

# EFFECTIVENESS OF LLM AND CONVENTIONAL SENTIMENT ANALYSIS MODELS IN UNDERSTANDING DIGITAL PUBLIC OPINION

ASSEGAF INSANI<sup>1</sup>, TANTY OKTAVIA<sup>1</sup>

<sup>1</sup>BGP, Master of Information System Management, Bina Nusantara University, Jakarta, Indonesia

E-mail: <sup>1</sup>assegaf.insani@binus.ac.id, <sup>1</sup>toktavia@binus.edu

## ABSTRACT

Sentiment analysis has become an important approach for understanding digital public opinion expressed through social media, particularly in response to government policies. In Indonesia, discussions related to issues such as the Indonesian National Armed Forces Bill (RUU TNI), the Criminal Code (KUHP), the House of Representatives (DPR), and other public policy topics are widely discussed on X (Twitter), including through hashtags such as #IndonesiaGelap that reflect socio-political criticism and public dissatisfaction. Rather than focusing on a single issue, this study analyzes sentiment patterns across multiple government policy topics. This research compares the effectiveness of Large Language Models (LLMs), namely ChatGPT and LLaMA, with conventional sentiment analysis models, including Support Vector Machine (SVM) and IndoBERT, in classifying Indonesian digital public opinion. The study adopts the CRISP-DM framework covering business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Quantitative evaluation was conducted using accuracy, precision, recall, and Macro F1-score metrics, while prompt sensitivity and stability analysis were performed on LLaMA using zero-shot and few-shot prompting approaches. The results show that SVM achieved the best performance with an accuracy of 0.8658 and Macro F1-score of 0.7651, followed by IndoBERT with an accuracy of 0.8446 and Macro F1-score of 0.7440. In contrast, LLaMA showed lower and less stable performance across prompting configurations. Qualitative analysis further indicates that ChatGPT demonstrates stronger contextual reasoning and better interpretation of ambiguous and implicitly critical texts. These findings suggest that supervised models remain more reliable for large-scale quantitative sentiment classification, while ChatGPT provides complementary value as an interpretative and validation tool for complex digital public opinion analysis.

**Keywords:** *SVM, IndoBERT, LLM, ChatGPT, LLaMA*

## 1. INTRODUCTION

The rapid growth of digital public opinion expressed through social media platforms has transformed the way governments, organizations, and researchers understand societal responses to public policies. In Indonesia, public discourse surrounding government regulations such as the Indonesian National Armed Forces Bill (RUU TNI), the Criminal Code (KUHP), and the role of the House of Representatives (DPR) is increasingly articulated through online platforms. These discussions are often accompanied by the use of hashtags such as #IndonesiaGelap, which represent expressions of public criticism and dissatisfaction. Rather than reflecting isolated reactions to a single issue, such digital conversations illustrate broader patterns of public sentiment across multiple government policy domains.

Beyond serving as a communication channel, social media has evolved into an alternative public sphere where citizens express opinions, concerns, and evaluations of governmental performance. This phenomenon is particularly relevant in contexts where public trust in political institutions remains fragile. This condition reflects a gap between public expectations and policy outcomes, which often triggers increased expression of negative sentiment in digital spaces. Several studies and public surveys have indicated that, despite high levels of political engagement, confidence in legislative and oversight institutions in Indonesia remains relatively low. As a result, social media platforms such as X (Twitter) have become a primary space for articulating criticism, negotiating public narratives, and shaping collective perceptions of policy decisions. Consequently, digital public opinion does not merely capture individual

sentiment but also reflects underlying socio-political dynamics that are critical for policy evaluation and public communication strategies.

To analyze these large-scale and rapidly evolving textual data, sentiment analysis has been widely adopted as a computational approach for identifying public attitudes embedded in unstructured text. Conventional sentiment analysis techniques, including supervised machine learning models such as Support Vector Machine (SVM) and transformer-based architectures like IndoBERT, have demonstrated strong performance in various applications [1]. However, these approaches remain constrained by several limitations, particularly when applied to social media data. Challenges such as informal language, slang, code-mixing, sarcasm, and context-dependent expressions often reduce classification reliability. Moreover, supervised models typically require large labeled datasets and domain-specific fine-tuning, which can limit their adaptability to emerging policy issues.

The emergence of large language models introduces a new paradigm in sentiment analysis by enabling zero-shot and few-shot learning without task-specific retraining. Models such as ChatGPT are capable of processing long and complex textual inputs while capturing semantic context, pragmatic meaning, and implicit emotional cues. Through zero-shot and few-shot prompting mechanisms, LLMs can perform sentiment classification without task-specific retraining, making them highly adaptive for dynamic social media environments [2]. This capability offers substantial potential for analyzing digital public opinion, particularly in dynamic policy environments where labeled data are limited or rapidly evolving. However, recent studies also highlight important limitations, including output instability and sensitivity to prompt formulation, where minor variations in prompts may lead to inconsistent classification results [3]. These challenges indicate that the effectiveness of LLMs should not be evaluated solely based on accuracy, but also in terms of stability, reproducibility, and contextual interpretation.

Despite the growing interest in LLMs, there is a significant literature gap regarding their empirical comparison against fine-tuned supervised models in low-resource and highly informal linguistic contexts, such as Indonesian social media discourse characterized by slang and code-mixing. Existing research often focuses primarily on performance metrics such as accuracy, without sufficiently examining aspects such as contextual interpretation, stability of results, sensitivity to prompt variation, and reasoning consistency [4].

Furthermore, the role of sentiment analysis as part of a broader information system for supporting public policy analysis and decision-making remains underexplored [5]. To address these gaps, this study formulates the following research questions: (1) How do conventional models (SVM, IndoBERT) and LLMs (ChatGPT, LLaMA) perform in classifying public opinion on government policies? (2) How does supervised learning compare to zero-shot/few-shot learning approaches? (3) To what extent does prompt variation affect the stability of LLM classifications? and (4) What are the specific advantages and limitations of LLMs compared to conventional models in this context?.

To address these questions, this study adopts the CRISP-DM framework to systematically compare SVM, IndoBERT, LLaMA, and ChatGPT. The new knowledge created in this study establishes a critical paradigm shift in sentiment analysis. Rather than producing incremental knowledge that simply validates LLMs as superior, this study empirically proves a crucial trade-off: while LLMs possess profound semantic and interpretative capabilities, they remain highly unstable and prompt-sensitive for large-scale unstructured Indonesian text. Consequently, this study proposes a new best practice for developing public opinion Decision Support Systems (DSS), advocating for a hybrid architecture where conventional models (like SVM or IndoBERT) handle quantitative sentiment aggregation for stability, while LLMs (like ChatGPT) are utilized as qualitative tools for contextual validation and interpreting ambiguous, implicit criticisms.

## 2. RELATED WORK

To establish the foundation of this research and identify the existing gaps, a systematic literature screening process was conducted. The literature screening criteria focused on peer-reviewed articles and empirical studies published within the last five years, primarily sourced from major academic databases such as Scopus, IEEE Xplore, and Google Scholar. The selection specifically prioritized research addressing Large Language Models (LLMs) performance in Natural Language Processing (NLP), prompt engineering sensitivity, and sentiment analysis applied to low-resource languages or Indonesian social media discourse. Studies that solely focused on traditional lexicon approaches without machine learning or LLM

comparisons were excluded to maintain relevance to the current technological paradigm shift.

## 2.1 Overview of Sentiment Analysis Methods and Challenges

Sentiment analysis is a major research area within text mining and natural language processing (NLP) that aims to identify opinions, emotions, or attitudes expressed in unstructured text. Early approaches primarily relied on rule-based and lexicon-based methods that determine sentiment polarity based on predefined lists of positive and negative words. As research progressed, these approaches evolved into supervised machine learning models such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression, which learn sentiment patterns from labeled datasets.

The development of deep learning and transformer architectures, including BERT and IndoBERT, marked a significant advancement by enabling models to capture semantic context between words within sentences. However, several challenges remain, particularly in the Indonesian language context. Social media texts frequently contain slang, sarcasm, and code-mixing, which complicate sentiment interpretation. Local studies indicate that although models such as IndoBERT improve classification accuracy, their performance remains constrained by the need for large labeled datasets and limited generalization across diverse social contexts. These limitations motivate the exploration of large language models such as ChatGPT, which offer zero-shot capabilities and adaptive interpretation across linguistic variations [6].

## 2.2 Large Language Models and Cross-Task Sentiment Performance

The emergence of large language models has shifted the sentiment analysis paradigm from task-specific training toward models with broad cross-domain generalization. LLMs such as GPT, PaLM, and Claude leverage large-scale transformer architectures that enable deep semantic understanding across multiple tasks. This approach differs fundamentally from fine-tuned transformer models like IndoBERT, which require retraining on domain-specific datasets to achieve optimal performance.

Zhang et al. (2024) conducted a comprehensive evaluation of LLM capabilities across various sentiment analysis tasks, including polarity classification, aspect-based sentiment analysis (ABSA), and subjectivity detection. Their findings indicate that LLMs excel in tasks requiring

cross-domain generalization and long-context understanding, while fine-tuned models remain competitive or superior when sufficient labeled data are available for specific domains. This trade-off highlights the contrast between flexibility and stability, where LLMs excel in cross-domain generalization, while fine-tuned models remain more stable in domain-specific tasks [7].

## 2.3 Prompt Engineering and Sensitivity in LLM-Based Sentiment Analysis

A defining characteristic of LLM-based sentiment analysis is its reliance on prompt-based interaction. Prompt engineering has emerged as a critical research area that examines how variations in instruction formulation influence model outputs. In sentiment analysis tasks, minor variations in prompt wording, structure, or instruction format can lead to substantial differences in classification results, particularly for ambiguous or sarcastic texts.

Gozzi and Di Maio (2024) investigated the impact of prompting strategies on LLM performance across NLP tasks, including sentiment analysis. Their study demonstrates that prompt quality significantly affects model stability and that small phrasing variations can alter performance outcomes, especially for subjective tasks. These findings underscore the importance of evaluating not only accuracy but also prompt sensitivity and reproducibility when assessing LLM-based sentiment analysis systems [8].

## 2.4 Evaluation of LLMs in Domain-Specific and Social Media Contexts

Several studies have examined the application of LLMs in domain-specific sentiment analysis, particularly in health-related social media discourse. He et al. (2024) evaluated LLM performance using zero-shot, few-shot, and fine-tuning approaches on social media data related to public health issues. Their results show that LLMs perform well on general and relatively formal text, but experience performance degradation when applied to noisy, slang-rich, or domain-specific data without calibration [9].

These findings suggest that the effectiveness of LLMs such as ChatGPT cannot be universally generalized across domains or languages. Testing LLM performance on Indonesian-language social media data, which is characterized by code-mixing and informal expressions, is therefore necessary to understand the boundaries of LLM adaptability beyond English-centric datasets.

## 2.5 Aspect-Based Sentiment Analysis and Indonesian-Specific Models

Aspect-Based Sentiment Analysis (ABSA) provides a more granular framework for understanding public opinion by identifying sentiment toward specific aspects rather than overall polarity. In the Indonesian context, ABSA research highlights the importance of language-specific modeling due to morphological variation and informal expression.

Jazuli et al. (2025) developed an ABSA model using IndoBERT for Indonesian-language datasets and demonstrated that fine-tuned IndoBERT outperformed traditional lexicon-based approaches in extracting aspects and assigning sentiment labels. The ability of IndoBERT to capture local linguistic nuances, including code-mixing, positions it as a stable and reliable benchmark for comparison with zero-shot LLM approaches such as ChatGPT [10].

## 2.6 ChatGPT in Digital Public Opinion Analysis

Recent studies increasingly focus on ChatGPT as both an analytical tool and a subject of public discourse. Fu et al. (2024) demonstrated that ChatGPT achieved competitive performance compared to deep learning-based models, particularly in low-resource language settings [11]. Studies analyzing public discourse surrounding ChatGPT also reveal diverse emotional and ethical responses toward AI adoption in digital society [12].

Comparative studies indicate that ChatGPT can generate more stable sentiment classifications for ambiguous and unstructured texts compared to several traditional machine learning approaches [13]. Other studies also report that conventional models such as SVM and CNN still require extensive training and parameter tuning to achieve optimal performance [14]. Similar findings demonstrated that ChatGPT can achieve competitive sentiment classification performance even without extensive supervised training [15]. Research on academic perceptions also indicates generally positive sentiment toward ChatGPT's utility, accompanied by ethical concerns regarding reliability and misuse [16].

## 2.7 Literature Synthesis and Research Positioning

The reviewed literature highlights several key insights. First, sentiment analysis has evolved rapidly alongside the development of LLMs, particularly in addressing linguistic challenges such as slang, sarcasm, and code-mixing in Indonesian social media data [6]. Second, comparative studies

reveal that ChatGPT exhibits strong cross-domain generalization and contextual understanding, but its performance is sensitive to prompt configuration and domain characteristics [7]. Third, domain-specific and language-specific models such as IndoBERT continue to demonstrate strong and stable performance, particularly for ABSA and Indonesian-language tasks [10]. Recent studies also explore hybrid and ensemble-based approaches that combine transformer embeddings with traditional machine learning techniques to improve robustness and reduce output variance in sentiment classification tasks, particularly for noisy social media data [17]. These findings suggest that integrating conventional supervised models with contextual LLM capabilities may provide a more balanced approach for analyzing complex digital public opinion.

Based on these findings, this study positions itself at the intersection of conventional supervised models and LLM-based approaches by systematically comparing supervised models such as SVM and IndoBERT with Large Language Model approaches including LLaMA and ChatGPT. The research aims to evaluate not only classification performance but also prompt sensitivity and result stability in analyzing Indonesian digital public opinion across multiple government policy issues.

## 3. METHODOLOGY

### 3.1 Research Design and Framework

This study employs a quantitative comparative research design within the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as the primary methodological structure. CRISP-DM is widely adopted in text mining and sentiment analysis research due to its systematic, iterative, and reproducible nature. The framework consists of six phases from Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, ensuring methodological rigor from problem formulation to result interpretation. Figure 1 shows proposed research framework based on the CRISP-DM methodology for comparative sentiment analysis using supervised models and large language models (LLMs).

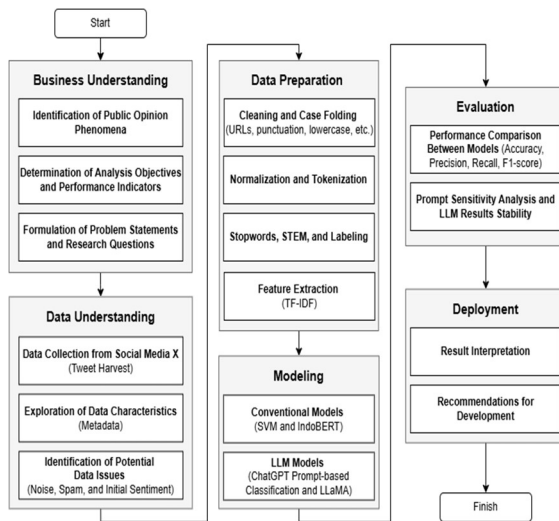


Figure 1. Proposed research framework based on the CRISP-DM

In this research, CRISP-DM enables the integration of conventional sentiment analysis models and large language models within a unified analytical workflow. Importantly, different roles are assigned to each model category to reflect their methodological strengths: supervised models are evaluated quantitatively, while ChatGPT is positioned as a qualitative interpretative and validation tool.

### 3.2 Data Source and Collection

The dataset consists of Indonesian-language public posts collected from the X (Twitter) platform over a one-year period. Data collection focused on discussions related to government policy issues, captured through predefined keywords and hashtags such as #IndonesiaGelap, RUU TNI, KUHP, and DPR. These topics were selected to capture diverse expressions of digital public opinion related to national policy discourse.

A purposive sampling strategy was applied to ensure that the collected data were topically relevant, linguistically appropriate (Indonesian language), and representative of public policy discussions. The dataset consisted of approximately 10,000 Indonesian-language posts collected over a one-year period. To address computational limitations, quantitative evaluation of LLaMA was conducted on a selected test subset, while supervised models were trained on the full labeled dataset. Only publicly accessible content was used, and ethical considerations were addressed by anonymizing user information and complying with platform data usage policies.

### 3.3 Data Preprocessing

Given the noisy and informal nature of social media text, extensive preprocessing was conducted to ensure data quality and consistency for downstream modeling tasks. The preprocessing pipeline included data cleaning, case folding, tokenization, stopword removal, normalization of slang and non-standard words, and stemming using Indonesian-language NLP tools.

For supervised models, the processed text was transformed into numerical representations using TF-IDF feature extraction. Sentiment labels (positive, negative, neutral) were assigned to enable training and evaluation of machine learning and deep learning models.

### 3.4 Sentiment Analysis Models

Four sentiment analysis approaches were implemented to represent the evolution of NLP techniques:

1. Support Vector Machine (SVM) was applied as a conventional supervised baseline model using TF-IDF feature vectors. Its robustness in high-dimensional text classification makes it suitable for benchmarking sentiment performance. Hyperparameter tuning was applied to supervised models to improve classification performance and generalization capability.
2. IndoBERT, a transformer-based model pre-trained for the Indonesian language, was fine-tuned on the labeled dataset to capture contextual and semantic nuances specific to Indonesian social media discourse.
3. LLaMA (LLM – Quantitative Evaluation) was employed as an open-source Large Language Model and evaluated quantitatively using zero-shot and few-shot prompting without task-specific fine-tuning. Its performance was assessed using standard classification metrics to examine the feasibility of LLM-based sentiment classification without task-specific fine-tuning.
4. ChatGPT (LLM – Qualitative Evaluation) was not treated as a conventional classifier but as a qualitative evaluation and interpretative model. It was used to analyze contextual understanding, reasoning consistency, sensitivity to ambiguous discourse, and output stability under prompt variation, rather than for direct quantitative classification.

This separation ensures methodological clarity and avoids unfair comparisons between fundamentally different modeling paradigms.

### 3.5 Evaluation Strategy

Model evaluation was conducted using a mixed evaluation approach:

- Quantitative evaluation was applied to SVM, IndoBERT, and LLaMA using accuracy, precision, recall, and Macro F1-score, with Macro F1 emphasized to address class imbalance in sentiment labels. For LLaMA, additional stability evaluation was conducted by comparing multiple prompt variants using agreement analysis, Cohen's Kappa, and entropy-based metrics to measure sensitivity and reproducibility across prompt configurations.
- Qualitative evaluation was conducted for ChatGPT to assess contextual interpretation, robustness to prompt variation, and alignment with human judgment in policy-related discourse.

This dual evaluation strategy enables a comprehensive assessment of both classification performance and interpretative capability, which is critical for analyzing complex digital public opinion.

## 4. RESULT AND DISCUSSION

### 4.1 Quantitative Performance Comparison

The quantitative evaluation reveals clear performance differences between conventional sentiment analysis models and large language models, highlighting a consistent gap between supervised approaches and prompt-based LLM classification in the context of Indonesian social media data. This is shown in Table 1 which contains the results of quantitative evaluation of several models.

Table 1. Performance Comparison Models

| Models         | Accuracy | Macro F1 | Precision | Recall |
|----------------|----------|----------|-----------|--------|
| SVM            | 0.8658   | 0.7651   | 0.7820    | 0.7531 |
| Indo BERT      | 0.8446   | 0.7440   | 0.7707    | 0.7223 |
| LLaMA Zeroshot | 0.6904   | 0.3576   | 0.3360    | 0.3852 |
| LLaMA Fewshot  | 0.5241   | 0.3989   | 0.5005    | 0.4350 |

The Support Vector Machine (SVM) achieved the highest overall performance, with an accuracy of 0.8658, Macro F1-score of 0.7651, precision of 0.7820, and recall of 0.7531, indicating strong robustness despite class imbalance. The

stability of SVM was further supported by cross-validation results, which showed consistent Macro F1-scores across multiple folds. Figure 2 present the confusion matrix of the SVM model well a strong concentration along the diagonal, indicating that most predictions match the true labels, with only minor misclassification across classes.

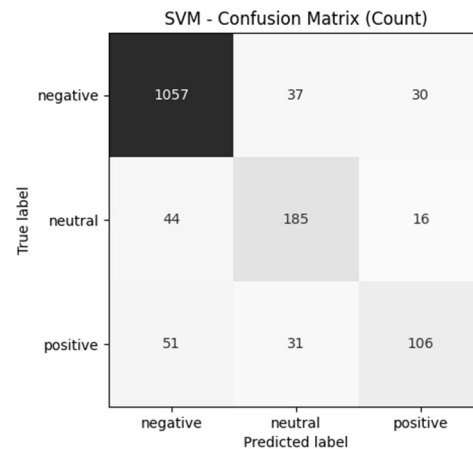


Figure 2. The Confusion Matrix of SVM Model

The confusion matrix visualization further confirms this performance, showing a strong concentration along the diagonal, which indicates accurate classification across sentiment classes with minimal misclassification. Hyperparameter tuning contributed to slight improvements in generalization performance, reinforcing SVM's effectiveness as a baseline classifier for Indonesian social media sentiment analysis. The confusion matrix visualization further confirms this performance, showing a strong concentration along the diagonal, which indicates accurate classification across sentiment classes with minimal misclassification.

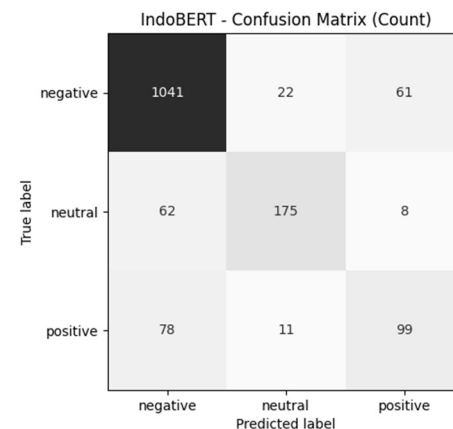


Figure 3. The Confusion Matrix of IndoBERT Model

The IndoBERT model also demonstrated competitive performance, achieving an accuracy of 0.8446, Macro F1-score of 0.7440, precision of 0.7707, and recall of 0.7223. As a transformer-based model pretrained specifically for the Indonesian language, IndoBERT was able to capture contextual and semantic nuances more effectively than traditional machine learning approaches. Class-wise analysis indicates that IndoBERT performed particularly well in identifying negative sentiment, which is prevalent and linguistically explicit in policy-related social media discourse.

However, IndoBERT exhibited limitations in distinguishing between neutral and positive sentiments, especially in texts containing implicit criticism or mixed emotional cues. This pattern is also reflected in the confusion matrix, where misclassification between neutral and positive classes is more frequent. Figure 3 show the confusion matrix of the IndoBERT model well in identifying negative sentiment, misclassification between neutral and positive classes is more noticeable, reflecting challenges in distinguishing subtle sentiment differences.

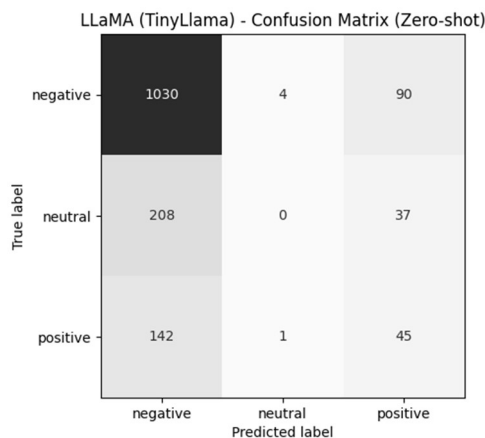


Figure 4. The Confusion Matrix of LLaMA Zero-shot Model

In contrast, LLaMA, evaluated using zero-shot and few-shot prompting strategies, showed substantially lower quantitative performance. In the zero-shot configuration, LLaMA achieved an accuracy of 0.6904, Macro F1-score of 0.3576, precision of 0.3360, and recall of 0.3852. The relatively low Macro F1-score indicates that the model struggled to consistently classify minority sentiment classes, particularly neutral and positive sentiments. Although LLaMA possesses broad contextual language understanding, the zero-shot approach without domain adaptation proved less effective for Indonesian social media texts

characterized by informal expressions, slang, sarcasm, and socio-political context. Figure 4 present the predictions are more dispersed compared to SVM and IndoBERT, indicating lower classification consistency and difficulty in distinguishing sentiment classes.

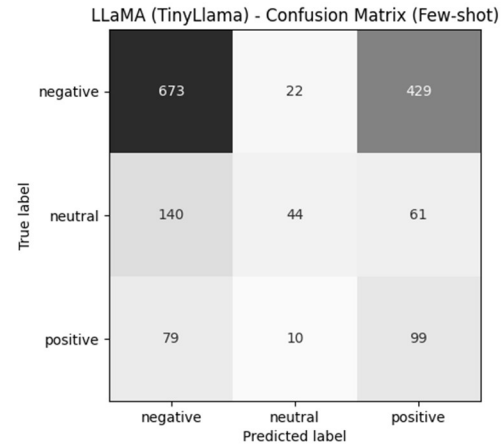


Figure 5. The Confusion Matrix of LLaMA Few-shot Model

When few-shot prompting was applied, the accuracy decreased to 0.5241, while the Macro F1-score slightly increased to 0.3989, with precision reaching 0.5005 and recall 0.4350. This performance degradation suggests that the inclusion of limited examples does not necessarily improve classification quality when prompts fail to adequately represent linguistic and contextual characteristics of Indonesian social media text. Figure 5 show that although some improvement in class distribution can be observed, significant misclassification remains, particularly in distinguishing negative and positive sentiments.

This also reflects the difficulty of adapting prompt-based LLMs to informal and context-rich Indonesian social media text. The confusion matrix further reveals a more dispersed prediction pattern, indicating frequent misclassification across sentiment classes and reduced discriminative capability. Overall, these findings indicate that prompt-based LLM classification without task-specific fine-tuning remains less stable for quantitative sentiment analysis.

## 4.2 Comparative Analysis Across Modeling Paradigms

The comparative analysis highlights a fundamental trade-off between stability and flexibility across different modeling paradigms. Supervised models such as SVM and IndoBERT benefit from structured training data and well-defined decision

boundaries, resulting in higher reproducibility and more reliable quantitative performance. These characteristics make supervised approaches suitable for large-scale sentiment classification tasks that require consistent and measurable outputs.

Conversely, the lower quantitative performance of LLaMA indicates that general-purpose large language models are not yet capable of replacing supervised models for sentiment classification of Indonesian social media data. The observed decline in performance under few-shot prompting further suggests that prompt-based learning is highly sensitive to example selection and prompt design. Without standardized prompting strategies or domain adaptation, LLM-based sentiment classification remains difficult to reproduce and scale.

Beyond numerical performance, these findings also reflect how different models respond to the linguistic characteristics of Indonesian social media data, which often include informal language, sarcasm, and context-dependent expressions. While supervised models provide stable classification boundaries, they may struggle to capture deeper contextual meaning. In contrast, LLMs offer flexibility in interpretation but lack consistency when applied to structured classification tasks.

### 4.3 Prompt Sensitivity and Output Stability in LLaMA

An important finding of this study concerns the sensitivity of LLaMA to prompt variations. Minor changes in prompt wording, structure, or sentiment label ordering resulted in noticeable differences in classification outcomes. This issue was particularly evident in neutral and informational texts, where sentiment polarity is often ambiguous or context-dependent.

Table 2. Prompt Sensitivity and Stability

| Aspect      | Few_v1 | Zero_v1 | Zero_v2  | Zero_v3 |
|-------------|--------|---------|----------|---------|
| Acc.        | 0.5241 | 0.6904  | 0.4669   | 0.1619  |
| Macro F1    | 0.3989 | 0.3576  | 0.2858   | 0.1194  |
| Avg. Aggre. | 0.3711 | 0.3952  | 0.1752   | 0.4603  |
| Avg. Cohen  | 0.0937 | 0.0737  | 0.0081   | 0.0559  |
| Stability   | Mod.   | Mod.    | Very Low | Low     |

Table 2 demonstrates that variations in prompt formulation significantly influenced LLaMA classification performance, particularly in zero-shot configurations. This finding indicates that prompt engineering plays a critical role in determining the stability and reproducibility of LLM-based sentiment analysis. Although few-shot prompting improved output consistency in some cases, the results remained unstable across different prompt configurations. This instability is reflected in the variation of predicted labels for the same input text, indicating a lack of deterministic behavior in prompt-based classification. As a result, LLaMA struggled to maintain consistent sentiment predictions, particularly for texts containing implicit meaning or mixed emotional signals.

Table 3. Overall Stability Metrics

| Metric                                     | Value  |
|--------------------------------------------|--------|
| Average Entropy (Base-2)                   | 1.0806 |
| Normalized Entropy                         | 0.6579 |
| Full Agreement Proportion                  | 0.0122 |
| Strong Majority Proportion ( $\geq 0.75$ ) | 0.4200 |

Table 3 indicates varying levels of agreement across prompting strategies, suggesting that LLaMA predictions remain sensitive to contextual instruction changes. Higher entropy values further reflect instability in sentiment classification consistency across prompts. This instability limits the applicability of open-source LLMs for policy-oriented sentiment analysis systems that require consistent and reproducible outputs over time. These findings underscore the need for careful prompt engineering and standardized evaluation protocols when employing LLMs for sentiment analysis.

### 4.4 Qualitative Evaluation of ChatGPT

Unlike other models in this study, ChatGPT was evaluated using a qualitative approach rather than direct quantitative benchmarking. The qualitative analysis demonstrates that ChatGPT exhibits strong contextual reasoning capabilities, particularly when interpreting ambiguous, implicitly critical, or informational texts commonly found in public policy discussions. ChatGPT was able to distinguish more effectively between factual narratives and emotional expressions, reflecting its capacity to process

pragmatic meaning beyond surface-level sentiment cues.

Furthermore, ChatGPT showed greater output stability across prompt variations compared to LLaMA. This consistency suggests that reinforcement learning from human feedback (RLHF) contributes to more reliable interpretative behavior. ChatGPT was also more capable of maintaining consistent reasoning when handling context-rich or socially nuanced expressions.

Table 4. Qualitative Comparison of LLaMA Few-shot and ChatGPT

| Sample Tweet | Initial  | LLaMA Few-shot | ChatGPT  |
|--------------|----------|----------------|----------|
| p1           | Negative | Positive       | Positive |
| p2           | Negative | Positive       | Negative |
| p3           | Negative | Positive       | Neutral  |
| p4           | Positive | Negative       | Negative |
| p5           | Positive | Negative       | Neutral  |
| p6           | Positive | Negative       | Negative |
| p7           | Neutral  | Negative       | Neutral  |
| p8           | Positive | Negative       | Negative |
| p9           | Positive | Negative       | Negative |
| p10          | Neutral  | Negative       | Negative |

p1–p10 denote anonymized and abbreviated sample tweets

Table 4 presents several examples of qualitative sentiment differences between LLaMA few-shot and ChatGPT predictions. The comparison shows that ChatGPT generally produced sentiment classifications that were more contextually aligned with the original tweet meaning, particularly for ambiguous, implicitly critical, or informational statements. In contrast, LLaMA few-shot frequently generated inconsistent predictions, often misclassifying contextual criticism as positive sentiment. These findings further support the argument that ChatGPT demonstrates stronger contextual reasoning and interpretative stability in understanding Indonesian public policy discourse. Although ChatGPT is not suitable for large-scale quantitative sentiment classification due to its non-deterministic nature, its ability to provide contextual explanations and sentiment rationales offers

valuable complementary support as an interpretative and validation tool.

#### 4.5 Implications for Digital Public Opinion Analysis

The overall findings indicate that no single sentiment analysis model is sufficient to fully capture the complexity of Indonesian digital public opinion. Supervised models such as SVM and IndoBERT remain the most reliable choices for quantitative sentiment classification, particularly in large-scale analyses requiring stable and reproducible results. However, their limitations in handling implicit meaning, contextual ambiguity, and nuanced socio-political expressions highlight the added value of large language models.

When positioned as a qualitative analytical component rather than a direct classifier, ChatGPT enhances interpretability and supports deeper understanding of public discourse. This complementary integration aligns sentiment analysis with broader information system objectives, enabling more nuanced insights for public policy evaluation and decision-making in the Indonesian context.

These results suggest that an effective sentiment analysis system for digital public opinion should not rely on a single model, but instead adopt a hybrid approach that combines the quantitative strength of supervised models with the contextual interpretative capabilities of LLMs.

#### 4.6 Limitations and Critique of the Study

Before drawing conclusions, it is important to critically evaluate the limitations of this study. First, the implementation of open-source LLMs (LLaMA) in this research did not involve domain-specific fine-tuning due to computational resource constraints, which may have contributed to its lower quantitative performance compared to the fully fine-tuned IndoBERT. Second, the reliance on prompt-based inference demonstrated a high sensitivity to prompt formulation slight linguistic variations caused significant fluctuations in LLM predictions, exposing a vulnerability in reproducibility. Lastly, from a deployment perspective, the current framework was tested on static, pre-collected datasets. It lacks real-time API integration, which is necessary for a fully functional early warning system in dynamic public policy environments. These limitations suggest that the findings represent a baseline comparison, and future iterations require architectural optimizations to handle real-time, large-scale data streams efficiently.

## 5. CONCLUSION

This study critically evaluated the effectiveness of conventional sentiment analysis models versus Large Language Models (LLMs) in understanding Indonesian digital public opinion. In light of the research objectives, the results demonstrate that supervised learning approaches, particularly SVM and IndoBERT, successfully achieved the highest quantitative reliability. SVM proved to be the most robust model for handling imbalanced sentiment data, while IndoBERT effectively captured Indonesian semantic nuances. Conversely, the evaluation revealed that relying solely on prompt-based LLMs (LLaMA) for quantitative classification is currently inadequate for this domain. The significant performance degradation and high entropy in zero-shot and few-shot settings indicate that LLMs remain highly sensitive to prompt design and struggle with consistency in informal, noisy text environments without task-specific fine-tuning.

However, the qualitative evaluation highlights that models utilizing Reinforcement Learning from Human Feedback (RLHF), such as ChatGPT, provide profound value in interpreting complex, ambiguous, and implicitly critical public discourse. Therefore, this study concludes that an effective sentiment analysis framework for public policy cannot rely on a single algorithm; it necessitates a hybrid deployment where supervised models drive quantitative aggregation and LLMs facilitate qualitative contextual reasoning.

Regarding open research issues and future work, several areas remain unattended based on the limitations of this study. First, the development of hybrid ensemble architectures that programmatically integrate the speed of conventional machine learning with the interpretative depth of LLMs requires further exploration. Second, there is a critical need to investigate standardized, automated prompt engineering frameworks to mitigate the output instability observed in LLMs. Third, future studies should focus on implementing real-time data streaming via social media APIs to test the scalability of these models in live Decision Support Systems (DSS). Finally, integrating Explainable AI (XAI) techniques into LLM predictions is essential to ensure that sentiment analysis results remain transparent and interpretable for non-technical policymakers.

## ACKNOWLEDGMENT

No potential conflict of interest was reported by the authors. No funding is sourced for this research. Author contributorship: Assegaf Insani - conceptualization, formal analysis, investigation, methodology, resources, visualization, writing and editing; Tanty Oktavia - conceptualization, formal analysis, project administration, validation and review. Data openness and transparency. To facilitate further research, we have made the data used in their study publicly available. Data supporting this study is available from kaggle at <https://www.kaggle.com/datasets/insaniassegaf/indonesian-political-discourse-2025> to the data is subject to approval and a data sharing agreement due to respondent company confidentiality surrounding sensitive operational and project data. For access request, please contact the corresponding author's email address for approval.

## REFERENCES:

- [1] N. A. Semary, W. Ahmed, K. Amin, P. Plawiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Front. Hum. Neurosci.*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1292010.
- [2] Z. Wang, Q. Xie, Y. Feng, Z. Ding, Z. Yang, and R. Xia, "Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study," Feb. 2024, [Online]. Available: <http://arxiv.org/abs/2304.04339>
- [3] T. Ouyang, A. P. MaungMaung, K. Konishi, Y. Seo, and I. Echizen, "Stability Analysis of ChatGPT-Based Sentiment Analysis in AI Quality Assurance," *Electronics (Switzerland)*, vol. 13, no. 24, Dec. 2024, doi: 10.3390/electronics13245043.
- [4] C. Leiter *et al.*, "ChatGPT: A meta-analysis after 2.5 months," *Machine Learning with Applications*, vol. 16, p. 100541, Jun. 2024, doi: 10.1016/j.mlwa.2024.100541.
- [5] Z. Li, F. Y. Chan, C. Gao, C. Tong, and Q. Ye, "Sentiment analysis and summarization with ChatGPT: implications for sales prediction," *Journal of Digital Management*, vol. 1, no. 1, May 2025, doi: 10.1007/s44362-025-00008-3.
- [6] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," Apr. 01, 2024, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2024.102048.

- [7] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment Analysis in the Era of Large Language Models: A Reality Check," 2024.
- [8] M. Gozzi and F. Di Maio, "Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts," *Electronics (Switzerland)*, vol. 13, no. 23, Dec. 2024, doi: 10.3390/electronics13234712.
- [9] L. He, S. Omranian, S. Mcroy, and K. Zheng, "Using Large Language Models for sentiment analysis of health-related social media data: empirical evaluation and practical tips," 2024. Accessed: Oct. 29, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12099340/>
- [10] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Applied Sciences (Switzerland)*, vol. 15, no. 1, Jan. 2025, doi: 10.3390/app15010172.
- [11] Z. Fu, Y. C. Hsu, C. S. Chan, C. M. Lau, J. Liu, and P. S. F. Yip, "Efficacy of ChatGPT in Cantonese Sentiment Analysis: Comparative Study," *J. Med. Internet Res.*, vol. 26, no. 1, Jan. 2024, doi: 10.2196/51069.
- [12] D. Heaton, J. Clos, E. Nichele, and J. E. Fischer, "'The ChatGPT bot is causing panic now – but it'll soon be as mundane a tool as Excel': analysing topics, sentiment and emotions relating to ChatGPT on Twitter," *Pers. Ubiquitous Comput.*, vol. 28, no. 6, pp. 875–894, Dec. 2024, doi: 10.1007/s00779-024-01811-x.
- [13] I. Hussain *et al.*, "Sentiment Analysis Classification of ChatGPT Tweets Using Machine Learning and Deep Learning Algorithms," 2024, doi: 10.56979/801/2024.
- [14] N. Widaad and D. Anggraini, "SENTIMENT ANALYSIS OF CHATGPT APP USER REVIEWS USING SVM AND CNN METHODS," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, pp. 1687–1700, Dec. 2024, doi: 10.52436/1.jutif.2024.5.6.4010.
- [15] G. Jeffson Sagala and Y. T. Samuel, "Sentiment Analysis on ChatGPT App Reviews on Google Play Store Using Random Forest Algorithm, Support Vector Machine and Naïve Bayes," 2024. [Online]. Available: <https://ijebss.ph/index.php/ijebss>
- [16] Y. Mamo, H. Crompton, D. Burke, and C. Nickel, "Higher Education Faculty Perceptions of ChatGPT and the Influencing Factors: A Sentiment Analysis of X," *TechTrends*, vol. 68, no. 3, pp. 520–534, May 2024, doi: 10.1007/s11528-024-00954-1.
- [17] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Applied Computational Intelligence and Soft Computing*, vol. 2024, 2024, doi: 10.1155/2024/2826773.