International Journal Of Humanities Education And Social Sciences ***(IJHESS)*** *E-ISSN:* **2808-1765**
Volume 6, Number 1, August 2026, *Page. 261 – 270*
Email : editorijhess@gmail.com

YouTube Sentiment Analysis on Felt Earthquake News Using LSTM and IndoBERT

Oktifar Tri Bandono[1)]*, Agung Budi Susanto[2)], Makhsun[3)]
[1,2,3)] Informatics Engineering, Computer Science Faculty, Universitas Pamulang, Indonesia

*Corresponding Author
Email: oktifartb@gmail.com

Abstract

This study analyzes public sentiment in YouTube comments on felt-earthquake news in Indonesia and compares a bidirectional Long Short-Term Memory implementation (LSTM) with IndoBERT. An experimental quantitative design was used. Comments were collected through the YouTube Data API v3 using official earthquake-event references from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) and keyword-based video searches covering January 2021 to August 2025. The acquisition stage produced 51,870 comments from 1,626 unique videos. Data were processed through duplicate removal, text cleaning, case folding, slang normalization, tokenization, and stopword removal, resulting in 49,041 clean comments. Positive, negative, and neutral labels were assigned by aggregating word-polarity scores from the Indonesian Sentiment Lexicon (InSet), which served as weak supervision. Stratified sampling divided the dataset into 80% training data and 20% testing data. The LSTM model used a 100-dimensional embedding, a 64-unit bidirectional LSTM layer, global max pooling, and early stopping; IndoBERT was fine-tuned from indobenchmark/indobert-base-p2 for four epochs. Performance was assessed with accuracy, macro precision, macro recall, macro F1-score, and confusion matrices. Positive sentiment accounted for 41.2% of the corpus, negative sentiment for 35.5%, and neutral sentiment for 23.2%. IndoBERT achieved 91.50% accuracy and a 91.03% macro F1-score, outperforming LSTM at 90.91% accuracy and a 90.34% macro F1-score. IndoBERT provided the strongest contextual classification, while LSTM remained a competitive and substantially lighter option for resource-constrained monitoring.

Keywords: Sentiment Analysis, Earthquake, Youtube, Lstm, Indobert, Disaster Communication

INTRODUCTION

Indonesia is one of the world's most seismically active countries because its territory lies within a complex convergence zone involving the Indo-Australian, Eurasian, and Pacific plates as well as several smaller crustal blocks (Bird, 2003). Consequently, felt earthquakes occur frequently and can affect communities across widely separated islands. Their effects are not limited to physical shaking or infrastructure damage. Earthquake events also produce rapid psychological and social responses, including anxiety, requests for information, criticism of institutions, expressions of grief, prayer, and solidarity. Understanding these responses is relevant to disaster communication because public interpretation of an event can influence trust, preparedness, and the circulation of official or unverified information.

Digital platforms make these reactions observable soon after an earthquake is felt. YouTube is particularly important in Indonesia because national and regional media channels upload reports that summarize magnitude, location, damage, eyewitness accounts, and statements from public authorities. The comment area then becomes a parallel public forum in which viewers respond to both the disaster and the way it is communicated. Compared with questionnaires, comments are unsolicited and can be collected at much larger scale. They therefore provide useful, although noisy, evidence of public emotion and information needs. Studies of earthquake-related social media have similarly shown the value of unstructured text for spatial and public-response analysis (Nursiyono & Gibran, 2023).

Sentiment analysis is a Natural Language Processing task that identifies evaluative orientation in text, commonly expressed as positive, negative, or neutral polarity. It has been applied to product reviews, political discussion, health communication, and disaster-related

messages (Medhat et al., 2014; Safitri et al., 2025). In disaster communication, polarity must be interpreted carefully. A positive label does not mean that an earthquake is regarded favorably; it may represent a prayer, hope for safety, gratitude for assistance, or solidarity with victims. Negative sentiment may include fear, grief, anger, or criticism, while neutral comments often contain factual questions. The value of sentiment analysis lies in summarizing these patterns at scale while retaining an interpretation grounded in the disaster context.

Indonesian YouTube comments present substantial modeling challenges. Users frequently employ abbreviated spellings, non-standard words, regional vocabulary, code-mixing, repeated characters, religious phrases, irony, and short questions with little explicit emotion. Lexicon-based methods are transparent and efficient for generating labels from large corpora, but their reliance on word-level polarity can miss negation and context (Taboada et al., 2011). These difficulties motivate the use of deep neural models that learn representations from sequences or contextual relationships rather than depending only on surface word counts.

Long Short-Term Memory networks were designed to reduce the vanishing-gradient problem in recurrent sequence learning and to retain information over longer dependencies (Hochreiter & Schmidhuber, 1997). For short social-media text, an LSTM can learn ordered lexical patterns with a relatively compact model. Transformer models instead use self-attention to connect every token with other tokens in a sequence (Vaswani et al., 2017). BERT extends this approach through bidirectional pretraining and task-specific fine-tuning (Devlin et al., 2019). IndoBERT is pretrained on Indonesian corpora and has demonstrated the benefit of language-specific contextual representation for Indonesian NLP benchmarks (Koto et al., 2020).

Previous work has applied conventional classifiers and deep learning to Indonesian social-media and disaster text. However, several gaps remain for felt-earthquake news on YouTube. First, many studies use Twitter/X or a small set of videos, whereas YouTube comments can be longer and more narrative. Second, direct comparisons between a lightweight recurrent architecture and an Indonesian pretrained transformer are limited in this domain. Third, the cultural meaning of prayers and solidarity can complicate the interpretation of automatically generated sentiment labels. A larger, multi-year corpus is therefore needed to examine both classification performance and the practical meaning of the resulting sentiment distribution.

This study contributes: (1) a corpus of 49,041 cleaned Indonesian comments derived from 1,626 videos covering multiple felt-earthquake events between 2021 and 2025; (2) a reproducible pipeline combining API-based acquisition, Indonesian text normalization, InSet-based weak labeling, and stratified evaluation; (3) an empirical comparison of LSTM and IndoBERT under the same three-class task; and (4) an error analysis that connects model behavior with informal language and disaster communication. The use of weak supervision makes large-scale experimentation feasible, while the comparison clarifies the tradeoff between contextual accuracy and computational efficiency.

The study addresses three questions: How can Indonesian YouTube comments on felt-earthquake news be preprocessed and labeled for sentiment analysis? How do LSTM and IndoBERT perform according to accuracy, precision, recall, F1-score, and confusion matrices? How can the two algorithms be implemented for this domain, and what practical tradeoffs follow from their results? The objective is not only to identify the highest-scoring model but also to determine whether a smaller recurrent model remains suitable for resource-constrained, near-real-time disaster monitoring.

RESEARCH METHODS

Research design and data source

The study used an experimental quantitative design in which two supervised text-classification models were trained and tested on the same labeled corpus. The unit of analysis was one public YouTube comment. The target population comprised comments on videos reporting felt earthquakes in Indonesia, while purposive inclusion criteria required that a video report an actual felt-earthquake event, that a comment be relevant and primarily Indonesian, and that duplicate or spam-like text be excluded. BMKG event records were used as the factual reference for event dates and locations. The observation period ran from January 2021 through August 2025.

Data collection used the YouTube Data API v3 and two complementary acquisition modes. Mode A began with 77 manually curated videos linked to 21 significant BMKG-referenced earthquake events and returned 14,559 comments. Mode B used the Search API with 20 earthquake-related queries to identify 1,549 additional videos and returned 37,311 comments. Fields retained during acquisition included the video identifier, channel title, publication time, comment text, like count, and reply count. The two modes produced 51,870 comments from 1,626 unique videos after video-level deduplication and period filtering.

Table 1. Summary of the YouTube comment acquisition

Acquisition mode	Unique videos	Comments
Mode A: manually curated BMKG-linked videos	77	14,559
Mode B: keyword-based Search API videos	1,549	37,311
Combined dataset	1,626	51,870

Source: Processed research data, 2026

Text preprocessing and sentiment labeling

The preprocessing pipeline was applied in a fixed order. Exact duplicate comments were removed first. Cleaning then deleted URLs, user mentions, HTML tags, punctuation, numbers, emojis, and non-text symbols. Case folding converted text to lowercase. A dictionary containing 4,347 non-standard-to-standard word pairs normalized common Indonesian abbreviations and slang, such as 'gak' to 'tidak' and 'smoga' to 'semoga'. Tokenization divided each comment into word tokens, and Indonesian stopwords were removed using an NLTK-based list supplemented with context-specific social-media terms. The process reduced 49,899 unique comments to 49,041 clean comments.

Sentiment labels were generated with the Indonesian Sentiment Lexicon (InSet). For each preprocessed comment, token polarity weights were summed. A total score greater than zero produced a positive label, a score below zero produced a negative label, and a score of zero produced a neutral label. The resulting labels were treated as weak supervision rather than as infallible human judgments. A manual review of 500 randomly selected comments was used as a consistency check and to identify ambiguous expressions for the later qualitative error analysis. This design enabled the full corpus to be labeled consistently while acknowledging that sarcasm, negation, and religious expressions may exceed word-level rules.

Table 2. Examples of the preprocessing transformation

Original comment	After cleaning and normalization	Final clean text
turut berduka cita untuk korban yg meninggal smoga amal ibadahnya diterima amin	turut berduka cita untuk korban yang meninggal semoga amal ibadahnya diterima amin	berduka cita korban meninggal semoga amal ibadah diterima amin
ya allah padahal di kalsel banjir sekarang gempa sulbar semoga selamat semua yg ada di sana	ya allah padahal di kalsel banjir sekarang gempa sulbar semoga selamat semua yang ada di sana	allah kalsel banjir gempa sulbar semoga selamat
musik penutupnya gak ngerasain orang lagi kesusahan @kompastv	musik penutupnya tidak merasakan orang lagi kesusahan	musik penutup merasakan orang kesusahan

Source: Processed research data, 2026

Dataset splitting and model specification

The weakly labeled corpus was divided using stratified sampling so that the positive, negative, and neutral proportions were identical in the training and testing subsets. Eighty percent of the data (39,232 comments) were assigned to training and 20% (9,809 comments) to the held-out test set. For the LSTM experiment, 10% of the training subset was used as internal validation data; the held-out test set was not used for model selection. Stratification was important because the neutral class was smaller than the two polar classes.

Table 3. Stratified training and testing split

Sentiment	Training (80%)	Testing (20%)	Total
Positive	16,169 (41.2%)	4,043 (41.2%)	20,212
Negative	13,945 (35.5%)	3,487 (35.5%)	17,432
Neutral	9,118 (23.2%)	2,279 (23.2%)	11,397
Total	39,232	9,809	49,041

Source: Processed research data, 2026

LSTM experiment

The recurrent experiment used a bidirectional LSTM, referred to as LSTM in the comparison. The maximum input length was 22 tokens, corresponding to the 95th percentile of comment length, and the vocabulary was limited to 10,000 tokens. The architecture combined a 100-dimensional embedding layer, SpatialDropout1D at 0.2, a 64-unit bidirectional LSTM layer, GlobalMaxPool1D, a 64-unit dense ReLU layer, dropout at 0.3, and a three-unit softmax output. It contained 1,092,931 trainable parameters. Training used Adam with a learning rate of 1 x 10^-3, sparse categorical cross-entropy, batch size 64, and a maximum of 15 epochs. Early stopping monitored validation accuracy with patience of four epochs and restored the best weights.

Table 4. LSTM architecture used in the experiment

Layer	Output shape	Parameters	Function
Embedding	(None, 22, 100)	1,000,000	100-dimensional token representation

Layer	Output shape	Parameters	Function
SpatialDropout1D	(None, 22, 100)	0	Feature dropout at 0.2
Bidirectional LSTM	(None, 22, 128)	84,480	64 units in two directions
GlobalMaxPool1D	(None, 128)	0	Strongest sequence features
Dense + ReLU	(None, 64)	8,256	Nonlinear classification features
Dropout	(None, 64)	0	Regularization at 0.3
Dense + softmax	(None, 3)	195	Three sentiment probabilities
Total	-	1,092,931	Trainable parameters

Source: Processed research data, 2026

IndoBERT experiment

The transformer experiment fine-tuned indobenchmark/indobert-base-p2 with a BertForSequenceClassification head containing three outputs. WordPiece tokenization and a maximum sequence length of 128 were used. The optimization settings were a learning rate of 2 x 10^-5, batch size 16, 100 warmup steps, weight decay 0.01, and four epochs. Fine-tuning ran on an NVIDIA Tesla T4 GPU. Unlike the LSTM, which learned task-specific representations from the experiment corpus, IndoBERT began with pretrained Indonesian contextual representations and adapted them to the earthquake-comment labels.

Evaluation procedure

Both models were evaluated on the same 9,809 held-out comments. Accuracy measured the proportion of all correct predictions. Precision, recall, and F1-score were calculated for each class, and macro averages gave equal weight to positive, negative, and neutral categories. Macro F1-score was emphasized because it balances precision and recall without allowing the largest class to dominate the summary (Sokolova & Lapalme, 2009). Confusion matrices were inspected to identify which class boundaries produced the most errors.

The comparison proceeded at three levels: overall metrics, class-level metrics, and qualitative review of misclassified examples. Training and test results were kept distinct. LSTM validation history was used only for early stopping, whereas final model claims were based on the held-out test set. Cross-model error agreement was also counted to determine whether one architecture consistently corrected the other's errors.

RESULTS AND DISCUSSION

Data preparation and sentiment distribution

The collection stage returned 51,870 comments. Exact duplicate removal eliminated 1,971 entries (3.80%), leaving 49,899 unique comments. Cleaning, normalization, tokenization, and stopword removal produced 49,041 usable comments. The mean cleaned length was 8.0 words, the median was 5 words, and the 95th percentile was 22 words. These statistics supported the LSTM maximum length of 22 and confirmed that most items were short.

InSet scoring assigned 20,212 comments to the positive class, 17,432 to the negative class, and 11,397 to the neutral class. The class proportions were not equal, but no class was rare enough to prevent stratified training and evaluation. Table 5 reports both the numerical distribution and the dominant form of expression observed during qualitative inspection.

Table 5. Distribution of lexicon-based sentiment labels

Sentiment	Comments	Percentage	Typical expression
Positive	20,212	41.2%	Prayer, hope, gratitude, solidarity
Negative	17,432	35.5%	Grief, anxiety, criticism, concern
Neutral	11,397	23.2%	Facts, locations, magnitude, questions
Total	49,041	100%	-

Source: Processed research data, 2026

Training behavior

LSTM training stopped after seven epochs. Training accuracy rose from 79.77% to 97.81%, while the highest validation accuracy, 90.80%, occurred at epoch 3. Validation loss declined through epoch 2 and then increased as training accuracy continued to improve, indicating mild overfitting. The early-stopping procedure restored the epoch-3 weights before test evaluation.

Table 6. LSTM training history

Epoch	Train acc.	Val. acc.	Train loss	Val. loss	Note
1	79.77%	89.07%	0.5076	0.3090	
2	92.94%	90.57%	0.2110	0.2708	
3	95.19%	90.80%	0.1460	0.2988	Best epoch
4	96.06%	90.77%	0.1160	0.3166	
5	96.69%	90.29%	0.0973	0.3841	
6	97.25%	90.57%	0.0804	0.4102	
7	97.81%	89.98%	0.0647	0.4682	Early stop

Source: Processed research data, 2026

IndoBERT validation accuracy increased across the four fine-tuning epochs and reached 92.50% at epoch 4. No comparable divergence between training progress and validation accuracy was observed during the scheduled run. The final test accuracy was 91.50%, while the restored LSTM model achieved 90.91%.

Overall and class-level test performance

Table 7. Overall test-set performance

Model	Accuracy	Macro precision	Macro recall	Macro F1-score
LSTM	90.91%	90.41%	90.28%	90.34%
IndoBERT	91.50%	91.76%	90.48%	91.03%
Difference	0.59 pp	1.35 pp	0.20 pp	0.69 pp

Source: Processed research data, 2026

IndoBERT exceeded LSTM on every overall metric. The largest difference was 1.35 percentage points in macro precision, while the smallest was 0.20 percentage points in macro recall. Thus, the ranking was consistent but the absolute gap was modest. Both models exceeded 90% accuracy and macro F1-score on the weakly labeled test set.

Table 8. Test performance by sentiment class

Class	Model	Precision	Recall	F1-score
Negative	LSTM	90.56%	91.31%	90.93%
Negative	IndoBERT	90.88%	92.60%	91.73%
Neutral	LSTM	87.64%	86.49%	87.06%
Neutral	IndoBERT	93.24%	84.16%	88.47%
Positive	LSTM	93.03%	93.05%	93.04%
Positive	IndoBERT	91.16%	94.68%	92.89%

Source: Processed research data, 2026

The neutral class had the lowest F1-score for both models. IndoBERT produced substantially higher neutral precision than LSTM (93.24% versus 87.64%) but lower neutral recall (84.16% versus 86.49%). Positive comments had the highest LSTM F1-score, while IndoBERT obtained its highest recall on the positive class. IndoBERT also improved negative recall from 91.31% to 92.60%.

Confusion matrices and error agreement

Table 9. Confusion matrices for LSTM and IndoBERT

Actual class	LSTM: Neg.	LSTM: Neu.	LSTM: Pos.	IndoBERT: Neg.	IndoBERT: Neu.	IndoBERT: Pos.
Negative	3,184	152	151	3,229	79	179
Neutral	177	1,971	131	169	1,918	192
Positive	155	126	3,762	155	60	3,828

Source: Processed research data, 2026

The matrices show that IndoBERT reduced negative-to-neutral errors from 152 to 79 and positive-to-neutral errors from 126 to 60. Its principal increase was neutral comments predicted as positive, which rose from 131 to 192. Correct positive predictions increased from 3,762 to 3,828, and correct negative predictions increased from 3,184 to 3,229. The number of correct neutral predictions was lower for IndoBERT, consistent with its lower neutral recall.

Table 10. Cross-model error agreement

Error category	Comments	Share of test set
Incorrect in LSTM, correct in IndoBERT	442	4.51%
Incorrect in IndoBERT, correct in LSTM	384	3.91%
Incorrect in both models	450	4.59%
Total LSTM errors	892	9.09%
Total IndoBERT errors	834	8.50%

Source: Processed research data, 2026

The error sets were not identical. IndoBERT corrected 442 comments that LSTM missed, whereas LSTM corrected 384 comments missed by IndoBERT. Both models failed on 450 of the same comments. This overlap indicates a persistent subset with linguistic ambiguity, while the non-overlapping errors show that the architectures captured partly complementary signals.

Discussion

Interpretation of the model comparison

The main experimental finding is that IndoBERT provided the best overall classification, but its advantage over LSTM was small. The 0.59-point accuracy gain and 0.69-point macro F1 gain are consistent with the different representation strategies. IndoBERT can connect words across an entire comment through self-attention and begins with Indonesian knowledge acquired during pretraining. This is useful when a comment contains mixed emotional cues or when the decisive word is separated from its target. LSTM learns the sequential patterns of the task corpus and therefore remains strong when polarity follows recurring local expressions. The result supports the wider observation that pretrained language models improve context-sensitive text classification while recurrent networks can remain competitive on focused domains with substantial training data (Zhang et al., 2022).

The comparison should not be interpreted solely as a leaderboard. IndoBERT has roughly 110 million parameters, whereas the LSTM implementation has 1.09 million. The transformer is therefore about two orders of magnitude larger and requires more memory and computational support. If batch processing on a GPU is available and semantic accuracy is the priority, IndoBERT is the preferred model. If a monitoring service must run continuously on limited hardware, LSTM offers a practical alternative with less than one percentage point loss in accuracy. Model selection should accordingly consider latency, infrastructure, update frequency, and the consequences of classification errors in addition to a single aggregate score.

Class-specific and linguistic effects

Neutral sentiment was the most difficult category. Many neutral comments consisted of short questions such as requests for location or magnitude, while others were factual statements containing disaster words that also appear in negative comments. IndoBERT's high neutral precision shows that it was conservative: when it predicted neutral, it was usually correct, but it failed to retrieve a larger portion of the neutral set. This precision-recall tradeoff matters operationally. A dashboard intended to identify information requests may prefer higher neutral recall, whereas a dashboard that displays only highly reliable neutral messages may prefer IndoBERT's precision profile.

Qualitative examples clarify why contextual modeling helped. The comment 'kemarin teriak teriak aceh merdeka, ini buktinya disaat aceh tertimpa bencana ... negara republik indonesia berusaha sekuat tenaga menolong' was labeled positive. LSTM predicted it as negative, apparently responding to negative language in the opening, while IndoBERT captured the later meaning of assistance. Likewise, 'subhanallah ... inilah teguran ummat manusia lupa sang pencipta' was labeled negative; LSTM predicted positive, but IndoBERT recognized the warning expressed by the full sentence. These cases illustrate why word polarity alone or early sequence cues may be insufficient.

LSTM nevertheless corrected 384 IndoBERT errors, particularly among short comments with familiar lexical patterns. Religious expressions provide a representative difficulty. A phrase such as 'innalillahi' commonly signals grief, but in a longer Indonesian disaster comment it can coexist with prayer, acceptance, and hope. The label can therefore depend on both cultural usage and the chosen annotation rule. Hybrid IndoBERT-LSTM systems have been explored for Indonesian text in other domains (Yefferson et al., 2024), and the complementary errors observed here suggest that an ensemble is a reasonable direction for later work.

Meaning for disaster communication

Positive sentiment was the largest class, but this must be read as solidarity rather than approval of a disaster. Prayers for victims, hopes for safety, gratitude for survival, and offers of support contain positive lexical and contextual signals. Their prevalence indicates a form of digital social capital that disaster communicators can recognize and reinforce. At the same time, the substantial negative share captures anxiety, sadness, criticism, and concerns about response

or information clarity. Monitoring changes in these categories after an event could help BMKG, BNPB, and media organizations identify when official clarification or more empathetic communication is needed.

A deployable system should not treat sentiment labels as definitive psychological diagnoses. Instead, model outputs can be aggregated by time, event, video, or theme and then reviewed by analysts. Sudden increases in negative or information-seeking comments may signal uncertainty, rumors, or gaps in public messaging. Neutral questions can be especially actionable because they point directly to missing details. The system is therefore best understood as decision support: it prioritizes large comment streams for human attention rather than replacing judgment.

Limitations and future work

The study has four principal limitations. First, the ground-truth labels were generated by a lexicon and are therefore pseudo-labels; strong test performance partly measures agreement with that labeling rule. Second, InSet cannot fully resolve sarcasm, negation scope, regional-language use, code-mixing, or the contextual polarity of religious expressions. Third, purposive video discovery and API availability may underrepresent deleted, disabled, or difficult-to-find comment threads. Fourth, the data came only from YouTube and may not generalize to the shorter interaction patterns of X, TikTok, Instagram, or private messaging platforms.

Future work should create an expert-annotated disaster benchmark to estimate label validity independently of the lexicon, develop a disaster-specific Indonesian lexicon, and evaluate event-level or time-based splits. Multilingual models may improve handling of regional languages and code-mixing. Calibration, inference latency, and robustness to new earthquake events should also be measured before operational deployment. Finally, an ensemble that combines IndoBERT's contextual strength with LSTM's performance on short lexical patterns could be tested against the 826 non-overlapping errors identified in this experiment.

CONCLUSION

This study developed a large-scale sentiment-analysis pipeline for Indonesian YouTube comments on felt-earthquake news. The YouTube Data API v3, BMKG-referenced event selection, six-stage preprocessing, InSet-based weak labeling, and stratified evaluation converted 51,870 raw comments from 1,626 videos into 49,041 clean labeled comments. Positive sentiment was the largest class at 41.2%, followed by negative sentiment at 35.5% and neutral sentiment at 23.2%. In this domain, the positive class mainly represented prayer, hope, and solidarity rather than a positive evaluation of the earthquake itself.

IndoBERT achieved the best overall result with 91.50% accuracy and a 91.03% macro F1-score, compared with 90.91% accuracy and a 90.34% macro F1-score for LSTM. IndoBERT was particularly effective at reducing negative-to-neutral and positive-to-neutral errors, while LSTM remained highly competitive with far fewer parameters. IndoBERT is therefore recommended when contextual accuracy and suitable compute are available; LSTM is a credible alternative for lighter real-time monitoring. The results support sentiment monitoring as an aid to disaster communication, provided that weak labels, cultural expressions, and model uncertainty are interpreted with human oversight.

REFERENCES

Bird, P. (2003). An updated digital model of plate boundaries. Geochemistry, Geophysics, Geosystems, 4(3), 1027. https://doi.org/10.1029/2001GC000252

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186). https://doi.org/10.18653/v1/N19-1423

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Koto, F., Rahmaningtyas, F. D., & Louvan, S. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In Proceedings of AACL-IJCNLP 2020 (pp. 843-857). https://arxiv.org/abs/2009.05387

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Nursiyono, J. A., & Gibran, R. K. (2023). Natural language processing for unstructured data: Earthquakes spatial analysis in Indonesia using Twitter. Innovatics, 5(1), 22-29.

Safitri, Y. D., Faisal, M. R., Kartini, D., Saragih, T. H., Abadi, F., & Bachtiar, A. M. (2025). Automatic analysis of natural disaster messages on social media using IndoBERT and multilingual BERT. Telematika, 18(2), 105-120. https://doi.org/10.35671/telematika.v18i2.3140

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 267-307. https://doi.org/10.1162/COLI_a_00049

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://arxiv.org/abs/1706.03762

Yefferson, D. Y., Lawijaya, V., & Girsang, A. S. (2024). Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news. IAES International Journal of Artificial Intelligence, 13(2), 1913-1924. https://doi.org/10.11591/ijai.v13.i2.pp1913-1924

Zhang, L., Wang, S., & Liu, B. (2022). Deep learning for sentiment analysis: A survey. WIREs Data Mining and Knowledge Discovery, 12(1), e1435. https://doi.org/10.1002/widm.1435