Sentiment Analysis of Public Comments on Platform X Regarding the Issue of Worker Burnout Using the IndoBERT Base Model

Nindita Fatha Dwi Lestari[1], Fajri Rakhmat Umbara[2], Melina[3]
[1,2,3]Universitas Jenderal Achmad Yani, Cimahi, Indonesia
Email: hello.nindita@gmail.com

Abstract

Burnout is a mental health problem frequently experienced by workers due to high work pressure, and Platform X users actively share opinions and emotions about it. This study aimed to analyze public sentiment on Platform X regarding worker burnout using IndoBERT Base P2. Data were scraped between June 2023 and June 2025, yielding 2,599 Indonesian-language tweets after filtering. Of these, 300 tweets were labeled manually by the researcher and a psychologist (Cohen's Kappa 0.8578), and 1,567 more were pseudo-labeled with a confidence threshold of 0.90. Class imbalance was handled with back translation augmentation of the training data (2,769 samples). A grid search of eight experiments identified five epochs and a learning rate of 2e-5 as the best configuration (validation F1-score 0.9797). On the same test set (374 samples), the weighted F1-score increased from 92.88% for the model without semi-supervised learning to 95.50%, together with augmentation and additional training data. Because most of the labeled dataset (1,567 of 1,867 tweets; 83.9%) consisted of pseudo-labels, the accuracy of 95.45% mainly reflects agreement with those labels. Negative sentiment dominated the dataset at 71.34% (a combination of manual labels, pseudo-labels, and model predictions), followed by neutral at 14.93% and positive at 13.74%.

Keywords: *Sentiment Analysis, Burnout, IndoBERT, Semi-Supervised Learning, Pseudo-Labeling.*

A. INTRODUCTION

Burnout is a condition of physical and emotional exhaustion experienced by workers when job demands exceed the resources available to the individual. This phenomenon is associated with declining work motivation, the emergence of negative attitudes toward work, and effects on job satisfaction and the intention to change jobs (Abadi et al., 2025). The literature shows that burnout is exacerbated by continuous work pressure and a lack of organizational support (Azizah et al., 2024). The urgency of this problem is reflected in the 2025 survey by the Society for Human Resource Management (SHRM) of 16,000 workers in 16 countries, including Indonesia, which concluded that work can affect mental health both positively and negatively; in the United States, 31% of respondents often or always felt work pressure, while excessive workload was one of the factors contributing to work stress, with a proportion reaching 47% (*Society for Human Resource Management (SHRM), "SHRM 2025 Insights Workplace Mental Health," 2025.*, 2025).

On the other hand, Platform X (formerly Twitter) has developed into a digital public space where people express opinions and respond to social issues openly and

in real time (Fadhillah, 2023). Public comments on this platform can be used to identify trends in public perception of an issue, although they are not intended as a representation of an individual's clinical psychological condition (Azmi & Al-Ghadir, 2024). The large and unstructured volume of social media data makes manual analysis infeasible, so sentiment analysis (classifying text as positive, negative, or neutral) is commonly used to map public perception (Dewi Salma et al., 2025).

Several previous studies have applied sentiment analysis to employment issues. On the issue of layoffs (PHK), IndoBERT outperformed SVM, random forest, and decision tree, and negative sentiment dominated (Nida Nur Aini Aryanti & Ozzi Suria, 2025). On the official DANA account on Platform X, IndoBERT was more accurate than traditional methods such as Naïve Bayes in understanding informal language (Jeevallucas et al., 2025). IndoBERT Base outperformed IndoBERT Lite after fine-tuning (Wildan Amru Hidayat & Nastiti, 2024), and a study of Generation Z working hours on Platform X also found a dominance of negative sentiment (Daffa et al., 2024).

Nevertheless, transformer-based research is generally still limited to policy issues, digital services, or specific working conditions, whereas studies that specifically analyze public comments on worker burnout in Indonesia using IndoBERT Base remain very limited. In addition, the availability of labeled data in Indonesian-language NLP research is limited, which makes manual annotation of an entire dataset difficult. This study addresses this through semi-supervised learning using self-training and pseudo-labeling, which makes use of labeled and unlabeled data simultaneously (Amini et al., 2025). IndoBERT was chosen because it was pre-trained on a large Indonesian-language corpus (Koto et al., 2020). The objectives of this study are (1) to evaluate the performance of IndoBERT Base P2 in classifying the sentiment of public comments on Platform X regarding worker burnout, and (2) to identify the distribution of the classified sentiment as an overview of public perception tendencies in the collected data. Its contributions include a two-stage semi-supervised pipeline with expert validation, three comparison experiments, and an analysis of subtopics and classification errors.

B. METHODS

1. Research Design and Flow

This study used a quantitative approach with a computational experimental design, namely a series of text-processing experiments using an existing pre-trained model without building a system or application. The research flow is shown in Figure 1, starting from data collection, EDA, manual validation, preprocessing, semi-supervised learning, data splitting, augmentation, second-stage fine-tuning with grid search, to final evaluation.



Figure 1. Research Flow

2. Population, Sample, and Data Collection

The study population was all Indonesian-language public comments (tweets) on Platform X discussing burnout, work stress, or work fatigue among Indonesian workers. Because tweets continue to accumulate, the population is dynamic and was therefore limited to the period from June 2023 to June 2025. The sample was drawn by purposive sampling based on four criteria: (1) relevance of the topic to the research keywords, (2) predominance of Indonesian, in both standard and non-standard forms, (3) relevance of the context to work, rather than education, social relationships, or sports, and (4) data cleanliness, namely free of spam, promotion, and duplication.

Data were collected through scraping using Tweet Harvest version 2.6.1 with an X authentication token. Formal keywords were determined from previous studies, namely burnout (Azizah et al., 2024), fatigue (Rahim Kamil et al., 2023), work life balance (Arifudin Arifudin et al., 2024), and work stress (Afiansyah & Adi, 2024), supplemented by informal keywords such as "*capek kerja*", "*tekanan kerja*",

"overwork", "*lembur*", "resign", "toxic office", and "toxic workplace". Keywords with neutral and positive nuances, such as "*motivasi kerja*" and "*kepuasan kerja*", were added to diversify the sentiment. Each keyword yielded 111 to 288 tweets, for a total of 3,626 raw tweets; after filtering spam or promotional tweets, tweets written entirely in a foreign language, and duplicates, 2,599 clean tweets were obtained with the attributes full_text and sentiment label. Of the attributes obtained from scraping (including usernames and tweet URLs), the analysis used only the tweet text (full_text) and the sentiment label, and all mentions were removed during the cleaning stage. Examples of raw scraping data, which still contain non-standard language, abbreviations, and a mixture of Indonesian and foreign languages, are shown in Table 1.

Table 1. Examples of raw scraping data

No	Raw Tweet
1	*Gen Z dibenci karena mereka berani bersuara. Bilang gak suka kerja rodi dibilang malas. Bilang gaji UMR gak cukup dibilang kurang bersyukur. Bilang bos toxic dibilang lemah mental. Padahal cuma mau hidup layak tanpa dijajah sistem.*
2	*emang harus banyak2 bersyukur dari hal kecil tadi pas pulang kerja dimotor aku nangis mikir kedepannya gimana ya ya Allah... sbg WNI cari kerja enggak segampang itu tapi aku gamau kesehatan mental dan psikis aku rusak cuma demi gaji 50k/sehari 14 jam kerja dgn toxic bos*

3. Data Preprocessing

The scraped text contained noise and was therefore processed in the following three stages (examples of the results at each stage are shown in Table 2).

a. *Case folding*: all letters were converted to lowercase to reduce redundancy caused by differences in capitalization.
b. *Cleaning*: removal of URLs, mentions, hashtags, irrelevant numbers and symbols, and emoji, because the analysis focused on text content.
c. *Normalization*: standardization of non-standard words, abbreviations, and slang using a normalization dictionary compiled manually by the researcher. Tweets dominated by English were detected with the langdetect library and then translated manually into Indonesian.

Table 2. Examples of the results of each data preprocessing stage

Original Text	Case Folding	Cleaning	Normalization
@username Gpp klo kata pekerja genZ. Saatnya cari tempat kerja yang nyaman untuk mental health	*@username gpp klo kata pekerja genz. saatnya cari tempat kerja yang nyaman untuk mental health*	*gpp klo kata pekerja genz saatnya cari tempat kerja yang nyaman untuk mental health*	*tidak apa apa kalau kata pekerja genz saatnya cari tempat kerja yang nyaman untuk kesehatan mental*

4. Data Labeling and Reliability Testing

Sentiment was categorized as positive (enthusiasm, support, or positive experiences toward work), negative (complaints, stress, or negative emotions), and neutral (informative without emotional expression). A total of 300 tweets were labeled

manually by two annotators, namely the researcher and a psychologist, based on written guidelines. When the labels differed, the final decision followed the psychologist's label because of the psychologist's higher scientific authority in detecting indications of burnout. Inter-annotator consistency was measured with Cohen's Kappa (Chamid et al., 2024), which compares the observed agreement with chance agreement as in Equation (1), implemented with the cohen_kappa_score function in Scikit-Learn and interpreted using the Landis and Koch scale (Ramadhani et al., 2021).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where P_o is the proportion of actual agreement and P_e is the proportion of agreement occurring by chance. The value of κ ranges from −1 to 1; values in the range 0.81–1.00 are categorized as almost perfect.

5. Semi-Supervised Learning: Self-Training and Pseudo-Labeling

Self-training was carried out in five steps (Amini et al., 2025): (1) training a bootstrap model on the manually labeled data, (2) predicting all unlabeled data, (3) selecting predictions with confidence above a threshold as pseudo-labels, (4) combining them with the original labeled data, and (5) retraining the model on the combined data. The bootstrap model was trained on 300 labeled samples (240 training and 60 validation, with a stratified 80:20 ratio) for five epochs with a batch size of 16, 50 warmup steps, a weight decay of 0.01, and class weighting. The confidence threshold was explored at values from 0.75 to 0.90 to balance the number and the reliability of labels. Tweets that did not meet the threshold were not used for training but were still labeled by the final model so that the sentiment distribution of the entire dataset could be obtained.

6. Data Splitting and Augmentation

The dataset resulting from semi-supervised learning was split 80:20 in a stratified manner (random_state = 42) into training and test data. The training data were then split 90:10 into clean training data and validation data before augmentation so that the validation set would not be contaminated by augmentation variants. Class imbalance in the training data was handled with back translation (Yoga Pratama et al., 2025), namely translating the text into an intermediate language and then back into Indonesian. Five intermediate languages were used in turn (English, Japanese, German, French, and Dutch), with a random deletion fallback mechanism with a probability of 0.15 when the back-translated result was identical to the original text. Augmentation was applied only to the minority-class training data, followed by duplicate removal, case folding, and repeated cleaning. Examples of the back translation process are shown in Table 3.

Table 3. Examples of the back translation process

Original Text (Indonesian)	Translation into Another Language	Back Translation into Indonesian
bos toxic masih bisa ditangani tapi kalau toxic npd semoga lekas sembuh karyawannya	(Inggris): a toxic boss can still be handled but if toxic npd hopefully the employees will recover soon	Bos yang toksik masih bisa ditangani, tetapi jika toksik npd semoga karyawannya segera pulih.
	(Jerman): ein toxischer Chef kann noch bewältigt werden aber wenn toxisches npd hoffentlich erholen sich die Mitarbeiter bald	Bos yang toksik masih bisa diatasi, namun jika npd toksik semoga karyawannya sembuh.

7. IndoBERT Tokenizing

Tokenization converts text into a sequence of subwords and numeric IDs that IndoBERT can process, with a maximum input length of 128 tokens. The special token [CLS] is added at the beginning as a marker of the classification input, [SEP] at the end of the sentence, and [PAD] to equalize the input length (Wildan Amru Hidayat & Nastiti, 2024). Examples of the results are shown in Table 4.

Table 4. Examples of IndoBERT tokenization results

Input	Tokenized	Encode
aku aku tetap mengundurkan diri dengan gaji kenaikan 45 tidak mau banget bertahan di lingkungan kerja toxic dan pekerjaan yang bikin kerja 7 hari seminggu	['[CLS]', 'aku', 'aku', 'tetap', 'mengundurkan', 'diri', 'dengan', 'gaji', 'kenaikan', '45', 'tidak', 'mau', 'banget', 'bertahan', 'di', 'lingkungan', 'kerja', 'tox', '##ic', 'dan', 'pekerjaan', 'yang', 'bikin', 'kerja', '7', 'hari', 'seminggu', '[SEP]']	[2, 304, 304, 830, 13645, 745, 79, 4340, 4418, 3449, 119, 422, 2174, 3878, 26, 1182, 494, 28519, 341, 41, 1367, 34, 2999, 494, 680, 406, 6452, 3]

8. Model and Fine-Tuning

BERT is a Transformer encoder-based model that learns word context bidirectionally (Devlin et al., 2019). IndoBERT Base P2 (indobenchmark/indobert-base-p2) has 12 transformer layers, 12 attention heads, and approximately 110 million parameters, and was pre-trained on an Indonesian corpus of more than 220 million words (Devlin et al., 2019). The P2 variant was trained on more complex data than P1 (Anugerah Simanjuntak et al., 2024). For classification, the 768-dimensional representation vector of the [CLS] token is passed to a classification layer that produces three logits (positive, negative, neutral), which the softmax function then converts into probabilities. Second-stage fine-tuning was performed through a grid search over epochs {3, 5, 10} and learning rates {2e-5, 3e-5, 4e-5}, with a fixed batch size of 16 and class weighting (positive 1.2819; negative 0.8866; neutral 0.9157). The best model was selected based on the highest validation F1-score.

9. Comparison Experiments and Evaluation

Three comparison experiments were evaluated on the same test data (374 samples): (1) IndoBERT Base P1 versus P2 with identical training data and configuration (5 epochs, learning rate 2e-5, batch size 16); (2) a model without semi-supervised learning (only the 300 manually labeled samples) versus the semi-supervised model under the same configuration; and (3) a lexicon-based baseline using the InSet Lexicon (3,609 positive words and 6,609 negative words) (Homepage et al., 2021), in which positive, negative, and zero polarity scores yield the positive, negative, and neutral classes. Performance was evaluated with a confusion matrix, accuracy, precision, recall, and F1-score (Grandini et al., 2020) as in Equations (2) to (5), with the weighted average as the overall metric because the class distribution is imbalanced.

$$Akurasi = \frac{\sum_k TP_k}{N} \quad (2)$$

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \quad (3)$$

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (4)$$

$$F1_k = 2 \times \frac{Precision_k \times Recall_k}{Precision_k + Recall_k} \quad (5)$$

where TP_k, FP_k, and FN_k are the numbers of true positive, false positive, and false negative predictions for class k, respectively, and N is the total number of test samples.

C. RESULTS AND DISCUSSION

1. Data Characteristics

A descriptive analysis of the 3,626 raw tweets showed an average tweet length of 28.22 words (about 178.17 characters), with a median of 29 words, a minimum length of 1 word, and a maximum of 60 words. The word "*kerja*" was the most dominant, followed by "*toxic*", "*resign*", "*mental*", and "*kantor*", while the most frequent bigrams were "*lingkungan kerja*" (239 times), "*tempat kerja*" (199 times), and "*mental health*" (190 times). Noise identification found 618 tweets containing URLs, 498 mentions, 103 hashtags, and 3,014 tweets with special symbols or characters, so preprocessing was necessary. Among the 2,599 clean tweets, 43 tweets (1.65%) were dominated by English and translated manually, while non-standard words were handled by the normalization dictionary, for example "*drpd*" became "*daripada*" and "*dijadiin*" became "*dijadikan*".

2. Reliability of Manual Labeling

Of the 300 tweets labeled, the annotators agreed on 275 tweets and disagreed on 25 tweets, so the observed agreement reached 0.9167 and Cohen's Kappa was 0.8578. This value falls in the almost perfect category so that the 300 manual labels are suitable to serve as the ground truth for training the bootstrap model. The details of the agreement are shown in Table 5.

Table 5. Contingency matrix of the labeling results by the psychologist and the researcher

Psychologist/Researcher	Negative	Neutral	Positive	Psychologist Total
Negative	160	13	6	179
Neutral	1	60	3	64
Positive	0	2	55	57
Total	**161**	**75**	**64**	**300**

The largest disagreement (13 tweets) occurred when the researcher judged a tweet as neutral but the psychologist judged it as negative, generally on tweets with implicit or ambiguous meaning that the psychologist recognized as indications of burnout, so the involvement of an expert is important. After resolution following the psychologist's labels, the ground truth consisted of 179 negative tweets (59.67%), 64 neutral (21.33%), and 57 positive (19.00%).

3. Pseudo-Labeling Results

The bootstrap model trained on 240 samples achieved an accuracy of 0.750 and an F1-score of 0.736 on the 60 validation samples. This performance is adequate because its function is only to generate pseudo-labels. The model predicted 2,299 unlabeled tweets, and the effect of the confidence threshold is shown in Table 6.

Table 6. Comparison of confidence threshold values in pseudo-labeling

Threshold	Passed	Not passed	Positive	Negative	Neutral
0,75	1.948	351	205	1.507	236
0,80	1.857	442	183	1.452	222
0,85	1.738	561	150	1.380	208
0,90	1.567	732	115	1.268	184

The higher the threshold, the fewer data pass but the higher the label confidence. A threshold of 0.90 was chosen to minimize the propagation of label errors, so that 1,567 tweets were accepted as pseudo-labels (negative 1,268 or 80.9%, neutral 184 or 11.7%, positive 115 or 7.3%) and 732 tweets did not pass. The dominance of the negative class reflects that the bootstrap model was more confident on the majority class. Tweets that did not pass were generally of mixed sentiment, in the form of prayers or hopes ("*semoga*", "*ya Allah*"), informative, or in the form of questions. The combination of 300 manual labels and 1,567 pseudo-labels formed a final dataset of 1,867 tweets.

4. Data Splitting and Augmentation

The data distribution at each stage is shown in Table 7. Before augmentation, the clean training data of 1,343 samples were highly imbalanced (negative 77.5%, neutral 13.3%, positive 9.2%). Back translation added 863 neutral and 917 positive samples, but 340 duplicates were removed, plus 14 new duplicates (8 neutral and 6 positive) after case folding and repeated cleaning. The final training data totaled 2,769 samples with a composition of 37.6% negative, 36.4% neutral, and 26.0% positive.

Table 7. Distribution of training, validation, and test data by sentiment class

Class	Final dataset	Training before augmentation	Training after augmentation	Validation	Test
Negative	1.447	1.041	1.041	116	290
Neutral	248	178	1.008	20	50
Positive	172	124	720	14	34
Total	**1.867**	**1.343**	**2.769**	**150**	**374**

The positive class originated from only 124 unique texts, so 315 back-translated results were duplicates and this class remained the smallest (720 samples); the remaining imbalance was handled with class weighting.

5. Hyperparameter Grid Search Results

Eight fine-tuning experiments were evaluated on 150 validation samples, with the best result of each experiment shown in Table 8.

Table 8. Best results of the eight grid search experiments on the validation data

Experiment	Epoch	Learning Rate	Validation accuracy	Validation F1-score
Exp1	3	2e-5	0,9467	0,9442
Exp2	5	2e-5	0,9800	0,9797
Exp3	10	2e-5	0,9533	0,9520
Exp4	3	3e-5	0,9333	0,9269
Exp5	5	3e-5	0,9600	0,9579
Exp6	10	3e-5	0,9533	0,9485
Exp7	3	4e-5	0,9467	0,9442
Exp8	5	4e-5	0,9533	0,9521

Experiment 2 (maximum of 5 epochs, learning rate 2e-5) produced the highest validation F1-score, namely 0.9797, with the best checkpoint at the fourth epoch. Figure 2 shows the trend of the validation F1-score against the number of epochs for each learning rate. At a learning rate of 2e-5, the F1-score increased from epoch 3 to epoch 5 (0.9442 to 0.9797) and then decreased at epoch 10 (0.9520). A similar pattern was seen at 3e-5 (0.9269; 0.9579; 0.9485), while at 4e-5, epoch 5 (0.9521) was also higher than epoch 3 (0.9442). Three epochs were not enough to converge, while 10 epochs produced a fluctuating validation loss and a declining F1-score after the peak, an indication of overfitting. A learning rate of 2e-5 produced a validation F1-score equal to or higher than 3e-5 and 4e-5 at every number of epochs that could be run (equal to 4e-5 at 3 epochs, namely 0.9442), so a conservative weight update appears more suitable for fine-tuning IndoBERT Base P2.

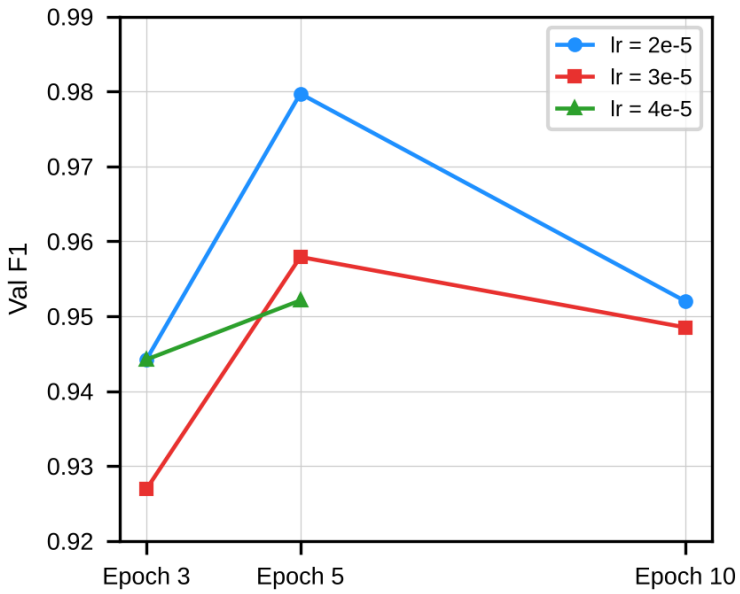


Figure 2. Trend of Validation F1-Score Per Learning Rate in the Grid Search Results

6. Final Evaluation on the Test Data

The best model was evaluated on 374 test samples that were not used in the training or validation of the final model. The results are shown in Table 9: an accuracy of 95.45% and a weighted F1-score of 95.50%. This value is slightly lower than the validation F1-score (0.9797), a difference of 0.0247. It should be noted that the labels of most of the test data came from pseudo-labels with a confidence of ≥ 0.90 (1,567 of 1,867 tweets in the dataset before splitting, or 83.9%). Accordingly, the values above should be interpreted mainly as the agreement of the model with those labels.

Table 9. Evaluation results of IndoBERT Base P2 on the test data

Sentiment class	*Precision*	*Recall*	*F1-score*	*Support*
Positive	0,78	0,82	0,80	34
Negative	0,97	0,97	0,97	290
Neutral	1,00	0,94	0,97	50
Accuracy			0,9545	374
Macro average	0,92	0,91	0,91	374
Weighted average	0,96	0,95	0,9550	374

The negative and neutral classes both obtained an F1-score of 0.97, and the neutral class achieved perfect precision (1.00). The positive class obtained the lowest F1-score (0.80), consistent with its smallest amount of source data, so that the model's exposure to the linguistic patterns of this class was limited. The confusion matrix in Figure 3 shows 17 errors out of 374 test samples: 14 occurred at the positive–negative boundary (6 positive predicted as negative and 8 negative predicted as positive) and 3 on neutral predicted as negative, while there were no errors between positive and neutral. The model predicted 291 negative tweets (77.8%), 47 neutral (12.6%), and 36

positive (9.6%), very close to the actual distribution of 290 (77.5%), 50 (13.4%), and 34 (9.1%).

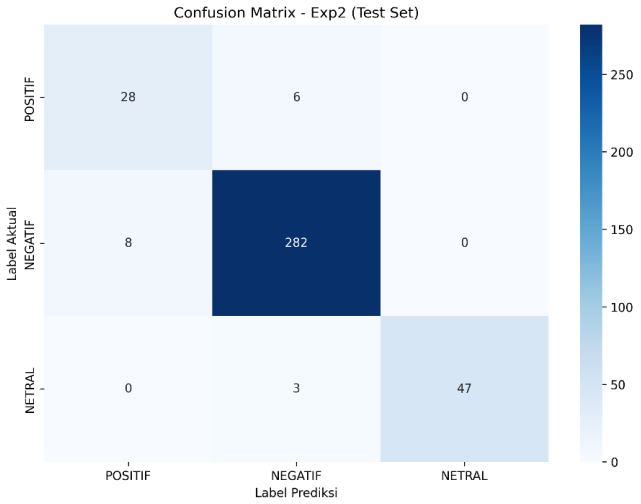


Figure 3. Confusion Matrix of the Best Model on the Test Data

7. Comparison with Other Approaches

The results of the three comparison experiments are summarized in Table 10. The lexicon-based approach reached an accuracy of only 66.04% and a weighted F1-score of 63.90%. All 50 neutral tweets were misclassified (43 as negative and 7 as positive; F1-score 0.00) because the dictionary-based approach does not understand context, sarcasm, or mixed sentiment. Compared with the lexicon-based approach, IndoBERT Base P2 with semi-supervised learning increased accuracy by 29.41 percentage points and the weighted F1-score by 31.60 percentage points.

Table 10. Performance comparison of the four approaches on the test data (374 samples)

Model	*Accuracy* (%)	*F1-score weighted* (%)	F1 positive	F1 negative	F1 neutral
Lexicon-based (InSet)	66,04	63,90	0,24	0,80	0,00
IndoBERT Base P2 tanpa SSL (300 data)	92,25	92,88	0,92	0,95	0,78
IndoBERT Base P1 + SSL	93,85	94,24	0,74	0,96	0,96
IndoBERT Base P2 + SSL	**95,45**	**95,50**	**0,80**	**0,97**	**0,97**

IndoBERT Base P2 obtained higher scores than P1 on all metrics (accuracy increased by 1.60 and the weighted F1-score by 1.26 percentage points; the accuracy difference is equivalent to 6 of the 374 test tweets, without a significance test; the configuration was taken from the P2 grid search results), especially on the positive class (F1-score 0.74 to 0.80; the positive precision of P1 was only 0.64). P1 was also less stable: the validation loss jumped from 0.2380 to 0.5120 at the third epoch and the F1-score fell from 0.9533 to 0.9136, while P2 improved steadily up to the fourth epoch.

The model without semi-supervised learning reached only a validation F1-score of 0.7464, a test accuracy of 92.25%, and a weighted F1-score of 92.88%. Semi-supervised learning with augmentation increased the test accuracy by 3.20 and the

weighted F1-score by 2.62 percentage points, with the largest improvement on the neutral class (0.78 to 0.97) because the neutral training data increased from 51 to 1,008 samples. The only decrease occurred in the positive class (0.92 to 0.80), which is presumed to be related to the dominance of negative pseudo-labels at 80.9%. It should be noted that this comparison mixes the effects of semi-supervised learning, augmentation, and the amount of training data (300 versus 2,769 samples), so the improvement cannot be attributed to semi-supervised learning alone. In addition, the test weighted F1-score of the model without semi-supervised learning (92.88%) was much higher than its validation F1-score on manual labels (0.7464), which indicates that the test data were dominated by relatively easy tweets. Overall, these results are consistent with the finding that combining labeled and unlabeled data improves accuracy (van Engelen & Hoos, 2020) and that IndoBERT excels on Indonesian-language social media text (Anugerah Simanjuntak et al., 2024).

Figure 4 compares the confusion matrices of the four approaches on the same 374 test samples. For P1, 15 negative tweets were predicted as positive, compared with 8 tweets for P2. The model without semi-supervised learning correctly classified all 50 neutral tweets, but 24 negative tweets and 4 positive tweets were incorrectly predicted as neutral; negative-to-positive errors in this model were only 1 tweet, compared with 8 tweets for P2. For the lexicon-based approach, no neutral tweets were predicted correctly (43 became negative and 7 positive) and 54 negative tweets were predicted as positive.

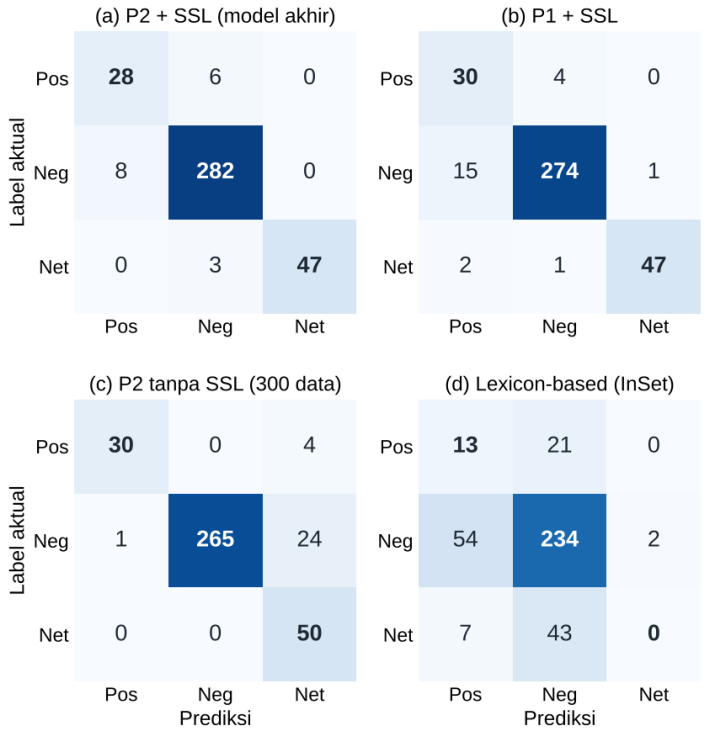


Figure 4. Confusion matrices of the four approaches on the test data: (a) IndoBERT Base P2 with SSL, (b) IndoBERT Base P1 with SSL, (c) IndoBERT Base P2 without SSL, and (d) lexicon-based (InSet)

8. Sentiment Distribution and Subtopics

The final model was applied to the 732 tweets that did not pass the threshold, which were then combined with the 1,867 labeled tweets so that the sentiment distribution of all 2,599 tweets could be obtained (Table 11). Negative sentiment dominated at 71.34%. This figure is a combination of three label sources: 300 manually labeled tweets (179 negative, or 59.67%), 1,567 pseudo-labels (1,268 negative, or 80.9%), and 732 tweets predicted by the final model (407 negative, 140 neutral, and 185 positive, or 55.6% negative; calculated from the difference between Table 11 and the final dataset in Table 7). The proportion of negative tweets in the whole data was lower than in the test data (77.8%) because only 55.6% of the 732 tweets that did not pass the threshold were predicted as negative.

Table 11. Sentiment distribution of the entire dataset

Sentiment class	Number of tweets	Percentage (%)
Negatif	1.854	71,34
Netral	388	14,93
Positif	357	13,74
Total	**2.599**	**100,00**

A word-frequency analysis of the 1,854 negative tweets, after stopword removal, showed dominant words that were grouped qualitatively into five subtopics (Table 12): *lingkungan kerja toxic* (toxic work environment), *pengunduran diri* (resignation), *dampak psikologis* (psychological impact), *beban dan jam kerja berlebihan* (excessive workload and working hours), and *kompensasi* (compensation). Some of these dominant words (for example *toxic*, *lembur*, *capek*, *stres*, and *kelelahan*) were also data collection keywords, so this grouping describes the themes that stand out in the collected data.

Table 12. Negative Sentiment Subtopics Based on Keywords and Their Frequencies

Subtopic	Keywords (frequency)
Lingkungan kerja *toxic*	*toxic* (548), kantor (338), lingkungan (201), bos (138)
Pengunduran diri	mengundurkan (382)
Dampak psikologis dan kelelahan	mental (416), stres (253), capek (250), kelelahan (247), kesehatan (159)
Beban dan jam kerja berlebihan	jam (345), berlebihan (210), lembur (146)
Kompensasi kerja	gaji (207)

9. Classification Error Analysis

A qualitative analysis of the 17 misclassified tweets showed three main patterns. First, narrative tweets that recount bad experiences in the past but are positive in meaning because they show adaptation, so that the model picked up their negative chronology. Second, tweets with a negative tone that end with a reassuring or motivational sentence, so that the model emphasized the closing tone. Third, sarcastic, ambiguous, or argumentative tweets containing words with strong emotional connotations, for example "*stres*" in an expression of gratitude or "*kelelahan*

mental" in an informative tweet. Examples of misclassified tweets for each pattern are shown in Table 13.

Table 13. Examples of Misclassification on the Test Data

Actual → predicted label	Tweet excerpt	Error pattern
Positive → Negative	"…perlahan mulai nyatu sama ritme kerjaan yang banyak walau masih belum sempurna"	Narasi: kronologi negatif di awal tweet mendominasi, padahal maknanya adaptasi
Negative → Positive	"…kelelahan mental dan mau istirahat dulu… katanya tidak apa apa… liburan dulu aja"	Penutup menenangkan menutupi isi utama berupa kelelahan
Negative → Positive	"…bekerja 4 tahun tanpa kenaikan gaji… kuat yuk kuat…"	Penutup motivasional menutupi keluhan gaji
Positive → Negative	"ya allah terimakasih lingkungan kerja aku bikin ketawa dan senang terus no stres stres"	Kata "stres" yang berulang memicu prediksi negatif
Neutral → Negative	"ciri ciri pekerja kelelahan mental 1 sampai kantor semenit sebelum jam kerja…"	Frasa "kelelahan mental" pada tweet informatif bernada humor
Neutral → Negative	"…kerja dengan durasi selama ini termasuk salah satu bentuk perbudakan modern…"	Istilah emosional kuat pada tweet argumentatif

Table 13 shows that the model sometimes still relies on words or phrases with strong emotional connotations without considering the overall context of the sentence.

D. CONCLUSION

Based on the results of the study, it can be concluded that: (1) IndoBERT Base P2 fine-tuned with a semi-supervised learning approach was able to classify the sentiment of public comments on Platform X regarding worker burnout with an accuracy of 95.45% and a weighted F1-score of 95.50% on test data of which most were pseudo-labeled, using the best configuration of five epochs and a learning rate of 2e-5 (validation F1-score 0.9797); (2) manual labeling by the researcher and a psychologist was reliable (Cohen's Kappa 0.8578) and pseudo-labeling with a confidence threshold of 0.90 produced 1,567 additional labels; (3) IndoBERT Base P2 obtained a higher weighted F1-score than IndoBERT Base P1 (94.24%), the model without semi-supervised learning (92.88%), and the lexicon-based baseline (63.90%); (4) the sentiment distribution of the 2,599 tweets was dominated by negative sentiment at 71.34% (a combination of manual labels, pseudo-labels, and model predictions on tweets collected with certain keywords), with five dominant-word subtopics, namely *lingkungan kerja toxic*, *pengunduran diri*, *dampak psikologis*, *beban kerja berlebihan*, and *kompensasi kerja*; and (5) classification errors mainly occurred at the boundary between positive and negative sentiment on narrative, mixed-tone, or ambiguous tweets. Future research is suggested to add manually labeled data for the positive and neutral classes, explore generative model-based augmentation, and apply adaptive pseudo-labeling with a different confidence threshold for each class.

ACKNOWLEDGMENTS

The authors would like to thank the Informatics Study Program, Faculty of Science and Informatics, Universitas Jenderal Achmad Yani, the examiners, and the psychologist who helped validate the data labeling so that this study could be completed.

REFERENCES

1. Abadi, R. R., Abadi, R. R., Ikhwan, M., Haeruddin, M., Fadilah Aswar, N., & Aslam, A. P. (2025). Burnout pada Tenaga Kerja Milenial: Menemukan Harapan Keseimbangan Kehidupan-Kerja. *Jurnal ARASTIRMA Universitas Pamulang*, *5*(1), 302–313.
2. Afiansyah, M. R., & Adi, A. N. (2024). Pengaruh Beban Kerja dan Stres Kerja Terhadap Burnout Pada Saat COVID-19. *Jurnal Kewirausahaan dan Inovasi*, *2*(2), 543–552. https://doi.org/10.21776/jki.2023.02.2.20
3. Amini, M. R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., & Maximov, Y. (2025). Self-training: A survey. *Neurocomputing*, *616*, 128904. https://doi.org/10.1016/j.neucom.2024.128904
4. Arifudin, A., Khansa, F. B., Mozza, J., Anisa, F., & Khumaerahtuz, Z. (2024). Keseimbangan Kehidupan Kerja Dan Burnout: Implikasinya Terhadap Kepuasan Kerja Karyawan di You N Me Store Cirebon. *Jurnal Ekonomi dan Pembangunan*, *2*(2), 203-217. https://doi.org/10.61132/jepi.v2i2.573
5. Aryanti, N. N. A., & Suria, O. (2025). Analisis sentimen terhadap pemutusan hubungan kerja di Indonesia: Komparasi IndoBERT dengan SVM, Random Forest, dan Decision Tree dengan optimasi TF-IDF. *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, *10*(2), 1158-1176. https://doi.org/10.36341/rabit.v10i2.6364
6. Azizah, M., Pamungkas, R., & Merdiaty, N. (2024). Kajian Literatur: Burnout & Strategi Penanganannya. *Jurnal Psikologi dan Bimbingan Konseling*, *9*. https://doi.org/10.6734/LIBEROSIS..V2I2.3027
7. Azmi, A. M., & Al-Ghadir, A. I. (2024). Using Twitter as a digital insight into public stance on societal behavioral dynamics. *Journal of King Saud University - Computer and Information Sciences*, *36*(5). https://doi.org/10.1016/j.jksuci.2024.102078
8. Chamid, A. A., Widowati, & Kusumaningrum, R. (2024). Labeling Consistency Test of Multi-Label Data for Aspect and Sentiment Classification Using the Cohen Kappa Method. *Ingenierie Des Systemes d'Information*, *29*(1), 161–167. https://doi.org/10.18280/isi.290118
9. Daffa, M., Fahreza, A., Luthfiarta, A., Rafid, M., Indrawan, M., & Nugraha, A. (2024). Analisis Sentimen: Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z. *Journal of Applied Computer Science and Technology (JACOST)*, *5*(1), 2723–1453. https://doi.org/10.52158/jacost.715
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational*

linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186). https://github.com/tensorflow/tensor2tensor

11. Fadhillah, N. (2023). Cyberspace Sebagai Ruang Publik Baru: Analisis Naratif Opini Publik Pada Isu Larangan Thrifting di Media Sosial Twitter. *Jurnal Ilmu Komunikasi Pattimura*, *02*(01).
12. Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. http://arxiv.org/abs/2008.05756
13. Homepage, J., Musfiroh, D., Khaira, U., Eko, P., Utomo, P., Suratno, T., Studi, P., Informasi, S., Sains, F., & Teknologi, D. (2021). Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon. *MALCOM: Indonesian Journal of Machine Learning and Computer Science 1*, 24–33.
14. Jeevallucas, P., Surya Gautama, J., Wibowo, A., Siang, J. J., Informasi, S., Informasi, T., Kristen, U., & Wacana, D. (2025). Analisis Sentimen Pengguna Terhadap Akun X/Twitter Resmi "Dana" dengan Algoritma Indobert Sentiment Analysis of Users Toward the Official "Dana" Account on X/Twitter Using the Indobert Algorithm. *Journal of Information Technology and Computer Science (INTECOMS)*, *8*(2).
15. Kamil, A. R., Lianda, I. C., Silaban, N. Y., & Wulansari, N. (2023). Identifikasi Burnout Syndrome Perawat Di Lingkungan Aliansi RS Islam Jakarta. *Jurnal Ilmiah Keperawatan IMELDA*, *9*(2), 115-121.
16. Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020, December). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757-770). http://arxiv.org/abs/2011.00677
17. Pratama, A. Y., Sanjaya, G. A., Lubis, N. K., Aditya, M. R., & Yennimar, Y. (2025). Analisis Sentimen Publik Terkait Danantara Menggunakan Algoritma IndoBERT pada Platform Media Sosial. *Metik Jurnal*, *9*(1), 92-100. https://doi.org/10.47002/metik.v9i1.1055
18. Ramadhani, R., Juandi, D., & Nurlaelah, E. (2021). A Meta-Analysis on The Effect of Inquiry Learning Model on Students' Mathematical Problem-Solving Skills. *Indonesian Journal of Science and Mathematics Education*, *4*(3), 302–312. https://doi.org/10.24042/ijsme.v4i3.9730
19. Salma, T. D., Kurniawan, M. F., Darmawan, R., & Basri, A. (2025). Analisis Sentimen Berbasis Transformer: Persepsi Publik terhadap Nusantara pada Perayaan Kemerdekaan Indonesia yang Pertama. *vol*, *9*, 757-764.
20. Simanjuntak, A., Lumbantoruan, R., Sianipar, K., Gultom, R., Simaremare, M., Situmeang, S., & Panggabean, E. (2024). Research and analysis of indobert hyperparameter tuning in fake news detection. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, *13*(1), 60-67. https://doi.org/10.22146/jnteti.v13i1.8532
21. *Society for Human Resource Management (SHRM), "SHRM 2025 Insights Workplace Mental Health," 2025.* (2025).

22. van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6
23. Wildan, A. H., & Nastiti, V. R. S. (2024). Perbandingan kinerja pre-trained IndoBERT-base dan IndoBERT-lite pada klasifikasi sentimen ulasan TikTok Tokopedia Seller Center dengan model IndoBERT. *Jsii (Jurnal Sist. Informasi)*, *11*(2), 13-20. https://doi.org/10.30656/jsii.v11i2.9168