Proceedings of
The Sunan Ampel International Conference of Political and Social Sciences (SAICoPSS),
The Faculty of Social and Political Sciences, UIN Sunan Ampel Surabaya
Volume 4, 2026, 1-19
DOI: https://doi.org/10.15642/saicopss.2026.4.1.1-19

ALGORITHMIC BLIND SPOTS IN DIGITAL CRISIS SENSING: HUMAN VALIDATION OF SOCIAL MEDIA SENTIMENT ON INDONESIA'S FREE NUTRITIOUS MEALS PROGRAM

Bayu Aulia Priyantomo
[1]Universitas Airlangga, Surabaya, Indosnesia
bayu.aulia.priyantomo.fisip-2026@student.unair.ac.id

ABSTRACT

Digital social media monitoring offers governments a rapid means of detecting public concerns, but automated sentiment classification may overlook policy-crisis signals expressed in descriptive or context-dependent language. This study examines algorithmic blind spots in sentiment analysis of Indonesia's Free Nutritious Meals (MBG) Program through a comparative content-analysis design combining automated classification with manual validation. The study used 100 high-engagement TikTok and Instagram video posts collected between January and June 2026. After duplicate removal, relevance screening, and ambiguity checks, 66 posts remained for analysis. Automated classification labelled 51 posts as neutral, 10 as positive, and five as negative, whereas manual validation identified 37 negative, 20 positive, and nine neutral posts. Exact agreement occurred in only 19 posts (28.79%). Notably, 29 of the 37 manually validated negative posts were automatically classified as neutral. The findings reveal a pattern of algorithmic neutralization, in which substantively critical events may lose their policy significance when expressed in factual or journalistic language. The study argues that effective digital crisis sensing requires a socio-technical approach that combines computational scale with structured human validation to ensure semantic relevance, contextual interpretation, risk recognition, and appropriate governmental response. This hybrid approach strengthens quality, accountability, and policy usefulness in government monitoring

Keywords: Algorithmic Blind Spots, Digital Crisis Sensing, Human Validation, Sentiment Analysis, Free Nutritious Meals Program

INTRODUCTION

Digital transformation has changed the way governments understand society. Public aspirations are no longer conveyed only through surveys, formal complaints, public hearings, or mass media coverage, but also through posts, comments, videos, and various forms of interaction on social media. For governments, these digital traces offer an opportunity to identify public problems before they develop into policy crises. This practice is often referred to as social media listening or digital crisis sensing, namely the ability to detect changes in attention, emotions, complaints, and risks through digital

data. Buzoianu S Bîră (2021) show that social media listening has become an important part of crisis communication and management, while Acikara et al. (2023) emphasize the contribution of social media analytics to the effectiveness of disaster response. Nevertheless, the availability of large amounts of data does not automatically produce accurate knowledge. Digital monitoring systems would assist governments only if the data analyzed are relevant, consistent, and classified correctly.

Sentiment analysis is widely used to simplify public conversations into positive, negative, and neutral categories. Its advantages lie in its speed, reach, and ability to display trends that are difficult to identify through manual reading. Birjali et al. (2021) and Wankhade et al. (2022) demonstrate the development of sentiment analysis methods, ranging from lexicon-based approaches and machine learning to deep learning and the use of language models. In the context of public administration, Verma (2022) shows that sentiment analysis has the potential to help governments evaluate services, understand public needs, and support data-driven decision-making. However, these benefits are often accompanied by the assumption that algorithmic outputs are direct representations of public opinion. In fact, sentiment labels are the result of technical decisions influenced by category definitions, data characteristics, linguistic context, labelling processes, and the classification models used.

The main issue is not only whether a model achieves a high level of accuracy, but also whether the resulting categories are able to capture the actual policy meaning. Hartmann et al. (2023) show that the accuracy of sentiment analysis tools varies according to the data context and the purpose for which they are used. After evaluating eleven sentiment analysis tools on seven social media datasets, He et al. (2022) even warn that these tools do not always perform consistently across different data. Recent studies in the field of public health have also found that interpretability, bias, linguistic complexity, class imbalance, and data quality remain major obstacles to the use of sentiment analysis for decision-making (Gandy et al., 2025; Villanueva-Miranda et al., 2025). Thus, analytical failure may occur when technical measures appear satisfactory, but the system is unable to recognize events that substantively contain criticism, danger, rejection, or demands for accountability.

This vulnerability is even greater on video-based social media. The meaning of a post is not always contained in its written caption. The uploader's position may be conveyed through intonation, facial expressions, images, on-screen text, music, irony, or excerpts from events. Das & Singh (2023) emphasize that multimodal sentiment analysis requires the integration of textual, audio, and visual information. If a system reads only the post caption, content concerning unfit food, suspected poisoning, distribution accidents, or demonstrations may be labelled neutral because the sentences used take the form of descriptive reports. Yet, from the perspective of crisis management, such posts carry risk signals that require government attention. Text cleaning and preprocessing are indeed important for improving classification quality

(Duong & Nguyen-Thi, 2021), but language processing cannot fully replace an assessment of the social context, events, and policy substance.

The issue of relevance arises even before the sentiment classification process is conducted. Keywords may be polysemous, change meaning, or be used in trends that are entirely unrelated to the policy object. Duplicate posts may inflate the volume of conversation, while differences in the metadata of the same content may alter the aggregation results. If these errors are not examined, monitoring dashboards may display a false level of public attention. In a government environment, the consequences are more serious than mere statistical errors. Communication resources may be directed towards the wrong issues, genuinely urgent problems may be handled too late, and decisions may be justified by invalid indicators. Zuiderwijk et al. (2021) and Madan & Ashok (2023) identify data quality, transparency, organizational capacity, accountability, and human oversight as central issues in the application of artificial intelligence in the public sector.

The need for human validation is consistent with the human-in-the-loop approach. Mosqueira-Rey et al. (2023) explain that human involvement may be incorporated into the processes of model labelling, learning, evaluation, and correction. In digital crisis sensing, humans are not merely final reviewers, but also bearers of domain knowledge who are able to distinguish linguistic sentiment from the substantive orientation of an event. A news report may be written without emotional words yet still indicate a governance failure. Conversely, a highly emotional expression may not necessarily be related to the policy being monitored. Human validation therefore needs to include duplicate checking, semantic relevance, contextual adequacy, label consistency, and issue coding. Human involvement should not be understood as a weakness of automation, but rather as a quality-control mechanism that allows algorithmic outputs to be used responsibly.

The Free Nutritious Meals Program (*Makan Bergizi Gratis*, MBG) provides an important context in which to examine this issue. As a large-scale public programmed that intersects with child nutrition, schools, public budgets, food distribution, employment, and governance, MBG has generated diverse public experiences and assessments. Digital conversations may contain support from beneficiaries, but also criticism concerning food quality, food safety, financing, distribution delays, political legitimacy, and institutional transparency. In such a situation, the government requires a system that not only counts polarity, but is also able to distinguish operational complaints, the need for information clarification, and indications of a crisis of trust. Bulut & Poth (2022) show that sentiment analysis can be used to assess the consistency of public health communication. Meanwhile, Lohiniva et al. (2022) demonstrate that social listening can help governments adapt information to the needs of diverse social groups. Nevertheless, these benefits continue to depend on data validity and the accuracy of the initial interpretation.

Studies specifically concerning the MBG Program have begun to develop, but their orientation is still dominated by the development and comparison of classification models. Triningsih et al. (2025) applied machine-learning algorithms to conversations about the MBG Program on the social media platform X. Julianto & Nurpajar (2025)) used automated labelling assisted by DeepSeek AI and VADER before conducting classification with the K-Nearest Neighbor algorithm. These studies are important because they demonstrate the feasibility of sentiment analysis for examining responses to the MBG Program. However, their focus on model accuracy and label distribution has not fully answered more fundamental questions, namely whether the data collected are genuinely relevant to the programmed, whether the same content is counted more than once, and whether linguistically neutral labels contain signals of a policy crisis. Errors in the reference data may make model performance appear convincing even though its substantive interpretation is problematic.

To map the position of this study, a limited systematic literature review was conducted by following the PRISMA 2020 transparency principles developed by Page et al. (2021). The preliminary search covered journal articles published between 2021 and September 2026 through publisher websites, PubMed, and academic search engines. The keywords used combined the terms “sentiment analysis”, “human validation”, “human-in-the-loop”, “social media listening”, “crisis sensing”, “public sector”, “public health”, and “*Makan Bergizi Gratis*”. Articles were included if they discussed at least one of four areas: the limitations of sentiment classification, human validation, crisis or policy monitoring through social media, and sentiment analysis of the MBG Program. Conference proceedings, theses, preprints, articles without verifiable metadata, and technical studies unrelated to social or policy interpretation were excluded. Based on these criteria, 19 articles were used in the narrative synthesis to establish the theoretical and empirical position of the study.

The literature synthesis reveals three research clusters that have not been adequately connected. The first cluster discusses the development of algorithms along with challenges related to language, bias, data, and evaluation. The second cluster discusses social media listening, crisis communication, and the application of artificial intelligence in government. The third cluster applies sentiment analysis algorithms to conversations about the MBG Program. Literature in the first cluster rarely translates classification errors into risks to state capacity. The second cluster tends to accept analytical results as inputs that are ready for use, whereas the third cluster more frequently tests models on platform X without conducting a layered audit of duplication, term relevance, contextual ambiguity, and the substantive orientation of video posts. Few studies have empirically demonstrated how descriptive language may cause crisis signals to be categorized by algorithms as neutral sentiment. Therefore, a gap exists between computational accuracy and policy usefulness. A system may successfully classify linguistic form but fail to recognize the crisis meaning required by the government to determine a response.

This gap is evident in the present study, which uses 100 rows of data from high-engagement video posts on TikTok and Instagram between January and June 2026. The initial examination identified 21 groups of duplicate posts, resulting in 23 repeated rows. After the duplicates were removed, the 100 rows represented only 77 unique posts. The subsequent relevance check identified eight posts that used the abbreviation "MBG" to refer to "*Mas Bahlil Ganteng*" and a song trend, rather than the Free Nutritious Meals Program. These eight posts generated 4,998,375 engagements, or 13.03% of the total engagement in the unique dataset, as well as 66,209,877 views, or 12.41% of all views. If they had not been excluded, this content could have improperly inflated the level of attention directed towards the MBG Program. In addition, three posts had captions that were too brief for their relevance and sentiment orientation to be determined from the text alone. After all screening stages had been completed, 66 relevant posts remained, consisting of 53 TikTok posts and 13 Instagram posts.

A further anomaly was identified when the automated classification results were compared with the manual validation The automated system labelled 51 posts as neutral, ten as positive, and only five as negative. By contrast, the manual validation identified 37 posts with a negative orientation, 20 as positive, and nine as neutral. Exact agreement occurred in only 19 of the 66 posts, or 28.79%. More importantly, 29 of the 37 substantively critical posts were automatically classified as neutral, while three other critical posts were categorized as positive. The posts that were missed addressed issues of governance, political legitimacy, food safety, budgets, implementation, distribution, and operational risks. This difference cannot be treated as an ordinary technical variation because misclassification may cause the government to fail to recognize early-warning signals.

These initial findings show that automated sentiment analysis tends to read linguistic polarity, whereas manual validation can assess the substantive orientation of an event. Captions concerning student demonstrations, demands for audits, alleged corruption, unfit food, accidents involving distribution vehicles, or suspected poisoning may be written using informative and descriptive sentences. Linguistically, these sentences may not contain strong emotional words. Substantively, however, the events reported clearly carry risks to safety, programmed integrity, government legitimacy, and public trust. In this study, the tendency of algorithms to convert reports of critical events into neutral labels is referred to as algorithmic neutralization. This concept explains one form of algorithmic blind spot in which a calm writing style is equated with the absence of a policy problem.

Based on this gap, the novelty of this study lies in four aspects. First, the study distinguishes textual sentiment from the substantive orientation of an issue. A sentence may be neutral in its writing style but indicate a negative event as a policy signal. Second, the study positions data validation, including deduplication, relevance screening, ambiguity checking, label consistency, and issue coding, as part of state capacity in conducting digital crisis sensing. Third, the study provides empirical

evidence of algorithmic neutralization, namely the tendency of systems to convert reports of critical events into neutral labels because criticism is conveyed through descriptive language or does not explicitly contain emotional vocabulary. Fourth, the study proposes a hybrid monitoring principle, in which automation is used to obtain speed and scale, while human assessment is used to ensure meaning, relevance, level of risk, and policy context.

This study is not intended to generalize the sentiment of the entire Indonesian population or to assess the overall success of the MBG Program. The data used constitutes a purposive sample of high-engagement posts, and the results are therefore more appropriately used to understand the characteristics of content that attract considerable attention. The initial audit is also still limited to post captions and does not include the complete audio, visual elements, or user comments. Therefore, this study is directed at evaluating the consistency and limitations of automated sentiment classification in detecting policy-crisis signals and explaining the forms of error that emerge when machine results are compared with human interpretation. More specifically, this study asks to what extent automated classification agrees with manual validation, what types of critical content are most frequently classified as neutral, and how human validation can be integrated into government social media monitoring systems. Within these limitations, this conference paper makes a clear methodological contribution to the issue of algorithmic blind spots and human validation without taking on the entire contribution of the main journal article concerning the development of DPOG, longitudinal mapping, and the governance of digital public opinion on a broader scale.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

a. Social Media Listening and Digital Crisis Sensing

Social media provides governments with continuously updated information about public concerns, policy experiences, and emerging risks. Social media listening involves collecting, interpreting, and responding to digital conversations, while digital crisis sensing focuses on identifying early indications of service failure, misinformation, safety risks, and declining institutional trust. Studies show that social media listening can strengthen crisis communication and support rapid responses to emerging conditions (Buzoianu & Bîră, 2021; Acikara et al., 2023). However, monitoring becomes institutionally meaningful only when public expressions are interpreted and connected to government action (Macnamara, 2019).

b. Automated Sentiment Analysis and Algorithmic Blind Spots

Sentiment analysis classifies social media content as positive, negative, or neutral and enables large volumes of data to be processed rapidly (Birjali et al., 2021; Wankhade et al., 2022). Nevertheless, its accuracy varies according to datasets, language, context,

labelling procedures, and classification models (He et al., 2022; Hartmann et al., 2023). This limitation is particularly significant for video-based content, whose meaning may depend on visual and audio elements (Das & Singh, 2023). This study defines the failure to recognise substantively critical events as an algorithmic blind spot, while the classification of critical reports as neutral is termed algorithmic neutralization.

c. Human Validation and the Human-in-the-Loop Approach

The human-in-the-loop approach incorporates human judgement into data labelling, evaluation, model correction, and decision-making, particularly when automated systems encounter ambiguity or insufficient context (Mosqueira-Rey et al., 2023). In digital crisis sensing, human validation includes detecting duplicates, verifying semantic relevance, assessing contextual adequacy, checking label consistency, and coding substantive policy issues. Human involvement therefore functions as a quality-control mechanism rather than a weakness of automation. It improves the reliability and policy usefulness of automated analysis while strengthening accountability, transparency, and responsible artificial intelligence in government (Zuiderwijk et al., 2021; Madan & Ashok, 2023).

d. Socio-Technical Systems Theory

This study applies Socio-Technical Systems Theory, which argues that organizational performance depends on the interaction between technical and social subsystems (Trist & Bamforth, 1951). Automated sentiment classification represents the technical subsystem, while human validation represents the social subsystem. Automation provides speed and scalability, whereas humans provide contextual and policy understanding. This perspective is complemented by Easton's (1965) political systems theory, in which social media conversations function as policy inputs and feedback. Digital crisis sensing therefore connects digital data, algorithmic classification, human interpretation, and governmental response.

METHOD

This study employed a comparative content-analysis design that integrated descriptive quantitative analysis with qualitative manual validation. The study aimed to identify discrepancies between automated sentiment classification and human interpretation of the substantive policy orientation contained in social media posts related to Indonesia's Free Nutritious Meals (*Makan Bergizi Gratis*, MBG) Program. Manual validation was used as a quality-control procedure to examine whether automated sentiment labels adequately captured the policy meaning and potential crisis signals contained in the posts.

The data were obtained from NoLimit Indonesia and initially consisted of 100 high-engagement video posts purposively selected from TikTok and Instagram between January and June 2026. The initial dataset comprised 78 TikTok posts and 22 Instagram posts. The use of purposive sampling was intended to focus the analysis on highly

visible content with substantial public engagement rather than to estimate the overall distribution of sentiment across all social media conversations concerning the MBG Program.

A sequential data-validation process was conducted before the comparative analysis. First, duplicate records were identified by examining post links, posting accounts, and captions. This process identified 23 duplicate rows, leaving 77 unique posts. Second, the remaining posts were screened for relevance to the government's MBG Program. Eight posts were excluded because the abbreviation "MBG" referred to unrelated content, including "*Mas Bahlil Ganteng*" and a song-related trend. Third, three additional posts were excluded because their captions were too brief or ambiguous to determine their relevance and sentiment orientation reliably. Following these stages, the final analytical sample consisted of 66 relevant posts, comprising 53 TikTok posts and 13 Instagram posts. This sequence is consistent with the screening process reported in the Results section.

Each retained post was then manually validated by examining its written caption and available contextual information. The validation focused on the substantive policy orientation of the post rather than solely on its linguistic tone. Posts were classified into three sentiment categories: positive, negative, and neutral. Negative orientation included content indicating criticism, rejection, implementation failure, safety concerns, governance problems, or other policy risks, even when expressed in descriptive or journalistic language. Positive orientation referred to content expressing support, perceived benefits, improvements, or favorable experiences, while neutral orientation referred to posts primarily conveying information or clarification without a clear positive or negative policy orientation.

After sentiment validation, the substantive issues contained in the posts were mapped inductively into thematic categories. These categories included political legitimacy and protest; governance, corruption, and transparency; menu quality and food safety; budgeting and financing; implementation, access, and distribution; SPPG operations and safety; public benefits; employment; programmed communication; food-quality improvement; strengthened oversight; and information clarification. This issue-mapping process was used to identify the types of policy signals associated with each sentiment orientation.

The automated sentiment classifications were subsequently compared with the manually validated classifications using frequency distributions, percentages, cross-tabulations, and exact agreement rates. Exact agreement represented the proportion of posts for which the automated system and manual validation assigned the same sentiment category. The comparison was used descriptively to identify patterns of correspondence and misclassification rather than to measure reliability between independent human coders. Engagement was calculated as the combined number of likes, shares, and comments associated with each post. The temporal distribution of

validated posts from January to June 2026 was also examined descriptively to identify changes in sentiment orientation across the observation period.

The analysis was subject to several limitations. First, because the sample consisted of purposively selected high-engagement posts, the findings cannot be generalized to the entire digital conversation surrounding the MBG Program. Second, the final analytical sample included only TikTok and Instagram posts. Third, the validation relied primarily on written captions and available contextual information and did not systematically incorporate the complete audio, visual, on-screen text, or user comments associated with each video. Finally, manual validation was conducted as a substantive quality-control procedure rather than through multiple independent coders; therefore, the study did not assess intercoder reliability. These limitations should be considered when interpreting the findings.

RESULTS AND DISCUSSIONS

Results

a. Data Validation and Analytical Sample

The initial dataset comprised 100 high-engagement video posts: 78 from TikTok and 22 from Instagram. Manual validation identified 21 duplicate groups containing 23 repeated rows. After these rows were removed, 77 unique posts remained. Seven duplicate groups carried different sentiment labels despite having identical content and links, while one group contained inconsistent engagement values.

The relevance assessment identified eight posts in which the abbreviation ‘MBG’ referred to ‘*Mas Bahlil Ganteng’* or a song trend rather than the Free Nutritious Meals Program. These irrelevant posts generated 4,998,375 engagement interactions and 66,209,877 views. Three additional posts were excluded because their captions were too brief or ambiguous to establish relevance and sentiment orientation. The final sample comprised 66 posts: 53 from TikTok (80.30%) and 13 from Instagram (19.70%). Figure 1 presents the screening process.

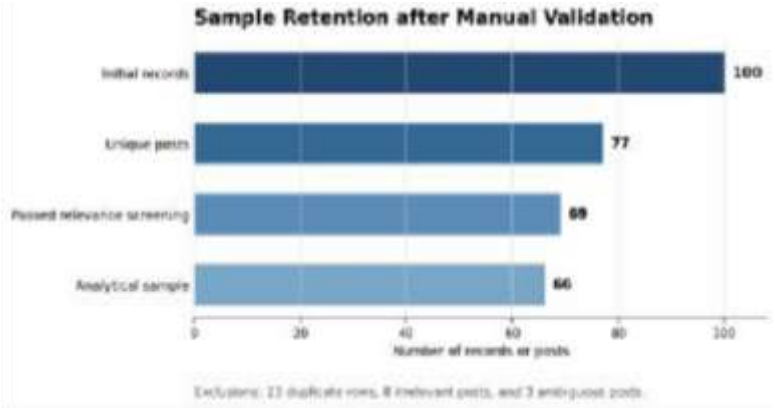


Figure 1. Sample retention after manual validation.

b. Engagement and Conversation Visibility

The 66 relevant posts generated 32,459,385 engagement interactions, comprising 26,236,150 likes (80.83%), 5,427,536 shares (16.72%), and 795,699 comments (2.45%). TikTok accounted for 26,731,911 interactions (82.35%), whereas Instagram accounted for 5,727,474 (17.65%). The posts accumulated 463,933,052 views; however, most originated from TikTok because many Instagram posts had either zero or unavailable view counts.

Manually Validated sentiment	Posts	Share	Engagement	Engagement share	Views
Negative	37	56.06%	19,304,425	59.47%	268,743,503
Positive	20	30.30%	9,334,290	28.76%	143,417,966
Neutral	9	13.64%	3,820,670	11.77%	51,771,583
Total	66	100%	32,459,385	100%	463,933,052

Table 1. Distribution, engagement, and views by manually validated sentiment

The highest-engagement post was published by BBC News Indonesia and concerned a student demonstration criticizing several government policies, including the MBG Program. It generated 1,883,300 engagement interactions and 19,700,000 views. The second-highest post reported a school principal's protest meals deemed unfit for consumption, generating 1,510,099 interactions and 25,326,502 views. Among the positive posts, a mother's expression of gratitude after her children received Ramadan MBG packages generated 1,123,793 interactions. A post about rejecting an MBG delivery because of menu quality generated 1,119,000 interactions and recorded the highest number of views, at 32,300,000.

c. Comparison of Automated Classification and Manual Validation

To assess the consistency of the automated sentiment classification, manual validation was conducted on a sample of 66 posts. The comparison focused on whether the sentiment labels generated automatically were consistent with the interpretation obtained through manual review. The automated system classified 51 posts as neutral, 10 as positive, and five as negative. In contrast, manual validation identified 37 posts as negative, 20 as positive, and nine as neutral.

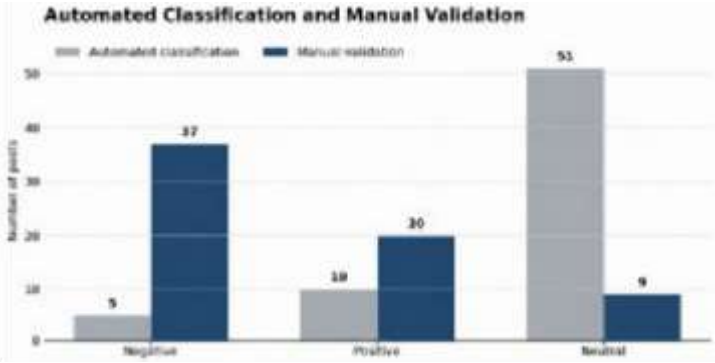


Figure 2. Comparison between automated classification and manual validation

Overall, the automated classification and manual validation assigned the same sentiment label to 19 of the 66 posts, representing an exact agreement of 28.79%. The remaining 47 posts (71.21%) received different sentiment classifications. The distribution indicates a notable difference between the automated results and the manual assessment, particularly in the classification of negative and positive content.

Manual Validation	Auto negative	Auto positive	Auto neutral	Total
Negative	5	3	29	37
Positive	0	6	14	20
Neutral	0	1	8	9
Total	5	10	51	66

Table 2. Cross-tabulation of automated classification and manual validation

The cross-tabulation shows that the largest discrepancy occurred among posts manually identified as negative. Of the 37 posts classified as negative through manual validation, only five (13.51%) received the same negative label from the automated system. Three posts (8.11%) were automatically classified as positive, while 29 posts (78.38%) were classified as neutral. This pattern indicates that a substantial proportion of posts containing negative evaluations according to manual review were categorized as neutral by the automated system.

A similar pattern was observed for positive sentiment. Of the 20 posts identified as positive through manual validation, six (30.00%) were also classified as positive by the automated system, while 14 posts (70.00%) were classified as neutral. None of the manually validated positive posts were automatically classified as negative.

The highest level of consistency was found in the neutral category. Of the nine posts identified as neutral through manual validation, eight (88.89%) were also classified as neutral by the automated system, while one post (11.11%) was classified as positive. Thus, compared with the positive and negative categories, the automated system showed greater consistency in identifying neutral content.

Taken together, the manual validation indicates that the automated system had a strong tendency to assign neutral labels to posts that were interpreted manually as containing evaluative sentiment. Of the 51 posts automatically classified as neutral, 29 were manually identified as negative, 14 as positive, and only eight remained neutral after manual validation. Therefore, the predominance of neutral sentiment generated by the automated classification should be interpreted with caution.

The manual validation in this study was employed as a descriptive quality check of the automated sentiment classification rather than as an assessment of reliability between human coders. Accordingly, the comparison was used to identify patterns of correspondence and discrepancy between automated classification and manual interpretation. These validation findings were subsequently taken into consideration when interpreting the aggregate sentiment results.

d. Issue Mapping by Sentiment Orientation

Within the negative category, politics, protest, legitimacy and governance, corruption, and transparency each comprised ten posts. Together, these categories accounted for 20 of the 37 negative posts. Menu quality and food safety comprised only five posts but generated 80,815,134 views, the highest total among the negative issue categories.

Orientation	Issue category	Posts	Engagement
Negative	Politics, protest, and legitimacy	10	5,420,178
Negative	Governance, corruption, and transparency	10	4,960,749
Negative	Menu quality and food safety	5	3,442,472
Negative	Budget, economy, and financing	5	2,337,691
Negative	Implementation, access, and distribution	4	1,754,934
Negative	SPPG operations and safety	3	1,388,401
Positive	Public benefits and support	10	5,473,607
Positive	Employment and SPPG personnel	3	1,135,463
Positive	Implementation and expanded access	3	1,119,888
Positive	Programme legitimacy and communication	2	738,570
Positive	Food-quality improvement	1	598,400
Positive	Strengthened oversight	1	268,362
Neutral	Budget and alternative financing	4	1,709,151
Neutral	SPPG operations	2	1,132,268
Neutral	Governance and partner information	1	419,900
Neutral	Information clarification	1	292,320
Neutral	Implementation challenges	1	267,031

Table 3. Issue categories identified through manual validation

Within the positive category, public benefits and support accounted for ten of the 20 posts and generated 5,473,607 engagement interactions. Other issues included employment and SPPG personnel, implementation and expanded access, programme communication, improvements in food quality, and strengthened oversight. Meanwhile, four of the nine neutral posts addressed the budget, temporary programmed suspension, and alternative financing arrangements.

e. Longitudinal Patterns

In January, five posts were manually validated as negative and six as positive. February comprised five negative, three positive, and four neutral posts, while March comprised two negative, four positive, and three neutral posts. April and May contained only four and two relevant posts, respectively. The greatest concentration occurred in June: of the 28 relevant posts, 22 were negative, four positive, and two neutral. Posts published in June generated 13,864,679 engagement interactions, representing 42.71% of all engagement in the sample.

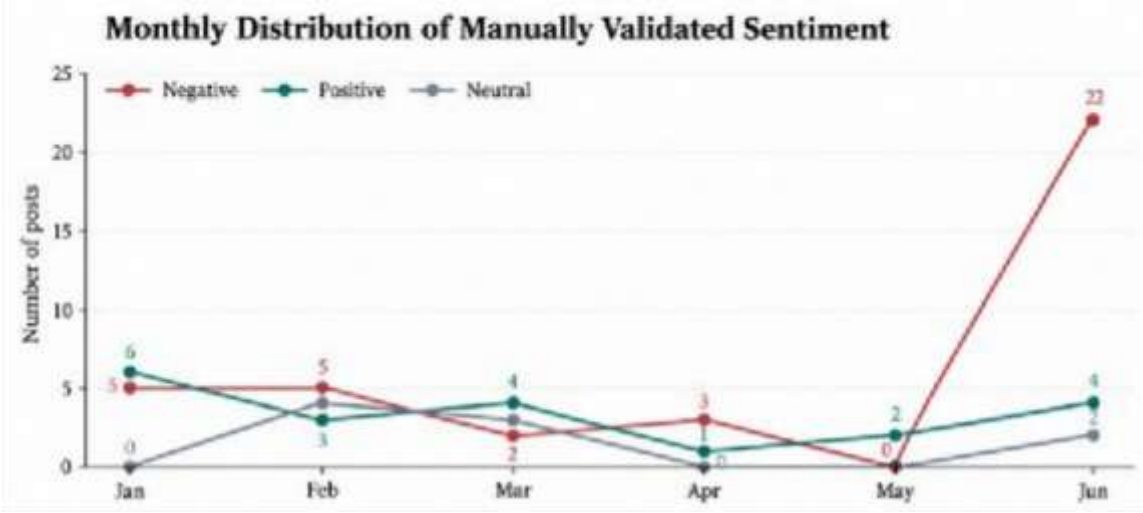


Figure 3. Monthly distribution of manually validate sentiment, January–June 2026.

Discussion

a. Algorithmic Neutralization as a Blind Spot in Crisis Sensing

The principal finding is not merely the low exact-agreement rate but a systematic pattern of discrepancy. The automated system assigned 77.27% of the posts to the neutral category, whereas manual validation showed that 56.06% contained criticism, rejection, risk, or implementation failure. By classifying 29 of the 37 negative posts as neutral, the system not only misread emotional tone but also reduced the perceived urgency of events that warranted government attention. This pattern is termed algorithmic neutralization: substantively critical events lose their policy significance because they are reported in descriptive or journalistic language.

This finding is consistent with van Atteveldt et al. (2021), who show that the validity of sentiment analysis depends heavily on the method employed and must be tested against human annotation. Gilardi et al. (2023) demonstrate that language models can perform very well on certain annotation tasks; nevertheless, such performance does not eliminate the need for human benchmarks and context-specific evaluation. In the case of MBG, the problem lies in the distinction between linguistic polarity and policy meaning. A sentence reporting unfit meals, suspected food poisoning, a distribution accident, a demonstration, or alleged corruption may sound factual, yet the event itself remains a negative signal concerning programmed delivery.

The limitations of text-based analysis are especially important because every unit of analysis was a video post. Gandhi et al. (2023) and Zhu et al. (2023) emphasize that multimodal sentiment emerges from the interaction of text, sound, facial expression, and visual content. By examining only written captions, a system may miss intonation, on-screen text, the condition of the food, recipient reactions, irony, and other visual cues that shape meaning. A neutral label assigned to a video caption therefore cannot be treated as evidence that the post contains no policy risk.

b. Human Validation as a Socio-Technical Capacity

The validation results show that blind spots emerge at the earliest stages of data processing. Duplicates artificially inflate volume, unrelated uses of the abbreviation MBG contaminate the topic, and extremely brief captions force the system to decide without sufficient context. The eight irrelevant posts alone generated nearly five million engagement interactions. Had they been retained, the government could have overestimated the level of public attention directed towards the programmed. The quality of crisis sensing therefore depends not only on the classification model but also on data governance before classification begins.

From a Socio-Technical Systems perspective, algorithms and humans perform distinct yet interdependent functions. Automation provides speed, consistency, and large-scale processing, whereas humans assess semantic relevance, ambiguity, institutional context, and policy consequences. Mikalef et al. (2022) show that AI capabilities in government agencies arise not from technology alone but from organizational conditions, regulatory arrangements, and institutional support. Neumann et al. (2024) likewise find that the importance of technological and organizational factors varies across stages of AI adoption. Successful digital monitoring therefore depends on an organizational design that specifies who reviews the outputs, how errors are corrected, and where validated signals are directed.

Mergel et al. (2023) identify implementation, rather than adoption alone, as the central challenge of public-sector AI. This observation is directly relevant to the MBG case: a dashboard may function technically yet still fail administratively if its classifications cannot guide action. Maragno et al. (2023) stress that AI implementation is shaped by configurations of enabling factors, use opportunities, and organizational constraints. Giest & Klievink (2024), meanwhile, show that AI changes bureaucratic roles differently across organizational contexts. Human validation should consequently be institutionalized as a routine procedure— with a validation protocol, risk thresholds, escalation mechanisms, and an audit trail—rather than applied as an incidental correction after a crisis has intensified.

c. Social Media as Governmental Input and Feedback

According to Easton's political systems theory, societal demands and support enter the political system as inputs are processed by institutions and produce policies that are subsequently evaluated by the public. In the digital sphere, posts about MBG may function as inputs and feedback concerning food quality, distribution, budgets, programmed benefits, and governance integrity. However, such inputs acquire institutional value only when they can be recognized, interpreted, and directed to the competent unit. When substantive criticism is neutralized by an algorithm, citizens' voices remain in the database but lose their capacity to influence decision-making.

van Noordt & Misuraca (2022) show that public-sector AI in the European Union offers potential benefits, but its impact depends on how the technology is embedded in governmental processes. Haug et al. (2024) similarly emphasize that digitally induced change in the public sector is organizational change, not merely technology installation. In this context, listening to digital publics requires a connection among signals, validation, authority, response, and reassessment. A government cannot be said to listen merely because it counts positive, negative, and neutral posts without demonstrating

how the detected issues shape institutional action.

d. Implications for Government Monitoring and Response

The issue map indicates at least three response pathways. First, information problems such as rumors, financing alternatives, or food safety clarification require rapid verification and evidence-based communication. Second, operational problems such as menu quality, delays, cold meals, accidents, and unequal distribution require field inspection, corrective action, and resolution tracking. Third, governance problems such as alleged corruption, conflicts of affiliation, procurement, and SPPG transparency require audits, document disclosure, and institutional accountability. These pathways cannot be determined by the word polarity alone.

Signal type	Examples of MBG issues	Primary response	Follow-up unit
Information	Rumors, clarification, financing	Verification and evidence-based explanation	Public communication
Operational	Menu quality, distribution, safety	Inspection, correction, and resolution tracking	Programmed implementer/SPPG
Governance	Corruption, procurement, transparency	Audit, disclosure, and accountability	Internal/external oversight

Table 4. Recommended response pathways by signal type.

The longitudinal pattern reinforces the need for issue-based escalation. Conversations from January to March combined implementation criticism with accounts of programmed benefits. In June, attention shifted sharply towards corruption, audits, transparency, demonstrations, and legitimacy. A shift from service complaints to a crisis of trust demands a different response. Communication alone is insufficient when a problem has entered the realm of institutional integrity; conversely, a remediable operational issue need not always be treated as a political crisis. Issue-based coding helps the government select a proportionate response

e. Study Limitations

The findings should be interpreted considering four limitations. First, the sample was selected purposively based on engagement and therefore does not represent the sentiment distribution across all 4,463,750 monitored conversations. Second, the final analytical sample included only TikTok and Instagram posts, which limit comparison across the broader set of monitored social media platforms. Third, the validation relied primarily on written captions and available contextual information and did not systematically incorporate complete audio, visual, on-screen text, or user comments. Fourth, the manual validation was conducted as a substantive quality-control procedure rather than through multiple independent validators. Consequently, the study focused on identifying discrepancies between automated classification and manual interpretation rather than assessing intercoder reliability.

Future research should employ larger and more representative samples across months and platforms, incorporate multiple independent validators to strengthen the robustness of manual assessment, and integrate textual, audio, visual, and on-screen information. Further studies should also compare conventional sentiment polarity

classification with policy-event or risk-oriented classification to determine whether critical policy signals can be detected more accurately. Nevertheless, this study demonstrates that, without relevance screening and contextual human interpretation, automated systems may understate crisis signals expressed through formal, factual, or descriptive language.

CONCLUSION

This study demonstrates that automated sentiment classification alone may be insufficient for reliable digital crisis sensing in the context of Indonesia's Free Nutritious Meals (*Makan Bergizi Gratis, MBG*) Program. The findings show that analytical limitations emerge not only at the classification stage but also during data preparation. From the initial 100 high-engagement posts, duplicate removal, relevance screening, and ambiguity assessment reduced the dataset to 66 relevant posts. This reduction highlights the importance of data validation before automated outputs are interpreted as indicators of public attention or policy sentiment.

The comparison between automated classification and manual validation revealed a substantial discrepancy between linguistic sentiment and substantive policy meaning. The automated system classified 51 of the 66 posts as neutral, whereas manual validation identified 37 posts as negative, 20 as positive, and nine as neutral. Exact agreement was found in only 19 posts, or 28.79%. Most importantly, 29 of the 37 posts manually identified as negative were classified automatically as neutral. This pattern indicates that critical policy signals can be obscured when automated systems rely primarily on linguistic polarity without sufficiently accounting for the substantive meaning of the events being reported.

The study conceptualizes this pattern as algorithmic neutralization, a form of algorithmic blind spot in which substantively critical events are assigned neutral labels because they are expressed through factual, descriptive, or journalistic language. Posts concerning food safety, programmed implementation, governance, budgeting, distribution, political legitimacy, and accountability may therefore appear linguistically neutral while still representing significant policy risks. The findings suggest that a neutral automated label should not automatically be interpreted as the absence of criticism, risk, or crisis relevance.

These findings reinforce the importance of a socio-technical approach to government social media monitoring. Automation remains valuable for processing large volumes of digital data rapidly and consistently, but human validation is necessary to assess semantic relevance, contextual ambiguity, substantive policy orientation, and potential risk. Human involvement should therefore be incorporated as a structured quality-control mechanism within digital monitoring processes rather than used only as an occasional correction after automated classification.

For government practice, effective digital crisis sensing should integrate automated monitoring with systematic procedures for deduplication, relevance screening, manual validation, issue identification, and escalation of critical signals. Such an approach can reduce the risk that important public concerns remain visible in the data yet lose their policy significance through inappropriate classification. In this sense, the usefulness of sentiment analysis should not be judged solely by its ability to assign positive, negative,

or neutral labels, but by its capacity to support the recognition and interpretation of issues that require institutional attention.

These conclusions should nevertheless be interpreted within the study's methodological limitations. The purposive sample consisted only of high-engagement TikTok and Instagram posts and therefore does not represent the entire digital conversation surrounding the MBG Program. The analysis also relied primarily on written captions and available contextual information rather than a systematic examination of complete audio, visual, on-screen textual, and user-comment elements. Furthermore, manual validation was conducted as a substantive quality-control procedure rather than through multiple independent coders. Future research should therefore employ larger and more representative samples across platforms and time periods, incorporate multiple independent validators, and integrate textual, visual, and audio information to further examine the distinction between linguistic sentiment and substantive policy-event orientation. These limitations are consistent with those identified in the manuscript.

REFERENCES

Acikara, T., Xia, B., Yigitcanlar, T., & Hon, C. (2023). Contribution of Social Media Analytics to Disaster Response Effectiveness: A Systematic Review of Literature. *Sustainability*, *15*(11), 8860. https://doi.org/10.3390/su15118860

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/10.1016/j.knosys.2021.107134

Bulut, O., & Poth, C. N. (2022). Rapid assessment of communication consistency: sentiment analysis of public health briefings during the COVID-19 pandemic. *AIMS Public Health*, *9*(2), 293–306. https://doi.org/10.3934/publichealth.2022020

Buzoianu, C., & Bîră, M. (2021). Using Social Media Listening in Crisis Communication and Management: New Methods and Practices for Looking into Crises. *Sustainability*, *13*(23), 13015. https://doi.org/10.3390/su132313015

Das, R., & Singh, T. D. (2023). Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges. *ACM Computing Surveys*, *55*(13s), 1–38. https://doi.org/10.1145/3586075

Duong, H.-T., & Nguyen-Thi, T.-A. (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), 1. https://doi.org/10.1186/s40649-020-00080-x

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, *91*, 424–444. https://doi.org/10.1016/j.inffus.2022.09.025

Gandy, L. M., Ivanitskaya, L. V, Bacon, L. L., & Bizri-Baryak, R. (2025). Public Health Discussions on Social Media: Evaluating Automated Sentiment Analysis Methods. *JMIR Formative Research*, *9*, e57395. https://doi.org/10.2196/57395

Giest, S. N., & Klievink, B. (2024). More than a digital system: how AI is changing the role of bureaucrats in different organizational contexts. *Public Management Review*, *26*(2), 379–398. https://doi.org/10.1080/14719037.2022.2095001

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30). https://doi.org/10.1073/pnas.2305016120

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Haug, N., Dan, S., & Mergel, I. (2024). Digitally induced change in the public sector: a systematic review and research agenda. *Public Management Review*, *26*(7), 1963–1987. https://doi.org/10.1080/14719037.2023.2234917

He, L., Yin, T., & Zheng, K. (2022). They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. *Journal of Biomedical Informatics*, *132*, 104142. https://doi.org/10.1016/j.jbi.2022.104142

Julianto, I. T., & Nurpajar, D. S. (2025). User Sentiment Analysis X Towards Makan Bergizi Gratis Program Using Automatic Labeling Technique with Deepseek AI. *Journal of Intelligent Systems Technology and Informatics*, *1*(2), 37–44. https://doi.org/10.64878/jistics.v1i2.43

Lohiniva, A.-L., Sibenberg, K., Austero, S., & Skogberg, N. (2022). Social Listening to Enhance Access to Appropriate Pandemic Information Among Culturally Diverse Populations: Case Study from Finland. *JMIR Infodemiology*, *2*(2), e38343. https://doi.org/10.2196/38343

Madan, R., & Ashok, M. (2023). AI adoption and diffusion in public administration: A systematic literature review and future research agenda. *Government Information Quarterly*, *40*(1), 101774. https://doi.org/10.1016/j.giq.2022.101774

Maragno, G., Tangi, L., Gastaldi, L., & Benedetti, M. (2023). Exploring the factors, affordances and constraints outline the implementation of Artificial Intelligence in public sector organizations. *International Journal of Information Management*, *73*, 102686. https://doi.org/10.1016/j.ijinfomgt.2023.102686

Mergel, I., Dickinson, H., Stenvall, J., & Gasco, M. (2023). Implementing AI in the public sector. *Public Management Review*, 1–14. https://doi.org/10.1080/14719037.2023.2231950

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., Gupta, M., & Niehaves, B. (2022). Enabling AI capabilities in government agencies: A study of determinants for European municipalities. *Government Information Quarterly*, *39*(4), 101596. https://doi.org/10.1016/j.giq.2021.101596

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, *56*(4), 3005–3054. https://doi.org/10.1007/s10462-022-10246-w

Neumann, O., Guirguis, K., & Steiner, R. (2024). Exploring artificial intelligence adoption in public organizations: a comparative case study. *Public Management Review*, *26*(1), 114–141. https://doi.org/10.1080/14719037.2022.2048685

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D.,
Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Triningsih, E., Afdal, M., Permana, I., & Rozanda, N. E. (2025). Analisis Sentimen Terhadap Program Makan Bergizi Gratis Menggunakan Algoritma Machine Learning Pada Sosial Media X. *Building of Informatics, Technology and Science (BITS)*, *6*(4), 2240–2250. https://doi.org/10.47065/bits.v6i4.6534

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3), 101714. https://doi.org/10.1016/j.giq.2022.101714

Verma, S. (2022). Sentiment analysis of public services for smart society: Literature review and future research directions. *Government Information Quarterly*, *39*(3), 101708. https://doi.org/10.1016/j.giq.2022.101708

Villanueva-Miranda, I., Xie, Y., & Xiao, G. (2025). Sentiment analysis in public health: a systematic review of the current state, challenges, and future directions. *Frontiers in Public Health*, *13*. https://doi.org/10.3389/fpubh.2025.1609749

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, *95*, 306–325. https://doi.org/10.1016/j.inffus.2023.02.028

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577