JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

SELENIUM–INDOBERT PIPELINE FOR PSEUDO-LABELING SENTIMENT ANALYSIS OF INDONESIAN YOUTUBE COMMENTS

Fazli Nugraha Tambunan[1*], Heru Satria Tambunan [1], Doughlas Pardede [2]
[1]Computerized Accounting, Information System, STIKOM Tunas Bangsa
[2]Information Technology, university Deli Sumatera
email: *fazli@amiktunasbangsa.ac.id

Abstract: YouTube has become a major platform for public discourse in Indonesia, yet large-scale sentiment analysis of its comments remains challenging due to dynamic content, informal language, and limited labeled data. This study proposes a Selenium–IndoBERT pipeline for sentiment analysis of Indonesian YouTube comments using a pseudo-labeling approach. Data were collected from ten YouTube videos discussing the One Piece flag phenomenon, yielding 10,842 comments after preprocessing. Selenium was employed to extract comments from dynamic pages, while IndoBERT was fine-tuned on a small manually labeled dataset and used to generate pseudo-labels for unlabeled data. Model performance was evaluated using probabilistic metrics, including Coverage, Expected Calibration Error (ECE), and Brier Score. At a confidence threshold of 0.75, 78.5% of comments received pseudo-labels, with an ECE of 0.095 and a Brier Score of 0.174. Manual validation showed substantial agreement with human annotations (Fleiss' kappa = 0.72). The sentiment analysis revealed a polarized public response, with 42.3% positive, 34.8% negative, and 22.9% neutral comments, alongside temporal dynamics showing higher emotional intensity in early discussions. This study concludes that the proposed pipeline enables scalable and reliable sentiment analysis of Indonesian YouTube comments with minimal manual annotation, while contributing both methodological insights for semi-supervised NLP and empirical understanding of public discourse on culturally significant phenomena in digital spaces.

Keywords: IndoBERT; Indonesian language; pseudo-labeling; sentiment analysis; YouTube comments

Abstrak: YouTube telah menjadi salah satu platform utama dikursus publik di Indonesia, namun analisis sentimen komentar dalam skala besar menghadapi tantangan berupa konten dinamis, bahasa informal, dan keterbatasan data berlabel. Penelitian ini mengusulkan pipeline Selenium–IndoBERT untuk analisis sentimen komentar YouTube berbahasa Indonesia dengan pendekatan pseudo-labeling. Data dikumpulkan dari sepuluh video YouTube yang membahas fenomena bendera One Piece, menghasilkan 10.842 komentar setelah prapemrosesan. Selenium digunakan untuk mengekstraksi komentar dari halaman dinamis, sementara IndoBERT dilatih ulang menggunakan sejumlah kecil data berlabel manual dan diterapkan untuk menghasilkan pseudo-label pada data tidak berlabel. Evaluasi dilakukan menggunakan metrik probabilistik, meliputi Coverage, Expected Calibration Error (ECE), dan Brier Score. Pada ambang kepercayaan 0,75, sebanyak 78,5% komentar memperoleh pseudo-label dengan nilai ECE 0,095 dan Brier Score 0,174. Validasi manual menunjukkan tingkat kesepakatan yang substansial dengan anotator manusia (Fleiss' kappa = 0,72). Analisis sentimen mengungkap respons publik yang terpolarisasi, dengan 42,3% komentar positif, 34,8% negatif, dan 22,9% netral, serta dinamika temporal yang menunjukkan intensitas emosional lebih tinggi pada diskusi awal. Penelitian ini menyimpulkan bahwa pipeline yang diusulkan memungkinkan analisis sentimen komentar YouTube berbahasa Indonesia yang skalabel dan andal dengan anotasi manual minimal, sekaligus memberikan kontribusi metodologis untuk NLP semi-supervised dan pemahaman empiris tentang wacana publik terhadap fenomena budaya signifikan di ruang digital.

Kata kunci: analisis sentimen; bahasa Indonesia; IndoBERT; komentar YouTube; pseudo-labeling

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

The rapid growth of digital platforms in Indonesia has transformed public opinion expression, with YouTube emerging as a major space for public discourse [1]. Comment sections contain rich information reflecting public sentiment, yet their analysis poses challenges due to massive volume, unstructured content, and linguistic diversity of Indonesian netizens [2]. Manual analysis is impractical for large-scale data, while conventional automated approaches struggle with informal language characteristics including slang, abbreviations, and code-switching [3].

Recent advances in Natural Language Processing (NLP), particularly transformer-based architectures such as BERT, have demonstrated strong capabilities in capturing contextual semantic information [4]. For Indonesian, IndoBERT—a pretrained model trained on Indonesian corpora—has shown superior performance in various NLP tasks including sentiment classification [5]. However, its application to YouTube comment analysis remains underexplored, especially for culturally specific phenomena like the "One Piece flag" discourse in Indonesian online communities [6].

A primary challenge is limited labeled training data [7]. Creating large manually annotated datasets is time-consuming and costly, encouraging exploration of semi-supervised learning strategies, particularly pseudo-labeling [8]. When combined with transformer models, pseudo-labeling offers a solution for scaling sentiment analysis without extensive manual annotation. However, evaluating pseudo-label reliability requires probabilistic metrics such as Expected Calibration Error (ECE) and Brier Score [9], [10].

Previous studies have focused on individual components: web scraping techniques [11], transformer-based models for Indonesian text cla ssification [12], and pseudo-labeling methods [13]. However, few studies combine these into a unified pipeline for large-scale YouTube comment analysis in the Indonesian context [14]. Most Indonesian sentiment analysis research has concentrated on Twitter data using manually labeled datasets [15], leaving a gap in automated YouTube comment analysis through pseudo-labeling [16].

This study aims to address these gaps by proposing an integrated pipeline combining Selenium-based crawling for automated comment collection with IndoBERT-based pseudo-labeling for sentiment analysis. The research focuses on comments related to the "One Piece flag" phenomenon as a representative example of Indonesian netizen engagement with contemporary cultural symbols [17]. Unlike prior studies relying on accuracy-based evaluation, this research employs probabilistic metrics—Coverage, Average Confidence, ECE, and Brier Score—to assess pseudo-label quality and reliability [18].

The integration of Selenium and IndoBERT within a single pipeline represents a methodological contribution. Selenium enables effective extraction of dynamic JavaScript-based content, while IndoBERT provides state-of-the-art performance in capturing linguistic nuances of Indonesian text [19]. Together, these components facilitate scalable sentiment analysis without extensive manual annotation [20].

The objectives are: (1) to develop an automated pipeline for large-scale YouTube comment collection and sentiment analysis; (2) to implement and evaluate pseudo-labeling using In-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

doBERT; and (3) to assess pipeline effectiveness using probabilistic evaluation metrics.

METHOD

Data Source and Selection Criteria

The dataset was collected from ten publicly accessible Indonesian YouTube videos discussing the One Piece flag phenomenon. Selection criteria included topical relevance, audience engagement (comment volume), and diversity of content framing. Videos included news outlets, political talk shows, and independent commentary channels to capture variation in public discourse. Table 1 summarizes the analyzed videos.

Data Collection and Preprocessing

Selenium WebDriver automated comment extraction with dynamic scrolling and randomized delays (2-4 seconds). Preprocessing included lowercasing, URL removal, slang normalization using custom dictionary, and stopword removal.

Pseudo-Labeling with IndoBERT

IndoBERT-base-p1 was fine-tuned on 500 manually annotated comments (positive, negative, neutral) for 5 epochs using AdamW (learning rate 2×10^{-5}). The model then generated pseudo-labels for unlabeled comments with confidence threshold 0.75, retaining only predictions above this threshold.

Evaluation Metrics

Probabilistic metrics used:

Coverage = (N_Labeled / N_Total) × 100%

Expected Calibration Error (ECE): alignment between confidence and accuracy

Score: mean squared error between predicted probabilities and outcomes.

Experimental Setup

Experiments used Google Colaboratory with NVIDIA T4 GPU. Temporal split: first eight videos for development, remaining two for validation. Manual validation of 300 pseudo-labeled comments by three Indonesian-speaking annotators with Fleiss' kappa.

RESULT AND DISCUSSION

Data Collection and Preprocessing Outcomes

The pipeline retrieved 10,842 comments with 92.3% extraction rate. Preprocessing reduced dataset by 15.7%, removing non-textual entries and non-Indonesian comments while preserving informal language characteristics.

Sentiment Distribution and Coverage Analysis

Table 2 presents the sentiment distribution and coverage at various confidence thresholds. At threshold 0.75, 8,512 comments (78.5%) received pseudo-labels with distribution: 42.3% positive, 34.8% negative, 22.9% neutral. This polarized response suggests slightly dominant favorable sentiment with substantial critical engagement.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 1. YouTube Videos Analyzed in This Study

Video Topic (Shortened Title)	Video Topic (Shortened Title)	Video Topic (Shortened Title)
Tan Malaka and One Piece Flag	Commentary	3,775
Differing Views on One Piece Flag	News	770
Why Is the One Piece Flag Controversial?	Political Talk Show	1,621
One Piece Flag Before Independence Day	Political Talk Show	279
One Piece Flag Controversy	Commentary	821
One Piece Flag Is Not Treason	Political Talk Show	156
Palace Response to One Piece Flag	News	495
President's View on One Piece Flag	News	257
Gus Dur's Statement Revisited	Commentary	134
Legal Implications of One Piece Flag	News	115

Table 2. Sentiment Distribution and Coverage at Various Confidence Thresholds

Threshold	Coverage (%)	Positive (%)	Negative (%)	Neutral (%)	ECE	Brier Score
0.60	89.3	41.8	34.2	24.0	0.142	0.223
0.70	83.7	42.1	34.5	23.4	0.117	0.198
0.75	78.5	42.3	34.8	22.9	0.095	0.174
0.80	71.2	42.6	35.1	22.3	0.081	0.158
0.85	61.8	42.9	35.4	21.7	0.069	0.143
0.60	89.3	41.8	34.2	24.0	0.142	0.223

Lower thresholds increased coverage but reduced reliability (higher ECE, Brier Score); higher thresholds improved calibration at coverage expense. Threshold 0.75 provided optimal balance.

Probabilistic Evaluation and Model Reliability

Table 3 compares IndoBERT performance against baseline models. IndoBERT achieved ECE of 0.095 and Brier Score 0.174, indicating well-calibrated confidence estimates, consistently outperforming baseline models. This reinforces transformer-based architecture advantages for Indonesian sentiment analysis with informal language and limited labeled data.

Temporal and Source-Based Sentiment Patterns

Table 4 shows sentiment distribution by time period. Comments within first 24 hours showed higher emotional intensity (47.1% positive), declining to 38.9% in later stages, with negative sentiment increasing from 30.2% to 38.5%. This suggests initial expressive reactions followed by more critical discourse.

Table 5 presents sentiment distribution by content source. Entertainment-oriented channels displayed higher positive sentiment (51.2%), while news and political talk shows showed more critical engagement (33.7% positive), indicating source framing influences sentiment expression.

Validation with Human Annotation

Table 6 presents inter-annotator agreement results. Fleiss' kappa of 0.72 indicates substantial agreement between

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

model predictions and human judgments. Disagreements primarily involved ambiguous expressions with sarcasm or culturally implicit critique, where the model labeled as neutral but humans interpreted as negative.

Thematic Interpretation of Sentiment

Table 7 summarizes thematic patterns identified through qualitative analysis. Positive comments emphasized creative expression; negative comments focused on political concerns; neutral comments adopted informational tone, grounding quantitative distributions.

Summary of Contributions

Overall, the results validate the feasibility of an integrated Selenium–IndoBERT pipeline for sentiment analysis of Indonesian YouTube comments. The study contributes empirical insights into public discourse surrounding a culturally specific phenomenon while demonstrating a methodological framework applicable to broader digital sentiment analysis tasks. The alignment between automated predictions and human judgment supports the reliability of the approach for exploratory and analytical applications in Indonesian NLP research.

Table 3. Model Performance Comparison

Model	Accuracy	ECE	Brier Score
IndoBERT (pseudo-label)	0.83*	0.095	0.174
LSTM + FastText	0.74	0.187	0.281
SVM + TF-IDF	0.71	0.203	0.312
Naive Bayes + TF-IDF	0.68	0.226	0.335

*Validated on 300 manually annotated samples

Table 4. Sentiment Distribution by Time Period

Time Period	Positive (%)	Negative (%)	Neutral (%)
First 24 hours	47.1	30.2	22.7
Days 2-3	41.3	35.6	23.1
Day 4 and after	38.9	38.5	22.6

Table 5. Sentiment Distribution by Content Source

Source Category	Positive (%)	Negative (%)	Neutral (%)
Entertainment/Commentary	51.2	28.4	20.4
News	33.7	42.1	24.2
Political Talk Shows	35.8	40.3	23.9

Table 6. Inter-Annotator Agreement and Model-Human Comparison

Metric	Value
Fleiss' Kappa (among 3 annotators)	0.78
Fleiss' Kappa (model vs. annotators)	0.72
Agreement on Positive comments	84.3%
Agreement on Negative comments	81.7%

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. XII No 2, Maret 2026, hlm. 375 – 382 ISSN 2550-0201 (Online)
DOI: http://dx.doi.org/ 10.33330/jurteksi.v12i2.4415
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 7. Thematic Patterns by Sentiment Category

Sentiment	Dominant Themes	Example Characteristics
Positive	Creativity, cultural hybridity, freedom of expression	Appreciation of pop culture symbolism
Negative	National symbol appropriateness, political implications, trivialization	Concerns about serious issue dilution
Neutral	Informational, deliberative, balanced	Seeking clarification, presenting multiple viewpoints

CONCLUSION

This study demonstrates the feasibility of an integrated pipeline combining Selenium web crawling and IndoBERT pseudo-labeling for sentiment analysis of Indonesian YouTube comments. Empirical results from 10,842 comments on the One Piece flag phenomenon reveal polarized public sentiment: 42.3% positive, 34.8% negative, and 22.9% neutral. Probabilistic evaluation at confidence threshold 0.75 achieved 78.5% coverage with ECE of 0.095 and Brier Score 0.174, indicating well-calibrated predictions. Manual validation confirmed substantial agreement with human judgment (Fleiss' kappa = 0.72).

The pipeline addresses key challenges in analyzing large-scale, informal Indonesian online discourse without extensive manually labeled datasets. Temporal analysis showed sentiment dynamics shifting from emotional intensity to critical deliberation. Source-based variations indicated framing influences on sentiment expression. These findings contribute both methodological insights for semi-supervised NLP and empirical understanding of Indonesian digital discourse on culturally significant phenomena.

Limitations include YouTube's comment visibility constraints potentially affecting representativeness, exclusive focus on textual content excluding multimodal cues, and assumption of stable confidence-accuracy relationships across topics. Future research should incorporate multimodal features, improve adaptation to informal and dialectal Indonesian, and apply the pipeline to cross-platform or longitudinal studies.

This study provides a robust methodological framework for researchers and practitioners seeking to monitor and understand public sentiment in Indonesia's evolving digital environment.

BIBLIOGRAPHY

[1] H. Murfi, S. Theresia Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl. Soft Comput.*, vol. 151, pp. 1–15, 2024, doi: 10.1016/j.asoc.2023.111112.

[2] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based BERT model in sentiment analysis," *Appl. Comput. Eng.*, vol. 71, no. 1, pp. 14–20, 2024, doi: 10.54254/2755-2721/71/2024ma.

[3] P. Lison, J. Barnes, and A. Hubin, "skweak: Weak Supervision Made Easy for NLP," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Internatio*

nal Joint Conference on Natural Language Processing: System Demonstrations, Stroudsburg, PA, USA: Association for Computatio nal Linguistics, 2021, pp. 337–346. doi: 10.18653/v1/2021.acl-demo.40.

[4] P. Thota and E. Ramez, "Web Scraping of COVID-19 News Stories to Create Datasets for Sentiment and Emotion Analysis," *ACM Int. Conf. Proceeding Ser.*, pp. 306–314, 2021, doi: 10.1145/3453892.3461333.

[5] A. Namoun, M. A. Humayun, and W. Nawaz, "A Multimodal Data Scraping Tool for Collecting Authentic Islamic Text Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 12, pp. 219–227, 2024, doi: 10.14569/IJACSA.2024.0151224.

[6] V. Suter, M. Shahrezaye, and M. Meckel, "COVID-19 Induced Misinformation on YouTube: An Analysis of User Commentary," *Front. Polit. Sci.*, vol. 4, no. March, pp. 1–10, 2022, doi: 10.3389/fpos.2022.849763.

[7] K. Sharma and G. M. Borkar, "Comparative Analysis of Dynamic Web Scraping Strategies: Evaluating Techniques for Enhanced Data Acquisition," in *Advancements in Communication and Systems*, Soft Computing Research Society, 2024, pp. 241–252. doi: 10.56155/978-81-955020-7-3-22.

[8] A. Sahoo, R. Chanda, N. Das, and B. Sadhukhan, "Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data," in *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, IEEE, Aug. 2023, pp. 658–663. doi: 10.1109/ICSCC59169.2023.10335061.

[9] Fransiscus and A. S. Girsang, "Sentiment Analysis of COVID-19 Public Activity Restriction (PPKM) Impact using BERT Method," *Int. J. Eng. Trends Technol.*, vol. 70, no. 12, pp. 281–288, Dec. 2022, doi: 10.14445/22315381/IJETT-V70I12P226.

[10] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, p. 5641, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[11] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[12] U. Malik, S. Bernard, A. Pauchet, C. Chatelain, R. Picot-Clémente, and J. Cortinovis, "Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets," *IEEE Access*, vol. 12, pp. 15902–15916, 2024, doi: 10.1109/ACCESS.2024.3354705.

[13] J. Lai, X. Wang, Q. Xiang, W. Quan, and Y. Song, "A Semi-Supervised Stacked Autoencoder Using the Pseudo Label for

Classification Tasks," *Entropy*, vol. 25, no. 9, 2023, doi: 10.3390/e25091274.

[14] D. Marutho and V. G. Utomo, "Benchmarking IndoBERT and Transformer Models for Sentiment Classification on Indonesian E-Government Service Reviews," *J. Transform.*, vol. 23, no. 1, pp. 86–95, Jul. 2025, doi: 10.26623/transformatika.v23i1.12095.

[15] I. Mirpulatov, S. Illarionova, D. Shadrin, and E. Burnaev, "Pseudo-Labeling Approach for Land Cover Classification Through Remote Sensing Observations with Noisy Labels," *IEEE Access*, vol. 11, no. July, pp. 82570–82583, 2023, doi: 10.1109/ACCESS.2023.3300967.

[16] K. Huang, J. Geng, W. Jiang, X. Deng, and Z. Xu, "Pseudo-loss Confidence Metric for Semi-supervised Few-shot Learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2021, pp. 8651–8660. doi: 10.1109/ICCV48922.2021.00855.

[17] Z. Wang, Y. Luo, Z. Chen, S. Wang, and Z. Huang, "Cal-SFDA: Source-Free Domain-adaptive Semantic Segmentation with Differentiable Expected Calibration Error," in *Proceedings of the 31st ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2023, pp. 1167–1178. doi: 10.1145/3581783.3611808.

[18] D. Brahma and P. Rai, "A Probabilistic Framework for Lifelong Test-Time Adaptation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 3582–3591. doi: 10.1109/CVPR52729.2023.00349.

[19] A. Jazuli, Widowati, and R. Kusumaningrum, "Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback," *Appl. Sci.*, vol. 15, no. 1, pp. 1–28, 2025, doi: 10.3390/app15010172.

[20] W. J. Kusoema and I. Ibrahim, "Sentiment Analysis on the PT Pertamina Corruption Case using IndoBERT and RCNN Methods," *SISTEMASI*, vol. 14, no. 5, p. 2246, Sep. 2025, doi: 10.32520/stmsi.v14i5.5392.