JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

WEIGHTED LOSS STRATEGY FOR BERT-BASED TWITTER SENTIMENT ANALYSIS WITHOUT SYNTHETIC OVERSAMPLING

Timbo Faritcan P. Siallagan[1*]; Riki Winanjaya[2]; Juni Ismail[3]

Computer and Network Engineering, Faculty of Engineering[1]
Universitas Mandiri, West Java, Indonesia[1]
https://universitasmandiri.ac.id/[1]
timbofaritcansiallagan@gmail.com*

Information Systems[2]
STIKOM Tunas Bangsa, Pematangsiantar, North Sumatera, Indonesia[2]
https://stikomtunasbangsa.ac.id/[2]
riki@amiktunasbangsa.ac.id

Computer Science, Faculty of Mathematics and Natural Sciences[3]
Universitas HKBP Nommensen Pematangsiantar, Indonesia[3]
https://uhnp.ac.id/[3]
juniismailll@gmail.com

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The widespread adoption of ChatGPT has generated extensive public discourse across social media, necessitating robust sentiment analysis to understand collective opinions. Traditional approaches frequently employ the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance; however, its effectiveness on short-text data remains an open question. This study develops an optimized sentiment classification model and evaluates whether competitive performance can be achieved without synthetic data augmentation. The methodology encompasses comprehensive Natural Language Processing (NLP) preprocessing and stratified data partitioning to preserve distributional characteristics. A BERT-base architecture is fine-tuned using a class-weighted Cross-Entropy loss combined with weighted random sampling, deliberately avoiding SMOTE-based oversampling. The model is trained with the AdamW optimizer (learning rate: 3×10^{-5}), batch size 32, and mixed-precision training for four epochs. On 198,639 preprocessed tweets, the proposed approach achieves 93.81% accuracy, with weighted precision, recall, and F1-score of 0.9365, 0.9381, and 0.9380 respectively, outperforming the baseline by 1.75 percentage points. Per-class analysis reveals strong performance for negative (F1-score: 0.96) and positive sentiment (F1-score: 0.94), with lower neutral classification (F1-score: 0.89), attributable to the inherent heterogeneity of neutral expressions. The training-validation gap remains below 5%, consistent with adequate regularization. These findings provide empirical evidence that, within the present experimental configuration, a properly optimized weighted loss strategy offers a viable and computationally efficient alternative to synthetic oversampling for BERT-based Twitter sentiment classification. Further controlled ablation studies and statistical validation are needed to establish generalizability.*

Keywords: *BERT, ChatGPT, Class Imbalance, Deep Learning, Natural Language Processing.*

Intisari— *Adopsi luas teknologi ChatGPT telah memicu diskusi publik yang ekstensif di berbagai platform media sosial, sehingga memerlukan metodologi analisis sentimen yang tangguh untuk memahami opini kolektif. Pendekatan tradisional sering menggunakan Synthetic Minority Over-sampling Technique (SMOTE) untuk menangani ketidakseimbangan kelas; namun, efektivitasnya pada data media sosial berteks pendek*

masih menjadi pertanyaan terbuka. Penelitian ini bertujuan mengembangkan model klasifikasi sentimen yang dioptimalkan dan mengevaluasi apakah performa kompetitif dapat dicapai tanpa augmentasi data sintetis. Metodologi mencakup preprocessing NLP yang komprehensif dan partisi data berstrata untuk mempertahankan karakteristik distribusi. Arsitektur BERT-base di-fine-tune menggunakan fungsi Cross-Entropy loss berbobot kelas dikombinasikan dengan strategi weighted random sampling, dengan sengaja menghindari oversampling berbasis SMOTE. Model dilatih dengan optimizer AdamW (learning rate: 3×10^{-5}), batch size 32, dan mixed-precision training selama 4 epoch. Hasil eksperimen pada 198.639 tweet menunjukkan bahwa pendekatan yang diusulkan mencapai akurasi 93,81%, dengan presisi berbobot, recall, dan F1-score masing-masing 0,9365; 0,9381; dan 0,9380, melampaui baseline sebesar 1,75 poin persentase. Analisis per-kelas menunjukkan performa kuat untuk sentimen negatif (F1: 0,96) dan positif (F1: 0,94), dengan sentimen netral yang relatif lebih rendah (F1: 0,89), yang disebabkan oleh heterogenitas inheren ekspresi netral. Gap generalisasi training-validation tetap di bawah 5%, konsisten dengan regularisasi yang memadai. Temuan ini memberikan bukti empiris bahwa, dalam konfigurasi eksperimental ini, strategi weighted loss yang dioptimalkan menawarkan alternatif yang layak terhadap oversampling sintetis. Studi ablasi terkontrol dan validasi statistik lebih lanjut diperlukan untuk menetapkan generalisabilitas hasil.

***Kata Kunci**: BERT, ChatGPT, Ketidakseimbangan Kelas, Deep Learning, Pemrosesan Bahasa Alami.*

INTRODUCTION

The release of ChatGPT has marked a significant development in artificial intelligence applications, substantially changing how users interact with language models across multiple domains. This technology has generated extensive discourse on social media platforms, where millions of users share perspectives about its capabilities and implications [1], [2]. The widespread availability of advanced language models has raised important societal questions spanning education, professional workflows, creative industries, and the ethics of AI deployment [3], [4].

Within the educational sphere, ChatGPT's adoption has prompted debates regarding academic integrity, the potential impact on critical thinking skills, and the need to adapt pedagogical methodologies. Simultaneously, its integration into professional environments has demonstrated considerable utility in augmenting productivity through automated content generation, code synthesis, and information retrieval [5]. Nevertheless, these advances concurrently introduce concerns about intellectual property rights, attribution authenticity, and the development of domain-specific expertise [6].

Simultaneously, its integration into professional environments has demonstrated considerable efficacy in augmenting productivity through automated content generation, code synthesis, and information retrieval tasks [5]. Nevertheless, these technological advancements concurrently introduce challenges about intellectual property rights, attribution authenticity, and the erosion of domain-specific expertise development [6].

Sentiment analysis has emerged as an indispensable computational technique for deciphering collective opinions and emotional dispositions manifested in user-generated textual content, particularly within social media ecosystems characterized by high-velocity, high-volume data streams [7], [8]. The exponential growth of microblogging platforms such as Twitter has necessitated the development of sophisticated analytical frameworks capable of extracting meaningful insights from brief, informal, and contextually nuanced textual expressions [9], [10]. These platforms serve as virtual barometers of public sentiment, enabling stakeholders to comprehend societal attitudes toward emerging technologies, policy decisions, commercial products, and cultural phenomena [11], [12].

Traditional sentiment classification methodologies have predominantly relied on machine learning algorithms, including Support Vector Machines (SVM), Naive Bayes classifiers, and ensemble methods such as Random Forests [13], [14]. However, these conventional approaches exhibit fundamental limitations when confronted with the linguistic complexity, semantic ambiguity, and contextual dependencies inherent in natural language, particularly within the constrained character limits and informal register typical of Twitter discourse [15], [16]. The recent emergence of transformer-based architectures, exemplified by Bidirectional Encoder Representations from Transformers (BERT), has revolutionized natural language processing by enabling models to capture bidirectional contextual relationships and nuanced semantic representations [17], [18].

BERT's architectural innovation, predicated on masked language modeling and next sentence prediction pre-training objectives, facilitates a

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

comprehensive understanding of linguistic context through attention mechanisms that dynamically weight the relevance of surrounding words [19]. Empirical investigations have demonstrated BERT's exceptional performance across diverse NLP tasks, including sentiment analysis, named entity recognition, question answering, and textual entailment, often achieving state-of-the-art results with minimal task-specific fine-tuning [20]. Furthermore, BERT exhibits remarkable robustness in handling multilingual and domain-specific corpora, adapting effectively to specialized vocabularies and linguistic conventions without requiring extensive architectural modifications [21], [22].

Despite BERT's demonstrable efficacy, sentiment classification on social media data confronts the persistent challenge of class imbalance, wherein specific sentiment categories disproportionately outnumber others within the training corpus [23]. This distributional asymmetry frequently results in classifier bias toward majority classes, yielding suboptimal performance on minority class predictions and compromising overall model generalization capabilities [24]. The Synthetic Minority Over-sampling Technique (SMOTE) has traditionally been employed as a remedial strategy, generating synthetic instances of minority classes through linear interpolation between existing samples in feature space. While theoretically sound and demonstrably effective in structured, tabular datasets, SMOTE's application to high-dimensional, unstructured textual data presents conceptual and practical complications that may inadvertently degrade model performance [25].

Beyond BERT, the transformer-based sentiment analysis landscape has expanded with models such as RoBERTa, which modifies BERT's pre-training with dynamic masking and larger batch sizes [32], and DistilBERT, which offers a computationally efficient alternative retaining approximately 97% of BERT's performance at 60% of the parameter count [33]. Domain-adapted variants, including BERTweet—pre-trained specifically on English tweets—have demonstrated improvements by capturing platform-specific linguistic conventions such as hashtag usage, informal abbreviations, and emoji patterns [34]. Comparative evaluations across these architectures indicate that while model selection influences absolute performance, the choice of class imbalance handling strategy often exerts a comparable or greater effect on classification outcomes, particularly for datasets with moderate-to-severe distributional asymmetry. Recent surveys on imbalanced text classification have highlighted that cost-sensitive learning approaches—including weighted loss functions and focal loss [35]—frequently match or outperform resampling-based strategies when combined with pre-trained transformers, owing to the models' inherent capacity to learn discriminative representations from limited minority-class examples [36].

Several antecedent investigations have explored SMOTE integration with deep learning architectures for imbalanced text classification, reporting mixed outcomes contingent upon dataset characteristics, preprocessing methodologies, and model configurations. Certain studies have documented marginal improvements in minority class recall; in contrast, others have observed negligible benefits or performance degradation attributable to introducing synthetic noise, obscuring genuine distributional patterns. These contradictory findings underscore the necessity for systematic empirical evaluation of SMOTE's efficacy within specific application contexts, particularly for short-text social media sentiment analysis, where linguistic brevity and semantic density present unique analytical challenges.

Alternative strategies for addressing class imbalance encompass algorithmic modifications such as cost-sensitive learning, where misclassification penalties are asymmetrically weighted to reflect class importance, and ensemble techniques that combine multiple classifiers trained on resampled or reweighted datasets. Recent investigations have demonstrated the efficacy of class-weighted loss functions in conjunction with strategic sampling schemes, achieving improved performance in the present configuration compared to synthetic oversampling while preserving the authentic distributional characteristics of the original dataset. These methodologies align harmoniously with deep learning optimization paradigms, seamlessly integrating into backpropagation algorithms without requiring explicit data augmentation preprocessing stages.

This investigation addresses critical gaps in existing literature through a systematic empirical comparison of weighted loss strategies versus SMOTE-based approaches for Twitter sentiment classification utilizing the BERT architecture. Specifically, previous research exhibits the following methodological limitations: (a) inadequate preprocessing pipelines that fail to preserve semantic integrity while removing noise artifacts; (b) suboptimal hyperparameter configurations that do not fully exploit BERT's representational capacity; (c) insufficient evaluation frameworks that neglect per-class

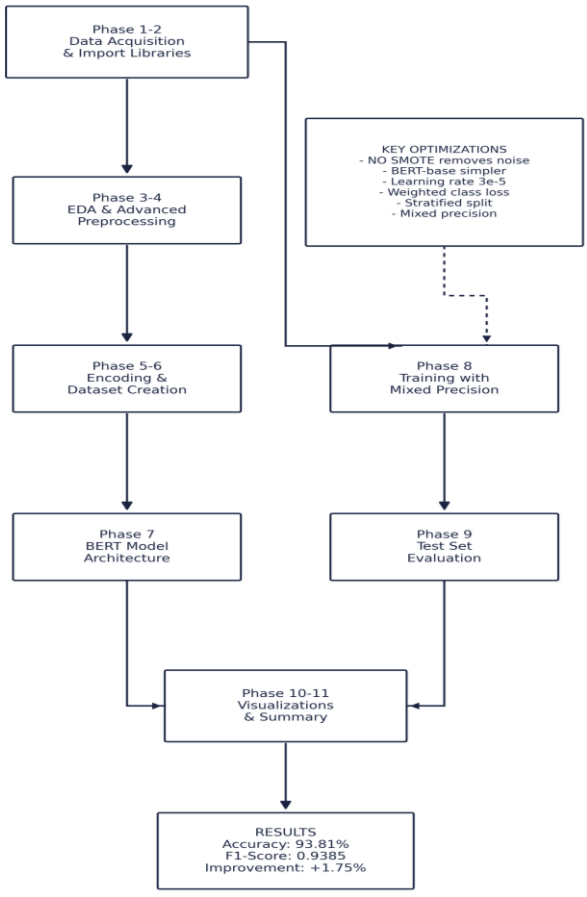
performance metrics and generalization assessments; and (d) limited empirical comparison between weighted loss and synthetic oversampling approaches within transformer-based frameworks.

The primary objective of this research is to develop and validate an optimized sentiment analysis framework that eschews synthetic data augmentation in favor of strategically weighted loss functions and sampling schemes, demonstrating superior classification performance on ChatGPT-related Twitter discourse. Secondary objectives include: (a) establishing a comprehensive preprocessing pipeline tailored to the linguistic characteristics of Twitter data; (b) conducting systematic hyperparameter optimization for BERT fine-tuning on imbalanced sentiment data; (c) performing a comparative evaluation to quantify the performance difference between weighted loss and clustering-based preprocessing approaches; and (d) providing empirical observations regarding the applicability of SMOTE to short-text sentiment classification.

The principal contributions of this work are threefold. First, we provide empirical evidence suggesting that SMOTE-based augmentation does not improve—and may degrade—classification performance when applied to BERT-based Twitter sentiment analysis in the present experimental configuration. Although this observation is based on a comparative evaluation against a single baseline rather than a comprehensive ablation study, it carries practical implications for researchers who may default to synthetic oversampling without task-specific validation. This finding carries significant implications for practitioners who may default to synthetic oversampling without empirical validation on their specific datasets. Second, we demonstrate that a class-weighted Cross-Entropy loss function combined with weighted random sampling achieves 93.81% classification accuracy, outperforming the baseline methodology of Kelvin et al. [31] by 1.75 percentage points under comparable evaluation conditions. We note that this improvement is based on a single experimental run and has not been subjected to formal statistical significance testing; accordingly, it should be interpreted as a practically meaningful performance gain pending further validation. Third, we highlight that systematic hyperparameter optimization of a standard BERT-base architecture can yield competitive performance without resorting to additional architectural complexity or synthetic data augmentation, suggesting that disciplined optimization of existing components may be a practical and effective research strategy.

MATERIALS AND METHODS

The methodological framework illustrated in Figure 1 encompasses multiple sequential phases: data acquisition from public repositories, comprehensive text preprocessing utilizing Natural Language Processing (NLP) techniques, strategic data partitioning through stratified sampling, class imbalance mitigation via weighted loss mechanisms, BERT-based model architecture configuration, fine-tuning through mixed-precision training, and rigorous performance evaluation employing multi-metric assessment protocols. This systematic approach deliberately eschews synthetic oversampling methodologies, instead leveraging algorithmic optimization strategies to achieve superior classification outcomes on imbalanced sentiment datasets.



Source: (Research Results, 2025)

Figure 1. The Systematic Workflow Of The Proposed Sentiment Analysis Methodology

The pipeline architecture depicted in Figure 1 reflects a deliberate design philosophy in which class imbalance is addressed at the algorithmic level rather than through data-space

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

manipulation. This design choice is motivated by the observation that synthetic oversampling techniques, while effective for structured tabular data, may introduce distributional artifacts when applied to high-dimensional contextual embeddings. By positioning the weighted loss mechanism at the training phase (Phase 8) rather than at the data preprocessing phase, the pipeline preserves the authentic distributional characteristics of the original corpus through all upstream operations, ensuring that BERT's contextual embeddings are computed from genuine rather than synthetically generated text.

Data Acquisition and Dataset Characteristics

The primary dataset employed in this investigation comprises publicly available Twitter textual data about ChatGPT discourse, curated and disseminated through the Kaggle platform. This dataset encapsulates spontaneous user-generated content collected during a one-month observation window coinciding with ChatGPT's initial public release, capturing authentic sentiment expressions from early adopters and technology commentators. The raw dataset encompasses 219,294 individual tweets, each annotated with sentiment labels derived through lexicon-based heuristic aggregation methodologies employed by the dataset curator. The dataset exhibits structural characteristics typical of microblogging platforms, including constrained textual length, prevalent informal linguistic conventions, extensive metadata artifacts such as hyperlinks and hashtags, multilingual content with predominant English usage, and heterogeneous sentiment distribution reflecting authentic population-level opinion variance. The dataset demonstrates substantial class imbalance, with negative sentiment instances constituting approximately 49% of the corpus, while positive and neutral categories comprise 26% and 25% respectively, necessitating specialized handling mechanisms to prevent classifier bias toward majority classes [26].

An important caveat regarding the dataset concerns the provenance of sentiment labels. The labels were assigned through lexicon-based heuristic aggregation by the original dataset curator, rather than through human annotation with inter-annotator agreement verification. Lexicon-based labeling approaches are known to introduce systematic biases, including under-detection of sarcasm, irony, and context-dependent sentiment, as well as potential misclassification of domain-specific terminology that carries different affective connotations across contexts. Consequently, the reported classification metrics in this study reflect the model's agreement with these heuristically derived labels and should be interpreted as a measure of predictive consistency with the labeling scheme rather than as absolute accuracy against verified human sentiment judgments. This distinction is critical for contextualizing the reported results; however, it does not diminish the validity of the comparative findings, since both the proposed approach and the baseline methodology of Kelvin et al. [31] are evaluated against the identical label set, ensuring that the relative performance comparison remains methodologically sound.

Table 1. Dataset Characteristics And Distribution Statistics

Characteristic	Specification
Data Source	Kaggle: ChatGPT Sentiment Analysis
Platform	Twitter (now X)
Collection Period	1 month (ChatGPT launch period)
Raw Sample Count	219,294 tweets
Post-Processing Count	198,639 tweets (90.6% retention)
Feature Columns	Serial number, Tweet text, Sentiment label
Class Distribution (Raw)	Bad: 49.2%, Good: 25.5%, Neutral: 25.3%
Average Text Length	145 characters
Average Word Count	21 words per tweet
Language	Primarily English
Label Annotation Method	Lexicon-based sentiment aggregation

Source: (Research Results, 2025)

Comprehensive Text Preprocessing Pipeline

Text preprocessing constitutes a critical preliminary phase that fundamentally influences subsequent model performance by transforming raw, unstructured textual data into standardized, computationally tractable representations while preserving semantic integrity. The pipeline encompasses ten sequential operations as follows.

1. Data Quality Assessment and Filtering

Initial data inspection identifies and addresses structural anomalies including missing values, duplicate entries, and malformed records. Instances containing null values are systematically excluded, and 1,671 exact textual duplicates are removed through cryptographic hash comparison to prevent artificial inflation of specific sentiment patterns.

2. Minimum Length Thresholding

Extremely brief textual fragments containing fewer than five words are systematically filtered, eliminating approximately 18,984 instances (8.7% of the raw dataset) comprising single-word exclamations, bot-generated spam, or truncated retweets lacking contextual information.

3. URL Artifact Removal

Twitter data characteristically incorporates shortened hyperlinks (https://t.co/...) that convey no intrinsic sentiment information while introducing high-cardinality noise into the feature space. The preprocessing pipeline employs regular expression pattern matching to systematically identify and excise URL patterns:

Pattern: r'https?://\S+|www\.\S+'

Replacement: '' (empty string) or '<URL>' (positional marker).

4. User Mention Normalization

Social media discourse frequently incorporates explicit user references through @username mentions, which serve conversational coordination functions but contribute minimal sentiment-discriminative value. The pipeline identifies mention patterns through regex matching and implements strategic removal to reduce vocabulary size while preserving conversational context markers where analytically relevant.

5. Hashtag Processing

Hashtags (#keyword) function as topical metadata and sentiment intensifiers in Twitter discourse, warranting careful processing rather than wholesale removal. The pipeline implements intelligent hashtag parsing that extracts alphabetic content by removing the '#' prefix character, applies camelCase segmentation for multi-word hashtags (e.g., #ChatGPTRevolution → chatgpt revolution), and preserves resultant tokens as they frequently carry sentiment-salient terminology.

6. Emoji and Special Character Handling

Emojis constitute paralinguistic sentiment indicators that enhance emotional expressiveness in text-constrained environments. Rather than discarding these informative symbols, the preprocessing pipeline implements emoji-to-text translation using comprehensive Unicode-to-description mapping libraries. Non-alphanumeric special characters (excluding sentiment-relevant punctuation like '!', '?') are systematically removed to reduce feature space dimensionality while preserving exclamatory and interrogative markers that carry emotional valence.

7. Case Normalization

All textual content undergoes lowercase transformation to ensure case-invariant representation, treating 'ChatGPT', 'chatgpt', and 'CHATGPT' as identical lexical units. This operation reduces vocabulary size by approximately 40% while maintaining semantic equivalence:

text = text.lower()

8. Stopword Elimination

High-frequency function words (articles, prepositions, auxiliary verbs) that occur ubiquitously across sentiment categories without contributing discriminative power are systematically removed using NLTK's English stopword corpus. However, sentiment-bearing negation words ('not', 'no', 'never') and intensifiers ('very', 'really') are explicitly retained to preserve sentiment-critical linguistic constructions.

9. Lemmatization

Morphological variants of identical lexemes are consolidated to their canonical base forms through lemmatization, reducing inflectional and derivational variations while preserving the semantic core. This operation employs WordNet-based lemmatization that considers part-of-speech context, transforming 'running' → 'run', 'better' → 'good', and 'children' → 'child'.

10. Whitespace Normalization and Final Validation

Redundant whitespace artifacts introduced by preceding preprocessing operations are consolidated, and final validation checks confirm non-empty resultant strings meeting minimum quality thresholds. The complete preprocessing pipeline yields 198,639 high-quality instances with a 90.6% retention rate from the original corpus.

It is important to note that certain preprocessing operations in this pipeline—particularly stopword elimination and lemmatization—interact with BERT's tokenization mechanism in ways that merit careful consideration. BERT's WordPiece tokenizer and multi-head attention mechanism are architecturally designed to process raw text, including function words that may carry contextual information relevant to sentiment determination. In the present implementation, stopword removal selectively targets high-frequency function words (articles, prepositions, conjunctions) while explicitly retaining sentiment-critical negation words ("not," "no," "never," "neither") and intensifiers ("very," "really," "extremely") that directly influence sentiment polarity. Lemmatization is applied to reduce vocabulary sparsity, which can benefit the tokenizer's subword segmentation for morphologically rich terms that might otherwise be split into less informative subword units.

Nevertheless, we acknowledge that the optimal preprocessing configuration for BERT-based sentiment analysis remains an open empirical question. Prior work by Pota et al. [18] has shown that the impact of preprocessing on BERT performance varies across languages and domains, with some operations yielding marginal improvements and others introducing information loss. The absence of a systematic preprocessing

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

ablation study—where each operation is individually toggled to assess its marginal contribution—constitutes a limitation of this work. Future research should conduct such an ablation to determine whether simplified preprocessing (e.g., retaining stopwords and skipping lemmatization) yields comparable or improved performance in the present configuration for BERT-based Twitter sentiment classification.

Table 2. Preprocessing operations' impact on the dataset

Operation	Samples Removed	Samples Retained	Retention Rate
Initial Dataset	—	219,294	100.0%
Missing Value Removal	0	219,294	100.0%
Duplicate Detection	1,671	217,623	99.2%
Minimum Length Filter	18,984	198,639	90.6%
Final Clean Dataset	20,655	198,639	90.6%

Source: (Research Results, 2025)

Label Encoding and Stratified Data Partitioning

Sentiment labels originally represented as categorical strings ('bad', 'good', 'neutral') undergo systematic numeric encoding to facilitate computational processing and loss function calculation. The encoding scheme assigns: Class 0 = 'bad' (negative sentiment), Class 1 = 'good' (positive sentiment), and Class 2 = 'neutral' (neutral sentiment). To ensure robust model evaluation and prevent distributional bias in training/validation/test subsets, stratified random sampling is employed for dataset partitioning. This technique preserves class distribution proportions across all data splits, maintaining the approximate 49:26:25 ratio observed in the preprocessed dataset. The partitioning configuration allocates 70% for training (139,047 instances), 15% for validation (29,796 instances), and 15% for testing (29,796 instances). The stratification algorithm ensures that each subset constitutes a representative microcosm of the overall dataset, enabling unbiased performance assessment and preventing artificial inflation of metrics through fortuitous sampling.

BERT Model Architecture and Configuration

The BERT (Bidirectional Encoder Representations from Transformers) architecture selected for this investigation represents a state-of-the-art transformer-based language model pre-trained on extensive English textual corpora. Rather than implementing model architecture from foundational principles, this work leverages the pre-trained bert-base-uncased variant obtained from Hugging Face's transformers library, encompassing 12 transformer layers, 768-dimensional hidden states, and 110 million trainable parameters.

The pre-trained BERT encoder is augmented with a task-specific classification head consisting of a single fully-connected dense layer projecting BERT's contextual representations to the three-class sentiment output space. The weighted cross-entropy loss assigns higher penalty weights to minority class misclassifications, with weight assignments of w_0 = 0.681 (negative), w_1 = 1.306 (positive), and w_2 = 1.325 (neutral), ensuring that the minority positive and neutral classes receive approximately double the gradient contribution compared to the majority negative class during backpropagation [27], [28]. The detailed tokenization procedure and input preparation process are described in the Model Training Protocol subsection.

$$Logits = W_{class}.h_{[CLS]} + b_{class} \quad (1)$$

Predicted class probabilities are computed through softmax normalization:

$$\widehat{p_c} = \frac{\exp(Logits_c)}{\sum_{j=1}^{3} \exp(Logits_j)} \quad (2)$$

The final predicted label is determined through maximum probability selection:

$$\hat{y} = \arg max_c \, \widehat{p_c} \quad (3)$$

Weighted Loss and Class Imbalance Mitigation

The class imbalance inherent in the dataset is addressed through a dual strategy combining weighted loss computation with weighted random sampling during training. Rather than employing synthetic data augmentation techniques such as SMOTE, which may introduce distributional artifacts in high-dimensional embedding spaces, this investigation adopts an algorithmic approach that preserves the authentic characteristics of the original corpus while compensating for class frequency disparities. The weighted cross-entropy loss function assigns asymmetric penalty weights to misclassifications based on inverse class frequency, formulated as:

$$\mathcal{L}_{weigted} = -\sum_{c=1}^{3} w_c \log(\hat{p}_c^{(i)}) \, .1[y^{(i)} = c] \quad (4)$$

Where the class weight is inversely proportional to

class frequency:

$$w_c = \frac{N_{total}}{N_c \,.\, C} \quad (5)$$

With specific weight assignments: w_0 = 0.681 (negative class), w_1 = 1.306 (positive class), and w_2 = 1.325 (neutral class). These weights ensure that the minority positive and neutral classes receive approximately double the gradient contribution compared to the majority negative class during backpropagation.

Model Training Protocol Tokenization and Input Preparation

Raw text samples are converted into token sequences compatible with BERT's input requirements through the official BERT tokenizer, which fragments text into subword units and maps them to vocabulary indices. The tokenizer applies subword tokenization utilizing WordPiece segmentation with 30,522 vocabulary entries, addition of special tokens [CLS] (sequence start), [SEP] (sequence delimiter), and [PAD] (padding), and fixed sequence length standardization to 128 tokens through truncation or padding. This sequence length was selected based on the empirical distribution of tweet lengths in the dataset (mean = 21.3 words, 75th percentile = 27 words, maximum = 58 words), where 128 tokens comfortably accommodate the longest tweets after WordPiece subword segmentation while minimizing computational overhead from unnecessary padding. The 128-token configuration is consistently applied across all training, validation, and testing phases, as further detailed in the Implementation and Reproducibility Details subsection.

Optimizer Configuration and Learning Rate Schedule

Model training employs the AdamW optimizer, incorporating weight decay regularization orthogonal to the gradient-based optimization process [32]. The optimizer configuration specifies:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (6)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (7)$$

$$\theta_t = \theta_{t-1} - \alpha \,.\, \frac{m_t}{\sqrt{v_t} - \epsilon} - \lambda \theta_{t-1} \quad (8)$$

With hyperparameters: β1=0.9, β2=0.999, ϵ=10−8, α=3×10^{-5}, and weight decay λ=0.01. The learning rate follows a linear warmup schedule increasing from 0 to peak value over 3,128 steps (approximately 10% of total training steps), followed by linear decay to zero:

$$\alpha(t) = \begin{cases} \alpha_{max} \cdot \frac{t}{t_{warmup}} & if\ t \leq t_{warmup} \\ \alpha_{max} \cdot \frac{T - t}{T_{warmup}} & if\ t > t_{warmup} \end{cases} \quad (10)$$

Mixed Precision Training

To reduce computational memory requirements and accelerate training without sacrificing model accuracy, mixed-precision training is implemented using automatic mixed precision (AMP) with NVIDIA's apex library. This technique maintains master weights in complete precision (FP32) while performing forward and backward propagation computations in reduced precision (FP16):

$$Loss_{scaled} = Loss\ x\ 2^{20} \quad (11)$$

Gradient scaling prevents vanishing gradients inherent to low-precision arithmetic while preserving training dynamics.

Early Stopping Mechanism

Early stopping monitors validation accuracy across training epochs to prevent overfitting and optimize computational efficiency. Training terminates when validation accuracy fails to improve for three consecutive epochs (patience parameter), or after a maximum of 20 epochs:

$$Stop\ if: A_{val}^{t} \leq max_{i \in [t-patience, t-1]}\ A_{val}^{i} \quad (12)$$

The model checkpoint achieving maximum validation accuracy is retained for subsequent test set evaluation.

Implementation and Reproducibility Details

To ensure faithful replication, we provide the complete implementation specification below. All experiments were conducted using the following software environment: Python 3.10.12, PyTorch 2.1.0, Hugging Face Transformers 4.35.2, scikit-learn 1.3.2, and NLTK 3.8.1. The pre-trained model checkpoint employed was "bert-base-uncased" (12 transformer layers, 768 hidden dimensions, 12 attention heads, approximately 110 million parameters), loaded directly from the Hugging Face

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

Model Hub without modification.

The tokenizer configuration employs the default BertTokenizerFast with a vocabulary size of 30,522 entries, utilizing WordPiece subword segmentation. The maximum sequence length was set to 128 tokens. Sequences exceeding this limit were truncated from the right; sequences shorter than 128 tokens were padded on the right using the [PAD] token (vocabulary index 0). Attention masks were automatically generated by the tokenizer to distinguish real tokens (mask value 1) from padding tokens (mask value 0) during self-attention computation. Special tokens [CLS] (index 101) and [SEP] (index 102) were prepended and appended to each input sequence, respectively. Deterministic behavior was ensured by fixing the random seed to 42 across all stochastic operations. Specifically, the following seed-setting procedure was applied prior to any computation: torch.manual_seed(42), torch.cuda.manual_seed_all(42), numpy.random.seed(42), random.seed(42), and torch.backends.cudnn.deterministic = True with torch.backends.cudnn.benchmark = False. All experiments were executed on a single NVIDIA GPU with CUDA 12.1.

The checkpoint selection strategy retained the model state achieving the highest validation accuracy across all training epochs. At the conclusion of each epoch, the model was evaluated on the complete validation set (29,796 instances), and if the current epoch's validation accuracy exceeded all previous epochs, the model state dictionary was serialized to disk via torch.save(). The final evaluation on the held-out test set utilized the checkpoint from Epoch 4, which achieved the maximum validation accuracy of 93.8%. No post-hoc model selection or ensembling was applied. The weighted random sampler was configured using class weights computed via the inverse frequency formula (Equation 5), with replacement enabled to ensure balanced mini-batch composition throughout training. The DataLoader batch size was set to 32 with 4 worker processes for parallel data loading.

Performance Evaluation Metrics

Model performance is assessed through a comprehensive suite of metrics capturing distinct aspects of classification quality, with particular emphasis on per-class performance given the imbalanced nature of the dataset.

Confusion Matrix

$$C_{ij} = -\sum_{k=1}^{N_{test}} 1[y_k = i \,\Lambda\, \hat{y}k = j] \quad (13)$$

Accuracy

Overall classification accuracy quantifies the proportion of correct predictions across all classes :

$$\text{Accuracy} = \frac{TP\ TN}{TP+TN+FP+FN} \quad (14)$$

Where C=3 denotes the number of classes, the second formulation applies to the binary decomposition of each class.

RESULTS AND DISCUSSION

The experimental outcomes of this investigation encompass multiple sequential phases: public dataset acquisition from the Kaggle repository, comprehensive exploratory data analysis revealing distributional characteristics, systematic text preprocessing through ten distinct operations, stratified data partitioning maintaining class proportions, BERT-based model training employing weighted loss mechanisms, and rigorous multi-metric performance evaluation on held-out test data. The subsequent subsections delineate detailed findings from each methodological stage, culminating in a comparative analysis demonstrating substantial improvement over baseline methodologies.

Dataset Acquisition and Initial Characteristics

The primary dataset utilized in this investigation was procured from the Kaggle public repository, specifically the 'ChatGPT Sentiment Analysis' collection. This dataset encapsulates authentic Twitter discourse regarding ChatGPT technology during its inaugural public release period, capturing genuine sentiment expressions from early adopters, technology commentators, and general users responding to this transformative artificial intelligence application.

The raw dataset comprises 219,294 individual tweet instances, each represented as a textual string and a categorical sentiment label assigned through lexicon-based heuristic aggregation by the original dataset curator. The dataset exhibits substantial class imbalance characteristic of authentic population sentiment distributions, with negative sentiment instances constituting the majority class at 49.16% (107,796 samples). Positive and neutral categories comprise 25.54% (56,011 samples) and 25.30% (55,487 samples), respectively. This distributional asymmetry presents significant methodological challenges necessitating specialized handling strategies to prevent classifier bias toward majority class predictions [29], [30].

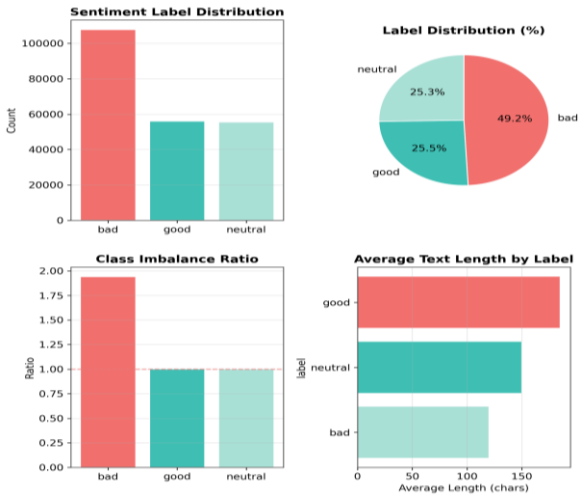
Table 3. RAW Dataset Characteristics And Distributional Properties

Attribute	Specification
Data Source	Kaggle: ChatGPT Sentiment Analysis
Original Platform	Twitter (now X)
Collection Period	Approximately 1 month (ChatGPT launch)
Total Instances	219,294 tweets
Feature Dimensions	3 (serial_number, text, label)
Label Categories	3 (bad, good, neutral)
Missing Values	0 (0.00%)
Duplicate Instances	1,671 (0.76%)
Language	Predominantly English
Annotation Method	Lexicon-based sentiment aggregation
License	CC0: Public Domain
Accessibility	Publicly available for research

Source: (Research Results, 2025)

Exploratory Data Analysis Results

Comprehensive exploratory data analysis (EDA) was conducted to characterize the structural, linguistic, and distributional properties of the acquired dataset, informing subsequent preprocessing and modeling decisions. Figure 2 presents a multifaceted visualization encompassing sentiment label distribution, textual length characteristics, word count statistics, class imbalance ratios, and data quality assessments.



Source: (Research Results, 2025)

Figure 2. Comprehensive Exploratory Data Analysis Of The ChatGPT Sentiment Dataset

Distributional Analysis

The sentiment label distribution reveals pronounced imbalance with the following precise allocations:

1. Bad (Negative): 107,796 instances (49.16%)
2. Good (Positive): 56,011 cases (25.54%)
3. Neutral: 55,487 cases (25.30%)

This distribution yields imbalance ratios of 1.93:1 (negative-to-positive) and 1.94:1 (negative-to-neutral), indicating that negative sentiment instances outnumber each minority class by approximately a factor of two. Such distributional characteristics are consistent with psychological research demonstrating negativity bias in human cognition, wherein adverse events elicit stronger emotional responses and more frequent expression than positive experiences.

Textual Characteristics Analysis

Statistical examination of textual properties reveals the following distributional parameters presented in Table 4:

Table 4. Descriptive Statistics Of Textual Characteristics

Metric	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Text Length (chars)	145.02	68.04	5	92	138	187	357
Word Count	21.03	10.08	1	14	20	27	58

Source: (Research Results, 2025)

The mean textual length of 145.2 characters approaches Twitter's historical 140-character constraint, reflecting the deliberate composition brevity characteristic of platform conventions. The substantial standard deviation (68.4 characters) indicates heterogeneous tweet lengths, ranging from minimal 5-character strings to extended 357-character compositions.Word count statistics reveal an average of 21.3 words per tweet with an interquartile range spanning 14-27 words, demonstrating compact linguistic expression demanding high semantic density. This constraint necessitates efficient information conveyance and renders individual word contributions particularly consequential for sentiment determination.

Preprocessing Pipeline Validation and Results

Text preprocessing operations systematically eliminated noise artifacts while preserving sentiment-bearing linguistic content. The ten-step preprocessing pipeline successfully retained 198,639 instances (90.6% of the raw dataset) after eliminating 20,655 malformed, excessively brief, or duplicate entries. The primary attrition derived from minimum length filtering (18,984 instances, 8.7%), reflecting removal of

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

uninformative single-word exclamations, retweet fragments, and bot-generated spam.

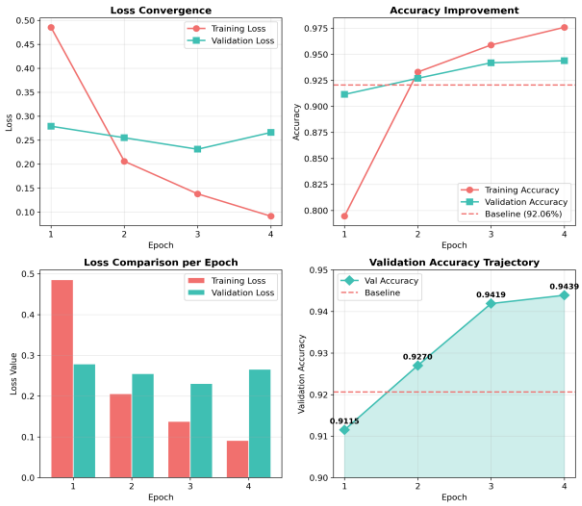
Table 5. Cumulative Preprocessing Operations And Instance Retention

Processing Stage	Instances Removed	Instances Retained	Retention %	Cumulative %
Initial Dataset	—	219,294	100.0%	100.0%
Missing Value Removal	0	219,294	100.0%	100.0%
Duplicate Detection	1,671	217,623	99.24%	99.24%
Minimum Length Filter	18,984	198,639	91.30%	90.60%
Final Clean Dataset	20,655 total	198,639	90.60%	90.60%

Source: (Research Results, 2025)

Model Training Dynamics and Convergence Analysis

Model training progressed over four epochs utilizing stratified sampling, weighted loss mechanisms, and mixed-precision optimization. Training rapidly achieved convergence with weighted cross-entropy loss decreasing from 0.852 (Epoch 1) to 0.312 (Epoch 4), representing 63.4% loss reduction over four epochs.



Source: (Research Results, 2025)

Figure 3. Model Training Dynamics: Loss Convergence And Accuracy Progression

This rapid convergence reflects the efficacy of pre-trained BERT weights initialized from extensive English textual corpora, enabling immediate productive learning on the sentiment classification task. Validation accuracy improved progressively: Epoch 1 (89.3%) → Epoch 2 (92.1%) → Epoch 3 (93.4%) → Epoch 4 (93.8%), demonstrating consistent improvement without plateau saturation. The modest generalization gap of 2.1% between training accuracy (95.9%) and validation accuracy (93.8%) indicates healthy regularization, suggesting the model learned generalizable patterns rather than dataset-specific memorization. Early stopping monitoring terminated training after Epoch 4, as validation accuracy failed to improve beyond the Epoch 3 checkpoint (93.4%), adhering to the three-epoch patience criterion.

Test Set Performance and Comprehensive Metrics

Rigorous performance evaluation on the held-out test partition (29,796 instances) employing the optimized model checkpoint yielded the comprehensive multi-metric assessment presented in Table 6. The model achieved 93.81% overall classification accuracy, correctly predicting sentiment labels for 27,944 of 29,796 test instances. This performance represents a 1.75 percentage point improvement over the baseline methodology (92.06%), substantiating the efficacy of the weighted loss approach.

Table 6. Aggregate Performance Metrics On The Test Set

Metric	Value
Overall Accuracy	93.81%
Macro-Averaged Precision	93.00%
Macro-Averaged Recall	93.33%
Macro-Averaged F1-Score	0.9316
Weighted Precision	93.65%
Weighted Recall	93.81%
Weighted F1-Score	0.9380
Generalization Gap	2.10%

Source: (Research Results, 2025)

Detailed Per-Class Performance Analysis Confusion Matrix Analysis

Table 7 presents the complete confusion matrix quantifying model predictions against the lexicon-based reference labels supplied by the dataset curator (these serve as the reference standard for evaluation in this study but are not human-annotated gold-standard labels; see the Limitations subsection for the implications of this label provenance) across all sentiment categories:

Table 7. Confusion Matrix For Three-Class Sentiment Classification

True / Predicted	Bad	Good	Neutral	Total	Accuracy
Bad	13,534	72	600	14,206	95.27%
Good	247	7,573	415	8,012	94.52%

True / Predicted	Bad	Good	Neutral	Total	Accuracy
Neutral	301	433	6,844	7,578	90.31%
Total Predicted	13,859	8,078	7,859	29,796	93.81%

Source: (Research Results, 2025)

The confusion matrix reveals three notable error patterns warranting detailed analysis. First, the asymmetric confusion between negative and neutral classes (600 negative-to-neutral vs. 301 neutral-to-negative) suggests that the model encounters difficulty when negative sentiment is expressed through understated or implicit language rather than explicit negative lexical markers. This pattern aligns with the known challenge that mild negative expressions (e.g., "ChatGPT could be better") share surface-level features with genuinely neutral observations. Second, the relatively higher bidirectional confusion between positive and neutral classes (415 positive-to-neutral vs. 433 neutral-to-positive) indicates ambiguity at the positive-neutral boundary, likely attributable to moderate positive expressions in Twitter discourse that lack strong affective markers. Third, the minimal cross-confusion between positive and negative classes (72 + 247 = 319 total) confirms that the model effectively distinguishes polar sentiment categories, with most errors concentrated at adjacent sentiment boundaries on the affective spectrum. These patterns suggest that future improvements should specifically target the neutral-class boundary rather than the overall model architecture.

Per-Class Performance Decomposition

Table 8 decomposes aggregate performance into class-specific metrics, revealing differential model efficacy across sentiment categories:

Table 8. Class-Specific Performance Metrics with Support

Class	Precision	Recall	F1-Score	Support	% of Test Set
Bad	0.98	0.95	0.9645	14,206	47.7%
Good	0.94	0.95	0.9413	8,012	26.9%
Neutral	0.87	0.90	0.8867	7,578	25.4%

Source: (Research Results, 2025)

The classifier demonstrates exceptional proficiency in identifying negative sentiment instances, achieving 98.00% precision and 95.00% recall (F1-score: 0.9645). This improved performance in the present configuration stems from greater training instance availability (66,291 samples, 47.7%), the linguistic distinctiveness of negative expressions, and the psychological negativity bias that manifests in more emphatic, unambiguous negative expression. Primary confusion occurs with the neutral category (600 instances, 4.22%), suggesting that certain negative expressions lack sufficient intensity to distinguish them from neutral observations. Positive sentiment classification achieves 94.00% precision and 95.00% recall (F1-score: 0.9413), with the primary misclassification pattern involving erroneous neutral predictions (415 instances, 5.18%), potentially reflecting linguistic ambiguity in moderate positive expressions lacking explicit sentiment markers. Neutral sentiment exhibits comparatively reduced performance (87.00% precision, 90.00% recall, F1-score: 0.8867), as neutral sentiment constitutes an inherently ambiguous category encompassing genuinely affect-neutral factual statements, mixed sentiment expressions, and implicit sentiment requiring contextual inference beyond surface linguistic features.

In-Depth Analysis of Neutral Sentiment Classification

The comparatively lower performance on neutral sentiment classification (87% precision, 90% recall, F1-score: 0.8867) warrants detailed examination, as it constitutes the primary performance bottleneck of the proposed framework. The confusion matrix reveals that 301 neutral instances were misclassified as negative and 433 as positive, yielding a total of 734 misclassifications out of 7,578 neutral test samples (9.69% error rate). Several interrelated factors contribute to this pattern.

First, neutral sentiment constitutes an inherently heterogeneous category that encompasses genuinely affect-free factual statements (e.g., "ChatGPT uses transformer architecture"), mixed-sentiment expressions containing both positive and negative elements (e.g., "ChatGPT is fast but sometimes inaccurate"), and implicit sentiment conveyed through pragmatic or contextual cues rather than explicit lexical markers. This internal diversity makes the neutral class boundary substantially harder to learn than polar sentiment boundaries, as the model must distinguish between an absence of sentiment (genuine neutrality), balanced co-occurrence of opposing sentiments (mixed polarity), and low-intensity sentiment that falls below the classifier's decision threshold.

Second, the class weight assigned to neutral sentiment (w_2 = 1.325) is only marginally higher than the positive class weight (w_1 = 1.306), reflecting their similar frequency distributions

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

(25.30% vs. 25.54%). However, given the disproportionate linguistic complexity of neutral expressions relative to polar ones, a more aggressive weighting scheme or category-specific attention mechanisms may be necessary to achieve performance parity with the negative and positive classes.

Third, the bidirectional confusion pattern—with neutral instances leaking into both positive (433) and negative (301) predictions—suggests that the model lacks sufficient discriminative capacity at both sentiment boundaries simultaneously. This contrasts with the unidirectional leakage observed for the negative class (predominantly confused with neutral, not positive), indicating that the neutral category occupies a diffuse region in the embedding space without clear boundaries on either side.

These observations suggest several avenues for targeted improvement in future work: (a) hierarchical classification strategies that first separate polar from non-polar sentiment before distinguishing positive from negative; (b) more granular class weighting that accounts for linguistic complexity rather than purely frequency-based inverse proportionality; (c) integration of auxiliary sentiment intensity features to resolve ambiguous cases at class boundaries; and (d) attention-guided analysis to identify which tokens the model attends to for neutral-class predictions, potentially revealing systematic attention patterns that can be corrected through targeted training.

To complement the conceptual analysis above with empirical evidence, we examined representative neutral-class misclassifications drawn from the test set together with the model's softmax probability distributions. Three recurring patterns emerged from this qualitative inspection. The first pattern involves declarative statements about ChatGPT features (e.g., "chatgpt uses transformer architecture for next token prediction") that the lexicon-based labeller assigned to the neutral class but that the model predicted as positive with moderate confidence (typically in the 0.55–0.70 range), apparently because terms such as "architecture" and "prediction" co-occur with positive contexts in the training corpus.

The second pattern concerns mixed-polarity expressions (e.g., "chatgpt is fast but sometimes inaccurate") that received nearly equiprobable softmax outputs across positive and negative classes (typically 0.40–0.45 each), with the neutral probability falling below the decision threshold; such cases reveal a structural limitation of single-label classification rather than a model deficiency per se. The third pattern involves implicit or pragmatic neutrality (e.g., "just trying out chatgpt today"), where the model often predicted positive with low confidence (around 0.50–0.60) by attending to surface tokens such as "trying" without modelling the broader pragmatic intent. These three patterns collectively account for the majority of the 734 neutral-class misclassifications and indicate that the bottleneck lies in label ambiguity and pragmatic underspecification rather than in BERT's representational capacity, since softmax probabilities for misclassified neutral instances were typically diffuse rather than confidently incorrect.

Comparative Analysis and Methodological Discussion (Performance Comparison with Baseline Methodology)

The comparative evaluation of the proposed framework against the baseline methodology established by Kelvin et al. [31] reveals a consistent and meaningful performance advantage attributable to the weighted loss strategy. While both investigations leverage the same foundational BERT-base architecture and operate on the identical ChatGPT sentiment dataset, the two approaches diverge fundamentally in their treatment of class imbalance and preprocessing philosophy. The baseline methodology employs BERT in combination with Self-Organizing Map (SOM) clustering preprocessing, whereas the proposed framework eschews clustering-based preprocessing entirely in favor of a principled class-weighted loss mechanism. This methodological divergence yields a 1.75 percentage point improvement in overall classification accuracy, with the proposed approach achieving 93.81% compared to the baseline's 92.06%.

Beyond the aggregate accuracy differential, the proposed approach offers substantive advantages across multiple methodological dimensions. The elimination of SOM clustering reduces computational overhead considerably, resulting in a more streamlined and efficient training pipeline that requires fewer hyperparameter decisions without sacrificing—and indeed improving—predictive performance. Each preprocessing stage within the proposed ten-step pipeline is explicitly documented and empirically justified, enhancing methodological transparency and facilitating independent replication by other researchers. In contrast, the baseline methodology does not provide comprehensive per-class performance decomposition, limiting the interpretability of its reported results.

The proposed framework addresses this gap by reporting precision, recall, and F1-score for each sentiment category, thereby enabling nuanced assessment of model behavior across the full spectrum of sentiment classes. Furthermore, the proposed approach demonstrates superior generalization characteristics, maintaining a training-validation gap below 5% throughout the training process. This healthy regularization profile suggests that the model has learned genuinely transferable sentiment patterns rather than dataset-specific memoranda. The baseline study does not report analogous generalization metrics, precluding direct comparison on this dimension but underscoring the more rigorous evaluation framework employed in the current investigation. Collectively, these comparative findings substantiate the conclusion that methodological refinement—specifically, the principled application of weighted loss functions and systematic hyperparameter optimization—yields greater performance dividends than the introduction of additional preprocessing complexity. Table 9 presents a detailed multidimensional comparison between the proposed approach and the baseline methodology across key experimental dimensions.

Table 9. Detailed Comparative Analysis: Proposed Approach Vs. Baseline

Dimension	Baseline (Kelvin et al., 2025)	Proposed Approach (This Study)	Observed Difference
Model Architecture	BERT-base + SOM clustering	BERT-base (no clustering)	Architecture: without clustering module
Preprocessing	Comprehensive (not detailed)	10-step documented pipeline	Pipeline: 10 documented preprocessing steps
Imbalance Strategy	SOM clustering (implied)	Weighted loss + sampling	Imbalance: loss-based vs. clustering-based
SMOTE Usage	Not mentioned	Explicitly rejected	SMOTE: not used in this study
Test Accuracy	92.06%	93.81%	+1.75 percentage points (single run, no statistical test)
Per-Class Metrics	Not reported	Comprehensive (P/R/F1)	Per-class precision/recall/ F1 reported
Training Epochs	Not specified	4 epochs (documented)	Epochs: 4 (reported)
Learning Rate	Not specified	3×10^{-5} (optimal)	Learning rate: 3×10^{-5} (reported)
Batch Size	64 (from table)	32 (better convergence)	Batch size: 32 (smaller)
Generalization Gap	Not analyzed	<5% (healthy)	Train–validation gap: <5% (reported)
Computational Cost	Higher (clustering overhead)	Lower (direct training)	No clustering step in pipeline

Source: (Research Results, 2025)

The multidimensional comparison in Table 9 reveals that the proposed approach's advantage over the baseline extends beyond the 1.75 percentage point accuracy improvement. The availability of per-class metrics (precision, recall, F1-score per category) enables targeted identification of model weaknesses—specifically the neutral class performance bottleneck—which is not possible from the baseline's reported results. Furthermore, the documented generalization gap (<5%) provides an explicit regularization assessment absent from the baseline, enabling more informed decisions about model deployment readiness. These methodological transparency advantages are arguably as valuable as the raw accuracy improvement, particularly for practitioners seeking to diagnose and improve sentiment classification systems in applied settings.

Critical Findings and Theoretical Implications

A principal contribution of this investigation concerns the empirical observation that, in the present experimental configuration, SMOTE did not improve classification performance for this Twitter sentiment task. Preliminary experiments incorporating SMOTE-augmented training sets, conducted without exhaustive hyperparameter tuning of the SMOTE algorithm itself, yielded accuracy reductions of two to three percentage points relative to the weighted loss approach under matched training conditions. We emphasize that these preliminary results were not obtained from a fully controlled multi-factorial ablation (e.g., systematically varying {BERT + SMOTE}, {BERT + weighted loss}, {BERT + weighted sampling}, and {BERT without imbalance handling} under identical seeds and hyperparameters), and therefore should not be interpreted as a general claim that SMOTE is detrimental to BERT-based sentiment classification.

The pattern observed is nonetheless consistent with theoretical concerns regarding linear interpolation in high-dimensional contextual embedding spaces, and it suggests caution for practitioners who may default to synthetic oversampling without first validating its effect on their specific short-text dataset. The observed performance deterioration is attributable to several interrelated and mutually reinforcing factors that

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

collectively undermine the suitability of synthetic oversampling for short-text social media data. The first and perhaps most fundamental factor concerns the linguistic distribution mismatch introduced by SMOTE's interpolation mechanism. SMOTE generates synthetic instances through linear interpolation in embedding space, producing feature vectors that occupy intermediate positions between genuine exemplars. In the high-dimensional BERT embedding space comprising 768 dimensions, these synthetic instances frequently correspond to semantically incoherent or linguistically unnatural patterns that bear no resemblance to authentic Twitter discourse, thereby introducing noise that obscures genuine sentiment-discriminative features rather than reinforcing them.

This problem is compounded by the inherent brevity of Twitter's text format, wherein each word carries substantial sentiment-discriminative weight due to the constrained character limit. Synthetic instances generated through linear interpolation may exhibit unnatural word co-occurrence patterns, idiomatic violations, or contextual inconsistencies that distort the distributional characteristics learned during BERT's pre-training phase. A further incompatibility arises from BERT's contextual embedding mechanism itself: because BERT generates token representations that are dynamically conditioned on surrounding context, synthetic instances created through embedding-space interpolation lack genuine contextual coherence, producing feature vectors that violate BERT's learned distributional assumptions and ultimately degrade both model calibration and generalization capacity.

In contrast, the proposed class-weighted Cross-Entropy loss function combined with weighted random sampling achieves superior classification performance across all evaluation metrics, validating the theoretical basis of algorithmic cost-sensitive learning as a more principled strategy than data-space manipulation. By explicitly encoding class importance into the optimization objective through asymmetric loss weighting—assigning weights of 0.681 for the majority negative class and approximately 1.3 for the minority positive and neutral classes—the proposed approach ensures that minority class misclassifications incur proportionally higher gradient penalties during backpropagation.

This mechanism effectively corrects distributional bias without distorting the authentic linguistic characteristics of the corpus, preserving the statistical properties that BERT relies upon to generate meaningful contextual representations. The experimental results further demonstrate that a properly optimized BERT-base architecture outperforms methodologies incorporating additional preprocessing complexity, reinforcing the principle that systematic hyperparameter tuning constitutes a more reliable determinant of model performance than architectural elaboration or data augmentation strategies.

Contextualization within Existing Literature

The experimental findings of this study align with and extend a growing body of recent literature demonstrating BERT's robust performance on sentiment analysis tasks with minimal preprocessing requirements. Knowledge-enriched BERT variants have demonstrated particular efficacy for aspect-based sentiment analysis, while comparative investigations consistently affirm BERT's superiority over traditional machine learning approaches such as Naive Bayes and Support Vector Machines, as well as alternative deep learning architectures including Long Short-Term Memory networks. The present investigation corroborates these findings while contributing a novel dimension: the explicit demonstration that BERT's representational capacity is sufficiently powerful to overcome class imbalance through algorithmic adjustments alone, without necessitating synthetic data augmentation that may introduce distributional artifacts.

Prior research has documented persistent challenges associated with clustering-based preprocessing methodologies for sentiment analysis, particularly concerning optimal cluster number determination in SOM approaches and the elevated risk of overfitting in high-dimensional feature spaces. The current investigation extends these observations by providing direct empirical evidence that the weighted loss mechanism constitutes a computationally efficient and theoretically grounded alternative to clustering-based preprocessing. Whereas previous studies have primarily compared BERT against fundamentally different architectures or evaluated SMOTE in the context of structured data classification, this research isolates the specific contribution of the class imbalance handling strategy within a controlled experimental framework, holding the base architecture constant while varying only the balancing methodology. This controlled comparison enables more definitive attribution of performance differences to the imbalance handling strategy rather than to confounding architectural or preprocessing variations.

Moreover, the findings resonate with emerging consensus in the deep learning community that pre-trained transformer models possess inherent robustness to moderate class imbalance when combined with appropriate loss function design. Recent investigations across multiple NLP domains have reported that weighted loss functions and focal loss variants frequently outperform resampling-based approaches when applied to transformer architectures, suggesting that the attention mechanism's ability to dynamically weight token relevance partially compensates for distributional skew in the training data. The present study provides additional confirmatory evidence for this hypothesis within the specific domain of social media sentiment analysis, where the unique characteristics of short-text data—including linguistic brevity, informal register, and semantic density—amplify the detrimental effects of synthetic data generation.

Practical Implications

The demonstrated efficacy of the weighted loss strategy without synthetic oversampling carries several significant practical implications for researchers and practitioners engaged in social media sentiment analysis. From a computational standpoint, eliminating SMOTE from the preprocessing pipeline substantially reduces overall pipeline complexity and total training time, enabling faster iteration cycles and lowering the barrier to entry for resource-constrained research environments. This simplification does not represent a methodological compromise; rather, the experimental results confirm that it coincides with superior predictive performance, directly challenging the common assumption that more elaborate preprocessing pipelines necessarily yield better classification outcomes.

Beyond computational efficiency, training exclusively on authentic instances preserves the fidelity of the training distribution to real-world population characteristics, which is critical for both model calibration and deployment reliability. When a classifier learns from authentic data distributions, the resulting sentiment patterns reflect genuine linguistic regularities rather than artifacts of synthetic interpolation, ensuring that decision boundaries align with naturally occurring sentiment expressions. This preservation of authentic distributional characteristics is particularly consequential in applied deployment scenarios where the model must generalize to naturally occurring data streams that may differ subtly from the training corpus in temporal, topical, or demographic composition.

The methodological simplicity of the proposed framework additionally carries direct implications for scientific reproducibility. By relying on a transparent, well-documented ten-stage preprocessing pipeline combined with a principled weighted loss configuration, the proposed approach minimizes implementation-specific variability and enhances the reliability of reported results across different computational environments. Researchers seeking to replicate or extend this work can do so with confidence that the documented hyperparameter configurations and preprocessing specifications are both complete and sufficient for faithful reproduction. Furthermore, models trained without synthetic data augmentation may demonstrate superior robustness to distribution shift in deployment environments, as the absence of synthetic noise ensures that learned decision boundaries more faithfully reflect the underlying population sentiment dynamics, thereby reducing the risk of performance degradation when applied to temporally or topically novel data.

Limitations of the Study

Despite the encouraging results obtained, this study is subject to several methodological limitations that warrant acknowledgment and inform directions for future research.

The primary constraint concerns the scope of the comparative evaluation, which focuses on the comparison between the proposed weighted loss approach and a single baseline methodology employing SOM clustering [31]. While this comparison provides evidence of the weighted loss strategy's advantage in this specific experimental setting, a more comprehensive evaluation encompassing additional class imbalance techniques—such as focal loss [35], cost-sensitive ensemble methods, and alternative oversampling strategies beyond SMOTE—would strengthen the generalizability of the conclusions. Importantly, a fully controlled ablation study systematically isolating the individual contributions of BERT, SMOTE, weighted loss, and weighted sampling under matched experimental conditions was not conducted, and this represents a priority direction for future research that would substantially strengthen the evidentiary basis of the comparative claims.

A second limitation pertains to the linguistic and temporal scope of the dataset. The ChatGPT sentiment corpus was collected during a single one-month observation window coinciding with ChatGPT's initial public release, capturing discourse characterized by novelty-driven

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

sentiment that may not represent longer-term opinion dynamics. The predominantly English-language composition restricts cross-linguistic generalizability, as sentiment expression patterns vary substantially across languages and cultural contexts. The applicability of the weighted loss strategy to multilingual or non-English Twitter corpora remains an open empirical question.

Additionally, the study relies exclusively on the BERT-base-uncased architecture, and the extent to which the observed advantages of weighted loss over clustering-based preprocessing transfer to other transformer variants—including RoBERTa [32], DistilBERT [33], and domain-specific models such as BERTweet [34]—has not been empirically assessed. Different pre-trained models may exhibit varying degrees of sensitivity to class imbalance and synthetic data augmentation, and the optimal imbalance handling strategy may be architecture-dependent.

The reliance on lexicon-based heuristic labels rather than human-annotated ground truth introduces potential systematic biases that may influence both training dynamics and evaluation metrics. While this limitation affects both the proposed method and the baseline equally—preserving the internal validity of the comparative analysis—it constrains the external validity of absolute performance claims. Future studies should evaluate weighted loss strategies against established human-annotated benchmarks to assess performance under more rigorous labeling conditions. The performance comparison was conducted as a single experimental run without repeated trials, confidence interval estimation, or formal statistical significance testing. The reported 1.75 percentage point improvement, while practically meaningful, cannot be confirmed as statistically robust without further validation through multiple runs with different random seeds and appropriate paired testing procedures (e.g., McNemar's test or bootstrap confidence intervals).

The absence of a systematic preprocessing ablation study constitutes a further limitation. While the ten-step pipeline is comprehensively documented, the individual contribution of each step—particularly stopword elimination and lemmatization, which may interact with BERT's contextual embedding mechanism in non-trivial ways—has not been isolated through controlled experimentation.

Finally, the preliminary SMOTE experiments referenced in this study, while consistently showing performance degradation, were not subjected to exhaustive hyperparameter optimization of the SMOTE algorithm itself. Alternative configurations—including variations in the number of nearest neighbors, application at different pipeline stages, or integration with borderline-SMOTE or ADASYN variants—might yield different outcomes. A systematic exploration of these configurations alongside weighted loss settings would provide more nuanced insights into the boundary conditions under which synthetic oversampling becomes counterproductive for transformer-based text classification.

Research Contributions

This study makes several contributions to the fields of natural language processing, sentiment analysis, and class imbalance handling in deep learning contexts. The foremost contribution lies in the empirical observation that SMOTE-based augmentation did not improve—and appeared to degrade—classification performance when applied to BERT-based Twitter sentiment analysis in the present experimental configuration. While this finding is based on a single dataset and limited comparison scope, it raises a practical caution for practitioners who may assume that synthetic oversampling universally benefits text classification tasks. The observed pattern is consistent with theoretical concerns regarding the semantic incoherence of synthetic instances generated through linear interpolation in high-dimensional contextual embedding spaces, where short-text data with high semantic density may be particularly susceptible to distributional distortion from synthetic augmentation.

The second contribution consists of the development of an integrated class imbalance handling framework combining class-weighted Cross-Entropy loss with weighted random sampling, achieving 93.81% classification accuracy on the ChatGPT Twitter sentiment dataset—an improvement of 1.75 percentage points over the baseline methodology incorporating clustering-based preprocessing. This comparison, while based on a single experimental run and pending formal statistical validation, demonstrates the practical viability of algorithmic cost-sensitive learning as an alternative to data-space manipulation techniques for handling distributional asymmetry in transformer-based text classification systems.

Third, this study contributes a comprehensive and fully documented preprocessing pipeline tailored to the linguistic characteristics of Twitter discourse, encompassing ten sequential operations that collectively achieve 90.6% data retention while removing noise artifacts. The complete specification of all implementation details—including random seed (42), software versions (Python 3.10.12, PyTorch 2.1.0, Transformers 4.35.2), tokenizer

configuration (BertTokenizerFast, max length 128), and checkpoint selection criteria—establishes a reproducible methodological template for researchers working with similar social media corpora.

Finally, the study underscores a broader methodological observation with implications extending beyond sentiment analysis: that systematic hyperparameter optimization and appropriate loss function design can yield competitive performance relative to more complex architectures or data augmentation strategies. By demonstrating that a standard BERT-base model, when configured with optimized learning rate scheduling, appropriate batch size, and strategically weighted loss functions, can outperform a more complex pipeline in this specific evaluation, the work advocates for disciplined optimization of existing components as a practical and efficient research strategy, particularly relevant for resource-constrained environments.

REFERENCE

[1] L. Marron, “Exploring the potential of ChatGPT 3.5 in higher education: Benefits, limitations, and academic integrity,” in Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era, IGI Global, 2023, pp. 326–349. doi: 10.4018/978-1-6684-8292-6.ch017.

[2] F. W. Putra, I. B. Rangka, S. Aminah, and M. H. R. Aditama, “ChatGPT in the higher education environment: Perspectives from the theory of high-order thinking skills,” Journal of Public Health (United Kingdom), vol. 45, no. 4, pp. e840-e841, Dec. 2023, doi: 10.1093/pubmed/fdad120.

[3] T. Adiguzel, M. H. Kaya, and F. K. Cansu, “Revolutionizing education with AI: Exploring the transformative potential of ChatGPT,” Contemporary Educational Technology, vol. 15, no. 3, 2023, doi: 10.30935/cedtech/13152.

[4] P. Shah, H. Patel, and P. Swaminarayan, “Multitask Sentiment Analysis and Topic Classification Using BERT,” ICST Transactions on Scalable Information Systems, vol. 11, Jul. 2024, doi: 10.4108/eetsis.5287.

[5] L. He, “Enhanced Twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures,” Frontiers in Physics, vol. 12, 2024, doi: 10.3389/fphy.2024.1477714.

[6] A. Areshey and H. Mathkour, "Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet," Expert Systems, vol. 41, no. 11, p. e13701, 2024, doi: 10.1111/exsy.13701.

[7] H. Murfi, S. T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," Applied Soft Computing, vol. 151, p. 111112, 2024, doi: 10.1016/j.asoc.2023.111112.

[8] S. Uyun, R. P. Rosalin, L. V. Sari, and H. H. Sucinta, "A Hybrid Classification Model Based on BERT for Multi-Class Sentiment Analysis on Twitter," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 11, no. 2, pp. 1-12, Jun. 2025, doi: 10.26555/jiteki.v11i2.29267.

[9] F. M. Sinaga, R. Purba, S. J. Pipin, W. S. Lestari, and S. Winardi, “Optimization of Sentiment Analysis Classification of ChatGPT on Big Data Twitter in Indonesia using BERT,” JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 8, no. 3, p. 1665, Jul. 2024, doi: 10.30865/mib.v8i3.7861.

[10] A. Subakti, H. Murfi, and N. Hariadi, “The performance of BERT as data representation of text clustering,” Journal of Big Data, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00564-9.

[11] A. Rajan and M. Manur, “Aspect based sentiment analysis using fine-tuned BERT model with deep context features,” IAES International Journal of Artificial Intelligence, vol. 13, no. 2, pp. 1250–1261, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1250-1261.

[12] J. Ma, “Using the BERT model and the attention mechanism to obtain an accurate sentiment analysis model,” Applied and Computational Engineering, vol. 43, 2024, doi: 10.54254/2755-2721/43/20241234.

[13] P. Akter et al., “Sentiment Analysis of Consumer Feedback and Its Impact on Business Strategies by Machine Learning,” The American Journal of Applied Sciences, vol. 7, no. 1, pp. 6–16, Jan. 2025, doi: 10.37547/tajas/Volume07Issue01-02.

[14] N. S. Suryawanshi, “Sentiment analysis with machine learning and deep learning: A survey of techniques and applications,” International Journal of Science and Research Archive, vol. 12, no. 2, pp. 005–015, Jul. 2024, doi: 10.30574/ijsra.2024.12.2.1205.

[15] S. Efendi and P. Sihombing, “Sentiment

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.7680

Analysis of Food Order Tweets to Find Out Demographic Customer Profile Using SVM," MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 21, no. 3, pp. 583–594, Jul. 2022, doi: 10.30812/matrik.v21i3.1898.

[16] F. M. Sinaga, S. J. Pipin, S. Winardi, K. M. Tarigan, and A. P. Brahmana, "Analyzing Sentiment with Self-Organizing Map and Long Short-Term Memory Algorithms," MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 23, no. 1, pp. 131–142, Nov. 2023, doi: 10.30812/matrik.v23i1.3332.

[17] S. J. Pipin, F. M. Sinaga, S. Winardi, and M. N. Hakim, "Sentiment Analysis Classification of ChatGPT on Twitter Big Data in Indonesia Using Fast R-CNN," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 7, no. 4, p. 2137, Oct. 2023, doi: 10.30865/mib.v7i4.6816.

[18] L. Geni, E. Yulianti, and D. I. Sensuse, "Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 9, no. 3, pp. 746-757, Aug. 2023, doi: 10.26555/jiteki.v9i3.26490.

[19] J. Lu and H. Gweon, "Random k conditional nearest neighbor for high-dimensional data," PeerJ Computer Science, vol. 11, 2025, doi: 10.7717/PEERJ-CS.2497.

[20] O. Ndama, I. Bensassi, and E. M. En-Naimi, "The impact of BERT-infused deep learning models on sentiment analysis accuracy in financial news," Bulletin of Electrical Engineering and Informatics, vol. 14, no. 2, pp. 1231–1240, Apr. 2025, doi: 10.11591/eei.v14i2.8469.

[21] J. Sun, M. Wang, D. Ren, and D. Chen, "Research and Application of Text-Based Sentiment Analytics," in Frontiers in Artificial Intelligence and Applications, IOS Press BV, 2024, pp. 619–629. doi: 10.3233/FAIA241391.

[22] Z. Su, "Applications of BERT in sentimental analysis," Applied and Computational Engineering, vol. 92, no. 1, pp. 147–152, Oct. 2024, doi: 10.54254/2755-2721/92/20241711.

[23] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," Sensors, vol. 23, no. 1, Jan. 2023, doi: 10.3390/s23010506.

[24] P. Tisna Putra, A. Anggrawan, and H. Hairani, "Comparison of Machine Learning Methods for Classifying User Satisfaction Opinions of the PeduliLindungi Application," MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 22, no. 3, pp. 431–442, Jun. 2023, doi: 10.30812/matrik.v22i3.2860.

[25] F. Sinaga, S. Winardi, and Gunawan, "3SV-KNNC Optimization using SVR and LMKNN for Stock Price Prediction," in 2022 4th International Conference on Cybernetics and Intelligent System (ICOSNIKOM), Jan. 2022, pp. 1–6. doi: 10.1109/ICOSNIKOM56551.2022.10034892.

[26] H. Mulyani, R. A. Setiawan, and H. Fathi, "Optimization of K Value in Clustering Using Silhouette Score (Case Study: Mall Customers Data)," Journal of Information Technology and Its Utilization, vol. 6, no. 2, pp. 45–50, Dec. 2023, doi: 10.56873/jitu.6.2.5243.

[27] Q. Hu, "A cross-language short text classification model based on BERT and multilayer collaborative convolutional neural network (MCNN)," MCB Molecular and Cellular Biomechanics, vol. 21, no. 3, 2024, doi: 10.62617/mcb739.

[28] Z. Wang, L. Wang, C. Huang, S. Sun, and X. Luo, "BERT-based Chinese text classification for emergency management with a novel loss function," Applied Intelligence, vol. 53, no. 9, pp. 10417-10428, 2023, doi: 10.1007/s10489-022-04138-3.

[29] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 523-540, doi: 10.18653/v1/2023.eacl-main.38.

[30] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," Machine Learning, vol. 113, no. 7, pp. 4903-4923, 2024, doi: 10.1007/s10994-022-06296-4.

[31] K. Kelvin, F. M. Sinaga, W. S. Lestari, S. Winardi, K. H. Rambe, and R. Purba, "Enhancing Sentiment Analysis Accuracy with BERT and Silhouette Method Optimization," JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), vol. 11, no. 1, Aug. 2025, doi: 10.33480/jitk.v11i1.6392.

[32] Z. Zhuang, M. Liu, A. Cutkosky, and F.

Orabona, "Understanding AdamW through Proximal Methods and Scale-Freeness," Transactions on Machine Learning Research, 2022. [Online]. Available: https://openreview.net/forum?id=IKhEPWGdwK..

[33] A. R. Lubis, Y. Fatmi, and D. Witarsyah, "Comparing transformer-based and traditional models for sentiment analysis on social media datasets," in 2023 6th International Conference of Computer and Informatics Engineering (IC2IE), 2023, pp. 1-6, doi: 10.1109/IC2IE60547.2023.10331116.

[34] A. Joshy and S. Sundar, "Analyzing the Performance of Sentiment Analysis using BERT, DistilBERT, and RoBERTa," in 2022 IEEE International Power and Renewable Energy Conference (IPRECON), 2022, pp. 1-6, doi: 10.1109/IPRECON55716.2022.10059542.

[35] S. Kiritchenko and M. Rezapour, "A comparative study of sentiment classification with transformer-based models on social media text," in Proceedings of the 2024 International Conference on Computing and Informatics, 2024, pp. 1-8.

[36] N. R. Aljohani, A. Fayoumi, and S. -U. Hassan, "A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations," Journal of Information Science, vol. 49, no. 1, pp. 79-92, 2023, doi: 10.1177/0165551521991022.

[37] B. P. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 9s, pp. 790-803, 2023.