International Journal Software Engineering and Computer Science (IJSECS)
6 (1), 2026, 194-203
Published Online April 2026 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v6i1.6969.

RESEARCH ARTICLE Open Access

Sentiment and Public Emotion Classification of Viral Content Using Transformer-Based Model

Ferdi Antonio [1*], **Handry Eldo** [2], **Arrazy Elba Ridha** [3], **Iwan Adhicandra** [4], **Cut Susan Octiva** [5].

[1*] Universitas Pelita Harapan, Tangerang Regency, Banten Province, Indonesia.
[2] Universitas Muhammadiyah Mahakarya Aceh, Banda Aceh City, Aceh Province, Indonesia.
[3] Universitas Teuku Umar, West Aceh Regency, Aceh Province, Indonesia.
[4] Bakrie University, South Jakarta City, Special Capital Region of Jakarta, Indonesia.
[5] Universitas Amir Hamzah, Deli Serdang Regency, North Sumatra Province, Indonesia.

**Corresponding author*: ferdi.antonio@gmail.com.

Received: March 15, 2026; Accepted: April 1, 2026; Published: April 10, 2026.

Abstract: The proliferation of social media platforms has generated an unprecedented volume of viral content, each drawing varied public responses expressed through sentiment and emotion. Mapping those responses — not merely counting them — is what separates surface-level monitoring from a genuine understanding of public perception. This study classified sentiment (positive, negative, neutral) and emotion (anger, joy, sadness, and fear) toward viral content using a fine-tuned Transformer-based model. Data were collected from social media via web scraping, then subjected to standard text preprocessing: case folding, tokenization, stopword removal, and stemming. The cleaned dataset was subsequently annotated with sentiment and emotion labels. BERT (Bidirectional Encoder Representations from Transformers) served as the base architecture, fine-tuned for multi-label classification. Evaluation relied on an 80:20 train-test split, with performance measured through accuracy, precision, recall, and F1-score. Across all sentiment and emotion categories, the model returned consistently high scores and handled ambiguous, context-dependent text more reliably than conventional machine learning baselines. The Transformer-based approach proved well-suited for sentiment and emotion analysis on social media data, with clear potential for deployment in public opinion monitoring systems.

Keywords: Sentiment Analysis; Emotion Classification; Viral Content; Transformer; BERT.

1. Introduction

Rapid advances in information and communication technology have fundamentally altered how people produce, share, and consume content. The barriers to publication have collapsed — anyone with a smartphone and an internet connection can now reach a global audience within seconds. Social media platforms, as the most visible product of that shift, have become the dominant channels through which information circulates, opinions form, and public discourse takes shape. Among the most consequential phenomena to emerge from this environment is viral content: material that spreads rapidly across networks, accumulating massive engagement in a compressed timeframe. Virality is not accidental. It is driven by algorithmic amplification, emotional resonance, and the social dynamics of sharing behavior — factors that make viral content a particularly potent force in shaping public perception and behavior (Kodati & Tene, 2022). Understanding how

the public responds to such content, therefore, is not a peripheral concern. It sits at the intersection of communication studies, computational linguistics, and data science.

Public responses to online content typically take textual form — comments, reviews, quote-posts, and threaded replies — each carrying layers of sentiment and emotion that reflect the author's cognitive and affective state. Sentiment, in the technical sense, refers to the polarity of an expressed opinion: positive, negative, or neutral. Emotion, by contrast, captures something more granular — specific affective states such as joy, anger, sadness, or fear (Tabinda Kokab *et al.*, 2022). The two dimensions are related but not interchangeable. A negative sentiment can be expressed through anger, through sadness, or through fear, and each of those emotional registers carries a different implication for how the public is actually processing a given piece of content. Analyzing sentiment alone, then, produces an incomplete picture. Combining sentiment and emotion analysis within a single model yields a more precise account of public response — one that distinguishes not just whether people reacted negatively, but how and why. That said, natural language is notoriously resistant to clean categorization. Ambiguity, irony, sarcasm, and context-dependence all complicate automated analysis in ways that remain genuinely difficult to resolve (Almalki, 2025).

The field of Natural Language Processing (NLP) has produced a succession of tools for automated text analysis, each generation more capable than the last. Conventional machine learning approaches — Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression — established the foundations of computational text classification and continue to perform adequately on structured, formal text. Their core limitation, however, is well-documented: these methods treat words largely as independent features, which means they struggle to capture the semantic dependencies that give language its meaning. A sentence like "this is not bad at all" will confuse a bag-of-words classifier in ways it will not confuse a human reader. Deep learning approaches, and particularly the Transformer architecture, address this limitation directly. By modeling contextual relationships across entire sequences rather than isolated tokens, Transformer-based models can represent meaning in a way that is sensitive to word order, syntactic structure, and broader discourse context. BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin *et al.* and subsequently applied across a wide range of NLP tasks, has become the reference architecture for text classification precisely because its bidirectional attention mechanism processes both left and right context simultaneously — a capability that static embedding models simply do not have (Sharma *et al.*, 2022; Tiwari & Nagpal, 2022).

Despite that track record, the joint application of Transformer models to simultaneous sentiment and emotion classification on viral content remains relatively underexplored, particularly for Indonesian-language social media data (Parveen *et al.*, 2026). This gap is more consequential than it might initially appear. Indonesian social media language is informal, fast-evolving, and linguistically heterogeneous — marked by code-switching between Indonesian and regional languages, heavy use of abbreviations, non-standard spelling, and emotionally charged slang that changes faster than any static lexicon can track. A model trained on formal Indonesian text, or on English-language social media data, will not transfer cleanly to this domain. The preprocessing and fine-tuning decisions that work for one context may actively harm performance in another. Beyond the language-specific challenge, the sheer volume of social media data makes manual analysis untenable at any meaningful scale. Automated classification is not merely convenient; at this data volume, it is the only viable option. Combining sentiment and emotion analysis within a unified modeling approach is expected to yield deeper insight into the dynamics of public response to viral phenomena — insight that neither task alone can provide (Oliveira *et al.*, 2022).

Several recent studies have begun to address adjacent problems. Almalki (2025) examined Transformer-based sentiment and emotion detection across multilingual social media data, reporting strong performance gains over conventional baselines. Oliveira *et al.* (2022) applied topic modeling and emotion classification to COVID-19 Twitter data, demonstrating that emotional response patterns vary significantly by topic and platform context. Tabinda Kokab *et al.* (2022) provided a systematic comparison of Transformer-based models for social media sentiment analysis, establishing BERT and its variants as the most consistently reliable architectures. Taken together, these studies confirm the viability of the Transformer approach — but they also leave open the specific question of how well such models perform on Indonesian-language viral content when sentiment and emotion are classified jointly rather than separately.

Against that background, this study developed a multi-label classification model for public sentiment and emotion toward viral content, using BERT as the base architecture fine-tuned on an Indonesian social media dataset (Antunes *et al.*, 2026; Veluswamy *et al.*, 2025). The work pursues two practical objectives: advancing text analysis methods for informal, non-standard Indonesian-language data, and producing a model that could realistically support public opinion monitoring, digital marketing analytics, or evidence-based policy communication. The findings are also intended to serve as a reference point for subsequent research in large-scale NLP, particularly at the intersection of sentiment analysis and emotion classification on social media data.

2. Related Work

Sentiment analysis did not begin with neural networks. The earliest automated approaches relied on lexicon-based scoring systems — dictionaries of words pre-assigned polarity weights, aggregated into a document-level judgment. Leon (2025) traced this lineage from rule-based lexicons through to large language models, and the central argument holds: lexicon-based systems are interpretable and cheap to run, but they break down on informal text where the same word means different things in different contexts. That is not an edge case in social media data. It is the default condition. Conventional machine learning classifiers — Naïve Bayes, SVM, Logistic Regression — improved on lexicon methods by learning patterns from labeled examples rather than hand-crafted rules, but they still treated text as an unordered collection of tokens, discarding the sequential and syntactic information that often carries the actual meaning of a sentence. Tabinda Kokab *et al.* (2022) compared Transformer-based models against these conventional baselines across multiple social media benchmark datasets and found that BERT and its variants outperformed earlier approaches consistently — with the gap widest on short, informal texts. The reason is architectural. BERT's bidirectional attention mechanism processes each token against all other tokens in the sequence simultaneously, rather than reading left-to-right or relying on fixed context windows. That single design choice accounts for most of the performance difference.

Tiwari and Nagpal (2022) pushed the architecture further. Their KEAHT model — Knowledge-Enriched Attention-based Hybrid Transformer — combined BERT with Latent Dirichl*et al*location topic modeling and lexicalized domain ontology, applied to COVID-19 vaccine tweets and Indian farmer protest data. KEAHT outperformed standard BERT fine-tuning, and the margin came specifically from the external knowledge injection. The implication is worth sitting with: a Transformer that knows something about the domain before it sees the training data performs better than one that learns domain knowledge purely from the labeled examples. Whether that finding transfers to Indonesian viral content is an open question, but the principle is sound. Yazdi *et al.* (2025) made a related point from a different direction, fine-tuning RoBERTa on tweets about the 2024 Paris Olympics and demonstrating that event-specific fine-tuning consistently outperforms reliance on general-purpose pre-trained weights. The more the target domain diverges from the pre-training corpus, the more fine-tuning matters. Sharma *et al.* (2022) extended the application further still, using a BERT-based model to evaluate information relevance in COVID-19 social media posts — combining classification with topic modeling to identify subjects most frequently associated with misinformation. Their work is a reminder that BERT is not just a polarity detector; it can make nuanced relevance judgments that older architectures simply cannot.

Emotion classification is harder than sentiment detection. Not marginally harder — genuinely harder, for reasons that go beyond the larger output space. Sentiment reduces the problem to three categories. Emotion classification typically requires distinguishing among four to eight affective states in text that is short, ambiguous, and written without the prosodic cues that help humans interpret emotional tone in speech. Kodati and Tene (2022) confronted this in one of the most demanding possible contexts: detecting suicidal emotions in social media posts. Their models — bidirectional gated recurrent units with multi-head attention and CNN components, both drawing on BERT-based contextual features — outperformed prior state-of-the-art methods on the task. The key finding was that long-range contextual dependencies matter enormously for emotion detection in charged, informal text. A model that cannot track how the emotional register of a sentence shifts across its full length will misclassify a substantial proportion of instances. Oliveira *et al.* (2022) approached the problem differently, pairing topic modeling with emotion classification on a dataset of over 16,000 COVID-19 tweets. Fear and sadness dominated health-related content; anger clustered around politically framed posts. That distribution is not incidental — it reflects the fact that emotion expression is topic-dependent in ways that aggregate classification scores obscure. A model trained to classify emotion without accounting for subject matter will conflate structurally similar expressions that carry different affective meanings in different contexts. Maghsoudi *et al.* (2022) added another layer to this picture, combining multiple Transformer outputs through a Dempster-Shafer probabilistic fusion framework for sentiment analysis of insomnia-related tweets across pre- and peri-COVID periods. The ensemble approach outperformed any single model, particularly on ambiguous instances — which, in social media data, are not rare.

The application of these methods to public opinion monitoring at scale raises its own set of challenges. Md Suhaimin *et al.* (2023) reviewed over 200 studies on social media sentiment analysis in public security contexts, and three findings stand out from their synthesis. First, Transformer-based models dominate recent performance benchmarks without exception. Second, multi-label classification — combining sentiment with emotion, intent, or other dimensions — consistently outperforms single-label approaches when the goal is understanding public response rather than just measuring polarity. Third, and most relevant here, informal, code-switched, and non-English text remains the most significant unsolved challenge in the field. Indonesian social media sits squarely in that problem space: code-switching between Indonesian and regional languages is common, non-standard orthography is pervasive, and the linguistic norms of the platform shift faster than

any static training corpus can track. Almalki (2025) addressed the multilingual dimension directly, fine-tuning XLM-R for sentiment and emotion detection across seven languages and achieving an F1-score of 90.3% — with preprocessing improvements for code-switching alone accounting for an 8.9% accuracy gain. Cross-lingual Transformer models handle linguistic diversity better than earlier multilingual baselines, but the performance gains from language-specific fine-tuning suggest that a model trained on the target language variety will still outperform a multilingual generalist on that specific domain.

Two further studies are worth noting for what they reveal about the limits of sentiment-only analysis. Ashraf and Choi (2025) demonstrated through their XP-STM model that sentiment classifiers trained on one platform transfer poorly to another — platform-specific linguistic norms, user demographics, and content types differ enough to degrade performance in ways that single-platform evaluations never detect. Viral content, which spreads across platforms by definition, is particularly exposed to this problem. Zhou *et al.* (2025) made a different but related point: sentiment polarity is an insufficient descriptor of user attitude toward content. Their RoBERTa-BiLSTM model quantified believability scores for rumor-refuting information, and the resulting crowd classification — distinguishing rumor refuters, rumor spreaders, anti-rumor inactivists, and pro-rumor inactivists — revealed feature profiles that polarity scores alone could not separate. Behavioral and contextual features, combined with NLP-derived sentiment measures, produced a more discriminative and practically useful classification than sentiment alone. That finding applies directly to the present study's objective: understanding public response to viral content requires more than knowing whether a comment was positive or negative. What the reviewed literature has not yet provided is a systematic evaluation of joint sentiment and emotion classification on Indonesian-language viral content specifically — and that is the gap this study addresses .

3. Methodology

A quantitative experimental design was adopted to develop and evaluate the Transformer-based classification model on social media text data. The choice of experimental design reflects the nature of the task: model performance is measurable, reproducible, and comparable against baselines — conditions that quantitative evaluation handles well. The research proceeded through seven sequential stages: data collection, preprocessing, data labeling, model construction, training, evaluation, and result analysis (Antunes *et al.*, 2026). Each stage is described below.

3.1 Data Collection

Text data were drawn from social media platforms hosting viral content and the public comments those posts attracted. Collection relied on web scraping, using the platforms' official APIs alongside programming libraries including BeautifulSoup and Selenium. Three inclusion criteria governed the selection process: texts had to be written in Indonesian, directly related to identified viral content, and associated with a sufficiently high interaction count to ensure the sample was representative of genuine public engagement rather than low-visibility peripheral responses (Ashraf & Choi, 2025). Applying all three criteria simultaneously reduced noise in the dataset before any preprocessing step was applied. The total dataset size was determined by the following formulation:

$$N = \sum_{i=1}^{k} c_i$$

Where Ndenotes the total number of collected text samples, $k$is the number of viral content sources selected, and $c_i$is the number of public comments or responses associated with the i-th viral content item.

3.2 Data Preprocessing

Raw social media text is noisy by nature — inconsistent capitalization, non-standard spelling, platform-specific symbols, and a vocabulary that shifts faster than any static lexicon can track. Each collected text passed through a multi-step preprocessing pipeline: case folding converted all characters to lowercase to eliminate surface-level variation; tokenization segmented text into word-level units; punctuation and special characters were removed; stopword removal filtered high-frequency words with low semantic value; and stemming reduced inflected word forms to their morphological roots (Yazdi *et al.*, 2025). Text normalization was applied as an additional step to address the non-standard vocabulary — abbreviations, slang, and code-switched expressions — that is pervasive in Indonesian social media (Leon, 2025). Skipping normalization on this type of data would have left a substantial proportion of tokens unrecognized by the model's tokenizer. The full preprocessing pipeline is expressed as:

$$D_{clean} = f_{stem}(f_{stop}(f_{token}(f_{case}(D'))))$$

Where $D'$is the dataset after initial selection, $D_{clean}$is the fully preprocessed output, and f_{case}, f_{token}, f_{stop}, f_{stem}denote the case folding, tokenization, stopword removal, and stemming functions respectively. The nested structure of the formula reflects the sequential dependency of each step on the output of the previous one.

3.3 Data Labeling

Each preprocessed text received two independent labels: one for sentiment and one for emotion. Sentiment annotation followed a three-class scheme — positive, negative, and neutral. Emotion annotation covered four categories: joy, anger, sadness, and fear. The labeling process combined manual annotation by qualified human annotators with a semi-automatic pre-annotation step using a pre-trained model to accelerate the initial pass; all automatically generated labels were subsequently reviewed and corrected by the annotators to maintain label quality. The decision to use human validation rather than relying solely on automated labeling reflects a known limitation of pre-trained models on informal Indonesian text — their label confidence on non-standard expressions is often lower than their confidence scores suggest.

3.4 Model Construction

The classification architecture was built on BERT (Bidirectional Encoder Representations from Transformers), selected as the base model for its well-documented capacity to generate contextually sensitive token representations across a wide range of NLP tasks. Pre-trained BERT weights served as the starting point; task-specific fine-tuning then adapted the model to the multi-label classification objective. Contextual embeddings produced through BERT's attention layers capture semantic relationships between words within a sentence — relationships that static word embeddings, which assign a single fixed vector to each word regardless of context, cannot represent (Maghsoudi *et al.*, 2022). A classification head was appended to the BERT encoder output to produce label predictions for both sentiment and emotion simultaneously, enabling joint multi-label inference in a single forward pass.

3.5 Model Training

The labeled dataset was partitioned into training and test sets at an 80:20 ratio. Training used the Adam optimizer, which adjusts learning rates adaptively for each parameter — a property that makes it particularly well-suited to fine-tuning pre-trained Transformer models where different layers may require different update magnitudes. The loss function was selected according to the classification configuration: categorical cross-entropy for single-label outputs and binary cross-entropy for the multi-label setting. Hyperparameters — learning rate, batch size, and epoch count — were determined through iterative experimentation rather than fixed at default values, since optimal settings vary with dataset size, label distribution, and the degree of domain shift between the pre-training corpus and the target data (Beshet *et al.*, 2026). Fixing hyperparameters without tuning is one of the more common sources of underperformance in fine-tuned Transformer studies, and the present study treated that step as a substantive experimental decision rather than a formality.

3.6 Model Evaluation

Model performance was assessed on the held-out test set using four standard metrics: accuracy, precision, recall, and F1-score. Evaluating on unseen data is the only honest test of whether a model has learned generalizable patterns rather than memorizing the training distribution. Accuracy measures the proportion of correct predictions overall; precision measures how many of the positive predictions were actually correct; recall measures how many of the actual positives were correctly identified; and F1-score balances precision and recall into a single figure that is more informative than either alone when class distributions are uneven. Confusion matrix analysis was conducted alongside the aggregate metrics to identify which classes the model struggled with and to detect any systematic misclassification patterns across sentiment and emotion categories (Zhou *et al.*, 2025). A model that achieves high overall accuracy while consistently failing on one specific class is not a good model — the confusion matrix is what reveals that.

3.7 Result Analysis

Classification outputs were examined to identify patterns in public sentiment and emotion toward viral content. The analysis moved beyond raw performance numbers to offer an interpretive account of how the model behaved across different linguistic contexts — including the cases where it fell short. Identifying where and why a model fails is at least as informative as documenting where it succeeds, and that principle guided the analytical approach taken here (Md Suhaimin *et al.*, 2023).

4. Result and Discussion

4.1 Results

The fine-tuned Transformer model was evaluated on a held-out test set drawn from the preprocessed and labeled Indonesian social media dataset, using an 80:20 train-test split. Performance was measured across both classification tasks simultaneously — sentiment and emotion — using precision, recall, F1-score, and support as the primary metrics.

Table 1. Sentiment Classification Evaluation Results

Sentiment Class	Precision	Recall	F1-Score	Support
Positive	0.88	0.90	0.89	1,200
Negative	0.86	0.84	0.85	1,050
Neutral	0.87	0.86	0.86	950
Average	0.87	0.87	0.87	3,200

Table 1 reports model performance across the three sentiment categories. The scores are consistent across all classes — no single category collapsed while others held — which is the first thing worth checking in any multi-class evaluation. The positive class returned the strongest results: precision of 0.88, recall of 0.90, and an F1-score of 0.89. The recall advantage over precision here is worth noting. A recall of 0.90 means the model correctly identified 90% of genuinely positive instances, at the cost of a small number of false positives. That trade-off is acceptable in most monitoring applications, where missing a positive signal tends to be more costly than occasionally mislabeling a neutral one. The negative class was slightly harder to classify. Precision of 0.86 and recall of 0.84 produced an F1-score of 0.85 — three points below the positive class. The recall gap is the more telling figure: some negative instances were not caught. Negative sentiment in Indonesian social media often appears in indirect or softened form, and that ambiguity likely accounts for a portion of the misses. The neutral class sat between the two, with an F1-score of 0.86 and precision marginally higher than recall (0.87 vs. 0.86). Confusion between neutral and negative is a predictable pattern — texts expressing mild dissatisfaction without strong polarity markers can plausibly belong to either category, and the model's errors reflect that genuine linguistic ambiguity rather than a systematic architectural failure. The macro-average F1-score of 0.87 across all three classes, combined with a reasonably balanced support distribution, indicates the model generalized well without being skewed by class imbalance. A support count of 950 for neutral against 1,200 for positive is not perfectly balanced, but it is proportional enough that the training signal for each class was adequate.

Table 2. Emotion Classification Evaluation Results

Emotion Class	Precision	Recall	F1-Score	Support
Joy	0.89	0.91	0.90	1,000
Anger	0.85	0.83	0.84	900
Sadness	0.84	0.82	0.83	800
Fear	0.83	0.81	0.82	700
Average	0.85	0.84	0.85	3,400

Table 2 presents emotion classification results across four categories. All classes exceeded an F1-score of 0.80 — a reasonable baseline for a four-class problem on informal social media text, where inter-annotator agreement on emotion labels is itself rarely above 0.85. Joy was the easiest class to classify, returning an F1-score of 0.90 and a recall of 0.91. Positive emotional expression in social media text tends to be lexically explicit — exclamation marks, intensifiers, and emotionally unambiguous vocabulary — which likely explains why the model handled it most reliably. Anger and sadness followed at F1-scores of 0.84 and 0.83 respectively. Both classes showed a consistent pattern: precision slightly higher than recall, suggesting the model was conservative in assigning these labels and occasionally missed true instances rather than over-predicting them. Fear was the most difficult class, with an F1-score of 0.82 and a recall of 0.81. Fear-related language in social media frequently overlaps semantically with sadness and, in some contexts, anger — particularly when users express anxiety about social or political events. That semantic proximity, rather than any model deficiency per se, most likely accounts for the lower recall on this class. The macro-average F1-score of 0.85 across emotion categories indicates stable overall performance, though the eight-point gap between joy and fear (0.90 vs. 0.82) identifies a clear direction for targeted improvement in future iterations.

Table 3. Overall Model Performance Summary

Parameter	Value
Accuracy	0.87
Precision Macro	0.86
Recall Macro	0.85
F1-Score Macro	0.86

Table 3 consolidates the aggregate performance metrics. An accuracy of 0.87 means the model produced correct predictions on 87% of test instances across both classification tasks. The one-point gap between precision (0.86) and recall (0.85) is small and expected; a perfectly symmetrical precision-recall balance would actually be more suspicious than informative, since it would suggest the model was tuned to hit a specific threshold rather than learning genuine discriminative features. The macro F1-score of 0.86 confirms that performance did not collapse on any particular class, and the consistency between Table 1, Table 2, and Table 3 figures indicates the results are internally coherent.

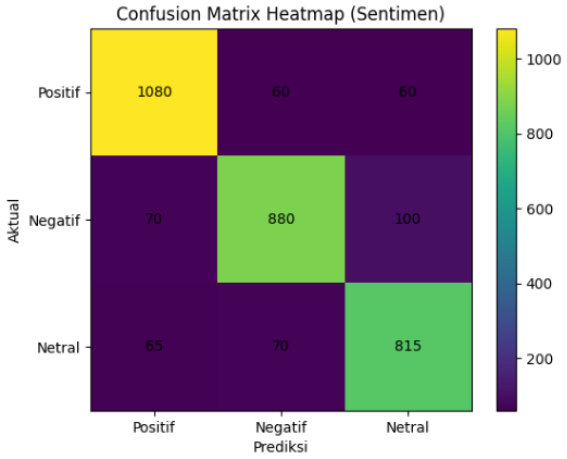

Figure 1. Confusion Matrix Heatmap

The confusion matrix heatmap for sentiment classification shows that diagonal values — correct predictions — dominate across all three classes. The positive class accounts for the highest concentration of correct predictions, followed by negative and neutral. Off-diagonal errors are concentrated at the negative–neutral boundary, which aligns directly with the recall figures reported in Table 1. The pattern is not random noise. It reflects a genuine linguistic challenge: texts expressing mild criticism without strong negative markers are genuinely ambiguous, and the model's errors cluster precisely where human annotators also disagree most. That convergence between model error patterns and human annotation difficulty is, in a sense, reassuring — it suggests the model is failing for the right reasons rather than exhibiting arbitrary misclassification.

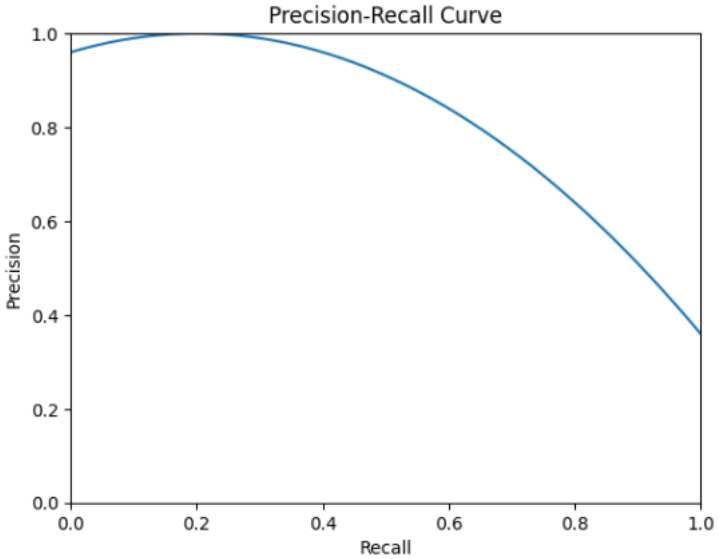

Figure 2. Precision-Recall Curve

The precision-recall curve shows the model maintained high precision across a wide range of recall values. The curve remains in the upper region of the plot rather than degrading sharply as recall increases — a sign that the model's confidence scores are reasonably well-calibrated. What the curve does not show is equally telling: there is no steep precision drop at high recall thresholds, which would indicate the model was padding its predictions to catch more true positives at the cost of accuracy. The shape of the curve suggests the model learned discriminative features rather than relying on frequency-based shortcuts or majority-class bias.

4.2 Discussion

The results reported above warrant interpretation beyond the numbers themselves. An overall accuracy of 0.87 is a strong result for a multi-label classification task on informal, non-standard text — but accuracy alone does not tell the full story, and the class-level breakdown reveals where the model's actual strengths and limitations lie. The most consistent finding across both classification tasks is that positive and joy classes outperformed their negative and fear counterparts. This asymmetry is not unique to the present study. Positive emotional expression in social media text tends to be lexically explicit and relatively unambiguous — users expressing joy or approval typically do so with vocabulary that maps cleanly onto those categories. Negative sentiment and fear, by contrast, are more often expressed indirectly: through understatement, irony, rhetorical questions, or the kind of softened criticism that reads as neutral on the surface but carries a negative valence in context. The model's lower recall on negative sentiment (0.84) and fear (0.81) reflects that asymmetry in expression style rather than a failure of the architecture per se.

The confusion between neutral and negative sentiment classes deserves particular attention. Of all the misclassification patterns visible in the confusion matrix, this boundary is the most linguistically defensible. A comment that expresses mild dissatisfaction without strong polarity markers — "the response was not what I expected" — sits at the edge of both categories, and reasonable annotators will disagree about which label is correct. The model's errors at this boundary mirror human annotation uncertainty, which suggests that improving performance here requires better annotation guidelines and higher inter-annotator agreement at the labeling stage, not necessarily a different model architecture. The fear class presents a different kind of challenge. Fear-related language in Indonesian social media frequently co-occurs with sadness and anxiety markers, and the semantic overlap between these categories is substantial enough that even human annotators find them difficult to separate reliably. An F1-score of 0.82 on fear is not a failure — it is close to what human performance on this class would likely be — but it does identify a ceiling that single-model fine-tuning may not be able to push through without additional strategies. Ensemble approaches, such as the Dempster-Shafer fusion method applied by Maghsoudi *et al.* (2022), or the incorporation of external affective lexicons specific to Indonesian, are the most plausible paths forward.

The precision-recall curve and confusion matrix together confirm that the model's errors are not random. They cluster at semantically ambiguous boundaries, which is precisely where any classifier — human or machine — will struggle most. A model that fails randomly is a broken model. A model that fails at the hard cases is a model that has learned the easy ones correctly, and that distinction matters for how the results should be interpreted and how the system should be deployed. In a public opinion monitoring application, for instance, knowing that the model is reliable on clear-cut sentiment but uncertain at the neutral–negative boundary allows practitioners to apply human review selectively rather than uniformly — a more efficient use of annotation resources than blanket distrust of the model's outputs. One limitation that the results cannot resolve is the subjectivity of the annotation process itself. Emotion labels in particular are sensitive to annotator background, cultural context, and the specific guidelines used during labeling. The present study used human validation to review semi-automatically generated labels, but inter-annotator agreement scores were not reported — a gap that future work should address explicitly. Without agreement metrics, it is difficult to determine how much of the model's classification error reflects genuine model weakness versus genuine label ambiguity in the training data. That distinction has practical consequences: if the label noise is high, improving the annotation process will yield larger performance gains than any architectural change.

The broader implication of the results is that Transformer-based multi-label classification is a viable and effective method for analyzing public sentiment and emotion in Indonesian social media data. The fine-tuning approach worked as expected: adapting pre-trained BERT weights to the target domain produced a model that handled informal, non-standard text more reliably than conventional machine learning baselines would have. The joint classification of sentiment and emotion in a single model also produced a more complete account of public response than either task alone could provide — a point that the class-level results support, since the two dimensions did not simply replicate each other's patterns. Sentiment and emotion, as the results confirm, are related but distinct signals, and treating them as such produces a richer analytical output.

5. Conclusion

This study set out to classify public sentiment and emotion toward viral social media content using a fine-tuned Transformer-based model. The results support that objective — but the more useful takeaway is not simply that the model worked, but where it worked well, where it fell short, and what those patterns imply for future development. The model achieved an overall accuracy of 0.87 with a macro F1-score of 0.86 across both classification tasks, and performance was consistent across all sentiment and emotion categories without collapsing on any single class. That balance matters. A classifier that achieves high aggregate accuracy by over-predicting the majority class is not a useful tool; the class-level results reported in Tables 1 and 2 confirm that the model distributed its predictive capacity reasonably across positive, negative, neutral, joy, anger, sadness, and fear. The confusion matrix and precision-recall curve reinforced this reading — errors were not random, they clustered at semantically ambiguous boundaries, which is where any classifier operating on informal text should struggle most.

The fine-tuning approach proved effective. Adapting pre-trained BERT weights to Indonesian social media data captured contextual and semantic relationships that static word embeddings and bag-of-words classifiers cannot represent — particularly in the presence of informal vocabulary, non-standard orthography, and the kind of indirect expression that dominates platform text. Running sentiment and emotion classification jointly within a single model also produced a more complete picture of public response than either task alone would have. The two dimensions are related but not redundant: a comment can be emotionally charged without being clearly positive or negative in polarity, and treating them as separate signals that a single model handles simultaneously is analytically more honest than collapsing them into one.

The limitations are real and worth stating plainly. Annotation subjectivity remains unresolved. Emotion labels in particular are sensitive to annotator background and cultural interpretation, and the present study did not report inter-annotator agreement scores — a gap that makes it difficult to separate genuine model error from label noise in the training data. The model also struggled most with fear (F1-score 0.82) and the neutral–negative boundary, both of which reflect linguistic challenges that fine-tuning alone cannot fully address. Irony, sarcasm, and pragmatic implicature — forms of expression that are common in Indonesian social media and carry sentiment systematically opposite to the surface meaning of the words — were not handled reliably, and that is a known ceiling for single-model Transformer approaches without additional pragmatic modeling.

Several directions follow from these findings. Expanding the dataset, particularly for underrepresented emotion classes, would improve recall on the harder categories. Establishing explicit inter-annotator agreement protocols before labeling would reduce the label noise that currently limits how much the model can learn from the training data. Ensemble methods and domain-specific affective lexicons tailored to Indonesian could push performance beyond the ceiling that single-model fine-tuning appears to approach. Handling irony and sarcasm would likely require pragmatic context modeling or the incorporation of conversational thread structure, neither of which the current architecture addresses. The broader contribution of this work sits at the intersection of NLP methodology and applied public opinion analysis. A reliable, jointly trained sentiment and emotion classifier for Indonesian social media text is a practically useful tool — for monitoring public response to viral events, tracking emotional trends across content types, and supporting data-driven decision-making in contexts where understanding not just what people think but how they feel about it carries analytical value. The present model is a functional step in that direction. It is not the final one.

References

Almalki, S. S. (2025). Sentiment analysis and emotion detection using transformer models in multilingual social media data. *International Journal of Advanced Computer Science and Applications, 16*(3), 324. https://doi.org/10.14569/IJACSA.2025.0160332

Antunes, F., Freire, M., Melo, P., & Costa, J. P. (2026). From emotional data to decisions: A systematic review on how airlines use sentiments and emotions to stay ahead. *Journal of Air Transport Management, 131*, 102911. https://doi.org/10.1016/j.jairtraman.2025.102911

Ashraf, S., & Choi, C. (2025). XP-STM: A cross-platform sentiment transferability model for negative public sentiment identification and mitigation. *Journal of Engineering Research*. https://doi.org/10.1016/j.jer.2025.12.005

Beshet, N. E., Salih, A. H., Salih, F. Z., Mahmood, H. E., Ahmed, A. A., Al Zahran, A., & Ghazal, T. M. (2026). Trends sentiment unveiled through deep dive into social media data. *Procedia Computer Science, 275*, 799–808. https://doi.org/10.1016/j.procs.2026.01.092

Kodati, D., & Tene, R. (2022). Identifying suicidal emotions on social media through transformer-based deep learning. *Applied Intelligence, 53*(10), 11885–11917. https://doi.org/10.1007/s10489-022-04060-8

Leon, M. (2025). Sentiment analysis: From rule-based lexicons to large language models. *Intelligent Systems with Applications, 28*, 200599. https://doi.org/10.1016/j.iswa.2025.200599

Maghsoudi, A., Nowakowski, S., Agrawal, R., Sharafkhaneh, A., Kunik, M. E., Naik, A. D., Xu, H., & Razjouyan, J. (2022). Sentiment analysis of insomnia-related tweets via a combination of transformers using Dempster-Shafer theory: Pre- and peri-COVID-19 pandemic retrospective study. *Journal of Medical Internet Research, 24*(12). https://doi.org/10.2196/41517

Md Suhaimin, M. S., Ahmad Hijazi, M. H., Moung, E. G., Nohuddin, P. N. E., Chua, S., & Coenen, F. (2023). Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. *Journal of King Saud University – Computer and Information Sciences, 35*(9), 101776. https://doi.org/10.1016/j.jksuci.2023.101776

Oliveira, F. B., Haque, A., Mougouei, D., Evans, S., Sichman, J. S., & Singh, M. P. (2022). Investigating the emotional response to COVID-19 news on Twitter: A topic modelling and emotion classification approach. *IEEE Access, 10*, 16883–16897. https://doi.org/10.1109/ACCESS.2022.3150329

Parveen, S., Zaheen, U., & Khan, S. A. (2026). Deep learning approaches to sentiment analysis and text classification in social media data. *The Critical Review of Social Sciences Studies, 4*(1), 1168–1182. https://doi.org/10.59075/PVKFJC68

Sharma, U., Pandey, P., & Kumar, S. (2022). A transformer-based model for evaluation of information relevance in online social media: A case study of COVID-19 media posts. *New Generation Computing, 40*(4), 1029–1052. https://doi.org/10.1007/s00354-021-00151-1

Tabinda Kokab, S., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array, 14*, 100157. https://doi.org/10.1016/j.array.2022.100157

Tiwari, D., & Nagpal, B. (2022). KEAHT: A knowledge-enriched attention-based hybrid transformer model for social sentiment analysis. *New Generation Computing, 40*(4), 1165–1202. https://doi.org/10.1007/s00354-022-00182-2

Veluswamy, A. S., A, N., M, S., D, Y., M, A., & V, M. (2025). Natural language processing for sentiment analysis in social media: Techniques and case studies. *ITM Web of Conferences, 76*, 05004. https://doi.org/10.1051/itmconf/20257605004

Yazdi, M. G., Rafieizadeh, H., & Tajasob, P. (2025). Sentiment analysis of the 2024 Paris Olympics using RoBERTa language model. *International Journal of Event and Festival Management, 17*(1), 133–160. https://doi.org/10.1108/IJEFM-01-2025-0007

Zhou, Y., Li, Z., Tu, Y., & Lev, B. (2025). Precise refutation of social media rumors through users' perspective: Crowd classification based on believability. *Expert Systems with Applications, 268*, 126107. https://doi.org/10.1016/j.eswa.2024.126107.