



Auditing Silver-standard Lexicon Labels for Indonesian Mental Health Sentiment on Twitter: A Leakage-safe Pipeline with Confidence-aware XAI

Fitri Marisa^{1*} Sharifah Sakinah Syed Ahmad² Deshinta Arrova Dewi³
 Agustinus Noertjahyana⁴ Anastasia L Maukar⁵ Istiadi¹

¹Informatics Engineering Department, Universitas Widya Gama Malang, Indonesia

²Faculty of Artificial Intelligence and Cyber Security, Universiti Teknikal Malaysia, Melaka, Malaysia

³Center for Data Science and Sustainable Technologies, INTI International University, Malaysia

⁴Informatics Department, Petra Christian University, Indonesia

⁵Industrial Engineering Department, President University, Indonesia

* Corresponding author's Email: fitrimarisa@widyagama.ac.id

Abstract: This study investigates sentiment classification in Indonesian social media discussions around mental health, focusing on three methodological challenges: class imbalance, the limitations of lexicon-based automatic labeling as a silver-standard proxy, and the need for auditable explanations. The dataset contains 5,184 lexicon-labeled tweets, preprocessed and represented using TF-IDF. To prevent data leakage, imbalance handling is performed via train-only resampling (SMOTE applied only to training data) within a leakage-safe pipeline. We evaluate several classical models using accuracy and macro metrics (macro-F1/macro-recall) to ensure fairer assessment across classes, and we add a modern Indonesian Twitter transformer baseline (IndoBERTweet) under the same repeated stratified split protocol. To validate label quality, we construct a Label Audit Subset (LAS) of 600 tweets (200 per lexicon class), annotated by two annotators and finalized through adjudication to produce gold labels. Although annotation reliability is high (92.5% agreement; Cohen's $\kappa = 0.874$), the audit reveals substantial lexicon-label noise (53.5% mismatch; $\kappa = 0.198$), particularly for Positive and Neutral labels, yielding more conservative performance estimates under gold-label evaluation. To improve auditability, we implement a confidence-aware LIME protocol that separates high-confidence, borderline, and hard-case predictions using probability, margin, and entropy. Token summaries exhibit drift when case selection is based on gold labels rather than lexicon labels ($\text{Jaccard}@20 = 0.176\text{--}0.290$), indicating that label noise can shift interpretive conclusions. Finally, token-drift signals are mapped into operational design hypotheses for further studies, including low-pressure gamification, trusted-ties social support (silaturahmi parameter, i.e., culturally grounded social-support ties), and friction mitigation via triage/escalation mechanisms in risk scenarios.

Keywords: Sentiment analysis, Label audit subset (LAS), Leakage-safe pipeline, Explainable artificial intelligence, Mental health.

1. Introduction

Mental health discourses on social media are becoming increasingly prominent [1]. Users often share emotional experiences and coping strategies in digital spaces, so conversational traces can be read as population signals to map patterns of anxiety, depression, and distress with caution regarding representativity, ethics, and privacy [1, 2]. Sentiment analysis is relevant for mapping trends in public

opinion and emotions, but requires adjustments to local language and cultural context to avoid bias in findings [3]. At the same time, the relationship between social media usage intensity and mental health outcomes is complex and not always linear; therefore, model output should be understood as a monitoring indicator, not a diagnosis, making model accountability a prerequisite [4].

The application of sentiment analysis to mental health issues is still limited by several methodological issues that can reduce the validity of

the findings. First, the limited availability of high-quality labels due to the high cost of manual annotation encourages the use of weak lexicon-based labeling (weak supervision), which is susceptible to noise (e.g., polysemy, irony, and pragmatic context) [5, 6]. Second, class imbalances are common and can mask the performance of minority classes, making accuracy-based aggregate metrics risky and misleading [6, 7]. Third, because model output can influence monitoring and follow-up priorities, interpretability is necessary for stakeholders to understand the rationale for predictions, rather than simply receiving a score [6-8]. Therefore, experimental designs should emphasize interclass fairness and replicable processes.

Methodologically, the literature often treats automatic labeling, imbalance handling, and interpretability as separate components. As a result, performance can appear high but be biased if resampling is not implemented in a leakage-safe manner [1], and auditing becomes difficult when explanations do not consider prediction stability [9]. To date, end-to-end approaches that explicitly combine silver-standard labeling, train-only resampling within a leakage-safe pipeline, and XAI that distinguishes between stable and ambiguous predictions for social media-based mental health monitoring are limited [10-13]. Furthermore, few studies have quantified lexicon-label noise using a subset of gold-label audits to assess its impact on evaluation and interpretation. Model outputs are positioned as discourse indicators for monitoring, not individual diagnoses, so interpretation requires human oversight.

Addressing this gap, we propose a leakage-safe pipeline for mental health sentiment analysis on Indonesian-language Twitter, equipped with confidence-guided LIME-based local explanations. The dataset is labeled with a lexicon (the silver standard) because most similar studies lack clinical ground truth [11]. The pipeline includes preprocessing, TF-IDF as an easily auditable baseline, and evaluation of several classical models using macro-F1 and macro-recall to address class imbalance [14, 15]. We also built a Label Audit Subset (LAS) to measure lexicon-label noise, generate constrained gold labels through adjudication, and test its impact on model evaluation and selection. In the XAI stage, predictions are mapped to high-confidence, borderline, and hard cases; token drift between label references is then derived into operational design hypotheses (gamification, social support, friction mitigation) for further study [16].

This paper is structured as follows. Section 2 reviews related Research and positions this study's

contribution within the literature on social media-based mental health sentiment analysis. Section 3 describes the end-to-end methodology used, including silver-standard labeling, Label Audit Subset (LAS) generation, a leakage-safe pipeline with train-only SMOTE, and a confidence-aware LIME-based XAI protocol. Section 4 presents the experimental results and analysis, while Section 5 discusses the implications of the findings, limitations, and future Research directions before concluding in Section 6.

2. Related works

2.1 Social media-based mental health monitoring

The use of social media data to understand mental health is gaining recognition, as it captures emotional expressions and psychological complaints in everyday situations. Systematic reviews report the widespread use of text-based approaches for symptom detection/monitoring, particularly depression, and mapping population trends. However, findings are often difficult to generalize due to heterogeneity in methods and varying label definitions [5, 11, 17, 18]. In crises such as the COVID-19 pandemic, social media is also useful for public health surveillance, although its psychological signals are sensitive to sampling strategies, label quality, and user population bias [11, 19]. Several studies have noted an increase in depression-related language during the pandemic [19, 20], and meta-analyses have found associations between intense social media use and symptoms of anxiety and depression [21, 22]. Therefore, strengthening methodologies and ethical guidelines remains necessary, including the availability of high-quality ground truth to reliably train advanced models [23-25]. If this is not done, the model output should be positioned as a discourse indicator, not a clinical judgment [18, 26].

2.2 Labeling: from “ground-truth” to silver-standard labels and weak supervision

Limited access to clinical labels encourages the use of indirect labels, such as self-disclosure, community membership, or lexicon [23]. Despite being effective at scale, these labels are proxy-based and necessitate validation for dependability. A rigorous examination of Twitter data revealed that only 13.4% of datasets provide ground truth, whereas the majority rely predominantly on keyword searches that may not accurately reflect mental processes [11]. Lexicon labeling facilitates the construction of

extensive corpora; nonetheless, it is susceptible to language ambiguity and community bias [27]. Research forecasting PTSD symptom trajectories using oral histories underscores the promise of AI/ML and emphasizes the necessity of clinical validation [28]. Investigating COVID-19 conversation requires methodological precision to enhance the credibility of results [19, 20]. Nonetheless, the integration of social media and clinical data remains problematic for construct validity and reliability [27]. Research forecasting PTSD symptom trajectories using oral histories underscores the promise of AI/ML, while emphasizing the necessity of clinical validation [28]. Research on COVID-19 discourse necessitates methodological rigor to bolster the legitimacy of findings [19, 20], whereas the integration of social media and clinical data continues to pose challenges regarding construct validity and reliability [27]. The implications of this gap for Research are the need for pipelines that acknowledge the silver-standard nature of labels from the outset, rigorously evaluate them, and provide auditable interpretability to mitigate the risk of bias/uncertainty [11, 19, 27].

2.3 Feature representation and model choice: classical vs. pre-trained vs. LLM era

The spectrum of mental health modeling methods ranges from classical baselines (TF-IDF + SVM/LR/NB) to transformer models [29]. Classical baselines are significant because of their high auditability and stability with small datasets, making them appropriate for reproducible, interpretable workflows [29]. Pre-trained and large language models often enhance performance; however, they entail computational costs, the potential for domain shift, and interpretability challenges [30, 31]. Empirical evidence indicates that transfer learning facilitates psychiatric screening on limited or imbalanced datasets [32] and that BERT is proficient for several NLP tasks, including sentiment analysis [33, 34]. Nonetheless, realistic implementations frequently require model reduction [31] or annotation methods such as multi-task active learning [30]. The identifiable gap is the need for a robust, auditable baseline with adequate interpretability before increasing model complexity.

2.4 Class imbalance, resampling practices, and the risk of leakage in evaluation

Class imbalance in social media-based mental health data can easily lead to biased performance

estimates when resampling is applied incorrectly. Oversampling is commonly used, but it should be applied after the split and only to the training data to prevent leakage and macro-F1/macro-recall inflation. SMOTE and its variants are widely used because they generate synthetic samples of the minority class [35, 36]. However, this can lead to overgeneralization and noise, leading to the development of more distribution-aware approaches such as SD-KMSMOTE and SMOTE-LMVDE [36, 37]. Another alternative is to use GANs to enrich the diversity of synthetic samples while maintaining the decision boundary [38-40]. Since oversampling does not always improve AUROC/AUPRC, the choice of evaluation metrics is also crucial [41]. Some studies incorporate Bayesian optimization/networks (e.g., LSSASMOTE, BOSME) to suppress noise and improve cost-aware learning [42].

2.5 XAI for mental health: from “feature explanations” to actionable, confidence-aware explanations

XAI is important in mental health applications (e.g., depression detection) because model outputs can influence intervention prioritization and case management. Therefore, the rationale for predictions needs to be explainable to stakeholders [40, 43]. Feature-based explanations (SHAP/LIME) are widely used for auditing and communication [40]. However, the literature emphasizes that explanations must also be sufficiently reliable for decision-making, making confidence-aware approaches relevant for distinguishing stable and borderline predictions [40, 44]. The TIFU framework promotes transparency in the probabilistic and multifactorial context of psychiatry [43], while the sociotechnical agenda emphasizes aligning XAI guidelines with practice needs [45] and evaluating model follow-up in the real world [44].

2.6 Synthesis of the current research position

The literature generally discusses silver-standard labeling, imbalance management, and XAI separately; however, all three must be integrated to achieve end-to-end system auditability. Silver-standard labeling helps with scalability while still incorporating scientific caveats, as demonstrated in automated labeling of medical images [9]. On the imbalance side, advanced oversampling such as LSSASMOTE and SMOTE-LMVDE aims to improve accuracy while controlling noise [36, 46]. On the interpretation side, XAI that considers

prediction confidence helps build trustworthiness in sensitive domains [26, 45, 47]. Therefore, this study makes contributions to a replicable pipeline that integrates all three, with a focus on evaluation validity, transparency, and interpretation prudence [6, 8, 45].

3. Research method

The Research data arises from a corpus of Indonesian-language text gathered from Twitter through a keyword-based query focused on mental health, utilizing the keywords: “#MentalHealth OR #MentalHealth OR #Anxiety OR #Depression OR #Stress OR #Burnout OR #Overthinking OR #AnxietyDisturbance” OR “anxiety” OR “stress” OR “depressi” OR “burnout” OR “overthinking.” The data spans November 6, 2025–January 26, 2026 (based on tweet timestamps, UTC). After retrieval, the data was filtered for Indonesian (lang=id), retweets were removed, and duplicates were removed. The raw text was stored in the text_raw column. From the initial 5,731 tweets, the deduplication and cleaning process yielded N=5,184 unique tweets for modeling.

3.1 Data and curation procedure

The curation stage included: (i) Indonesian language filtering using the lang attribute, (ii) retweet/RT removal, (iii) tweet_id-based deduplication. (and merging duplicate text when detected), and (iv) cleaning invalid/empty entries. For ethical and privacy reasons, the study used only public tweets, made no re-identification attempts, and reported results at the aggregate level; example excerpts were shortened without displaying usernames or full URLs. The final dataset for modeling consisted of N=5,184 documents; silver-standard sentiment labels are available in the lexicon_label column, while the text preprocessing results used as model input are available in the text_preprocessed column.

3.2 Lexicon-based sentiment labeling (silver-standard)

This study uses lexicon-based automatic labeling as weak supervision to form three sentiment classes (Positive, Neutral, Negative), given that large-scale manual annotation is expensive and time-consuming. The document sentiment score is calculated as: $score(d) = P(d) - N(d)$

where $n P(d) = \sum_{w \in d} I(w \in L_{pos})$ and $N(d) = \sum_{w \in d} I(w \in L_{neg})$, which is the number of tokens in the document that match the positive and negative

lexicons. Labeling rules: $score(d) > 0 \rightarrow \text{Positive}$; $score(d) = 0 \rightarrow \text{Neutral}$; $score(d) < 0 \rightarrow \text{Negative}$. Since these labels are derived from a lexicon, they should be treated as silver-standard proxy labels and may contain noise (e.g., vocabulary bias, irony/sarcasm, polysemy, and pragmatic context); therefore, findings are positioned as model performance against proxy labels rather than as performance against adjudicated ground truth. [1, 2].

3.3 Text preprocessing

Preprocessing is performed to suppress noise typical of social media and standardize text representation. It includes removing mentions (“@”), URLs, symbols/emojis, and non-alphanumeric characters; case folding; non-standard term normalization; space-based tokenization; Indonesian stopword removal; and Indonesian stemming. The preprocessing output is stored in the text_preprocessed column and used as modeling input.

3.4 TF–IDF feature extraction

Documents are represented using TF–IDF as a stable and easily auditable baseline. The TF–IDF weight is defined as $TF-IDF(t, d) = tf(t, d) \times \log\left(\frac{N}{df(t)+1}\right)$, where N is the number of documents and $df(t)$ is the frequency of documents containing term t; a constant 1 is added to avoid division by zero. The TF–IDF configuration uses unigram–bigram (ngram_range=(1,2)), min_df=2, and max_features=5918. To prevent leakage, the vectorizer is fit only to the training data and then applied (transform-only) to the test data, resulting in a 5918-dimensional feature space. The details of the feature extraction configuration are described in Table 1.

Table 1 is intended to serve as the single protocol specification for exact replication of the full pipeline. The subsequent method subsections (Sections 3.5–3.9) elaborate the rationale for each component, but all operational settings that affect reproducibility, splitting, train-only resampling, calibration, bootstrap reporting, bucket assignment, and LIME sampling are consolidated in Table 1 to minimize ambiguity.

3.5 Data splitting and leakage-safe pipeline

The evaluation used an 80:20 stratified split with random_state=42 (D_train=4147; D_test=1037) and repeated 10 times (repeated stratified split). At each iteration, TF–IDF was fit to the training data and then applied to the test data; train-only SMOTE was

Table 1. Single protocol table for exact replication of preprocessing, splitting, modeling, calibration, and confidence-aware XAI

Component	Parameter/Artifact	Value/Description
Text Preprocessing	Main steps	Remove mentions, URLs, emojis, and non-alphanumeric characters; apply case folding, non-standard term normalization, whitespace tokenization, Indonesian stopwords removal, and Indonesian stemming.
Text Preprocessing	Output column	text_preprocessed (final sparse-text input for TF-IDF modeling).
TF-IDF	Vectorizer	TfidfVectorizer(ngram_range=(1,2), min_df=2, max_features=5918); all other parameters follow scikit-learn defaults.
Leakage control	Fitting	The vectorizer is fit only on the training data in each split and is then used to transform the corresponding test data.
Data Split and Validation	Split	Stratified 80:20 split repeated 10× (StratifiedShuffleSplit; random_state = 42); metrics: Accuracy, Macro-F1, and Macro-Recall; 95% confidence intervals are percentile bootstrap intervals (B = 10,000) computed by resampling the 10 run-level scores with replacement and recomputing the mean.
Resampling	SMOTE (train-only)	SMOTE(k_neighbors = 5, random_state = 42) is applied only to X_train after TF-IDF transformation; X_test is never resampled (leakage-safe).
Model	Logistic Regression (LR)	LogisticRegression with scikit-learn defaults; compared under two conditions: train-only SMOTE and a no-SMOTE baseline using class_weight='balanced'.
Model	SVM (probabilistic output)	LinearSVC with scikit-learn defaults. For LAS/XAI analyses that require probabilities, CalibratedClassifierCV(method='sigmoid', cv=3) is fit only on the corresponding training split after feature preparation (and after train-only SMOTE in SMOTE variants), and the fitted calibrator is then applied unchanged to untouched test/LAS instances. This calibrated variant is reported as SVC_calibrated.
XAI	LIME	LimeTextExplainer is used within the text → TF-IDF → model pipeline for local token-level explanation.
Environment	Software	scikit-learn 1.6.1; supported libraries include numpy, pandas, imbalanced-learn, and lime. For exact replication, the full environment should be logged via pip freeze or requirements.txt.
Model (First Experiment)	DT / RF / k-NN / NB	All models follow scikit-learn defaults: DecisionTreeClassifier; RandomForestClassifier(n_estimators = 100); KNeighborsClassifier(n_neighbors = 5); MultinomialNB(alpha = 1.0).
Resources	Stopwords & stemmer	Stopword list: NLTK Indonesian stopwords; stemming: Sastrawi. If additional custom normalization or stopwords rules are used, they should be exported as reproducibility artifacts.
Transformer baseline	Model checkpoint	indolem/indobertweet-base-uncased.
Transformer input	Text variant	text_transformer (light cleaning only; no stopwords removal or stemming).
Transformer training	Fine-tuning config	max_len = 128; epochs = 3; lr = 2e-5; batch_size = 8 (effective batch size = 16 via gradient accumulation); optimizer = AdamW; seed = 42.
Imbalance handling (transformer)	Loss weighting	Class-weighted cross-entropy computed from the training split only (per run).
Evaluation protocol	Split reuse across baselines	Identical 10× repeated stratified 80:20 split indices are reused across TF-IDF baselines and IndoBERTweet (random_state = 42).
Sparse pipeline order	End-to-end sequence	Split → fit TF-IDF on training data → transform train/test → apply SMOTE only to X_train for SMOTE variants → fit classifier → if probabilistic SVM is required, fit the calibrator on the same training split only → predict on untouched test/LAS data → compute p_max, margin, and entropy → assign confidence bucket → run LIME on selected cases.
XAI reproducibility	Fixed settings	LimeTextExplainer is used in the text → TF-IDF → model pipeline; summary reporting uses num_features = 20 (top-k = 20); up to 20 instances per gold class per bucket are sampled; stability analysis uses 5 fixed seeds {11, 22, 33, 44, 55} and is summarized with Jaccard@20.

Confidence buckets	Numeric thresholds	Per gold class in LAS. Let $q_{0.80}^{(c)} = Q_{0.80}(p_{\max} \mid y_{\text{gold}} = c)$ and $t_{0.20}^{(c)} = Q_{0.20}(\text{margin} \mid y_{\text{gold}} = c)$. High-confidence correct = correct & $p_{\max} \geq q_{0.80}^{(c)}$; Borderline correct = correct & margin $\leq t_{0.20}^{(c)}$; Hard-case error = incorrect & $p_{\max} \geq q_{0.80}^{(c)}$. Sampling for token summaries: up to $n = 20$ instances per gold class per bucket (or all if < 20).
--------------------	--------------------	---

applied only to the training data. Scores are reported as means and 95% percentile bootstrap confidence intervals ($B = 10,000$), where the bootstrap resamples the 10 run-level metric values with replacement and recomputes the mean at each replicate for Accuracy, Macro-F1, and Macro-Recall. This design makes the reported estimates more robust to split variability while keeping the uncertainty calculation aligned across all baselines.[4].

3.6 Handling class imbalance: train-only SMOTE

The class distribution in the training data is imbalanced, so SMOTE was used to increase sensitivity to minority classes by generating synthetic samples. The implementation was rigorous: SMOTE was applied after the split and only to the TF-IDF features of the training data (train-only SMOTE). The training TF-IDF matrix was converted to a dense representation before resampling, as SMOTE operates on dense numeric data. After resampling, the training matrix size changes from (4,147, 5,918) to (10,533, 5,918), with a balanced class distribution (3,511 per class), while D_{test} retains its original distribution, ensuring the evaluation remains realistic and leak-free.

SMOTE generates minority samples by interpolating the feature space between near-neighbors, so the model's decision boundary is less "pulled" toward the majority class, and minority recall tends to increase. However, because synthetic samples can increase overlap among classes in bag-of-words-based features, the evaluation is still performed on the original test data (without resampling) to assess the trade-off between false alarms and missed detections realistically.

3.7 Candidate models and evaluation metrics

For a fair and reproducible comparison, six common classical models for TF-IDF text classification were evaluated during the screening stage: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Multinomial Naive Bayes (MNB). The main report then focused on the most competitive candidates (LR and LinearSVC) in a leakage-safe pipeline with train-

only SMOTE. The evaluation used Accuracy, macro metrics (Macro-F1 and Macro-Recall), and a confusion matrix to assess interclass error patterns.

Because the confidence-aware protocols ($p_{\max}$ /margin/entropy) and LIME require a probability output (`predict_proba`), the SVM model was operationalized as LinearSVC, calibrated using CalibratedClassifierCV (`method="sigmoid", cv=3`) on the training data. This variant is marked as SVC_calibrated in the LAS results and XAI analysis. In addition to the classical linear models, we also included a transformer-based baseline to assess whether a contextual pretrained model yields better class-balanced performance under the same evaluation protocol.

3.7.1. Transformer baseline (IndoBERTweet)

To complement the classical baselines with a stronger contextual model for Indonesian Twitter text, we included a modern transformer baseline by fine-tuning IndoBERTweet (`indolem/indobertweet-base-uncased`), a pretrained Indonesian language model specialized for Twitter data. Unlike the TF-IDF pipeline, the transformer input used lightly cleaned tweets with URL and user placeholders and whitespace normalization, but without stopword removal or stemming, in order to preserve contextual and sequential cues. For fair comparison, IndoBERTweet was evaluated under the same leakage-safe protocol as the classical baselines, namely a stratified 80:20 train-test split repeated 10 times with identical split indices across models (`random_state = 42`).

To address class imbalance without applying text-space oversampling, we used class-weighted cross-entropy, where class weights were computed exclusively from the training split in each run. Fine-tuning was performed using AdamW with a learning rate of $2e-5$, a maximum sequence length of 128, a fixed random seed of 42, and 3 epochs. This setup was selected to provide a strong yet practical transformer baseline while maintaining a comparable evaluation framework across model families. By keeping the split protocol identical and restricting imbalance handling to the training data only, the comparison between classical models and the transformer baseline remains methodologically

consistent and leakage-safe. In this study, the transformer baseline was used for comparative performance evaluation in Section 4.2, whereas the downstream LAS and LIME-based audit analyses remained anchored to the calibrated sparse linear pipeline to preserve auditability and methodological consistency. For exact replication, all probability-dependent analyses in LAS/XAI use the train-only calibrated SVC_calibrated variant defined in Table 1, rather than the raw margin outputs of uncalibrated LinearSVC.

3.8 Label audit and confidence-aware XAI protocol based on LIME

To test the reliability of silver-standard labels and their impact on evaluation and interpretation, this study constructed a Label Audit Subset (LAS) (600 tweets; 200 per lexicon class) annotated by two annotators and finalized through adjudication to a gold label. LAS was used to (i) estimate lexicon label noise and (ii) evaluate changes in model metrics/rankings when the evaluation reference shifted from the lexicon to the gold label.

From a measurement perspective, silver-standard labels can be viewed as noise-contaminated targets; this noise lowers the upper bound on observed performance and can shift token interpretations because the model is trained/tested against proxy labels. Therefore, κ and the noise rate per class in LAS were used to quantify the reliability of lexicon labels and to examine whether model rankings and LIME token summaries remained consistent when the evaluation reference shifted to the gold label.

In the interpretability phase, LIME was used to generate local explanations for instance-level predictions. To mitigate the risk of overconfidence, a confidence-aware protocol is implemented with three indicators: $\max_prob = \max_k p(y = k | x_i)$, $\text{margin} = p_{\text{top1}} - p_{\text{top2}}$, and $\text{entropy } H(p) = -\sum_k p_k \log p_k$ as an ambiguity indicator. Instances are grouped into: high-confidence correct (correct with the highest $\max_prob$), borderline correct (correct but with a small/ambiguous margin), and hard-case error (incorrect with a high $\max_prob$). LIME explanations then prioritize stable predictions and audit borderline/hard-case cases for safer monitoring, escalation, and label quality improvement in subsequent iterations.

3.8.1. LAS-informed noise-robust learning

To test whether LAS can be used not only for diagnosis but also to inform noise-aware learning, we implemented a simple LAS-informed baseline based

on class-reliability weighting. Let $\eta_c = \Pr(y_{\text{gold}} \neq y_{\text{lex}} | y_{\text{lex}} = c)$ be the empirical noise rate for lexicon class c estimated from LAS, and define class reliability as $r_c = 1 - \eta_c$. Using the LAS estimates (Table 4), the resulting reliability weights are $r_{\text{Negative}} = 0.815$, $r_{\text{Neutral}} = 0.305$, and $r_{\text{Positive}} = 0.275$. During training, each instance i is assigned a sample weight $w_i = r_{y_{\text{lex},i}}$, so that examples from noisier lexicon classes contribute less to the loss while preserving the full corpus (no relabeling). For leakage control, LAS items were excluded from the training pool, and evaluation was performed on LAS only (training pool $n = 4582$; LAS $n = 600$; overlap = 0). This strategy downweights training examples from classes with higher observed label noise, while retaining the full corpus and avoiding direct relabeling of the silver-standard dataset.

3.9 Bucket selection rules and confidence calibration check

For each tweet i in LAS, we compute three indicators from the calibrated probabilistic model: $p_{\max}(i) = \max_k p_k(i)$, $\text{margin}(i) = p_{(1)}(i) - p_{(2)}(i)$ (difference between the largest and second-largest class probabilities), and $\text{entropy}(i) = -\sum_k p_k(i) \log p_k(i)$, where $p_{(1)}(i) \geq p_{(2)}(i) \geq \dots$. Bucket thresholds are defined numerically and computed per gold class within LAS. Specifically, for each gold class c , let $q_{0.80}^{(c)}$ be the 80th percentile of $p_{\max}$ among instances with gold label c , and let $t_{0.20}^{(c)}$ be the 20th percentile of margin within the same subset. We then define: (1) High-confidence correct as correctly predicted instances with $p_{\max}(i) \geq q_{0.80}^{(c)}$; (2) Borderline correct as correctly predicted instances with $\text{margin}(i) \leq t_{0.20}^{(c)}$; and (3) Hard-case error as incorrectly predicted instances with $p_{\max}(i) \geq q_{0.80}^{(c)}$ (confident but wrong). These thresholds are computed separately for each gold class, rather than globally, to prevent the dominant Negative class from disproportionately determining the confidence cutoffs for the minority classes.

For token-drift summaries, we sample up to $n = 20$ instances per gold class per bucket (or all available instances if fewer than 20). Table 9 reports one representative instance per bucket (highest $p_{\max}$ for high-confidence/hard-case; smallest margin for borderline). As a calibration sanity check, we report Expected Calibration Error (ECE, 10 bins) and Brier score on LAS: ECE10 = 0.45–0.62 and Brier = 0.97–1.25 against lexicon labels, and ECE10 = 0.50–0.66

and Brier = 1.05–1.33 against gold labels. Accordingly, confidence in this study is used as an audit heuristic to prioritize review, rather than as a calibrated probability for high-risk decision-making.

4. Result

4.1 Data distribution, feature representation, and data splitting

The dataset was collected from Twitter (collection period: late 2025) using a mental health-themed query. From 5,731 tweets, deduplication yielded $N=5,184$ unique tweets (with no post-cleaning empty entries). Sentiment labeling was automated using a lexicon, so labels are silver-standard and potentially noisy (vocabulary/irony/context bias); results should therefore be interpreted as discourse indicators rather than clinical diagnoses. The label distribution was Negative=4,389, Neutral=462, and Positive=333, indicating substantial class imbalance with Negative predominating. Evaluation used a stratified 80:20 split (train=4,147; test=1,037) repeated 10 times. TF-IDF produced a 5,918-dimensional feature space (Table 1); the vectorizer was fit on training data and applied to test data (transform-only) to prevent leakage. A summary of splits and dimensions is provided in Table 2.

4.2 Model evaluation on a leakage-safe pipeline (repeated stratified split + 95% CI)

All results in this subsection should be interpreted as lexicon-referenced performance estimates, because the full dataset labels are treated as silver-standard proxy labels rather than adjudicated ground truth. Under the same leakage-safe evaluation protocol—namely a $10\times$ repeated stratified 80:20 split with performance summarized as means and 95% confidence intervals, we compared the strongest classical baselines (Logistic Regression and LinearSVC) with a modern transformer baseline, IndoBERTweet. For the classical models,

Table 2. Summary of TF-IDF Splits and Dimensions

Component	Value
Ratio (train:test)	80:20
Data-train Number	4147
Data-test Number	1037
Document Total (N)	5184
TF-IDF Feature Number	5918
TF-IDF (train) Dimension	(4147, 5918)
TF-IDF (test) Dimension	(1037, 5918)

we evaluated train-only SMOTE against a no-SMOTE baseline using class weighting, whereas IndoBERTweet used class-weighted loss computed from the training split only. As shown in Table 3, the highest Mean Accuracy was achieved by LinearSVC (class-weighted) at 0.912 [0.909, 0.914], while the strongest class-balanced performance was obtained by IndoBERTweet, which yielded the highest Mean Macro-F1 (0.750 [0.739, 0.761]) and Mean Macro-Recall (0.763 [0.749, 0.778]). Among the classical baselines, LR (train-only SMOTE) remained the strongest on the macro-level metrics (Mean Macro-F1 = 0.727 [0.714, 0.740]; Mean Macro-Recall = 0.719 [0.703, 0.734]), indicating that train-only resampling improves minority-class sensitivity relative to class weighting alone in the sparse TF-IDF setting. Overall, these results suggest a clear trade-off: the best sparse linear baseline still provides the strongest overall accuracy, whereas the transformer baseline offers the most favorable balance across classes under the same repeated, leakage-safe protocol.

4.3 Error analysis using a confusion matrix

To examine the remaining error structure in a directly interpretable sparse-feature setting, Figure 1 presents the confusion matrix from a representative LinearSVC + train-only SMOTE split (random_state = 42). Although IndoBERTweet achieved the highest Mean Macro-F1 and Mean Macro-Recall in the repeated-split comparison, and LinearSVC (class-weighted) achieved the highest Mean Accuracy,

Table 3. Comparison of classical and transformer baselines under the same leakage-safe protocol

Model	Mean Accuracy [95% CI]	Mean Macro-F1 [95% CI]	Mean Macro-Recall [95% CI]
LR (train-only SMOTE)	0.905 [0.901, 0.909]	0.727 [0.714, 0.740]	0.719 [0.703, 0.734]
LR (class-weighted)	0.906 [0.903, 0.909]	0.706 [0.695, 0.717]	0.646 [0.632, 0.659]
LinearSVC (train-only SMOTE)	0.906 [0.903, 0.910]	0.722 [0.712, 0.732]	0.695 [0.684, 0.708]
LinearSVC (class-weighted)	0.912 [0.909, 0.914]	0.726 [0.715, 0.737]	0.686 [0.673, 0.698]
IndoBERTweet (class-weighted loss)	0.908 [0.905, 0.912]	0.750 [0.739, 0.761]	0.763 [0.749, 0.778]

Note: All confidence intervals are 95% percentile bootstrap intervals ($B = 10,000$) computed by resampling the 10 run-level scores with replacement and recomputing the mean

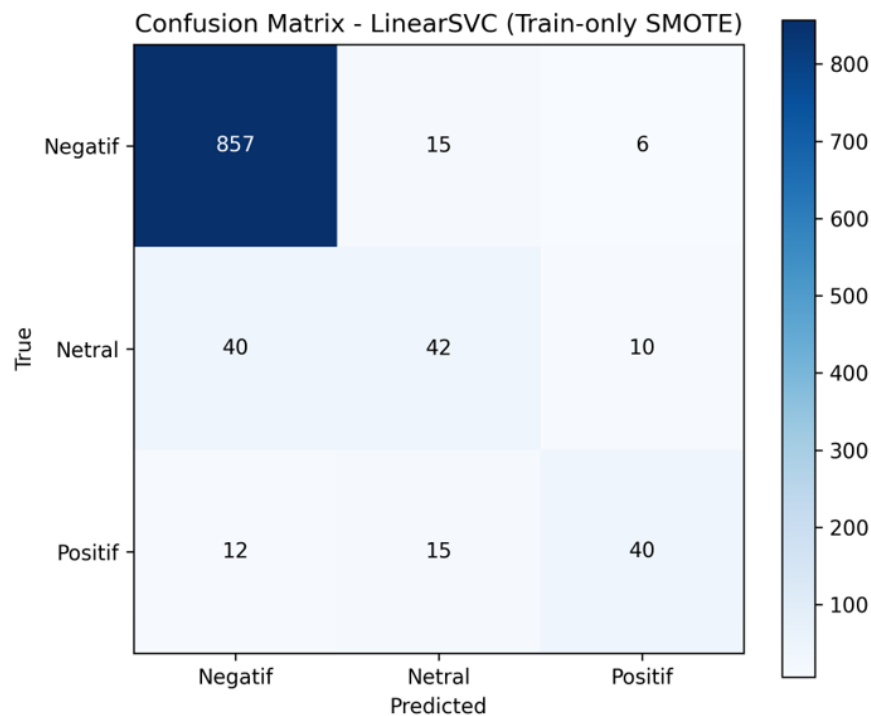


Figure. 1 Confusion matrix of LinearSVC after train-only SMOTE (representative split; test set n=1037)

Table 4. Results of the Label Audit and Noise Label Lexicon

Agreement and Cohen's κ (LAS)				
Subset	n	% Agreement	Cohen's κ	Notes
LAS (2 annotators)	600	92.5%	0.874	Very Strong
Borderline	184	96.7%	0.938	Very Strong
High-confidence	216	96.8%	0.938	Very Strong
Hard-case	200	84.0%	0.762	Strong
Confusion Lexicon vs Gold				
Lexicon \ Gold		Positive	Neutral	Negative
Positive		55	98	47
Neutral		22	61	117
Negative		7	30	163
Noise rate per class (based on LAS)				
Class (Lexicon)	n	Mismatch (n)	Noise rate	
Positive	200	145	72.5%	
Neutral	200	139	69.5%	
Negative	200	37	18.5%	
Overall	600	321	53.5%	

the confusion matrix of the sparse linear baseline remains useful for diagnosing class-level misclassification patterns in the same TF-IDF space used for the subsequent audit and explanation analyses. The dominant error pattern remains Neutral $\rightarrow$ Negative,

indicating that the Neutral class is the most difficult to separate when mental-health-related terms appear in informative, joking, quoted, or link-sharing contexts rather than in clearly affective statements. This pattern is consistent with the strong class imbalance in the full dataset and helps explain why improvements in macro-level performance are harder

to achieve than improvements in overall accuracy. Accordingly, Figure 1 should be read as a representative diagnostic view of residual error structure, rather than as evidence that this specific split or model variant is the best-performing model on every aggregate metric

4.4 Results of the label audit subset (LAS) and noise label lexicon

To test the reliability of silver-standard labels, a Label Audit Subset (LAS) was created containing 600 tweets (200 per lexicon class) annotated by two annotators and finalized through adjudication to

Table 5. Error Typology of LAS lexicon–gold mismatches (n = 321)

Error type	Operational definition	n	% of mismatches
Informative / quoted / link-sharing context	Distress-related keywords appear in informational, quoted, or link-sharing content rather than direct affective expression.	26	8.1%
Mixed-valence coping / recovery narrative	Distress tokens co-occur with coping, reassurance, gratitude, or recovery framing.	38	11.8%
Lexical ambiguity / polysemy	Mental-health-related keywords are used ambiguously, figuratively, or with shifted meaning.	23	7.2%
Humor / irony / sarcasm	Surface lexical polarity conflicts with intended pragmatic meaning.	38	11.8%
Code-switching / non-standard / OOV wording	Mixed language, slang, non-standard spelling, or OOV forms distort lexicon matching.	90	28.0%
Negation / scope / compositional polarity	Negation or compositional context reverses or alters apparent polarity.	102	31.8%
Other / unclear	Residual cases not adequately captured by the categories above.	4	1.2%
Total		321	100.0%

Table 6. LAS-informed noisy-label handling (class-reliability weighting) and its impact in LAS

Training strategy	LAS-lexicon Macro-F1	LAS-gold Macro-F1	LAS-lexicon Accuracy	LAS-gold Accuracy	LAS-gold Macro-Recall
Baseline LR (class_weight='balanced')	0.670	0.363	0.678	0.407	0.374
LAS-informed LR (balanced + reliability weight)	0.608	0.352	0.625	0.430	0.357

obtain a gold label. Annotation reliability was high (92.5% agreement; Cohen's $\kappa=0.874$). However, the lexicon vs. gold-label comparison revealed a 53.5% mismatch ($\kappa=0.198$), with noise levels varying across classes, most pronounced for Positive and Neutral (Table 4). These findings confirm that lexicon-based evaluations can inflate performance and distort interpretation.

4.4.1. Error typology of lexicon–gold mismatches

To move beyond aggregate mismatch rates, we manually coded the 321 LAS cases in which the lexicon label and the adjudicated gold label disagreed into a single dominant error type. Table 5 summarizes the resulting mismatch typology and its frequency distribution. The results show that lexicon-label noise is not random, but instead concentrates in recurrent linguistic and pragmatic patterns. The largest mismatch source was Negation / scope / compositional polarity (102 cases; 31.8%), followed by Code-switching / non-standard / OOV wording (90 cases; 28.0%). Together, these two categories account for 59.8% of all mismatches, indicating that surface-level lexicon rules are particularly vulnerable to polarity reversals and non-standard Twitter language. Additional mismatches arose from mixed-valence coping narratives, humor or sarcasm,

informative or quoted/link-sharing contexts, and lexical ambiguity. These findings help explain why Positive and Neutral lexicon labels exhibit substantially higher noise than Negative labels, and they further support interpreting lexicon-based results as proxy-label estimates rather than ground-truth performance.

To test whether LAS-derived noise information can be operationalized for noise-aware learning, we implemented a simple LAS-informed baseline based on class-reliability weighting, where training instances are weighted by class reliability estimated from LAS (Table 4). To keep the LAS evaluation leakage-free, LAS instances were excluded from the training pool (training pool $n = 4582$; LAS $n = 600$; overlap = 0). Table 6 reports the resulting LAS performance relative to the baseline LR (class_weight='balanced'). The LAS-informed weighting did not improve LAS-gold Macro-F1 (0.363 $\rightarrow$ 0.352) nor LAS-gold Macro-Recall (0.374 $\rightarrow$ 0.357), although LAS-gold Accuracy increased slightly (0.407 $\rightarrow$ 0.430). This outcome suggests that coarse class-level reliability weighting may trade minority-class sensitivity for overall accuracy, motivating future work on more fine-grained instance-level noise handling (e.g., mismatch-type-aware filtering or disagreement-driven reweighting) rather than uniform downweighting at the class level.

Table 7. LAS-only performance comparison under two evaluation references (LAS-lexicon vs LAS-gold) and aggregate ranking

Model	Macro-F1 (LAS-lex / LAS-gold) (Δ)	Accuracy (LAS-lex / LAS-gold) (Δ)	Macro-Recall (LAS-lex / LAS-gold) (Δ)	AvgRank (MF1+Acc)
LR_balanced	0.260/0.265 (+0.005)	0.357/0.315 (-0.042)	0.357/0.387 (+0.030)	1.25
SVC_calibrated	0.258/0.263 (+0.005)	0.355/0.315 (-0.040)	0.355/0.385 (+0.030)	2.00
KNN	0.177/0.185 (+0.008)	0.337/0.320 (-0.017)	0.337/0.344 (+0.008)	2.50

Note: LAS-lexicon = evaluation against lexicon-derived proxy labels within LAS; LAS-gold = evaluation against adjudicated gold labels within LAS; Δ = LAS-gold – LAS-lexicon. Aggregate rank summarizes the relative ordering across Macro-F1, Macro-Recall, and Accuracy

Table 8. LAS-only ranking by confidence bucket under LAS-lex vs LAS-gold evaluation references

Bucket (n)	Model	Macro-F1 (LAS-lex / LAS-gold) (Δ)	Macro-Recall / Accuracy (LAS-gold)	Rank (LAS-lex $\rightarrow$ LAS-gold) (Δ)
Borderline (184)	LR_balanced	0.242/0.235 (-0.007)	0.369/0.304	1 $\rightarrow$ 1 (+0)
Borderline (184)	SVC_calibrated	0.242/0.235 (-0.007)	0.369/0.304	1 $\rightarrow$ 1 (+0)
Borderline (184)	KNN	0.178/0.153 (-0.025)	0.333/0.299	3 $\rightarrow$ 3 (+0)
Hard-case (200)	SVC_calibrated	0.282/0.272 (-0.010)	0.347/0.320	1 $\rightarrow$ 1 (+0)
Hard-case (200)	LR_balanced	0.282/0.269 (-0.013)	0.342/0.315	2 $\rightarrow$ 2 (+0)
Hard-case (200)	KNN	0.196/0.206 (+0.010)	0.351/0.340	3 $\rightarrow$ 3 (+0)
High-confidence (216)	LR_balanced	0.215/0.229 (+0.014)	0.371/0.324	1 $\rightarrow$ 1 (+0)
High-confidence (216)	SVC_calibrated	0.207/0.198 (-0.009)	0.348/0.319	2 $\rightarrow$ 2 (+0)
High-confidence (216)	KNN	0.156/0.161 (+0.005)	0.333/0.319	3 $\rightarrow$ 3 (+0)

Note: Macro-F1 is reported as LAS-lexicon / LAS-gold, with Δ = LAS-gold – LAS-lexicon. Rank changes are shown as LAS-lexicon $\rightarrow$ LAS-gold, where Δ rank = 0 indicates unchanged model ordering within the bucket

4.5 Impact of label audit on model ranking in LAS

In the LAS-only audit evaluation ($n = 600$), comparing lexicon-referenced labels with adjudicated gold labels showed that performance estimates based on lexicon labels do not always reflect performance under audited labels. Table 7 summarizes the shifts in Macro-F1, Macro-Recall, and Accuracy for each model when the evaluation reference is changed from LAS-lexicon labels to LAS-gold labels. Model ranking based on Macro-F1 remained unchanged across the tested models (Spearman's $\rho = 1.000$), whereas Accuracy tended to decrease under gold-label evaluation, consistent with the substantial lexicon-label noise reported in Table 4. This result reinforces the need to report macro metrics and to distinguish clearly between lexicon-referenced full-dataset results and LAS-based gold-audit results, because the two evaluation references serve different inferential purposes. Furthermore, Table 8 shows that label-reference shifts primarily changed the magnitudes of the metrics rather than the relative ranking across models within the high-confidence, borderline, and hard-case buckets (Δ rank = 0).

4.6 XAI summary: confidence-aware LIME and token Drift

For instance-level auditing, LIME was used with a confidence-aware protocol that distinguishes between high-confidence correct, borderline correct, and hard-case errors based on maximum probability, margin, and entropy. A visual example of LIME explanations is shown in Figure 2, while representative examples per bucket and token summaries are presented in Table 9. Because the initial labels are silver-standard, tokens are treated as discourse patterns learned by the model rather than as clinical indicators. When case selection was based on gold labels rather than lexicons, the top-k token summaries changed significantly (Jaccard@20=0.176–0.290; Table 10). Table 10 reports point estimates from a fixed top-k summary, whereas Table 11 provides the large-sample controlled drift distribution requested by the reviewers (matched by predicted class and confidence bucket). The stability of LIME explanations across runs was also high (mean Jaccard ≥ 0.964 ; Table 12), making the reporting of top-k tokens reliable.

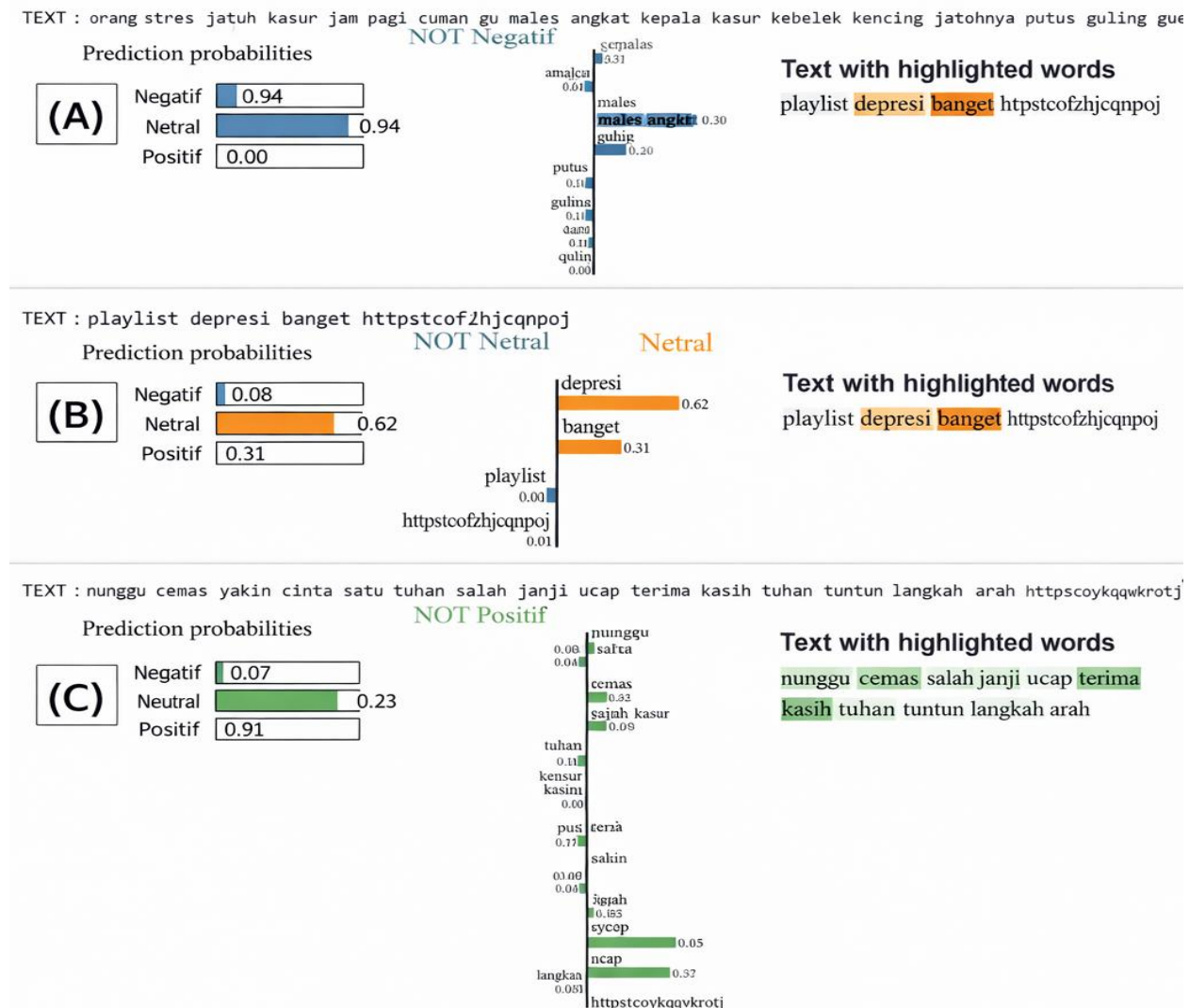


Figure. 2 Example LIME explanation per bucket (Original Indonesian tweet-translated for illustration): (A) high-confidence correct, (B) borderline correct, and (C) hard-case error

Calibration checks on the LAS showed that confidence tended to be overconfident (ECE10 0.50–0.66 against the gold label), so the $p_{\max}/\text{margin}$ indicator was used for audit priorities rather than as a calibrated probability for clinical decisions.

4.6.1. Robustness check: large-sample controlled token drift

To address potential confounds from bucket composition and class imbalance, we extended the token-drift analysis beyond point estimates by estimating the distribution of Jaccard@20 overlap between gold-referenced and lexicon-referenced token summaries under matched controls. Specifically, for each confidence bucket and predicted class, we drew matched-size bootstrap samples from the gold-referenced and lexicon-referenced pools, aggregated token importance across instances, and computed Jaccard@20 per

replicate. Table 11 shows that token drift persists under these controls for the dominant predicted class (Negative). In the high-confidence bucket, the controlled drift overlap is moderate (median Jaccard@20 = 0.333, IQR [0.290, 0.429], 5–95% 0.212–0.538), while drift becomes substantially stronger for borderline cases (median 0.143, IQR [0.143, 0.176]) and is strongest for hard-case errors (median 0.111, IQR [0.081, 0.111], 5–95% 0.053–0.176). These findings indicate that the token-level explanation summary can change materially when the evaluation reference shifts from lexicon proxy labels to audited gold labels, especially near the decision boundary and in confident-but-wrong cases. For the minority predicted classes (Neutral/Positive), controlled drift estimates were not computed when pool sizes were insufficient under the same matching constraints (Table 11, NA), which is consistent with the severe class imbalance and the bucket filtering process. While the present robustness check controls

Table 9. XAI examples (LIME) on the LAS (gold label): one representative instance per bucket

Bucket	Tweet snippet (abridged)	Gold→Pred (p_max)	Uncertainty (1-p_max)	+ Tokens (top-2)	- Tokens (top-2)
High-conf (correct)	But honestly, crying like this is really annoying. I've tried everything but I still cry, it's normal if I'm depressed.	Negative→Negative (1.000)	0.000	depresi / depression(+0.018); nangis / cry(+0.016)	jujur/honest (-0.009); dicoba/tried (-0.008)
Borderline (correct)	What's wrong, Jul? It's really bad. Don't worry, don't worry." / What's wrong, Jul? It's really bad. Don't worry, don't worry.	Negative→Negative (0.502)	0.498	Parah/Serious (+0.211); cemas (+0.188)	jangan / don't(-0.243); jangan / don't (-0.219)
Hard-case (error)	Mom, Dads Can Experience Postpartum Depression Too! http...	Neutral→Negative (0.999)	0.001	Depresi/depression (+0.016); Alami / Natural(+0.004)	https (-0.001); bun / mom(-0.001)

Note: Gold→Pred contains the gold label, prediction, and p_{max} ; uncertainty = $1 - p_{max}$. LIME tokens denote the words driving (+) and restraining (-). Text is shortened.

Table 10. Token drift Between Lexicon vs Gold (Jaccard@20)

Bucket	Jaccard@20 (gold vs lex)	n_toke n_gold	n_token _lex
High-confidence (correct)	0.250	20	20
Borderline (correct)	0.290	20	20
Hard-case (error)	0.176	20	20

for predicted class and confidence bucket through matched bootstrap resampling, additional sensitivity analysis across alternative LIME random seeds and different num_features settings remains a useful direction for future work.

5. Discussion

5.1 Leakage-safe pipeline (train-only SMOTE)

A key methodological contribution of this study is the train-only SMOTE strategy within a leakage-safe pipeline, which prevents resampling from contaminating the test data and yields more conservative performance reporting [48]. Under the lexicon-referenced full-dataset evaluation in Section 4.2, LinearSVC (class_weight='balanced') achieved the highest overall Accuracy, whereas IndoBERTweet produced the strongest Macro-F1 and Macro-Recall.

Table 11. Controlled token-drift distribution (Jaccard@20) between gold-referenced vs lexicon-referenced summaries (bootstrap; matched by predicted class and confidence bucket)

Bucket	Predicted class (yhat)	n_gold	n_lex	m	Jaccard@20 median [IQR]	Jaccard@20 (5-95%)
High-confidence (correct)	Negative	46	40	40	0.333 [0.290, 0.429]	0.212–0.538
High-confidence (correct)	Neutral	2	30	—	NA	NA
High-confidence (correct)	Positive	3	15	—	NA	NA
Borderline (correct)	Negative	34	34	34	0.143 [0.143, 0.176]	0.081–0.250
Borderline (correct)	Neutral	8	18	—	NA	NA
Borderline (correct)	Positive	0	9	—	NA	NA
Hard-case (error)	Negative	45	32	32	0.111 [0.081, 0.111]	0.053–0.176
Hard-case (error)	Neutral	20	3	—	NA	NA
Hard-case (error)	Positive	5	0	—	NA	NA

Note: NA indicates insufficient pool size for controlled resampling (one side < 10 instances). m denotes the matched bootstrap sample size per side and is set to $m = \min(n_{gold}, n_{lex})$

Table 12. LIME explanation stability (inter-run Jaccard mean)

Bucket	Mean Jaccard	Std. Jaccard	n
Borderline (correct)	0.964	0.081	5
Hard-case (error)	1.000	0.000	5
High-confidence (correct)	1.000	0.000	5

Within the classical sparse baselines, LR with train-only SMOTE remained competitive on macro-level metrics, indicating that train-only resampling can improve minority-class sensitivity relative to class weighting alone. Thus, the contribution of the leakage-safe design is not that one single balancing strategy universally dominates, but that model comparisons remain methodologically valid and interpretable under clearly stated label references.

5.2 Error pattern: neutral difficult to distinguish

The LinearSVC confusion matrix (Figure 1) shows a dominant Neutral→Negative classification trend. In social media text, Neutral often includes mental health terms in informative, humorous, or quote/link contexts (e.g., “*playlist depresi...*”), so strong tokens like “*depresi*,” “*stres*,” or “*overthinking*” push predictions toward Negative. This pattern aligns with mental health mining findings that content is heterogeneous and exhibits high pragmatic ambiguity in the absence of a richer context. In contrast, TF-IDF tends to capture lexical associations rather than authorial intent [48, 49]. Positive→Negative errors highlight challenges in coping narratives: words connoting distress (“*stress*,” “*cemas*”) persist but are followed by resolution/support (“*Alhamdulillah*,” “*tahan*,” “*terimakasih*”). Here, local explanation (LIME) serves as an audit tool to assess whether the model captures coping patterns or is simply biased toward distress tokens, which is relevant to the false alarm vs. missed detection trade-off in mental health monitoring[49, 50].

5.3 Silver-standard labeling caveats and their impact on model interpretation

Lexicon labeling is scalable, but it remains a silver standard and is therefore vulnerable to systematic noise [51, 52]. In the LAS audit, 321 of 600 cases (53.5%) differed between lexicon-derived proxy labels and adjudicated gold labels, confirming that lexicon-based results should be read as proxy-label estimates rather than as ground-truth

performance. The mismatch typology in Table 5 further shows that this noise is concentrated in specific discourse phenomena, especially Negation / scope / compositional polarity (31.8%) and Code-switching / non-standard / OOV wording (28.0%), which together account for 59.8% of all mismatches. Additional contributions come from mixed-valence coping narratives, humor or sarcasm, informative or quoted/link-sharing contexts, and lexical ambiguity. These findings indicate that lexicon-based evaluation can misstate both absolute performance and the token patterns later surfaced by LIME, particularly when polarity is altered by context or when Twitter language departs from standard lexical forms[49]. Accordingly, model outputs and local explanations should be interpreted as discourse-level indicators rather than as clinical claims. Although absolute metrics changed under LAS-gold evaluation, the Macro-F1-based model ranking remained stable, supporting the use of macro metrics plus label audits when making comparative model claims. This typology-based audit makes the label-noise argument more concrete by showing not only how much mismatch exists, but also which discourse patterns most often generate it.

5.4 Practical implications of a confidence-aware XAI protocol for intervention/monitoring

The main contribution of this study is the confidence-aware XAI protocol: explanations are not randomly selected but are arranged by predictive stability, for a more secure and targeted audit. The XAI analysis is divided into high-confidence, borderline, and hard-case categories, thereby separating safe cases for trend aggregation, ambiguous cases for human review/label audit, and the most informative “confident but wrong” cases for improving lexicon/weak supervision [54-56]. The LIME token also exhibits drift when selecting gold LAS-based examples over lexicon-based examples, so that interpretations are not locked into label noise. At the same time, the stability of explanations between runs remains high, supporting the auditability of the token summaries. The following section boils down the borderline/hard-case findings into operational design implications.

5.5 Connecting token drift to audit-informed design hypotheses

To avoid overstating the downstream implications of token drift, this study treats drift-to-design mapping as a hypothesis-generation step rather than as validated product guidance. Explanatory tokens from the confidence-aware LIME

audit were mapped to user-need signals (distress, coping/recovery, social support, and friction), then translated into design hypotheses across three dimensions: gamification, social support, and friction mitigation. For gamification, the patterns in borderline and hard-case cases suggest that low-pressure mechanics, such as micro-goals, non-punitive streaks, and reflective rewards, may be safer to test than competitive elements such as leaderboards or loss-based streaks. For social support, the observed patterns suggest that gradual prompts and trusted-ties-based support may be more appropriate for borderline cases than broader or more immediate social exposure. For friction mitigation, the drift patterns indicate that hard-case scenarios may benefit from testing crisis-oriented simplification, while borderline scenarios may benefit from reduced cognitive load through empathetic wording and fewer interface interruptions.

Thus, separating confidence strata allows XAI to function as an audit-informed mechanism for prioritizing which design hypotheses should be tested first. Measures such as micro-task completion, onboarding drop-off, time-to-help, and self-reported helpfulness are therefore proposed as future evaluation targets rather than validated effects of the present study.

5.6 Limitations

This work has several limitations. First, the initial sentiment labels are lexicon-based silver-standard proxies; although LAS helps quantify noise and guide audits, the generalizability of findings is still affected by vocabulary bias, irony, and pragmatic context. Second, confidence ($p_{\text{max}}/\text{margin}$) is used as an audit prioritization heuristic, not a calibrated probability estimate; therefore, confidence scores should not be directly interpreted as precise risk assessments for high-risk scenarios. Third, the determination of high-confidence/borderline/hard-case buckets is operational for audit purposes; the threshold used has not been validated against external outcomes. Fourth, TF-IDF was chosen as an easily auditable baseline, but it has limitations in capturing richer intent and context in social media text. Fifth, although we added a modern transformer baseline (IndoBERTweet) for a fair comparison, further work is needed to evaluate additional Indonesian transformers and context-aware explanation methods tailored to these models.

In addition, although Table 11 strengthens the robustness of the token-drift analysis under matched controls, the sensitivity of drift estimates to alternative LIME seeds and explanation lengths

(num_features) was not exhaustively examined in the present study and should be addressed in future work.

Conclusion and Future Work

This study shows that social media text-based mental health discourse monitoring can be made more replicable and auditable by using a TF-IDF baseline, multi-model evaluation, and train-only SMOTE within a leakage-safe pipeline. The LAS audit ($n=600$) showed 92.5% inter-annotator agreement (Cohen's $\kappa=0.874$), but only 46.5% agreement of lexicon labels to gold (53.5% mismatch; $\kappa=0.198$), suggesting that XAI performance estimates and interpretations may shift as label references are validated. Therefore, performance reporting should emphasize macro metrics and clearly state the label references used. The LIME confidence-aware protocol helps prioritize audits on borderline/hard cases, and token drift is used to derive design hypotheses for low-pressure gamification, social support based on trusted ties and friction mitigation through triage/escalation of assistance.

Moving forward, follow-up studies should gradually expand label audits, add probability calibration, and validate bucket thresholds before implementation. Follow-up implementations should also prioritize human-in-the-loop, aggregate analysis, and privacy/ethics principles to ensure safe and responsible use for monitoring/intervention. Furthermore, follow-up studies can expand the comparison to additional Indonesian transformer models and evaluate explanation methods designed for contextual encoders, complementing the auditable TF-IDF + LIME pipeline.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, F. Marisa; methodology, F. Marisa, and SSS. Ahmad; validation, F. Marisa; formal analysis, F. Marisa, and A. Noertjahyana; investigation, F. Marisa, and DA. Dewi; resources, F. Marisa, and AL. Maukar; data curation, F. Marisa, and I. Istiadi; writing original draft preparation, F. Marisa, and I. Istiadi; writing review and editing, F. Marisa, and DA. Dewi.

Acknowledgments

The authors would like to thank UWG, UTeM, INTI International University, Petra Christian

University, President University for research support and funding.

References

- [1] J. A. Naslund, A. Bondre, J. Torous and K. A. Aschbrenner, "Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice", *Journal of Technology in Behavioral Science*, Vol. 5, No. 3, pp. 245-257, 2020, doi: 10.1007/s41347-020-00134-x.
- [2] J. Licinio and M. Li, "Digital footprints as a new translational approach for mental health care: a commentary", *Discover Mental Health*, pp. 10-13, 2023, doi: 10.1007/s44192-023-00032-7.
- [3] S. Mangalik, *et al.*, "Robust language-based mental health assessments in time and space through social media", *NPJ Digital Medicine*, Vol. 7, No. 1, Art. 109, 2024, doi: 10.1038/s41746-024-01100-0.
- [4] S. Biglbauer and A. L. Korajlija, "Social Networks, Depression and Anxiety", *Socijalna psihijatrija*, Vol. 48, No. 4, pp. 404-425, 2020, doi: 10.24869/spsih.2020.404.
- [5] R. T.-Costa, A. Martinez-Millana, N. Azzopardi-Muscat and L. Lazeri, "Methodological and Quality Flaws in the Use of Artificial Intelligence in Mental Health Research: Systematic Review", *JMIR Mental Health*, Vol. 10, Art. e42045, 2023, doi: 10.2196/42045.
- [6] D. R. Rice and C. Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies", *Political Science Research and Methods*, Vol. 9, No. 1, pp. 20-35, 2021, doi: 10.1017/psrm.2019.10.
- [7] C. Yang, Y. Zhang, and S. Muresan, "Weakly-Supervised Methods for Suicide Risk Assessment: Role of Related Domains", In: *Proc. of 59th Annual Meeting of the Association for Computational Linguistics and 11th International Joint Conference on Natural Language Processing (Short Papers)*, pp. 1049-1057, 2021, doi: 10.18653/v1/2021.acl-short.133.
- [8] A. Barredo Arrieta, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82-115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [9] F. Angiulli, F. Fassetti and S. Nistico, "Improving local interpretable classifier explanations exploiting self-generated semantic features", *Neural Computing and Applications*, Vol. 37, pp. 13617-13641, 2025, doi: 10.1007/s00521-025-11227-0.
- [10] A. Meier and L. Reinecke, "Computer-Mediated Communication, Social Media, and Mental Health: A Conceptual and Empirical Meta-Review", *Communication Research*, Vol. 48, No. 8, pp. 1182-1209, 2021, doi: 10.1177/0093650220958224.
- [11] N. H. Di Cara, V. Maggio, O. S. P. Davis and M. A. Haworth, "Methodologies for Monitoring Mental Health on Twitter: Systematic Review", *Journal of Medical Internet Research*, Vol. 25, Art. e42734, 2023, doi: 10.2196/42734.
- [12] S. Mangalik, *et al.*, "Robust language-based mental health assessments in time and space through social media", *Machine Learning*, Vol. 112, pp. 5091-5114, 2023, doi: 10.1007/s10994-023-06342-9.
- [13] A. Bashar, R. Nayak, K. Luong and T. Balasubramaniam, "Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts", *Social Network Analysis and Mining*, Vol. 11, Art. 69, 2021.
- [14] Ş. Kayıkçı, "SenDemonNet: sentiment analysis for demonetization tweets using heuristic deep neural network", *Multimedia Tools and Applications*, Vol. 81, pp. 11341-11378, 2022, doi: 10.1007/s11042-022-11929-w.
- [15] M. L. Jamil, S. Pais, J. Cordeiro and G. Dias, "Detection of extreme sentiments on social networks with BERT", *Social Network Analysis and Mining*, Vol. 12, No. 1, Art. 55, 2022, doi: 10.1007/s13278-022-00882-z.
- [16] R. Jain, A. Kumar, A. Nayyar, K. Dewan and R. Garg, "Explaining sentiment analysis results on social media texts through visualization", *Multimedia Tools and Applications*, Vol. 82, pp. 22613-22629, 2023, doi: 10.1007/s11042-023-14432-y.
- [17] D. Liu, X. L. Feng, F. Ahmed, M. Shahid and J. Guo, "Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review", *JMIR Mental Health*, Vol. 9, No. 3, Art. e27244, 2022, doi: 10.2196/27244.
- [18] C. Bour, A. Ahne, S. Schmitz, C. Perchoux, C. Dessenne and G. Fagherazzi, "The Use of Social Media for Health Research Purposes: Scoping Review", *Journal of Medical Internet Research*, Vol. 23, No. 5, Art. e25736, 2021, doi: 10.2196/25736.
- [19] B. D. Davis, *et al.*, "Quantifying depression-related language on social media during the COVID-19 pandemic", *International Journal of*

- Population Data Science*, Vol. 5, No. 4, Art. 1716, 2022, doi: 10.23889/ijpds.v5i4.1716.
- [20] M. Bak, C. Chiu and J. Chin, "Mental Health Pandemic During the COVID-19 Outbreak: Social Media as a Window to Public Mental Health", *Cyberpsychology, Behavior, and Social Networking*, Vol. 26, No. 5, pp. 346-356, 2023, doi: 10.1089/cyber.2022.0116.
- [21] Y. Lee, Y. J. Jeon, S. Kang, J. I. Shin, Y. C. Jung and S. J. Jung, "Social media use and mental health during the COVID-19 pandemic in young adults: a meta-analysis of 14 cross-sectional studies", *BMC Public Health*, Vol. 22, Art. 995, 2022, doi: 10.1186/s12889-022-13409-0.
- [22] Q. Yang, J. Liu and J. Rui, "Association Between Social Network Sites Use and Mental Illness", *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, Vol. 16, No. 1, Art. 1, 2022, doi: 10.5817/CP2022-1-1.
- [23] J. Cha, S. Kim and E. Park, "A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community", *Humanities and Social Sciences Communications*, Vol. 9, No. 1, pp. 1-10, 2022, doi: 10.1057/s41599-022-01313-2.
- [24] A. Nadeem, M. Naveed, M. I. Satti, H. Afzal, T. Ahmad and K. Kim, "Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to Social Networking Data", *Sensors*, Vol. 22, No. 24, Art. 9775, 2022, doi: 10.3390/s22249775.
- [25] Y. Yu, Q. Li and X. Liu, "Automatic anxiety recognition method based on microblog text analysis", *Frontiers in Public Health*, Vol. 11, Art. 1080013, 2023, doi: 10.3389/fpubh.2023.1080013.
- [26] A. Meier and L. Reinecke, "Computer-Mediated Communication, Social Media, and Mental Health: A Conceptual and Empirical Meta-Review", *Communication Research*, Vol. 48, No. 8, pp. 1182-1209, 2021, doi: 10.1177/0093650220958224.
- [27] K. Saha, A. Yousuf, R. L. Boyd and J. W. Pennebaker, "Social Media Discussions Predict Mental Health Consultations on College Campuses", *Scientific Reports*, Vol. 12, Art. 123, 2022, doi: 10.1038/s41598-021-03423-4.
- [28] Y. Son, *et al.*, "World Trade Center responders in their own words: predicting PTSD symptom trajectories with AI-based language analyses of interviews", *Psychological Medicine*, 2021, doi: 10.1017/S0033291721002294.
- [29] D. Suresh, A. Naresh, K. Nagwani and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review", *Artificial Intelligence Review*, Vol. 56, No. 9, pp. 10345-10389, 2023, doi: 10.1007/s10462-023-10419-1.
- [30] G. Rotman and R. Reichart, "Multi-task Active Learning for Pre-trained Transformer-based Models", *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 1208-1225, 2022, doi: 10.1162/tacl_a_00515.
- [31] P. Ganesh, *et al.*, "Compressing Large-Scale Transformer-Based Models: A Case Study on BERT", *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 1061-1080, 2021, doi: 10.1162/tacl_a_00413.
- [32] H. Dai, C. Su, Y. Lee, Y. Zhang and C. Wang, "Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients", *Frontiers in Psychiatry*, Vol. 11, Art. 533949, 2021, doi: 10.3389/fpsy.2020.533949.
- [33] S. Aum, S. Choe, J. Kim and H. Ji, "srBERT: automatic article classification model for systematic review using BERT", *Systematic Reviews*, Vol. 10, Art. 285, 2021, doi: 10.1186/s13643-021-01763-w.
- [34] D. R. Hughes and J. Halberstadt, "Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models", *PLoS One*, Vol. 17, No. 10, Art. e0276367, 2022, doi: 10.1371/journal.pone.0276367.
- [35] M. Khushi, *et al.*, "A Comparative Analysis of Data Resampling Methods on Imbalance Medical Data", *IEEE Access*, Vol. 9, pp. 109960-109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [36] J. Li, "Synthetic Minority Oversampling Technique Based on Adaptive Local Mean Vectors and Improved Differential Evolution", *IEEE Access*, Vol. 10, pp. 74045-74058, 2022, doi: 10.1109/ACCESS.2022.3187699.
- [37] W. Yang, C. Pan and Y. Zhang, "An oversampling method for imbalanced data based on spatial distribution of minority samples SD-KMSMOTE", *Scientific Reports*, Vol. 12, Art. 17345, 2022, doi: 10.1038/s41598-022-21046-1.
- [38] J. Yun and J. Lee, "Learning from Imbalanced Data Using Triplet Adversarial Samples", *IEEE Access*, Vol. 11, pp. 34279-34290, 2023, doi: 10.1109/ACCESS.2023.3262604.
- [39] S. Ayoub, Y. Gulzar, J. Rustamov, A. Jabbari, F. A. Reegu and S. Turaev, "Adversarial Approaches to Tackle Imbalanced Data in Machine Learning", *Sustainability*, Vol. 15, No. 9, Art. 7097, 2023, doi: 10.3390/su15097097.

- [40] Y. N. Rao, "An Imbalanced Generative Adversarial Network-Based Approach for Network Intrusion Detection", *Sensors*, Vol. 23, No. 1, Art. 550, 2023, doi: 10.3390/s23010550.
- [41] M. Kim and K. H. Seo, "An empirical evaluation of sampling methods for the classification of imbalanced data", *PLoS One*, Vol. 17, No. 7, Art. e0271260, 2022, doi: 10.1371/journal.pone.0271260.
- [42] R. Delgado and J. D. Núñez-González, "Bayesian network-based over-sampling method (BOSME) with application to indirect cost-sensitive learning", *Scientific Reports*, Vol. 12, Art. 8724, 2022, doi: 10.1038/s41598-022-12682-8.
- [43] D. W. Joyce, A. Kormilitzin, K. A. Smith and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability", *NPJ Digital Medicine*, Vol. 6, Art. 6, 2023, doi: 10.1038/s41746-023-00751-9.
- [44] D. E. Ehrmann, "Making machine learning matter to clinicians: model actionability in medical decision-making", *NPJ Digital Medicine*, Vol. 6, Art. 7, 2023, doi: 10.1038/s41746-023-00753-7.
- [45] U. Ehsan, K. Saha, M. D. De Choudhury and M. O. Riedl, "Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI", In: *Proc. of the ACM on Human-Computer Interaction*, Vol. 7, Art. 1-32, 2023, doi: 10.1145/3579467.
- [46] Z. H. Wang and Q. Liu, "Imbalanced Data Classification Method Based on LSSASMOTE", *IEEE Access*, Vol. 11, pp. 32252-32260, 2023, doi: 10.1109/ACCESS.2023.3262460.
- [47] R. Dwivedi, D. Dave, O. Rana, P. Patel and B. Qian, "Explainable AI (XAI): Core Ideas, Techniques and Solutions", *ACM Computing Surveys*, 2022, doi: 10.1145/3561048.
- [48] A. Demircioğlu, "Applying oversampling before cross-validation will lead to high bias in radiomics", *Scientific Reports*, Vol. 14, Art. 11563, 2024, doi: 10.1038/s41598-024-62585-z.
- [49] A. N. Bazzano, T. Patel, E. Nauman and D. Cernigliaro, "Optimizing Telehealth for Diabetes Management in the Deep South of the United States: Qualitative Study of Barriers and Facilitators on the Patient and Clinician Journey", *Journal of Medical Internet Research*, Vol. 26, Art. e43583, 2024, doi: 10.2196/43583.
- [50] C. Saxena, M. Garg and G. Ansari, "Explainable Causal Analysis of Mental Health on Social Media Data", *Neural Information Processing*, pp. 172-183, 2023.
- [51] R. Tekumalla, "Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions", *Neural Computing and Applications*, 2021, doi: 10.1007/s00521-021-06614-2.
- [52] G. Fu, C. Song, J. Li, Y. Ma, P. Chen and R. Wang, "Distant Supervision for Mental Health Management in Social Media: Suicide Risk Classification System Development Study", *Journal of Medical Internet Research*, Vol. 23, No. 8, Art. e26119, 2021, doi: 10.2196/26119.
- [53] J. Geng, F. Cai, Y. Wang, H. Koepl, P. Nakov and I. Gurevych, "A Survey of Confidence Estimation and Calibration in Large Language Models", In: *Proc. of 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 6577-6595, 2024, doi: 10.18653/v1/2024.naacl-long.366.
- [54] S. Rai, "Key language markers of depression on social media depend on race", *Proceedings of the National Academy of Sciences*, Vol. 121, No. 14, Art. e2319837121, 2024, doi: 10.1073/pnas.2319837121.
- [55] D. Owen, A. J. Lynham, S. E. Smart, A. F. Pardiñas and J. Camacho Collados, "AI for Analyzing Mental Health Disorders Among Social Media Users: Quarter-Century Narrative Review of Progress and Challenges", *Journal of Medical Internet Research*, Vol. 26, Art. e59225, 2024, doi: 10.2196/59225.
- [56] N. Pourshahrokhi, M. Hunter, J. Farokhzadian and L. Ahmadian, "Identifying the content, capabilities, and design features of a mobile-based cognitive behavioral therapy intervention for managing menopausal symptoms", *International Journal of Medical Informatics*, Vol. 206, Art. 106163, 2026, doi: 10.1016/j.ijmedinf.2025.106163.