BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

Analyzing Generation Z's Sentiment on Working Hours and Mental Health Using IndoBERT: Evidence from TikTok Discussions

Alfilia Hilda Rahmatika*, Bella Okta Sari Miranda, Imam Adiyana, Isyiffah Falujjah Anugrah Putri

Faculty of Economics and Business, Telkom University Purwokerto, Purwokerto, Indonesia
Email: [1,*]alfiliahilda@telkomuniversity.ac.id, [2]bellaoktasarimiranda@telkomuniversity.ac.id, [3]imamadiyana@telkomuniversity.ac.id, [4]aizzifalujjah@student.telkomuniversity.ac.id
Correspondence Author Email: alfiliahilda@telkomuniversity.ac.id

Abstract−Working hours exceeding recommended limits may increase burnout, psychological stress, and sleep disturbances, particularly among Generation Z, who highly value mental health and work-life balance. Although public opinions on this issue are increasingly expressed on social media, most existing studies rely on conventional machine learning methods or lexicon-based labeling, which are less effective at capturing contextual meaning and linguistic nuance in informal social media text. To address this gap, this study proposes a transformer-based sentiment analysis framework that employs a fine-tuned BERT-based Multilingual model as the primary sentiment classifier, rather than merely as an automatic labeling tool, to analyze Generation Z's sentiment toward working hours and mental health based on TikTok comments. A total of 2,203 comments were collected through web scraping and processed using cleaning, normalization, tokenization, and BERT-based zero-shot sentiment labeling before being used to fine-tune the classification model. The model was evaluated using accuracy, precision, recall, and F1-score. The results indicate that negative sentiment dominated the dataset (41.13%), followed by positive (37.49%) and neutral (21.38%) sentiments. Frequently occurring terms, such as working hours, work, and resigning, suggest that users' concerns mainly relate to long working hours and the intention to leave their jobs. The fine-tuned model achieved excellent classification performance, although a gap between training and validation performance indicates the need for improved generalization. These findings demonstrate that BERT-based sentiment analysis can provide valuable insights for organizations, policymakers, and human resource practitioners in developing more flexible working-hour policies and mental health support programs tailored to Generation Z.

Keywords: BERT-base Multilingual; Generation Z; Mental Health; Sentiment Analysis; Working Hours

1. INTRODUCTION

Excessive working hours and their impact on employees' psychological well-being continue to be a major concern in modern workplaces. Although organizations increasingly recognize the importance of employee well-being, long working hours remain common across many industries. Prolonged exposure to excessive workloads has been associated with burnout, emotional exhaustion, chronic stress, anxiety, sleep disturbances, and lower job satisfaction. These negative outcomes not only reduce employees' quality of life but also affect organizational performance by increasing absenteeism and turnover intentions. The World Health Organization (WHO) [1] emphasizes that reasonable working hours, manageable workloads, and adequate workplace support are essential for maintaining employees' mental health. Likewise, the International Labour Organization (ILO) [2] recommends limiting working hours to approximately 40–48 hours per week because working beyond this threshold significantly increases the risk of burnout, psychological distress, and sleep disorders [3][4][5]. These recommendations demonstrate that excessive working hours should be viewed not only as a labor issue but also as an important public health concern.

Among today's workforce, Generation Z has become a particular focus because this generation places greater importance on mental health, personal well-being, and work–life balance than previous generations [6][7]. As Generation Z continues to enter the labor market, organizations are increasingly expected to provide healthier and more balanced working environments. However, current workplace conditions often do not reflect these expectations. According to the National Labor Force Survey published by Badan Pusat Statistik [8], approximately 25.47% of Indonesian employees work more than 49 hours per week, exceeding the standard working-hour limit established by labor regulations. Furthermore, a global survey conducted by Deloitte [9] found that many Generation Z employees experience work-related stress, with excessive workloads and long working hours identified as major contributing factors. This gap between recommended working-hour standards and actual workplace practices highlights the need to better understand how excessive working hours influence Generation Z's mental health.

The rapid growth of social media has created new opportunities to understand public perceptions of workplace issues. Unlike conventional surveys, social media allows individuals to express their opinions, experiences, and emotions more openly and spontaneously. Platforms such as X (formerly Twitter), TikTok, and other online communities have become spaces where Generation Z frequently discusses work stress, overtime, burnout, toxic workplace culture, and the importance of maintaining work–life balance. These discussions indicate increasing resistance toward organizational cultures that normalize excessive workloads and glorify overwork [10][11]. Rather than representing isolated personal experiences, these conversations reflect a broader collective concern regarding workplace practices that are perceived as harmful to employees' psychological well-being. Therefore, analyzing social media discussions provides valuable insights into how Generation Z perceives the relationship between working hours and mental health.

The growing availability of user-generated content has encouraged researchers to apply sentiment analysis techniques to investigate public opinions regarding workplace issues. Several previous studies have explored this topic using different machine learning approaches. Study [12] analyzed TikTok comments using Multinomial Naïve Bayes,

Support Vector Machine (SVM), and Multilayer Perceptron with Bag-of-Words and TF-IDF text representations. However, IndoBERT was employed only as an automatic labeling tool rather than as the primary sentiment classification model. Similarly, study [13] examined public opinions on X regarding the impact of working hours on Generation Z's mental health using Gaussian Naïve Bayes and SVM with TF-IDF feature extraction, while sentiment labels were generated using the InSet Lexicon. Study [14] similarly analyzed sentiment related to working hours and Generation Z's mental health but applied a Backpropagation Neural Network rather than a transformer-based model, again relying on non-contextual feature engineering. Study [15] proposed BERT-Fuse, a hybrid deep learning architecture that integrates BERT embeddings with LSTM, GRU, Transformer blocks, and CNN layers, further augmented with TF-IDF and Bag-of-Words features, achieving state-of-the-art accuracy of 97.05% for general mental health sentiment classification. While this demonstrates the substantial potential of transformer-based hybrids for detecting nuanced emotional cues in psychological texts, its scope remains broad and does not specifically address the intersection of excessive working hours, burnout culture, and Generation Z's mental health discourse, particularly within the Indonesian TikTok ecosystem. Study [16] approached the topic from a health-epidemiological perspective, examining night-shift work, tobacco dependence, and chronic disease risk among Generation Z overtime workers, rather than analyzing social media sentiment directly. Study [17] offered a broader, non-computational discussion of Generation Z's mental health challenges without employing any sentiment classification method. Collectively, these studies confirm that excessive working hours, burnout, and Generation Z's mental health have become increasingly discussed topics, and recent advances in hybrid BERT architectures have proven highly effective for mental health sentiment analysis in general. Yet, despite the proven power of transformer-based models, none of the existing studies applied a fine-tuned transformer-based model, specifically, a contextualized model optimized for the Indonesian language (IndoBERT), as the primary sentiment classifier for TikTok discourse on working hours and mental health in the Indonesian context. This underscores the persistent methodological and contextual gap that the current study aims to address.

Despite these contributions, important methodological limitations remain. Most previous studies rely on traditional text representation techniques, including Bag-of-Words, TF-IDF, or TF-IDF combined with lexicon-based sentiment labeling. These approaches mainly consider word frequency while overlooking contextual relationships among words, making them less effective in interpreting sarcasm, ambiguity, informal language, and emotional nuances that commonly appear in social media conversations. Recent advances in Natural Language Processing (NLP), particularly Bidirectional Encoder Representations from Transformers (BERT), provide a promising solution to these limitations. Through its bidirectional self-attention mechanism, BERT captures contextual relationships between words and produces richer semantic representations than conventional machine learning models. For Indonesian-language applications, IndoBERT has been pre-trained using large-scale Indonesian corpora, enabling it to better understand linguistic characteristics commonly found in Indonesian social media text [18]. Nevertheless, the use of IndoBERT as the primary sentiment classification model for analyzing Generation Z's perceptions of excessive working hours and mental health remains relatively limited.

Based on these considerations, this study proposes a comprehensive transformer-based sentiment analysis framework by employing IndoBERT as the primary sentiment classification model rather than merely using it for automatic sentiment labeling. The study aims to analyze public opinions expressed on social media regarding the relationship between excessive working hours, burnout-oriented workplace culture, and the mental health of Generation Z. By leveraging IndoBERT's ability to capture contextual meaning and semantic relationships, this study is expected to achieve more accurate sentiment classification than conventional machine learning approaches while providing deeper insights into the emotional and linguistic characteristics of online discussions. From an academic perspective, this study contributes to the growing literature on transformer-based sentiment analysis in the Indonesian language context. From a practical perspective, the findings are expected to provide valuable evidence for organizations, policymakers, and human resource practitioners in developing healthier working-hour policies and workplace environments that better support the mental well-being of Generation Z.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study employed a transformer-based sentiment analysis approach to examine Generation Z's opinions regarding working hours and mental health on social media. The research workflow is illustrated in Figure 1 and consists of four main stages: data collection, data preprocessing, data labeling, and model implementation and evaluation. This transformer-based approach builds on the BERT architecture introduced by [19] which uses a bidirectional self-attention mechanism to learn contextual word representation, and is conceptually related to IndoBERT [18], a variant pre-trained on large-scale Indonesian corpora. Unlike conventional feature-based methods such as Bag-of-Words or TF-IDF, which treat words as independent and context-free units, the self-attention mechanism in BERT allows each word to be represented in relation to every other word in a sentence, enabling the model to capture nuances such as negation and implicit sentiment that are common in informal social media text. IndoBERT extends this capability specifically for the Indonesian language, having been pre-trained on a large corpus of Indonesian text, making it particularly well suited for analyzing TikTok comments that frequently mix formal Indonesian, colloquial expressions, and regional slang. Each of the four stages in Figure 1 builds directly on the output of the previous one data collection gathers raw comments from

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

TikTok, and data preprocessing cleans and normalizes this text for analysis, data labeling assigns sentiment categories through a zero-shot labeling procedure, and model implementation and evaluation fine-tunes the Multilingual BERT-base model and assesses its performance using standard classification metrics.

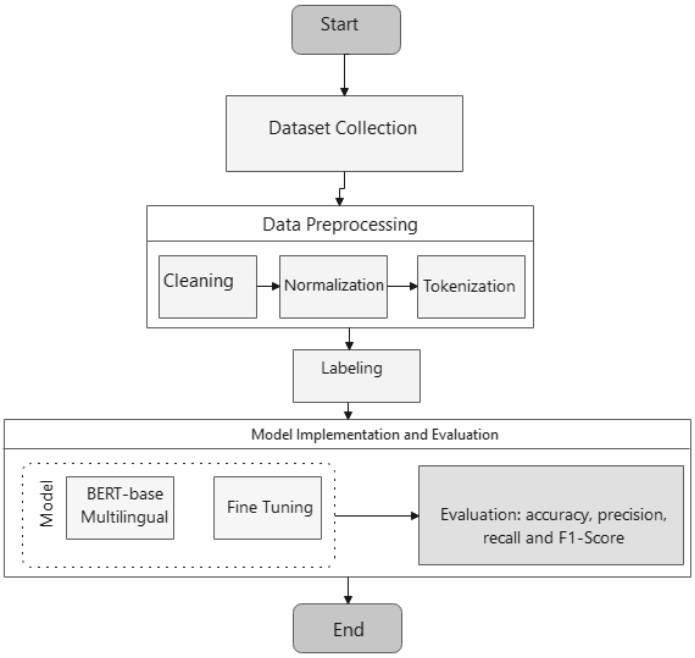

Figure 1. Research Process Flowchart

2.2 Data Collection

The first stage of this study was dataset collection, which involved gathering Generation Z's opinions from TikTok videos related to mental health. Data was collected automatically using a web scraping technique. The scraped data were then processed and stored in Comma-Separated Values (CSV) format to facilitate subsequent preprocessing and analysis. Data collection was designed to capture diverse user opinions, ensuring that the dataset represented a wide range of sentiments expressed in TikTok comments, including positive, neutral, and negative opinions. The resulting dataset served as the primary source for training and evaluating the Multilingual BERT-based sentiment classification model. Storing the data in CSV format facilitated efficient data processing using the Python programming language and ensured compatibility with various data analysis and machine learning libraries [20]. After the data collection stage was completed, the dataset proceeded to the data preprocessing stage, which included data cleaning, text normalization, and tokenization before being used for sentiment labeling and model training. This process is illustrated in Figure 2, which summarizes the data collection workflow used to gather TikTok comments for this study.

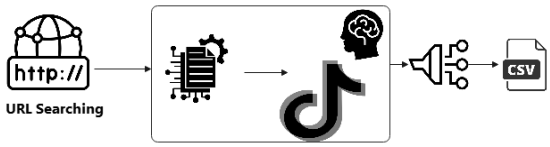

Figure 2. Data Collection

2.3 Data Preprocessing

The next stage was data preprocessing, which aimed to improve data quality before the model training process. This stage consisted of three main processes. The first process, cleaning, involved removing unnecessary elements such as URLs, emojis, symbols, irrelevant punctuation marks, specific numerical values, and excessive whitespace, ensuring that only meaningful textual information was retained [21]. The second process, normalization, standardized the text by converting all characters to lowercase and transforming non-standard words, abbreviations, and slang expressions into their standardized forms using a normalization dictionary [22]. The final process, tokenization, segmented the text into tokens using the Multilingual BERT WordPiece tokenizer [23]. During this stage, each token was converted into its corresponding token ID, and the special tokens [CLS] and [SEP] were appended to represent the input sequence required by the Transformer model [24].

2.4 Labeling

After the preprocessing stage was completed, all reviews were labeled using a BERT-based Zero-Shot Classification approach to generate the sentiment classes used as the target variable for model training [19][25]. This approach leverages

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

the capabilities of a pre-trained language model to classify sentiment without requiring manually labeled training data [26]. Each review was assigned to one of three sentiment categories: positive, neutral, or negative, based on the semantic similarity between the review content and the description of each sentiment label. The resulting labeled dataset was subsequently used as the ground truth for the fine-tuning and evaluation of the Multilingual BERT-base model [24].

2.5 Model Implementation and Evaluation

The final stage involved model implementation and performance evaluation. In this study, Multilingual BERT was employed as the primary model for sentiment classification of user reviews. This model is a pre-trained Transformer-based language model trained on a multilingual corpus, enabling it to capture rich semantic representations across multiple languages [27]. The model was adapted to the sentiment classification task through a fine-tuning process, in which its parameters were updated using the preprocessed and labeled review dataset [28]. This process enabled the model to adapt its previously learned linguistic representations to the specific characteristics of the study dataset, thereby improving its ability to generate more accurate sentiment predictions [29].

Model performance was assessed using four standard evaluation metrics commonly employed in sentiment classification: accuracy, precision, recall, and F1-score. Accuracy reflects the overall proportion of correctly classified instances among all testing samples [30]. Precision indicates how reliably the model assigns instances to the correct sentiment category, while recall measures the model's effectiveness in retrieving all relevant instances within each sentiment class [31]. To provide a more balanced evaluation, the F1-score was used by combining precision and recall through their harmonic mean. This metric is particularly valuable when the class distribution is imbalanced, as it considers both the correctness and completeness of the model's predictions rather than relying solely on overall accuracy [32]. Figure 3 presents the structure of the confusion matrix used to evaluate the model's classification performance.

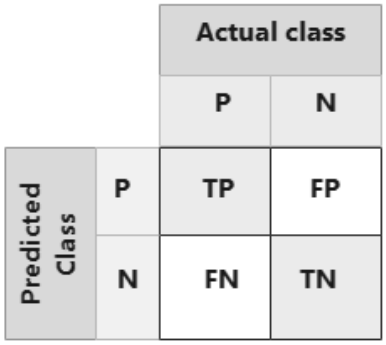

Figure 3. Confusion Matrix Structure

In addition, the classification results were analyzed using a confusion matrix as illustrated in Figure 3 to identify patterns of misclassification among sentiment classes and to evaluate the model's ability to distinguish positive, neutral, and negative reviews [33]. This analysis provides more detailed insights into the model's performance for each sentiment category, thereby serving as a basis for interpreting its strengths and limitations. Through this framework, the Multilingual BERT-base model leverages contextual representations generated by the self-attention mechanism to capture semantic relationships among words based on the overall sentence context [34]. This capability enables the model to effectively handle semantic ambiguity and linguistic variations commonly found in user reviews, resulting in superior sentiment classification performance compared with conventional approaches that rely solely on frequency- or statistical-based word representations. Consequently, the proposed model provides more accurate and reliable sentiment classification for analyzing user opinions.

3. RESULT AND DISCUSSION

This section presents the experimental results and analysis of the application of the BERT model for sentiment classification on a dataset of TikTok video comments discussing mental health. The discussion begins with an evaluation of the model's performance following the fine-tuning process using multiple classification metrics. It then proceeds with an analysis of the sentiment distribution to identify patterns in users' perceptions of mental health issues. Finally, the findings are comprehensively examined to evaluate the model's ability to capture contextual language representations and accurately classify sentiment in social media data.

3.1 Sentiment Distribution and Word Frequency Analysis

To provide an overview of the labeling results, this section presents the distribution of sentiment classes obtained from the TikTok comment dataset related to working hours and mental health. The dataset was manually labeled into three sentiment categories, namely negative, positive, and neutral, based on the opinions expressed by users. This distribution serves as the basis for understanding the overall sentiment tendency within the dataset before proceeding to the sentiment

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

classification process using the Multilingual BERT model. Figure 4 presents the distribution of sentiment classes obtained from the labeling process.

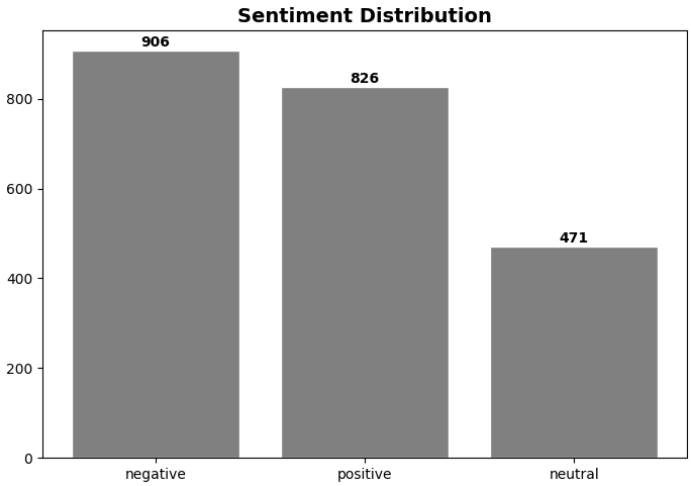

Figure 4. Sentiment Label Distribution

The results show that the dataset was categorized into three sentiment classes: negative, positive, and neutral. Based on the labeling results, the negative class contained the largest number of comments (906), followed by the positive class with 826 comments, while the neutral class had the fewest comments (471). Of the 2,203 comments analyzed, 41.13% were classified as negative, 37.49% as positive, and 21.38% as neutral. These results indicate that negative sentiment predominates in users' opinions regarding mental health–related content on TikTok, followed by positive sentiment, whereas neutral sentiment accounts for the smallest proportion.

The observed distribution also indicates the presence of class imbalance, particularly for the neutral class, which contains considerably fewer samples than the negative and positive classes. This imbalance may affect the model's learning process and classification performance. Therefore, model evaluation was not based solely on accuracy, but also incorporated precision, recall, and F1-score to provide a more comprehensive assessment of the model's ability to classify each sentiment category effectively.

In addition to the sentiment distribution, a word frequency analysis was also conducted to explore the dominant themes and topics discussed within the dataset. This analysis provides a preliminary understanding of the vocabulary most frequently used by users when discussing working hours and mental health on TikTok. To visualize the most frequently occurring terms, a word cloud was generated based on the preprocessed comment dataset, as presented in Figure 5.

Figure 5. Word Cloud of TikTok Comments Related to Working Hours

The most prominent words include "*kerja*", "*hidup*", "*gaji*", "*orang*", "*udah*", "*perang*", "*rumah*", and "resign". The prominence of these words suggests that users' discussions primarily revolve around employment, economic conditions, social life, and personal experiences associated with mental health. The word cloud provides an initial overview of the dominant themes within the dataset; however, it does not capture the semantic relationships or contextual meanings among words. To address this limitation, this study employs the BERT model, which can learn contextual representations and capture semantic dependencies between words, thereby enabling more accurate sentiment classification. In terms of interpretation, the word cloud indicates that users' comments are predominantly centered on employment, economic conditions, and social life, highlighting these topics as the primary context of sentiment related to mental health discussions on TikTok. Table 1 lists the ten most frequently occurring words identified after the preprocessing stage.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

Table 1. Top 10 Words After Preprocessing

No	Word	Count
1	*Jam*	845
2	*Kerja*	451
3	*Di*	369
4	*Pulang*	288
5	*Berangkat*	267
6	*Aku*	236
7	*Aja*	230
8	*Pagi*	176
9	*Lagi*	174
10	Resign	158

This preprocessing stage was performed to remove irrelevant elements, normalize the text, and generate more representative tokens for sentiment analysis. Word frequency was calculated based on the total number of occurrences of each token across the entire comment dataset. As shown in Table 1, the token "*jam*" appeared most frequently, with 845 occurrences, followed by "*kerja*" (451), "*di*" (369), "*pulang*" (288), "*berangkat*" (267), "*aku*" (236), "*aja*" (230), "*pagi*" (176), "*lagi*" (174), and "resign" (158). The predominance of these words indicates that most comments are related to work activities, daily routines, and time management, as reflected by the frequent occurrence of terms such as "*jam*," "*kerja*," "*berangkat*," "*pulang*," "*pagi*," and "resign." These findings suggest that employment-related issues and the pressures of everyday activities are among the topics most associated with discussions of mental health on TikTok. Although word frequency analysis provides an initial overview of the dominant themes in the dataset, it considers only the frequency of individual words without accounting for their contextual meaning within sentences. Consequently, this approach is unable to capture the semantic relationships underlying each comment. To address this limitation, this study employs Multilingual BERT-base, which can learn contextual representations and modeling semantic relationships among words, thereby enabling more accurate sentiment classification.

3.2 Model Performance Evaluation

To evaluate the performance of the fine-tuned model during training, the training and validation accuracy were monitored across five epochs. This evaluation is essential to observe how well the model learns from the training data while also assessing its ability to generalize to unseen validation samples, thereby helping to identify potential overfitting. Figure 6 presents the resulting accuracy curves for both the training and validation sets.

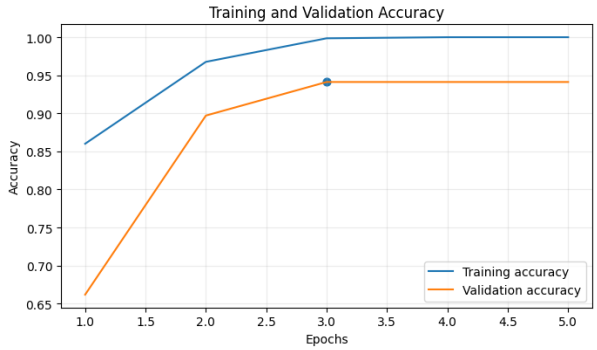

Figure 6. Training and Validation Accuracy

Figure 6 depicts the changes in training accuracy and validation accuracy throughout the five epochs of fine-tuning the Multilingual BERT-based model. During the initial epoch, the model achieved a training accuracy of approximately 86%, whereas the validation accuracy was around 66%. A considerable improvement was observed in the second epoch, with the training and validation accuracy increasing to approximately 97% and 90%, respectively. This rapid improvement indicates that the model quickly learned the underlying sentiment patterns from the training data. Beginning with the third epoch, the training accuracy reached 100% and remained unchanged for the rest of the training process. At the same time, the validation accuracy stabilized at approximately 94%, suggesting that additional training no longer produced substantial performance gains. This behavior indicates that the model had reached convergence while maintaining a satisfactory ability to generalize to unseen validation samples. Although the training accuracy exceeded the validation accuracy by roughly 6%, the model's overall performance was not evaluated based on accuracy alone. Additional metrics, including precision, recall, and F1-score, were also considered to provide a more balanced and reliable assessment of the classification results. Overall, the accuracy curves demonstrate that the Multilingual BERT-base model learned efficiently during the early stages of fine-tuning and achieved stable performance after the third epoch. This performance trend indicates that the model is capable of reliably classifying the sentiment of TikTok comments related to mental health.

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

In addition to accuracy, the training and validation loss were also monitored throughout the fine-tuning process to further assess the model's learning behavior. Loss reflects the magnitude of prediction error made by the model, where a lower loss value indicates better alignment between the predicted and actual sentiment labels. Monitoring this metric alongside accuracy provides a more comprehensive understanding of the model's convergence and its ability to generalize to unseen data. Figure 7 presents the resulting training and validation loss curves across the five training epochs.

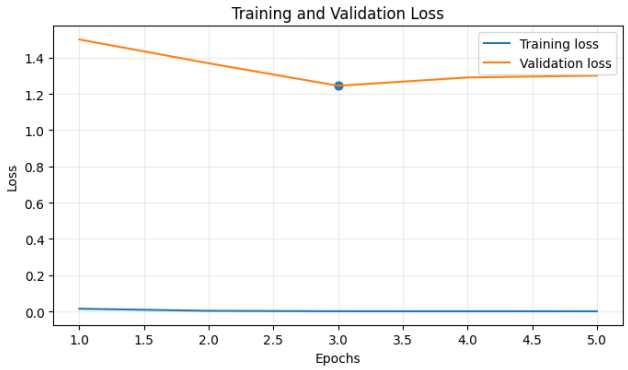

Figure 7. Training and Validation Loss

Figure 7 illustrates the changes in training loss and validation loss during the fine-tuning process of the Multilingual BERT-based model over five epochs. The training loss decreased steadily and approached zero, indicating that the model progressively learned the underlying patterns in the training data. Meanwhile, the validation loss declined from approximately 1.50 in the first epoch to 1.25 in the third epoch, after which it increased slightly and stabilized at approximately 1.30 until the end of training. The initial reduction in validation loss suggests an improvement in the model's ability to generalize to unseen validation data. However, the slight increase observed after the third epoch indicates that performance improvements had begun to plateau, suggesting that the training process was approaching convergence. Despite this minor increase, the variation in validation loss remained relatively small and was accompanied by consistently stable validation accuracy, indicating no evidence of significant overfitting. Overall, the loss curves demonstrate that the Multilingual BERT-based model effectively minimized prediction errors on the training data while maintaining stable performance on the validation dataset. These findings indicate that the fine-tuning process was successful and produced a model with strong generalization capability for the sentiment classification of TikTok comments related to mental health. Beyond accuracy and loss, the Macro F1-score was also monitored during the fine-tuning process to evaluate the model's performance across all sentiment classes more equally. Unlike accuracy, which can be biased toward the majority class, the Macro F1-score calculates the F1-score for each class independently and averages them, thereby providing a more balanced measure of performance, particularly important given the class imbalance observed in the neutral sentiment category. The Macro F1-score is computed by first calculating precision and recall independently for each of the three sentiment classes: negative, positive, and neutral, combining them into a class-specific F1 score, and then taking the unweighted average across all three classes. Because this calculation treats each class equally regardless of how many samples it contains, a model that performs well only on the majority classes while struggling with neutral comments would still show a lower Macro F1-score even if its overall accuracy remained high. Monitoring this metric therefore provides a more complete picture of whether the fine-tuned model is learning to distinguish all three sentiment categories, rather than simply favoring the classes with more training examples. Figure 8 presents the resulting training and validation Macro F1-score curves across the five training epochs.

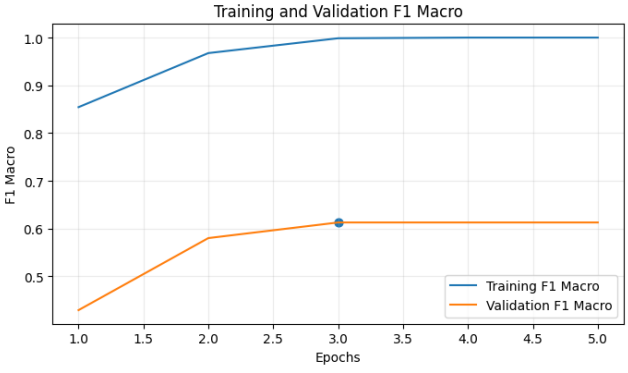

Figure 8. Training and Validation Macro F1-Score

Figure 8 illustrates the progression of the Training Macro F1-score and Validation Macro F1-score during the fine-tuning process of the Multilingual BERT-base model over five epochs. The Training Macro F1-score increased from approximately 0.86 in the first epoch to 1.00 by the third epoch and remained stable throughout the remaining training

BULLETIN OF COMPUTER SCIENCE RESEARCH
ISSN 2774-3659 (Media Online)
Vol 6, No 5, August 2026 | Hal 1909-1918
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v6i5.1292

epochs. In contrast, the Validation Macro F1-score improved from approximately 0.43 to 0.61, after which it remained relatively stable. The upward trend of both curves indicates that the model progressively improved its ability to classify the three sentiment classes during training. Nevertheless, the noticeable gap between the Training Macro F1-score and the Validation Macro F1-score suggests that the model performed better on the training data than on the validation data. This finding indicates that, although the model effectively learned the underlying patterns in the training dataset, its generalization capability on unseen data can still be improved, particularly for the minority class, which is more challenging to classify accurately. These findings confirm that the macro F1-score curves indicate that the Multilingual BERT-based model reached its optimal performance after the third epoch and maintained stable performance throughout the remaining training process. The stability of the Validation Macro F1-score demonstrates that the model achieved consistent classification performance, although there remains room for improvement in balancing performance across all sentiment classes. While the previous figures illustrate the overall trends in accuracy, loss, and Macro F1-score, the exact numerical values recorded at each epoch provide a clearer picture of the model's learning progress. These values are particularly useful for identifying the epoch at which the model began to converge and for confirming the absence of significant fluctuations during the fine-tuning process. Table 2 presents the training and validation loss values obtained throughout the five epochs of fine-tuning.

Table 2. Training and Validation Loss During BERT Fine-Tuning

Epoch	Training Loss	Validation Loss
1	1.472590	0.014702
2	0.300132	0.003259
3	0.007056	0.000975
4	0.001904	0.000690
5	0.000889	0.000636

A clear downward trend can be observed for both metrics, suggesting that the model continuously improved its predictions as the training process progressed. At the beginning of training, the model produced a training loss of 1.472590 and a validation loss of 0.014702. As additional epochs were completed, both values decreased markedly. By the third epoch, the training loss had been reduced to 0.007056, while the validation loss dropped to 0.000975. Further improvements were observed until the fifth epoch, where the training and validation losses reached 0.000889 and 0.000636, respectively. The steady decline in both loss values indicates that the fine-tuning procedure successfully optimized the model parameters, allowing the model to learn meaningful sentiment patterns while maintaining low prediction errors on unseen validation data. The consistently small validation loss during the final training stages further suggests that the model generalized well beyond the training dataset. Taken together, the results demonstrate that the Multilingual BERT-base model converged successfully during fine-tuning. The parallel reduction in both training and validation loss reflects stable learning behavior and indicates that the model achieved effective sentiment classification performance without exhibiting clear signs of substantial overfitting.

Complementing the loss values shown in Table 2, a more comprehensive evaluation of the model's classification performance was also conducted using four standard metrics: accuracy, precision, recall, and F1-score. These metrics provide a more detailed assessment of the model's ability to correctly classify each sentiment category, beyond what loss values alone can indicate. Table 3 presents the resulting evaluation scores recorded at each training epoch.

Table 3. Model Evaluation Results

Epoch	Accuracy	Precision	Recall	F1-Score
1	0.860056	0.895360	0.860056	0.854282
2	0.967651	0.970397	0.967651	0.967544
3	0.998594	0.998599	0.998594	0.998594
4	1.000000	1.000000	1.000000	1.000000
5	1.000000	1.000000	1.000000	1.000000

All evaluation metrics improved consistently as the number of training epochs increased, indicating that the model progressively learned the underlying sentiment patterns in the training dataset. In the first epoch, the model achieved an accuracy of 86.01%, precision of 89.54%, recall of 86.01%, and an F1-score of 85.43%. The model's performance improved substantially in the second epoch, with the accuracy increasing to 96.77%, and further approaching 99.86% in the third epoch. By the fourth and fifth epochs, all evaluation metrics reached 100%, indicating that the model correctly classified all training samples. The consistent improvement across all evaluation metrics demonstrates that the fine-tuning process was effective and that the model successfully converged during training. However, since the results reported in Table 3 were obtained from the training dataset, the near-perfect performance should be interpreted together with the validation or test results to ensure that the model generalizes well to unseen data and does not exhibit significant overfitting. In summary, the evaluation results indicate that the Multilingual BERT-base model achieved substantial performance improvements across successive training epochs and reached convergence by the fourth epoch. Although the model attained perfect scores on the training dataset, its overall effectiveness should ultimately be assessed based on its performance on the validation or test dataset, which provides a more reliable indicator of its generalization capability.

3.3 Discussion

Compared with earlier studies on this topic, the present findings both confirm and extend existing evidence. Studies [12] and [13] relied on conventional machine learning classifiers, namely Multinomial Naive Bayes, Support Vector Machine, and Multilayer Perceptron combined with Bag-of-Words, TF-IDF, or lexicon-based labeling, which primarily capture word frequency rather than contextual meaning. Study [14] applied a Backpropagation Neural Network, again relying on non-contextual feature engineering, while study [15] demonstrated the strong potential of hybrid BERT-based architectures for mental health sentiment analysis in general, achieving state-of-the-art accuracy, though its scope did not specifically address working hours, burnout culture, or the Indonesian TikTok context. In contrast, by fine-tuning Multilingual BERT-base as the primary classifier rather than using it only for automatic labeling, this study achieved near-perfect training performance and a stable validation accuracy of approximately 94%, suggesting that the contextual representations learned through self-attention provide a more reliable classification signal than frequency-based text representations, while remaining specifically tailored to this study's context. In terms of substantive findings, the predominance of negative sentiment and the frequent occurrence of terms related to working hours and resigning are consistent with the broader pattern reported in [16] and [17], which likewise identified excessive working hours and burnout as recurring concerns among Generation Z, even though those studies approached the topic from a health-epidemiological and a non-computational discussion perspective, respectively, rather than through sentiment analysis. This convergence across studies using different methods and disciplinary lenses strengthens confidence that the sentiment patterns identified are not an artifact of a single modeling approach. At the same time, the gap between training and validation performance observed in this study echoes a generalization challenge commonly reported in fine-tuned transformer-based sentiment analysis, indicating that, despite methodological improvements over prior lexicon-, frequency-, and non-transformer-based approaches, further work on data balancing and cross-platform validation remains necessary before these findings can be generalized beyond the TikTok dataset examined here.

4. CONCLUSIONS

This study demonstrates that a fine-tuned Multilingual BERT-base model can effectively classify Generation Z's sentiment toward working hours and mental health expressed in TikTok comments, achieving strong classification performance after converging in the later training epochs. Negative sentiment was found to be the most prevalent, followed by positive and neutral sentiment, with frequently occurring terms related to working hours, daily routines, and resigning indicating that Generation Z's online discussions center on long working hours and turnover intentions. For human resource management, these findings extend beyond a single organizational context and offer a broader lens through which HR practitioners and policymakers can understand the evolving expectations of Generation Z as they enter the workforce in increasing numbers. Public opinions expressed on social media can serve as an early warning signal of employee dissatisfaction, well before it surfaces through formal channels such as exit interviews or engagement surveys, enabling HR departments to shift from reactive to proactive talent retention strategies and to incorporate social listening as complementary tool alongside traditional workforce analytics. These findings can therefore inform the design of more flexible working-hour policies in line with international labour recommendations of 40–48 working hours per week, alongside mental health support programs such as privacy-preserving online counseling tailored to Generation Z's characteristics and values. More broadly, this study illustrates how sentiment analysis can serve as a methodological contribution to evidence-based human resource management, providing HR practitioners and organizational decision-makers with a scalable, data-driven approach to monitoring employee well-being that complements conventional HR instruments such as surveys and performance reviews. Nevertheless, the study is limited by its reliance on a single platform, an imbalanced sentiment distribution, and a gap between training and validation performance that leaves room for improved generalization; future research is therefore recommended to incorporate additional social media platforms, apply data-balancing techniques, compare Multilingual BERT with Indonesian-specific transformer models such as IndoBERT, and conduct qualitative analysis of negative sentiment comments to yield more targeted, evidence-based organizational policies that HR management and practitioners can act on with greater confidence.

REFERENCES

[1] W. H. O. World Health Organization, *WHO Guidelines on Mental Health at Work*. Geneva: World Health Organization (WHO), 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK586384/

[2] I. L. O. International Labour Organization, *Working Time and Work-Life Balance Around the World*, First edit. International Labour Organization (ILO), 2023. [Online]. Available: https://www.ilo.org/publications/working-time-and-work-life-balance-around-world

[3] S. U. Baek, Y. M. Lee, J. H. Yoon, and J. U. Won, "Long Working Hours, Work-life Imbalance, and Poor Mental Health: A Cross-sectional Mediation Analysis Based on the Sixth Korean Working Conditions Survey, 2020–2021," *J. Epidemiol.*, vol. 34, no. 11, pp. 535–542, 2024, doi: 10.2188/jea.JE20230302.

[4] A. S. Rivera, M. Akanbi, L. C. O'Dwyer, and M. McHugh, "Shift Work and Long Work Hours and their Association with Chronic Health Conditions: A systematic Review of Systematic Reviews with Meta-Analyses," *PLoS One*, vol. 15, no. 4, p. e0231037, Apr. 2020, doi: 10.1371/journal.pone.0231037.

[5] X. Niu, C. Li, and Y. Xia, "How does working time impact perceived mental disorders? New insights into the U-shaped relationship," *Front. Public Heal.*, vol. 2, 2024, doi: https://doi.org/10.3389/fpubh.2024.1402428.

[6] X. Chen, M. Masukujjaman, A. Al Mamun, J. Gao, and Z. K. M. Makhbul, "Modeling the significance of work culture on burnout, satisfaction, and psychological distress among the Gen-Z workforce in an emerging country," *Humanit. Soc. Sci. Commun.*, vol. 10, no. 1, p. 828, Nov. 2023, doi: 10.1057/s41599-023-02371-w.
[7] A. S. George, "The " Anti-Hustle " Ethos Among Generation Z Workers : An Investigation into Shifting Attitudes Towards Work-Life Balance Partners Universal Innovative Research Publication (PUIRP)," no. October, pp. 96–109, 2024, doi: 10.5281/zenodo.13993431.
[8] B. P. S. Badan Pusat Statistik, "Keadaan Pekerja di Indonesia Agustus 2025," Jakarta, 2025. [Online]. Available: https://www.bps.go.id/id/publication/2025/12/23/5fe58c9723acafc73fb95bff/keadaan-pekerja-di-indonesia-agustus-2025.html
[9] E. Codd, "World Mental Health Day 2025: Gen Zs and millennials on mental well-being at work," *Deloitte*, 2025. [Online]. Available: https://www.deloitte.com/global/en/issues/work/world-mental-health-day.html
[10] Z. Xueyun, A. Al Mamun, M. Masukujjaman, M. K. Rahman, J. Gao, and Q. Yang, "Modelling the significance of organizational conditions on quiet quitting intention among Gen Z workforce in an emerging economy," *Sci. Rep.*, vol. 13, no. 1, p. 15438, Sep. 2023, doi: 10.1038/s41598-023-42591-3.
[11] D. Listiorini, F. Feriandy, and W. Ningsih, "Work Life Balance and Mental Health Being of Generation Z in the Digital Industry," *Int. J. Asian Bus. Manag.*, vol. 4, no. 2, pp. 285–294, Apr. 2025, doi: 10.55927/ijabm.v4i2.170.
[12] I. Falujjah, A. Putri, A. H. Rahmatika, B. Okta, and S. Miranda, "Understanding Generation Z ' s Mental Health in Relation to Working Hours : A TikTok Sentiment Analysis Approach," *Multi Data Palembang Student Conf.*, vol. 5, no. 1, pp. 252–262, 2026.
[13] M. D. Al Fahreza, A. Luthfiarta, M. Rafid, and M. Indrawan, "Analisis Sentimen: Pengaruh Jam Kerja Terhadap Kesehatan Mental Generasi Z," *J. Appl. Comput. Sci. Technol.*, vol. 5, no. 1, pp. 16–25, Feb. 2024, doi: 10.52158/jacost.v5i1.715.
[14] N. Z. Farsya, A. Luthfiarta, Z. N. Maharani, and S. P. Ganiswari, "Optimizing Sentiment Analysis of Working Hours Impact on Generation Z's Mental Health Using Backpropagation," *J. MEDIA Inform. BUDIDARMA*, vol. 8, no. 3, p. 1555, Jul. 2024, doi: 10.30865/mib.v8i3.7827.
[15] M. M. Hossain, Sanjara, M. S. Hossain, S. Chaki, M. S. Rahman, and A. B. M. S. Ali, "Revolutionizing Mental Health Sentiment Analysis With BERT-Fuse: A Hybrid Deep Learning Model," *IEEE Access*, vol. 13, no. May, pp. 85428–85446, 2025, doi: 10.1109/ACCESS.2025.3568340.
[16] H.-L. Lin and W.-H. Liu, "The Impact of Night Shifts, Tobacco Dependence, Health Awareness, and Depression Risk on Chronic Disease Risk Among Generation Z Overtime Workers During the COVID-19 Pandemic," *Healthcare*, vol. 13, no. 5, p. 569, Mar. 2025, doi: 10.3390/healthcare13050569.
[17] A. M. Matilda, A. I. W. B. Prisca, and D. Darmanto, "UNDERSTANDING GEN Z'S MENTAL HEALTH CHALLENGES," *Phenom. Multidiscip. J. Sci. Res.*, vol. 3, no. 1, pp. 38–52, Feb. 2025, doi: 10.62668/phenomenon.v3i1.1402.
[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," 2020, doi: https://doi.org/10.48550/arXiv.2011.00677.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[20] S. K. Khatri and A. Srivastava, "Using Sentimental Analysis in Prediction of Stock Market Investment," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Sep. 2016, pp. 566–569. doi: 10.1109/ICRITO.2016.7785019.
[21] A.-M. Iddrisu, S. Mensah, F. Boafo, G. R. Yeluripati, and P. Kudjo, "A sentiment Analysis Framework to Classify Instances of Sarcastic Sentiments Within the Aviation Sector," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 2, p. 100180, Nov. 2023, doi: 10.1016/j.jjimei.2023.100180.
[22] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Comput. Sci.*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.
[23] L. Zhang, S. Wang, and B. Liu, "Deep Learning for Sentiment Analysis: A Survey," *WIREs Data Min. Knowl. Discov.*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1253.
[24] N. Darraz, I. Karabila, A. El-Ansari, N. Alami, and M. El Mallahi, "Integrated Sentiment Analysis with BERT for Enhanced Hybrid Recommendation Systems," *Expert Syst. Appl.*, vol. 261, p. 125533, Feb. 2025, doi: 10.1016/j.eswa.2024.125533.
[25] A. S. Talaat, "Sentiment analysis classification system using hybrid BERT models," *J. Big Data*, vol. 10, no. 1, p. 110, Jun. 2023, doi: 10.1186/s40537-023-00781-w.
[26] R. Obiedat, D. Al-Darras, E. Alzaghoul, and O. Harfoushi, "Arabic Aspect-Based Sentiment Analysis: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 152628–152645, 2021, doi: 10.1109/ACCESS.2021.3127140.
[27] K. Chakraborty, S. Bhattacharyya, R. Bag, and L. Mršić, "Sentiment Analysis on Labeled and Unlabeled Datasets Using BERT Architecture," *Soft Comput.*, vol. 28, no. 15–16, pp. 8623–8640, Aug. 2024, doi: 10.1007/s00500-023-08876-5.
[28] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on Sentiment Analysis: Evolution of Research Methods and Topics," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8469–8510, Aug. 2023, doi: 10.1007/s10462-022-10386-z.
[29] B. Wan, P. Wu, C. K. Yeo, and G. Li, "Emotion-cognitive Reasoning Integrated BERT for Sentiment Analysis of Online Public Opinions on Emergencies," *Inf. Process. Manag.*, vol. 61, no. 2, p. 103609, Mar. 2024, doi: 10.1016/j.ipm.2023.103609.
[30] M. P. Geetha and D. Karthika Renuka, "Improving the Performance of Aspect Based Sentiment Analysis Using Fine-Tuned Bert Base Uncased Model," *Int. J. Intell. Networks*, vol. 2, pp. 64–69, 2021, doi: 10.1016/j.ijin.2021.06.005.
[31] C. H. Miranda, G. Sanchez-Torres, and D. Salcedo, "Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture," *Data*, vol. 8, no. 6, p. 96, May 2023, doi: 10.3390/data8060096.
[32] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, p. 4157, May 2022, doi: 10.3390/s22114157.
[33] A. Borg and M. Boldt, "Using VADER Sentiment and SVM for Predicting Customer Response Sentiment," *Expert Syst. Appl.*, vol. 162, p. 113746, Dec. 2020, doi: 10.1016/j.eswa.2020.113746.
[34] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, "Sentiment Analysis of Consumer Reviews Using Deep Learning," *Sustainability*, vol. 14, no. 17, p. 10844, Aug. 2022, doi: 10.3390/su141710844.