality reduction, clustering, and class-based TF-IDF to generate semantically informed topic representations [15], [16].

Several recent studies have combined sentiment analysis and topic modeling to provide a more comprehensive understanding of public opinion [17], [18]. However, many of these studies still focus on sentiment labels and topic clusters without explicitly considering the reliability of model confidence when ranking issues. This creates an important methodological gap. If issue priority ranking uses model outputs as analytical indicators, confidence scores should not be treated as neutral or automatically reliable. Deep learning models and pretrained Transformers may produce overconfident probabilities, meaning that a high confidence score does not always correspond to a high likelihood of correctness [11], [12]. Therefore, probability calibration is needed when confidence scores are used as part of an issue-priority ranking mechanism.

Based on this gap, this study proposes a reliability-aware public issue priority mapping framework for analyzing public discussion of the MBG Program on platform X. The framework integrates three main components: validated sentiment labels, calibrated confidence scores, and BERTopic-based issue groups. Probability calibration is included not merely as a model reliability metric, but as a mechanism for making confidence scores more suitable for use in issue ranking. In this study, the Priority Score is constructed from Issue Share, Negative Ratio, and Average Calibrated Confidence. Issue Share represents the relative discussion volume of an issue, Negative Ratio represents the proportion of negative sentiment within that issue, and Average Calibrated Confidence represents the average calibrated probability confidence associated with the posts in the issue group.

This study makes two main contributions. First, it develops a reliability-aware NLP framework that integrates sentiment classification, probability calibration, and BERTopic for public issue priority mapping. Second, it demonstrates how digital public discussion about the MBG Program can be organized into preliminary monitoring input by combining issue volume, negative sentiment proportion, and calibrated confidence. The primary orientation of this study is an academic NLP framework for preliminary public-opinion monitoring, not an operational policy decision system or a definitive evaluation of MBG implementation.

This study aims to develop and evaluate an academic NLP framework for preliminary public-opinion monitoring of the MBG Program on platform X. Specifically, it evaluates

sentiment classification and calibration performance, identifies dominant issue themes using BERTopic, and integrates sentiment, calibrated confidence, and topic groups into a public issue priority map. The findings are intended to support cautious analysis of digital public discourse and should be triangulated before any formal policy use.

2. METHODS

This study employed a quantitative-computational NLP approach to develop a reliability-aware framework for preliminary public-opinion monitoring of the Free Nutritious Meal Program (MBG) on platform X. The framework integrated data collection, de-identification, preprocessing, relevance filtering, sentiment annotation, sentiment classification, probability calibration, BERTopic-based topic modeling, manual issue interpretation, and sentiment-topic integration into a Priority Score. The overall research workflow is presented in Figure 1.

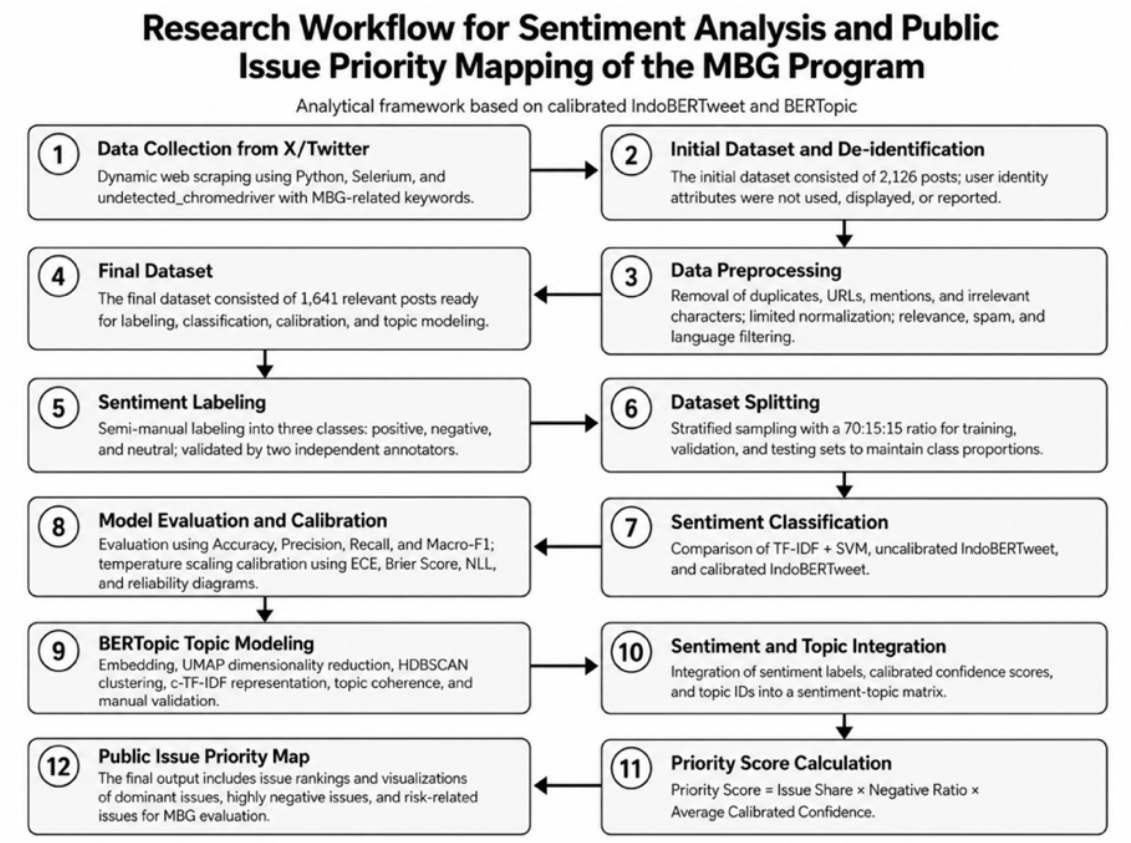

Figure 1. Research workflow based on calibrated IndoBERTweet and BERTopic for public issue priority mapping of the MBG Program

2.1. Data Collection and Preprocessing

Data were collected on June 4, 2026, from publicly visible posts on platform X using Python, Selenium, and undetected_chromedriver. Selenium was used to render dynamic search results [19]. The method accessed only publicly visible search results and did not collect private messages or restricted account information. Because platform access rules and technical restrictions may change, this collection approach is reported transparently as a methodological limitation rather than as official platform API access. The collected data were then cleaned, de-identified, annotated, and prepared for modeling on June 4-5, 2026. The search covered public posts from January 1, 2024, to April 30, 2026, using date-restricted query operators in the X search interface.

The search terms included "MBG", "makan siang gratis", "makan bergizi gratis", "program mbg", "maksi gratis", "susu gratis", "makan gratis", and "gizi gratis", combined with since: and until: operators. No explicit lang:id filter was applied; Indonesian-language and topic relevance were controlled through Indonesian keyword selection and subsequent manual relevance filtering. The scraper stored only username, post text, and search-date range; therefore, retweets, reposts, and quote posts could not be systematically distinguished based on post-type metadata. Exact duplicate texts were removed. Automated bot or spam detection was not applied because account-level metadata such as account age, follower count, posting frequency, verification status, and network behavior were not collected.

The initial dataset contained 2,126 posts. After removing 38 duplicates, 2,088 unique posts remained. Relevance filtering removed 418 posts referring to non-MBG contexts such as crypto tokens, Multibank, RWA, Mahjong, and other unrelated meanings of "MBG". A strict review removed 29 additional posts whose MBG context was unclear or not substantively relevant. The final experimental dataset consisted of 1,641 posts.

All direct identifiers, including usernames, account handles, post links, and other identifiable attributes, were removed before analysis. Each post was assigned an internal research ID, and results were reported only in aggregate form. Although the data were publicly visible, they were treated as digital traces requiring ethical caution [20], [21]. Preprocessing included removing URLs, mentions, HTML characters, excessive whitespace, non-substantive symbols, and selected non-standard word forms. Hashtag

symbols were removed while meaningful hashtag words were retained. Stopword removal and stemming were not used as main preprocessing steps because negation and contextual words are important for sentiment interpretation.

2.2. Sentiment Annotation and Dataset Splitting

Sentiment annotation used three categories: positive, negative, and neutral. Positive posts expressed support, appreciation, acceptance, or optimism toward the MBG Program. Negative posts expressed criticism, complaints, concerns, distrust, negative sarcasm, or dissatisfaction. Neutral posts were descriptive, informational, or lacked a clear evaluative tendency. Two annotators with backgrounds in information systems and computational text analysis labeled the dataset using a shared annotation guideline. For the final 1,641 posts, their pre-adjudication labels were compared before consensus. They initially agreed on 1,603 posts and disagreed on 38 posts, resulting in a pre-adjudication Cohen's Kappa of 0.9590 [22]. Disagreements were resolved through discussion and consensus adjudication by the two annotators; no third reviewer was involved. Thus, the final labels are consensus labels, and Kappa = 1.00 is not treated as an independent inter-annotator reliability statistic. Difficult annotation cases included sarcasm, political references, mixed sentiment, and indirect MBG mentions. Paraphrased examples used during adjudication covered positive support for child nutrition, negative criticism of meal-budget allocation, neutral reporting of an MBG event, sarcastic framing of free meals as political bait, and mixed support for children with concern about implementation. The final dataset consisted of 252 positive posts, 449 negative posts, and 940 neutral posts. Stratified sampling split the data into training, validation, and testing sets with a 70:15:15 ratio, producing 1,148, 246, and 247 posts, respectively. A random state of 42 was used for reproducibility following scikit-learn train_test_split [23].

2.3. Sentiment Classification and Probability Calibration

The sentiment classification experiment compared TF-IDF + SVM, uncalibrated IndoBERTweet, and calibrated IndoBERTweet. TF-IDF + SVM used unigram TF-IDF, min_df=2, Linear SVM, C=10, and class_weight=balanced to reduce class-imbalance effects. SVM probability calibration using Platt scaling or isotonic regression was not tested because the SVM was configured as a discriminative baseline; this comparison is recommended for future work.

IndoBERTweet was fine-tuned using indolem/indobertweet-base-uncased [10] on Tweet_Text_Clean with Final_Label as the target. The model used max sequence length 128, batch size 16, learning rate 2e-5, weight decay 0.01, three epochs, AdamW, a linear scheduler with 10% warm-up, gradient clipping at 1.0, and random seed 42. Stratified splitting handled class imbalance; class-weighted loss and focal loss were not used to keep fine-tuning stable and comparable with post-hoc calibration. Sarcasm was handled through annotation and error interpretation. The experiment ran in Google Colab with Python 3.12 and a CUDA-enabled GPU; the executable notebooks include package-version logging for scikit-learn, Transformers, PyTorch, BERTopic, sentence-transformers, UMAP, HDBSCAN, NumPy, pandas, and Matplotlib.

Probability calibration was applied to IndoBERTweet using temperature scaling because Transformer-based models may produce overconfident probability estimates [12], [24], [25], [26], [27]. The temperature parameter was optimized on the validation set by minimizing Negative Log-Likelihood, resulting in an optimal temperature of 0.929380. For a post x_i with logit vector z_i and class k, the calibrated probability is defined as shown in Equation 1:

$$p_T(y_i = k \mid x_i) = \frac{\exp(z_{i,k}/T)}{\sum_{j=1}^{K} \exp(z_{i,j}/T)} \quad (1)$$

where K = 3 represents positive, negative, and neutral classes. The calibrated confidence score was obtained shown in Equation 2:

$$c_i = \max_{k \in \{1,\ldots,K\}} p_T(y_i = k \mid x_i) \quad (2)$$

Calibration was used to improve the reliability of confidence scores, not to change predicted labels. Therefore, unchanged Accuracy and Macro-F1 after calibration were expected.

2.4. Model Evaluation

Classification performance was evaluated using Accuracy, Precision, Recall, and Macro-F1. Macro-F1 was emphasized because the dataset was imbalanced. Per-class metrics and confusion matrices were reported for TF-IDF + SVM and IndoBERTweet. Probability

reliability was evaluated using NLL, Brier Score, ECE, and reliability diagrams; ECE used 10 bins [27]. Because the testing set contained only 247 posts, the results are interpreted as controlled experimental findings, and strong population-level performance claims are avoided.

2.5. BERTopic Topic Modeling and Issue Interpretation

BERTopic was used to identify issue themes because it combines embeddings, dimensionality reduction, clustering, and class-based TF-IDF for semantic topic representation [15], [16]. The input text column was Tweet_Text_Clean, while Final_Label was retained for sentiment-topic integration. Embeddings were generated using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 with batch size 32 and normalized embeddings.

UMAP used n_neighbors=15, n_components=5, min_dist=0.0, metric=cosine, and random_state=42 [28]. HDBSCAN used min_cluster_size=15, min_samples=5, metric=euclidean, cluster_selection_method=eom, and prediction_data=True. Topic representation used CountVectorizer with Indonesian stopwords, ngram_range=(1,2), and min_df=2. BERTopic used top_n_words=10 and calculate_probabilities=True.

Topic coherence was calculated using the c_v metric [29]. BERTopic generated 19 non-noise topics, assigned 318 posts to Topic -1 as noise, and produced a global coherence score of 0.455244. Because X posts are short and informal, coherence was not the only validity criterion. Manual interpretation reviewed topic keywords, topic size, and anonymized representative posts. Topic -1 was excluded as noise, while Topic 0 and Topic 3 were excluded because they mainly represented general or informal discussions with low issue specificity. Topic -1, Topic 0, and Topic 3 contained 318, 524, and 67 posts, respectively; therefore, 909 posts were excluded from the main priority map and 732 posts were retained in seven substantive issue groups. The initial topic-to-issue mapping was reviewed by an additional coder from the author team, who assessed consistency among Topic IDs, keywords, representative themes, and assigned issue groups, including the exclusion of Topic -1, Topic 0, and Topic 3. LDA comparison, per-topic coherence, parameter robustness testing, and independent external policy-domain expert validation were not conducted and are acknowledged as limitations.

2.6. Sentiment–Topic Integration and Priority Score

The final stage integrated Final_Label, BERTopic topic ID, and calibrated confidence into a public issue priority map. Only substantive issue groups were included in the main ranking. Issue Share (IS) was defined shown in Equation 3:

$$IS_g = \frac{n_g}{N_s} \tag{3}$$

where n_g is the number of posts in issue group g, and N_s is the total number of substantive posts included in the priority map.

Negative Ratio (NR) was defined shown in Equation 4:

$$NR_g = \frac{n_{g,neg}}{n_g} \tag{4}$$

where $n_{g,neg}$ is the number of negative posts in issue group g. Average Calibrated Confidence (ACC) was defined shown in Equation 5:

$$ACC_g = \frac{1}{n_g} \sum_{i \in g} c_i \tag{5}$$

where c_i is the calibrated confidence score of post i. ACC was used as a reliability weight, not as a sentiment direction indicator. The final Priority Score (PS) was calculated shown in Equation 7:

$$PS_g = IS_g \times NR_g \times ACC_g \tag{7}$$

This formula ranks issue groups by discussion share, negative sentiment concentration, and calibrated confidence. It is an analytical ranking mechanism for initial public-opinion monitoring, not a definitive policy-risk score. Sensitivity analysis evaluates topic inclusion and alternative weighting assumptions.

3. RESULTS AND DISCUSSION

3.1. Dataset Filtering and Label Distribution

The data filtering process is summarized in Table 1. From the initial 2,126 collected posts, duplicate removal, relevance filtering, and strict review produced a final experimental dataset of 1,641 posts.

Table 1. Data filtering and relevance screening summary

Filtering Stage	Removed Posts	Remaining Posts	Description
Initial collected posts	-	2,126	Posts collected using MBG-related keywords
Duplicate removal	38	2,088	Exact duplicate texts were removed
Relevance filtering	418	1,670	Posts using "MBG" in unrelated contexts were removed
Strict relevance review	29	1,641	Posts with unclear or non-substantive MBG context were removed
Final experimental dataset	-	1,641	Dataset used for sentiment classification and topic modeling

The sentiment label distribution and stratified dataset split are presented in Table 2. The neutral class was the majority class, accounting for 57.28% of the dataset, followed by negative posts at 27.36% and positive posts at 15.36%.

Table 2. Sentiment label distribution and dataset splitting

Sentiment Label	Total Data	Percentage	Training Set	Validation Set	Testing Set
Positive	252	15.36%	176	38	38
Negative	449	27.36%	314	67	68
Neutral	940	57.28%	658	141	141
Total	1,641	100%	1,148	246	247

3.2. Sentiment Classification and Calibration

The overall classification performance of TF-IDF + SVM, uncalibrated IndoBERTweet, and calibrated IndoBERTweet is presented in Table 3. TF-IDF + SVM achieved the best classification performance, with an accuracy of 0.8138 and Macro-F1 of 0.7999.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

IndoBERTweet achieved an accuracy of 0.7611 and Macro-F1 of 0.7142. Temperature scaling did not change the predicted labels; therefore, the classification performance of calibrated IndoBERTweet remained identical to the uncalibrated version.

Table 3. Overall classification performance on the testing set

Model	Accuracy	Macro Precision	Macro Recall	Macro-F1
TF-IDF + SVM	0.8138	0.7926	0.8083	0.7999
IndoBERTweet	0.7611	0.7283	0.7034	0.7142
Calibrated IndoBERTweet	0.7611	0.7283	0.7034	0.7142

The per-class precision, recall, and F1-score are shown in Table 4. TF-IDF + SVM outperformed IndoBERTweet on the positive and neutral classes, while both models showed closer performance on the negative class. This result indicates that the Transformer-based model did not automatically outperform the classical baseline in this dataset.

Table 4. Per-class precision, recall, and F1-score

Model	Class	Precision	Recall	F1-score	Support
TF-IDF + SVM	Positive	0.7750	0.8158	0.7949	38
TF-IDF + SVM	Negative	0.7361	0.7794	0.7571	68
TF-IDF + SVM	Neutral	0.8667	0.8298	0.8478	141
IndoBERTweet	Positive	0.6563	0.5526	0.6000	38
IndoBERTweet	Negative	0.7313	0.7206	0.7259	68
IndoBERTweet	Neutral	0.7973	0.8369	0.8166	141

The confusion matrices of TF-IDF + SVM and IndoBERTweet are presented in Table 5. TF-IDF + SVM correctly classified 31 of 38 positive posts, 53 of 68 negative posts, and 117 of 141 neutral posts. IndoBERTweet correctly classified 21 of 38 positive posts, 49 of 68 negative posts, and 118 of 141 neutral posts. The main classification difficulty occurred in distinguishing evaluative positive or negative posts from neutral posts, especially for IndoBERTweet.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 5. Confusion matrices of TF-IDF + SVM and IndoBERTweet

	Actual / Predicted	Positive	Negative	Neutral
TF-IDF + SVM	Positive	31	1	6
	Negative	3	53	12
	Neutral	6	18	117
	Actual / Predicted	**Positive**	**Negative**	**Neutral**
IndoBERTweet	Positive	21	4	13
	Negative	2	49	17
	Neutral	9	14	118

Although TF-IDF + SVM produced better classification performance, IndoBERTweet was retained in the framework because it provides probability-based outputs that can be calibrated and used as confidence scores in the priority mapping stage. The calibration results are presented in Table 6. Temperature scaling modestly reduced NLL from 0.5239 to 0.5199 and ECE from 0.0700 to 0.0682, while the Brier Score remained nearly unchanged. Therefore, the calibration result should be interpreted as a small improvement in probability reliability, not as a substantial improvement in classification performance.

Table 6. Calibration metrics of IndoBERTweet before and after temperature scaling

Model	Accuracy	Macro-F1	NLL	Brier Score	ECE
IndoBERTweet without calibration	0.7611	0.7142	0.5239	0.3013	0.0700
Calibrated IndoBERTweet	0.7611	0.7142	0.5199	0.3013	0.0682

The reliability diagram in Figure 2 visualizes the confidence–accuracy alignment of IndoBERTweet before and after temperature scaling. The diagram supports the quantitative calibration metrics by showing that calibration slightly improved the alignment between predicted confidence and empirical accuracy, although the improvement was modest.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

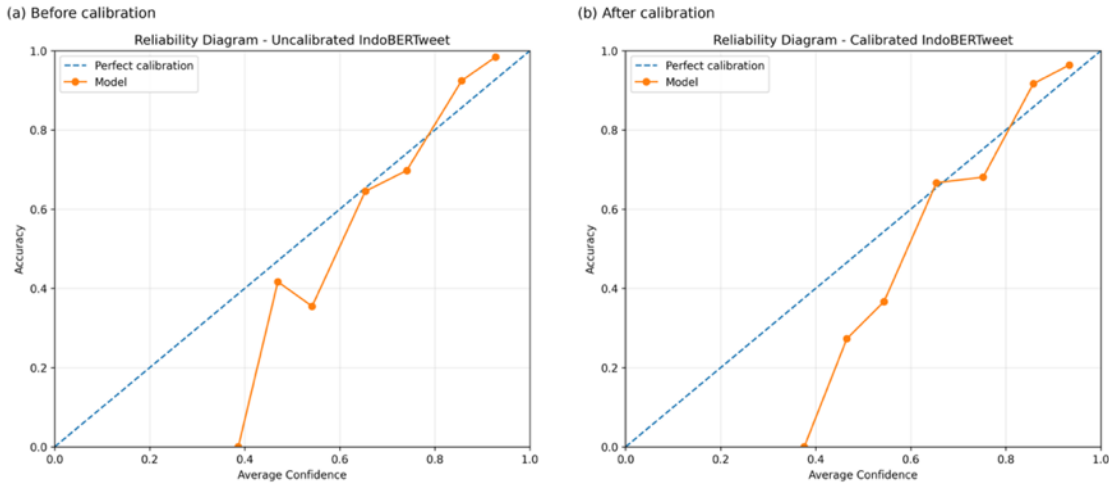

Figure 2. Reliability diagram of IndoBERTweet before and after temperature scaling

3.2.1. BERTopic and Manual Issue Grouping

After sentiment classification and calibration, BERTopic was used to identify issue themes in public discussion of the MBG Program. The BERTopic topic summary is presented in Table 7. The model generated 19 non-noise topics and assigned 318 posts to Topic -1 as noise. The global coherence score was 0.4552, indicating a moderate level of topic interpretability. Topic 0 and Topic 3 were not included in the main priority map because they mainly represented general or informal discussions rather than specific substantive policy issues.

Table 7. BERTopic topic summary

Topic ID	Count	Main Keywords (Indonesia)	Paraphrased Representative Theme	Status
-1	318	yg, siang, aja, rakyat	Unstable or mixed posts without a coherent issue cluster	Excluded
0	524	yg, aja, ya, ga	General or informal discussion about MBG	Excluded
1	151	indonesia, anak, free nutritious, nutritious	Support for MBG as a child nutrition program	Included
2	76	keracunan, kasus keracunan	Public concern about poisoning-related discourse	Included

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Topic ID	Count	Main Keywords (Indonesia)	Paraphrased Representative Theme	Status
3	67	siang, dapet siang, bagaimana	General informal comments about receiving meals	Excluded
4	65	pendidikan, sekolah, free	Discussion of MBG in relation to schools and education	Included
5	64	anak, lanjutkanmbg, sekolah	Supportive narratives about children and school meals	Included
6	57	dpr, pemerintah, anggaran, polri	Institutional and budget-related discussion	Included
7	44	presiden prabowo, prabowo, dukungan	Supportive discourse linked to presidential policy	Included
8	41	pajak, koruptor, rakyat, korupsi	Taxation, corruption, and governance concerns	Included
9	40	ayam, nutritious, meal	Menu and food component discussion	Included
10	31	02, voters, siang	Electoral and voter-related discourse	Included
11	25	siang, enggak, bang, gas	Informal political and public trust discourse	Included
12	25	keracunan, siswa, siswa keracunan	Student poisoning-related discussion	Included
13	21	susu, anak, kid, siang	Milk, children, and consumption quality discourse	Included
14	20	kuliah, pendidikan, mahasiswa	Education and student policy priority discussion	Included
15	19	efisiensi, daerah, budget, anggaran	Efficiency, regional budget, and fiscal concerns	Included

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Topic ID	Count	Main Keywords (Indonesia)	Paraphrased Representative Theme	Status
16	19	presiden, capek, asing	Political fatigue and public trust discourse	Included
17	17	kerja, gaji	Labor and operational workforce discussion	Included
18	17	siang, dikasih siang, gas	Informal policy and governance-related comments	Included

The manual grouping of BERTopic topics into substantive issue groups is shown in Table 8. Manual grouping was necessary because BERTopic produces numerical topic IDs that require contextual interpretation before they can be translated into policy-relevant issue categories. Before finalization, the topic-to-issue mapping was reviewed by an additional coder to assess the consistency between the topic keywords, representative themes, and assigned issue groups. The additional coder agreed that the included topics formed coherent substantive issue groups. Topic -1 was excluded as noise, while Topic 0 and Topic 3 were excluded because they did not form specific substantive issue groups and mainly represented general or informal discussion.

Table 8. Manual grouping of BERTopic topics

Manual Issue Group	TopicID	Status
Food safety, poisoning incidents, and consumption quality	2, 12, 13	Included in the issue priority map
Program support and benefits for children	1, 5, 7	Included in the issue priority map
Education and policy priorities in education	4, 14	Included in the issue priority map
Budget, efficiency, taxation, and governance	6, 8, 15, 18	Included in the issue priority map
Electoral politics and public trust	10, 11, 16	Included in the issue priority map
Program operations and labor	17	Included in the issue priority map

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Manual Issue Group	TopicID	Status
Food/menu narratives and consumption components	9	Included in the issue priority map
Noise	-1	Excluded from the main issue priority map
General discussions/informal comments	0, 3	Excluded from the main issue priority map

3.2.2. Public Issue Priority Mapping

The final stage integrated Final_Label, BERTopic topic ID, and calibrated confidence to generate a public issue priority map. Of the 1,641 posts, 732 posts were included in the seven substantive issue groups. The priority mapping results are presented in Table 9. The Priority Score was calculated based on Issue Share, Negative Ratio, and Average Calibrated Confidence.

Table 9. Priority Score of substantive issue groups

Rank	Issue group	Topic ID	Count	IS	NR	ACC	PS
1	Food safety/poisoning/consumption quality	2,12,13	122	16.67%	72.13%	0.8462	0.1017
2	Budget/efficiency/taxation/governance	6,8,15,18	134	18.31%	54.48%	0.7741	0.0772
3	Electoral politics/public trust	10,11,16	75	10.25%	28.00%	0.6784	0.0195
4	Education/policy priorities	4,14	85	11.61%	22.35%	0.7390	0.0192
5	Program support/children	1,5,7	259	35.38%	6.56%	0.7775	0.0181
6	Program operations/labor	17	17	2.32%	29.41%	0.7639	0.0052
7	Food/menu narratives/components	9	40	5.46%	5.00%	0.8749	0.0024

The ranking of public discussion themes based on Priority Score is visualized in Figure 3. A compact sensitivity analysis was also conducted to examine whether the ranking changed when confidence influence was reduced, when confidence was excluded, when negative ratio was emphasized, and when Topic 0 and Topic 3 were included.

Table 10. Sensitivity analysis for Priority Score and inclusion of Topic 0 and Topic 3

Sensitivity setting	Top issue	Score	Second issue	Interpretation
Baseline substantive topics	Food safety/poisoning/ quality	0.1017	Budget/tax/ governance	Top - two ranking retained.
Lower confidence influence	Food safety/poisoning/ quality	0.1106	Budget/tax/ governance	Top two unchanged.
Volume-negative only	Food safety/poisoning/ quality	0.1202	Budget/tax/ governance	Top two unchanged without ACC.
Higher negative influence	Food safety/poisoning/ quality	0.0864	Budget/tax/ governance	Top two unchanged when NR is emphasized.
Topic 0 and 3 included	General/informal MBG discussions	0.0599	Food safety/poisoning/consumption quality	Large volume can dominate, but low specificity supports exclusion from substantive map.

These results indicate that the most frequently discussed issue is not necessarily the highest-priority substantive issue. Program support and benefits for children had the largest share (259 posts; 35.38%) but ranked fifth because its Negative Ratio was only 6.56%. In contrast, food safety, poisoning-related discourse, and consumption quality ranked first because they combined substantial issue share, high negative concentration, and high calibrated confidence. Sensitivity analysis showed that the top two substantive issues remained stable when ACC was reduced, removed, or when NR was emphasized. When Topic 0 and Topic 3 were included as a general/informal group, their high volume

produced a competitive score, confirming that their exclusion narrows the map to substantive policy issues rather than all MBG-related discourse. Therefore, the Priority Score provides initial public-opinion monitoring input and should not be interpreted as verified evidence of field conditions or definitive policy evaluation.

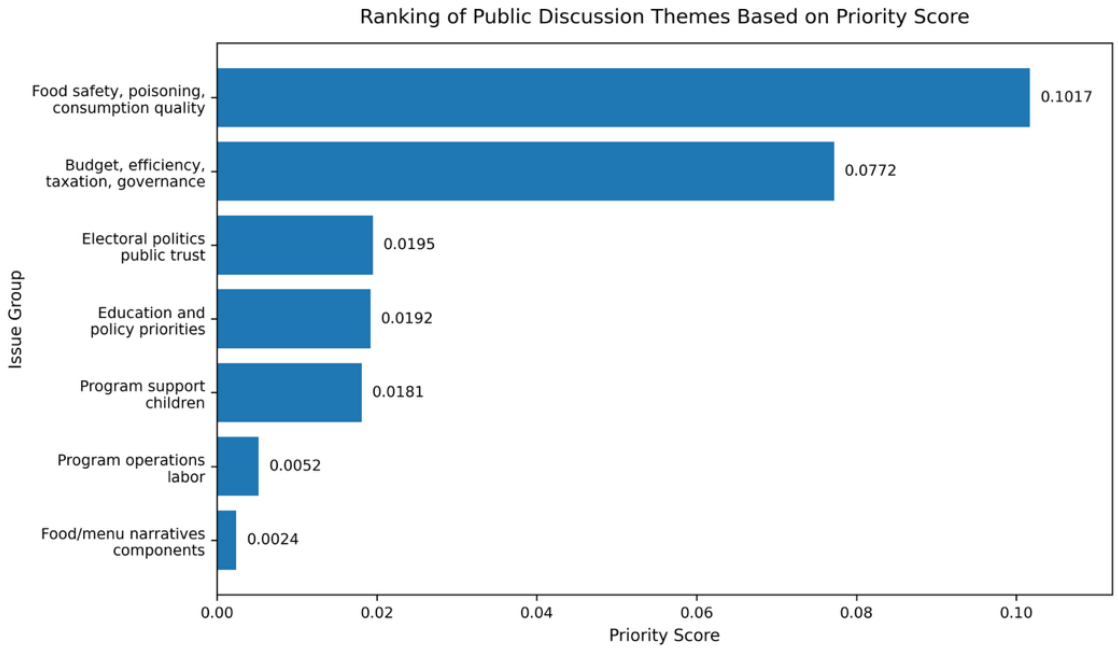

Figure 3. Ranking of Public Discussion Themes Based on Priority Score

3.3. Discussion

The first important finding concerns dataset validity. The reduction from 2,126 collected posts to 1,641 final experimental posts shows that relevance filtering was necessary because the acronym "MBG" was not always used to refer to the Free Nutritious Meal Program. If irrelevant posts had been retained, the sentiment classifier and BERTopic model could have learned patterns unrelated to the policy object. Therefore, duplicate removal, relevance filtering, strict review, and de-identification were not merely technical preprocessing steps, but also part of contextual validation.

The second finding concerns neutral-sentiment dominance. Neutral posts accounted for 57.28% of the dataset, indicating that many posts were descriptive, informational, or non-evaluative. This dominance affects priority mapping because large neutral or general-discussion clusters may appear important by volume alone. The formula addresses this

by including Negative Ratio; ACC increases priority only as a reliability weight, while sentiment direction is controlled by Negative Ratio.

The third finding is that TF-IDF + SVM outperformed IndoBERTweet on the testing set. This result is important because it shows that a Transformer-based model does not automatically outperform a classical baseline in every dataset. Several factors may explain this result. First, the dataset size was relatively limited for fine-tuning a Transformer model. Second, the class distribution was imbalanced, with the neutral class dominating the dataset. Third, platform X posts often contain informal language, implicit evaluation, political references, and sarcasm. Under these conditions, TF-IDF + SVM may capture discriminative lexical patterns more effectively than a fine-tuned Transformer model trained on limited data.

Nevertheless, IndoBERTweet remained important because the study aimed not only to maximize classification accuracy, but also to generate calibrated confidence scores for priority mapping. This decision has a limitation: confidence from a lower-performing model may still carry classification uncertainty. Therefore, confidence was used only as a reliability weight combined with Issue Share and Negative Ratio. Sensitivity analysis showed that top substantive issues remained stable even when confidence influence was reduced or removed, suggesting that ranking was not solely driven by IndoBERTweet confidence. Future work should compare this design with calibrated SVM using Platt scaling or isotonic regression.

The fourth finding concerns probability calibration. Temperature scaling produced only a modest improvement, reducing NLL from 0.5239 to 0.5199 and ECE from 0.0700 to 0.0682. This improvement should not be overstated. Calibration did not improve accuracy or Macro-F1 because temperature scaling does not change the predicted class order. Its value lies in making probability estimates slightly more reliable for downstream use. In this study, calibrated confidence was not treated as a sentiment indicator, but as a reliability weight in the priority mapping formula.

The fifth finding concerns topic modeling. BERTopic generated 19 non-noise topics with a global coherence score of 0.4552, showing that MBG discourse was multidimensional: food safety, education, budget efficiency, taxation, governance, electoral politics, public

trust, labor, and menu components. Sentiment analysis identifies polarity, while topic modeling clarifies the issue to which the sentiment refers.

The exclusion of Topic -1, Topic 0, and Topic 3 has methodological implications. Topic -1 represented noise or unstable clusters, while Topic 0 and Topic 3 mainly contained general or informal discussions without specific substantive policy issues. This exclusion reduced the main priority map to 732 substantive posts. The additional coder review supported the interpretation that these topics had lower issue specificity. Sensitivity analysis showed that including Topic 0 and Topic 3 would make general/informal MBG discussion highly competitive because of its volume but would reduce the interpretability of the map as a substantive issue-priority framework.

The sixth finding concerns the Priority Score ranking. Food safety, poisoning-related discourse, and consumption quality ranked first, even though this group did not have the largest number of posts. Its high rank was driven by a high Negative Ratio and high Average Calibrated Confidence. By contrast, program support and benefits for children had the largest number of posts but ranked fifth because its Negative Ratio was low. This result demonstrates that issue volume alone is not sufficient for priority mapping. A highly discussed issue is not necessarily a high-priority issue if the discussion is mostly supportive or neutral.

The finding that food safety and poisoning-related discourse ranked first should be interpreted carefully. It does not mean that the study verifies actual poisoning incidents or field-level implementation problems. Rather, it means that food safety, poisoning-related narratives, and consumption quality appeared as high-priority public discussion themes on platform X. This distinction is important because social media data reflect public discourse, perception, and attention, not direct evidence of factual implementation conditions.

The proposed framework contributes by integrating sentiment labels, BERTopic issue groups, and calibrated confidence into a single issue-ranking mechanism. Compared with sentiment analysis or topic modeling alone, it identifies substantive public concerns and adds a reliability-aware component through calibrated confidence, providing initial monitoring input for themes that may require closer attention.

Several limitations remain. The dataset came only from platform X and does not represent all Indonesian citizens. Bot/spam detection and coordinated-posting analysis were not conducted because account-level metadata were not collected, so political amplification and coordinated activity may affect the discourse. The test set was small, confidence intervals and repeated-seed evaluation were not included, and calibration improvement was modest. BERTopic grouping was reviewed by an additional coder from the author team but not by independent policy-domain experts. Finally, the Priority Score depends on formula and topic-inclusion choices; although sensitivity analysis was added, alternative weighting schemes may still change rankings. Overall, the results support a reliability-aware sentiment-topic framework for preliminary public-opinion monitoring. The framework is not a definitive policy evaluation system, operational early-warning tool, or causal evidence of MBG implementation outcomes. It organizes digital discourse into focused issue categories that require follow-up through official reports, field data, surveys, news verification, and expert assessment.

4. CONCLUSION

This study proposed a reliability-aware public issue priority mapping framework for analyzing MBG Program discussion on platform X by integrating sentiment analysis, probability calibration, and BERTopic. TF-IDF + SVM achieved stronger classification performance than IndoBERTweet (accuracy 0.8138; Macro-F1 0.7999 versus accuracy 0.7611; Macro-F1 0.7142). Nevertheless, IndoBERTweet remained useful because it provided probability-based confidence scores that could be calibrated and integrated into issue priority mapping. Temperature scaling did not change classification performance, but modestly reduced NLL from 0.5239 to 0.5199 and ECE from 0.0700 to 0.0682. BERTopic generated 19 non-noise topics, manually grouped into seven substantive issue groups. Food safety, poisoning-related discourse, and consumption quality ranked highest, followed by budget, efficiency, taxation, and governance. Sensitivity analysis showed that the top substantive issues remained stable under several weighting assumptions, while inclusion of Topic 0 and Topic 3 increased the influence of general/informal discussion because of their large volume. The main contribution is an initial public-opinion monitoring framework that combines issue volume, sentiment tendency, and probability reliability. The results are not definitive policy evaluation or verified evidence of MBG implementation. Future research should use multiple platforms, larger datasets,

independent policy-domain expert validation, alternative weighting schemes, calibrated SVM comparison, and triangulation with official reports, field data, news verification, and surveys.

REFERENCES

[1] Kementerian Keuangan Republik Indonesia, "Anggaran Makan Bergizi Gratis Rp71 triliun masuk postur RAPBN 2025." Accessed: Apr. 26, 2026. [Online]. Available: Kementerian Keuangan Republik Indonesia

[2] M. Lopreite, M. Misuraca, and M. Puliga, "Outbreak and integration of social media in public health surveillance systems: A policy review through BERT embedding technique," *Socioecon. Plann. Sci.*, vol. 95, Art. no. 101995, Oct. 2024, doi: 10.1016/j.seps.2024.101995.

[3] S. Verma, "Sentiment analysis of public services for smart society: Literature review and future research directions," *Gov. Inf. Q.*, vol. 39, no. 3, Art. no. 101708, Jul. 2022, doi: 10.1016/j.giq.2022.101708.

[4] A. Huwaidah, Adiwijaya, and S. Al Faraby, "Argument identification in Indonesian tweets on the issue of moving the Indonesian capital," *Procedia Comput. Sci.*, vol. 179, pp. 407–415, 2021, doi: 10.1016/j.procs.2021.01.023.

[5] A. Mohan, A. M. Nair, B. Jayakumar, and S. Muraleedharan, "Sarcasm detection using Bidirectional Encoder Representations from Transformers and Graph Convolutional Networks," *Procedia Comput. Sci.*, vol. 218, pp. 93–102, 2023, doi: 10.1016/j.procs.2022.12.405.

[6] R. Saputra, Y. Pristyanto, and I. N. Fajri, "Generative AI image sentiment analysis on social media X using TF-IDF and FastText," *J. Appl. Inform. Comput.*, vol. 9, no. 5, pp. 2509–2520, Oct. 2025, doi: 10.30871/jaic.v9i5.10627.

[7] E. Supriyadi and P. N. Makatita, "Sentiment analysis of TikTok user comments on QRIS adoption in Indonesia using IndoBERT," *Procedia Comput. Sci.*, vol. 269, pp. 121–130, 2025, doi: 10.1016/j.procs.2025.08.265.

[8] A. AlMasaud and H. H. Al-Baity, "On the robustness of Arabic aspect-based sentiment analysis: A comprehensive exploration of transformer-based models," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 10, Art. no. 102264, Dec. 2024, doi: 10.1016/j.jksuci.2024.102264.

[9] S. Taj, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Aspect-based sentiment analysis for software requirements elicitation using fine-tuned Bidirectional Encoder Representations from Transformers and Explainable Artificial Intelligence," *Eng. Appl. Artif. Intell.*, vol. 151, Art. no. 110632, Jul. 2025, doi: 10.1016/j.engappai.2025.110632.

[10] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization," in *Proc. 2021 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 10660–10668, doi: 10.18653/v1/2021.emnlp-main.833.

[11] M. H. Popescu, K. Roitero, and V. Della Mea, "Integrating confidence, difficulty, and language model calibration for better explainability in clinical documents coding: Applications of AI," *JMIR AI*, vol. 5, Art. no. e78764, Apr. 2026, doi: 10.2196/78764.

[12] J. Zhang, W. Yao, X. Chen, and L. Feng, "Transferable post-hoc calibration on pretrained transformers in noisy text classification," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, pp. 13940–13948, Jun. 2023, doi: 10.1609/aaai.v37i11.26632.

[13] D. Murthy *et al.*, "Categorizing E-cigarette-related tweets using BERT topic modeling," *Emerg. Trends Drugs Addict. Health*, vol. 4, Art. no. 100160, Dec. 2024, doi: 10.1016/j.etdah.2024.100160.

[14] C. C. Silva, M. Galster, and F. Gilson, "Applying short text topic models to instant messaging communication of software developers," *J. Syst. Softw.*, vol. 216, Art. no. 112111, Oct. 2024, doi: 10.1016/j.jss.2024.112111.

[15] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," arXiv:2203.05794, Mar. 2022.

[16] M. Grootendorst, "BERTopic," *BERTopic Documentation*. Accessed: Jun. 4, 2026. [Online]. Available: BERTopic Documentation

[17] M. D. Angelo, R. W. Harwenda, I. Budi, A. B. Santoso, and P. K. Putra, "Sentiment analysis and topic modeling of public opinion on Indonesia new capital city development policies," *Eduvest J. Univ. Stud.*, vol. 5, no. 5, pp. 5951–5969, May 2025, doi: 10.59188/eduvest.v5i5.51234.

[18] C. Li and X. Hu, "Medical artificial intelligence in scholarly and public perspective: BERTopic-based analysis of topic-sentiment collaborative mining," *Data Sci. Informetr.*, vol. 5, no. 1, pp. 33–42, Mar. 2025, doi: 10.1016/j.dsim.2025.05.001.

[19] Selenium, "WebDriver," *Selenium Documentation*. Accessed: Jun. 4, 2026. [Online]. Available: Selenium WebDriver Documentation

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[20] A. S. Franzke, A. Bechmann, M. Zimmer, and C. M. Ess, *Internet Research: Ethical Guidelines 3.0*. Association of Internet Researchers, 2020.

[21] J. Taylor and C. Pagliari, "Mining social media data: How are research sponsors and researchers addressing the ethical challenges?" *Res. Ethics*, vol. 14, no. 2, pp. 1–39, Apr. 2018, doi: 10.1177/1747016117738559.

[22] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[23] Scikit-learn Developers, "train_test_split," *scikit-learn 1.9.0 Documentation*. Accessed: Jun. 4, 2026. [Online]. Available: scikit-learn train_test_split documentation

[24] S. Desai and G. Durrett, "Calibration of pre-trained transformers," in *Proc. 2020 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Online, Nov. 2020, pp. 295–302, doi: 10.18653/v1/2020.emnlp-main.21.

[25] M. Minderer *et al.*, "Revisiting the calibration of modern neural networks," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 15682–15694, 2021.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, Sydney, NSW, Australia: PMLR, 2017, pp. 1321–1330.

[27] M. Chidambaram and R. Ge, "Reassessing how to compare and improve the calibration of machine learning models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[28] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Source Softw.*, vol. 3, no. 29, Art. no. 861, Sep. 2018, doi: 10.21105/joss.00861.

[29] H. Rahimi, D. Mimno, J. L. Hoover, H. Naacke, C. Constantin, and B. Amann, "Contextualized topic coherence metrics," in *Findings Assoc. Comput. Linguistics: EACL 2024*, St. Julian's, Malta, Mar. 2024, pp. 1760–1773, doi: 10.18653/v1/2024.findings-eacl.123.