computation

MDPI

Article

From Chaos to Gleichgewicht: An Application-Specific BiLSTM Framework with SMOTE, SHAP, and TF-IDF-Based Safety Support for Proxy Risk Screening on Noisy Social Media

Akas Bagus Setiawan [1], Hendra Yufit Riskiawan [1,*], Taufiq Rizaldi [1], Hermawan Arief Putranto [1], Rachmad Andri Atmoko [2], Andi Besse Firdausiah Mansur [3], Mohannad Alharthi [4] and Ahmad Hoirul Basori [4]

[1] Department of Information Technology, Politeknik Negeri Jember, Jember 68101, Indonesia; akasbagus_s@polije.ac.id (A.B.S.); taufiq_r@polije.ac.id (T.R.); hermawan_arief@polije.ac.id (H.A.P.)
[2] Faculty of Vocational Studies, Universitas Brawijaya, Malang 65145, Indonesia; ra.atmoko@ub.ac.id
[3] Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh 21911, Saudi Arabia; abmansur@kau.edu.sa
[4] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh 21911, Saudi Arabia; malharthi1@kau.edu.sa (M.A.); abasori@kau.edu.sa (A.H.B.)
* Correspondence: yufit@polije.ac.id

Abstract

Background: Social-media-based risk screening is promising, yet deployment quality is constrained by noisy language, class imbalance, low explainability, and uncertainty handling. Objective: This work develops an application-specific BiLSTM-centered proxy risk screening framework that combines chaos-regularized learning, SMOTE balancing, SHAP explanation, and TF-IDF similarity fallback. Methods: Experiments used an annotated Twitter corpus (>31,000 English posts) with neutral, hate, and offensive labels, where hate/offensive content was treated only as a proxy risk indicator rather than as evidence of depression, self-harm, or clinical psychological distress. The workflow covered text normalization, tokenization, fixed-length sequence modeling, imbalance correction, and thresholded inference. The model stack used GloVe embeddings with recurrent layers and chaos-based dropout, trained with Adam plus adaptive stopping/scheduling. Operational safety was supported through lexicon checks and a balanced TF-IDF reference set for low-confidence cases. Results: Validation performance reached an F1-score of 0.86, accuracy 0.79, and AUC 0.93 under the reported pipeline. Component-level evidence further indicated that SHAP, thresholding, TF-IDF retrieval, and lexicon checks add explanation, uncertainty handling, reference context, and conservative escalation beyond the underlying classifier, although their marginal effects were not isolated through rerun ablation. Conclusion: The proposed system offers a practical tradeoff between predictive quality, interpretability, and safety-aware operation for proxy risk screening on noisy social media. The contribution is positioned as an integrated application pipeline rather than a new NLP algorithm, and future work should validate the framework with clinically grounded labels, leakage-safe resampling, ablation studies, and transformer baselines.

Keywords: proxy risk screening; social media; BiLSTM; SMOTE; SHAP; TF-IDF; chaos theory; deep learning; explainable AI; uncertainty-aware decision-support

Academic Editor: Francesco Cauteruccio

Received: 19 June 2026
Revised: 20 July 2026
Accepted: 22 July 2026
Published: 24 July 2026

1. Introduction

As a major public-health concern, depression is increasingly studied through digital traces from social platforms [1,2]. Short and informal posts often contain emotional cues that are useful for large-scale early-risk screening and population-level monitoring [2–4]. Even so, social-media language is not equivalent to clinical evidence, so model outputs must be interpreted as risk indications rather than diagnostic conclusions [1,2].

From an implementation perspective, proxy risk classification on social text remains challenging due to noisy tokens, sparse semantics, and severe label imbalance [2,5,6]. Slang, abbreviations, sarcasm, and broken syntax reduce feature reliability for standard classifiers [4,6,7]. In realistic corpora, high-risk language typically appears as a minority pattern, which increases false-negative risk when imbalance treatment is weak [8–10].

Deep learning advances, including transformer models, have improved benchmark scores across mental-health and harmful-content NLP tasks [11–13]. However, deployment environments require more than raw accuracy: decisions must be interpretable, uncertainty-aware, and computationally efficient [1,14,15]. Prior work on uncertainty-aware XAI and classification with rejection shows that confidence information, explanation, and human oversight should be treated as connected parts of a decision-support process rather than as independent add-ons [16,17]. For operational screening systems, transparent outputs and fallback logic are necessary when confidence drops or context is ambiguous [14,18,19].

This conceptual framing is also aligned with the idea of "chaos" and "equilibrium" (gleichgewicht), reflected in our design that combines chaos-based regularization and class-balancing strategies for robust proxy risk detection.

This study does not pursue a clinical depression diagnosis. Instead, it models linguistic risk proxies derived from neutral versus hate/offensive annotations. This choice imposes a substantial construct-validity boundary: hostility or offensiveness is not equivalent to depression, self-harm, or psychological distress. In noisy online discourse, distress, hostility, and self-directed negativity may overlap in lexical and emotional markers; however, that overlap is only sufficient for conservative proxy screening, not for clinical inference.

Addressing these gaps, our approach presents an efficient and explainable BiLSTM-centered application pipeline for proxy risk detection in noisy social text. The pipeline combines chaos-based dropout for regularization stability [20], SMOTE for imbalance mitigation [8–10], and SHAP for feature-level explanation [14,15,21]. It also integrates TF-IDF similarity retrieval to supply reference context under uncertainty [14], plus a lexicon-based high-risk phrase check as an additional safety layer [4,18].

The primary contribution is a unified application-specific pipeline that jointly combines predictive modeling, interpretability, uncertainty handling, and conservative safety support under practical deployment constraints. Using a large annotated Twitter corpus, hate/offensive labels are treated as proxy indicators of risk-associated language, and the system is evaluated through classification performance (including F1-score and AUC) together with interpretability evidence. This framing positions the work as a deployable proxy-screening system while acknowledging that clinically oriented use cases require richer labels, stronger contextual modeling, and direct validation against clinically meaningful outcomes [2,11,12].

The novelty claim is therefore intentionally conservative. This work does not propose a new NLP methodology or algorithmic advance over BiLSTM, GloVe, SMOTE, SHAP, TF-IDF retrieval, or rule-based safety checks. Its contribution is the application-specific integration and documentation of these established components for proxy risk screening, with explicit uncertainty-aware fallback and human-review support.

2. Literature Review

The conceptual basis of this study draws on behavioral and affective theories of online self-disclosure. Previous studies indicate that distress expression on social media is often indirect, context-dependent, and influenced by platform norms [1,2,4]. As a result, linguistic cues and short sequential patterns remain central for proxy risk detection in noisy posts [2,6,7]. This evidence supports sequence modeling enriched with explicit uncertainty handling rather than reliance on a single confidence score [1,3,19].

From a modeling standpoint, BiLSTM is still practical when the goal is to balance classification effectiveness and implementation simplicity [11,12,22]. Label imbalance remains a structural issue in risk-related corpora, and SMOTE is repeatedly validated as an effective balancing strategy [8–10]. In parallel, chaos-based nonlinear regularization has been explored to stabilize training behavior under noisy inputs [20]. Together, these findings support our architecture selection for operational proxy risk screening.

For interpretability and safety, SHAP offers a principled attribution mechanism to explain token-level contribution patterns [14,15,21]. The more recent uncertainty-aware XAI literature further argues that explanations should be evaluated together with uncertainty communication and selective human escalation. Marx et al. [16] frame uncertainty as a core dimension of explanation reliability, while Thuy and Benoit [17] connect uncertainty estimation with classification rejection and human-in-the-loop decision-making. These studies are directly relevant to our confidence-based fallback design, but they also show that our pipeline follows an established uncertainty-aware decision-support pattern rather than introducing a new theoretical XAI method.

Because the present study relies on proxy signals extracted from social media, the construction and interpretation of such statistics also require caution. Vilenchik [23] demonstrates that seemingly simple social-media statistics can be highly sensitive to sampling and data-collection methodology. This concern applies to our hate/offensive-to-risk mapping: the labels are useful for studying operational proxy screening, but they cannot be treated as direct measurements of depression or psychological distress.

TF-IDF retrieval remains efficient for reference-based contextual support when prediction confidence is low [19]. Lexicon rules also function as practical safeguards for high-risk phrase detection [4,18]. Collectively, this literature supports a hybrid pipeline that unifies prediction, explanation, and conservative fallback behavior while keeping the claims bounded to application-specific proxy screening.

3. Methods

This section presents the full experimental pipeline, including data preparation, model construction, explainability, uncertainty handling, and reproducibility support. The process begins with preprocessing and imbalance correction, then proceeds to BiLSTM training and threshold selection, followed by SHAP interpretation and TF-IDF-based fallback logic. For readability, we organize the method into five linked components (A–E): data/exploration, model/training, explanation/safety, operational workflow, and reproducibility.

As shown in Figure 1, the workflow moves from dataset preparation and exploratory profiling (A) to model development and training (B), then to explanation and reference-based safety modules (C). Component (D) describes runtime inference behavior, while component (E) documents implementation and reproducibility controls. This structure aligns offline experimentation with deployment-time decision safeguards.

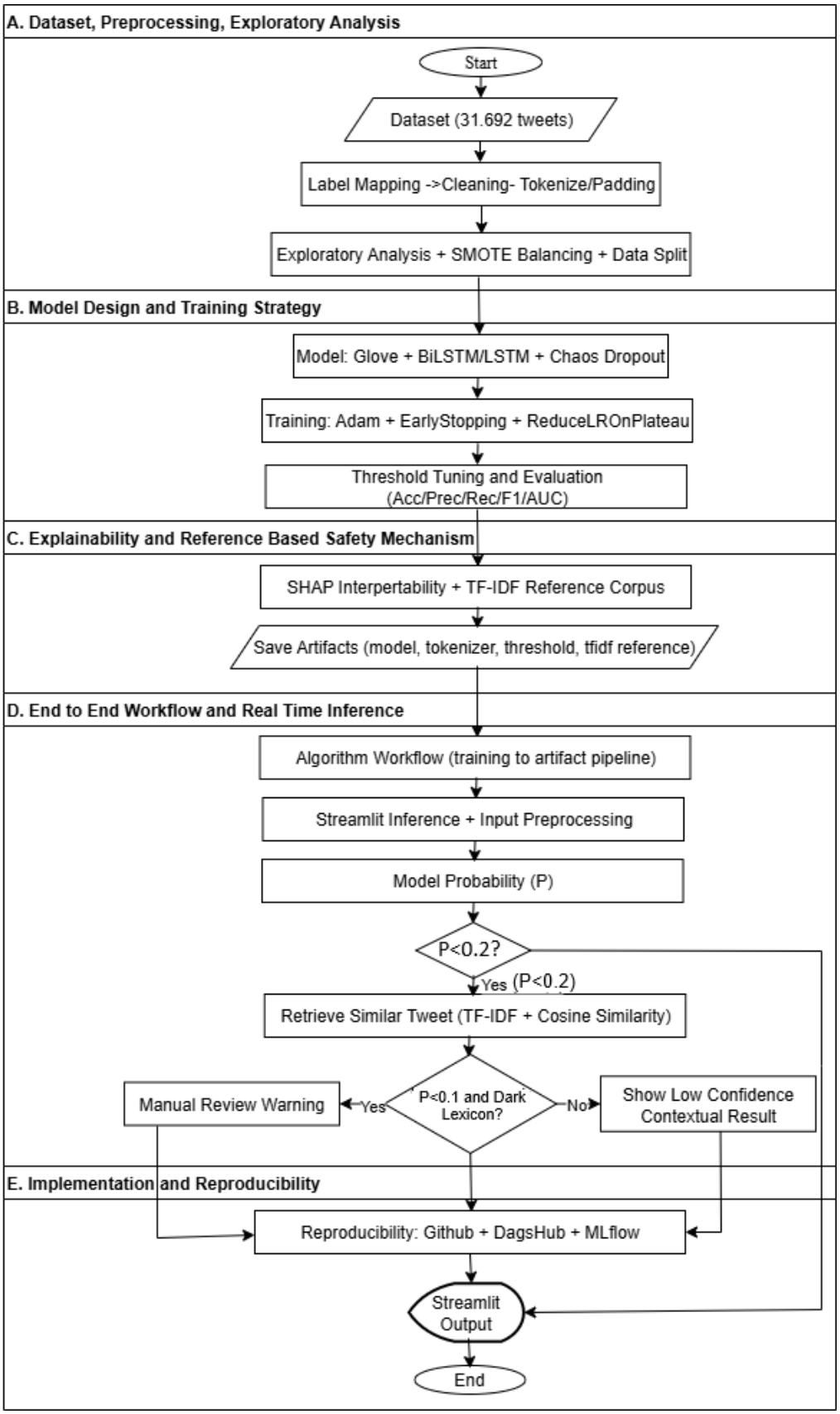


Figure 1. Research methodology flowchart.

3.1. Dataset, Preprocessing, and Exploratory Analysis

This study employs a large annotated Twitter dataset containing more than 31,000 English tweets with the labels neutral, hate speech, and offensive as describe in Table 1. We mapped

hate/offensive into a proxy risk class and neutral into a non-risk class. This mapping is a research-design choice for proxy screening only. It is not a diagnostic label transformation, and the risk-proxy class should be interpreted as hostile or offensive language that may share surface-level affective cues with psychological-risk discourse. Our preprocessing pipeline normalized whitespace and converted text to lowercase, removed noise (URLs, mentions, hashtags, numbers, and non-letter characters), applied Keras tokenization (20,000-vocabulary with OOV token), padded sequences to 50 tokens, and then addressed class imbalance.

A TF-IDF reference set of 200 tweets (100 risk-proxy and 100 non-risk) was prepared to support similarity-based fallback during uncertain predictions. We also used a curated dark-expression lexicon (e.g., "I want to die", "suicide") as an additional safety layer. The data source is the Kaggle Twitter Depression Dataset (https://www.kaggle.com, accessed on 1 February 2026). In our pipeline, `train.csv` contains 31,962 tweets organized into three fields (`id`, `label`, `tweet`), and `test.csv` contains 17,197 tweets with `id` and `tweet`. Before resampling, the training distribution was 29,720 instances for label `0` and 2242 for label `1`, confirming severe imbalance.

Table 1. Sample records from the dataset used in this study.

ID	Label	Tweet (Excerpt)
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction.
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx.
14	1	@user #cnn calls #michigan middle school 'build the wall' chant " #tcot
15	1	no comment! in #australia #opkillingbay #seashepherd #helpcovedolphins #thecove
35	1	it's unbelievable that in the 21st century we'd need something like this again. #neverump #xenophobia

To characterize the pre-training condition, we analyzed class imbalance, tweet-length behavior, and lexical sparsity. The original distribution was strongly skewed (29,720 neutral vs. 2242 risk-proxy), which can bias fitting and reduce minority sensitivity. The before/after profiles are summarized in Figure 2. Tweet lengths covered a wide range with a small set of extreme outliers. We detected 9 abnormal-length tweets, reflecting residual noise. Frequency inspection also identified 45,879 singleton terms, confirming high lexical sparsity typical of social-media corpora.

To mitigate imbalance, we applied SMOTE with an explicit leakage-control caveat. In a leakage-safe pipeline, the train/validation split must be performed before oversampling, and SMOTE must be fitted only on the training partition. If synthetic minority instances are generated before splitting, near-neighbor synthetic samples can cross the train/validation boundary and inflate validation performance. Therefore, the reported validation scores should be read as pipeline evidence rather than as definitive leakage-audited generalization results unless the split-before-SMOTE condition is verified in the experiment script. The resulting balanced class counts became symmetric ([29,720, 29,720]) as presented in Figure 2. Importantly, the non-padding token-length profile remained comparable after resampling, suggesting that balancing did not distort sequence-length characteristics. Overall, these checks indicate a transition from noisy, skewed data to a more trainable distribution.

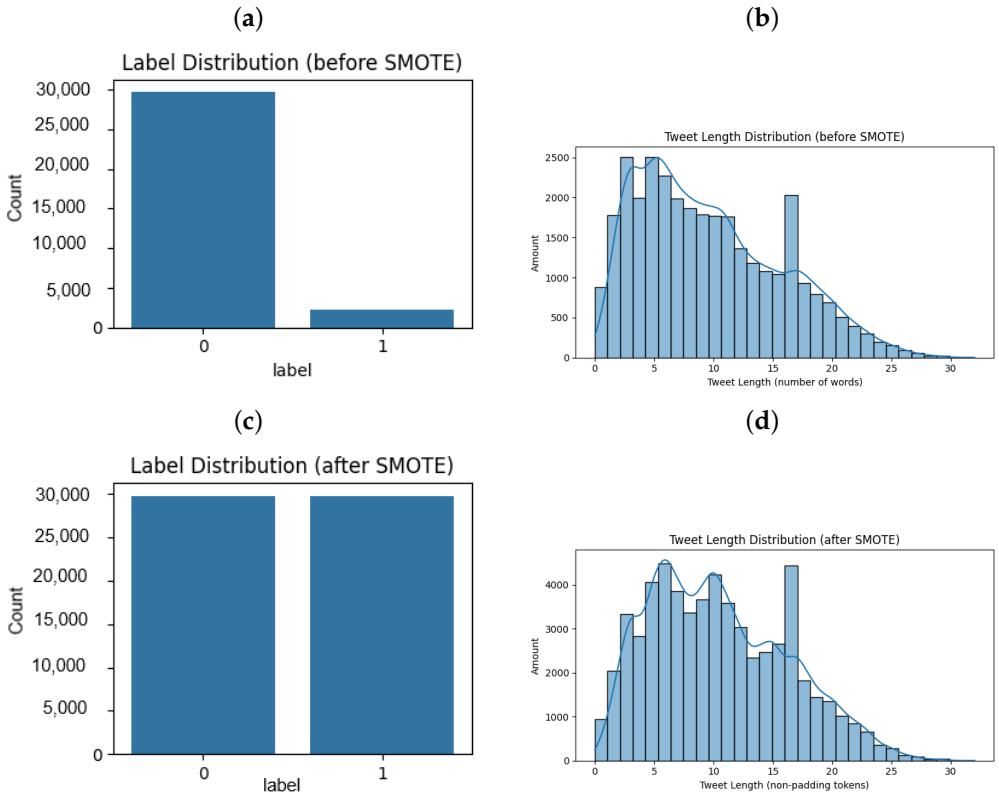


Figure 2. Dataset profiles before and after SMOTE: (**a**) label distribution before SMOTE; (**b**) tweet length before SMOTE; (**c**) label distribution after SMOTE; (**d**) tweet length after SMOTE.

Figure 2 indicates that SMOTE effectively corrected severe class imbalance while preserving the global tweet-length profile. However, because SMOTE interpolates between padded token-index sequences, the resulting synthetic vectors should be interpreted as numerical training aids rather than linguistically meaningful new tweets. Safer alternatives for future experiments include class weighting, focal loss, stratified sampling, back-translation or lexical augmentation before tokenization, and interpolation in embedding space rather than in discrete token-index space.

3.2. Model Design and Training Strategy

The chaos-based dropout sequence is generated using the logistic map equation:

$$x_{n+1} = r x_n (1 - x_n), \tag{1}$$

where r is the chaos parameter (e.g., 3.9) and x_0 is the initial value. The dropout rate is clipped to the range [0.2, 0.5].

SMOTE addresses class imbalance by constructing synthetic minority examples rather than duplicating existing observations [8–10]. In this study, SMOTE is used pragmatically as an imbalance treatment for fixed-length numerical sequence representations. This use has an important limitation: interpolation between token-index vectors does not guarantee that the resulting vector corresponds to a linguistically valid tweet. Therefore, SMOTE should be interpreted as a training regularization/balancing device, not as semantic text generation. A synthetic instance is constructed through linear interpolation between a minority sample x_i and one of its k-nearest minority neighbors x_{nn}:

$$x_{\text{new}} = x_i + \lambda (x_{nn} - x_i), \tag{2}$$

where λ is a random value in [0, 1].

Our architecture includes a trainable 100d GloVe embedding layer, a 64-unit BiLSTM (with dropout and recurrent dropout), a 32-unit LSTM, BatchNormalization, chaos-based dropout, a 32-unit ReLU dense block with additional chaos dropout, and a final sigmoid classifier (Dense(1)).

Training used the Adam optimizer with binary cross-entropy loss, EarlyStopping, and ReduceLROnPlateau callbacks, and evaluation metrics including accuracy, precision, recall, F1-score, and AUC. The best threshold was selected by maximizing the F1-score and saved as `threshold.txt`. The model and tokenizer were saved immediately after training, and the TF-IDF reference corpus was saved for application use.

SHAP (SHapley Additive exPlanations) is employed to interpret model predictions and quantify the contribution of individual features [14,15,21]. For a feature i, the SHAP value is defined as

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Big(f_{S \cup \{i\}}(x) - f_S(x) \Big), \tag{3}$$

where F denotes the complete feature set, S represents a subset of features excluding i, and $f_S(x)$ is the model output conditioned on the feature subset S.

3.3. Explainability and Reference-Based Safety Mechanism

To improve safety behavior and interpretability under uncertainty, we implemented TF-IDF similarity retrieval. After preprocessing, we built a balanced reference bank (100 risk-proxy, 100 non-risk) from training data and stored it for runtime use. Using `TfidfVectorizer`, each incoming tweet is projected into the same feature space; when model confidence is low (e.g., <0.2), cosine similarity is used to retrieve the nearest reference sample and display its label/score context. For very low confidence (<0.1) combined with dark-expression matches, the interface escalates to manual-review warning while retaining reference output.

TF-IDF for term t in document d is computed as

$$\text{tfidf}(t, d) = \text{tf}(t, d) \cdot \log\left(\frac{N}{\text{df}(t)} \right), \tag{4}$$

where $\text{tf}(t, d)$ is term frequency in document d, $\text{df}(t)$ is the number of documents containing t, and N is the total number of documents.

Cosine similarity between a query vector q and a reference vector r is computed as

$$\cos_{\text{sim}}(\mathrm{q}, \mathrm{r}) = \frac{\mathrm{q} \cdot \mathrm{r}}{\|\mathrm{q}\| \ \|\mathrm{r}\|}. \tag{5}$$

3.4. End-to-End Workflow and Prototype Inference

Training Workflow and Artifact Preparation is described as follow:

1. Initialization: Import libraries, define artifact paths, and utility functions (`clean_text`, `load_glove`, `chaos_dropout_sequence`).
2. Load and preprocess data: Load `train.csv` and clean tweet text.
3. Tokenization and save: Initialize Keras Tokenizer, fit on text data, then save tokenizer to `artifacts/tokenizer.pkl`.
4. Padding and split: Convert text to sequences, pad, and divide the original data into training and validation partitions.
5. Training-only SMOTE: Fit and apply SMOTE only on the training partition to reduce information-leakage risk; keep the validation partition untouched.
6. Prepare embedding: Load GloVe embeddings into the embedding matrix.

7. Build and compile model: Build the BiLSTM architecture and compile with Adam optimizer.
8. Train model: Train using training and validation data, with EarlyStopping and ReduceLROnPlateau callbacks.
9. Compute and save threshold: Predict probabilities on validation data, compute F1-score for various thresholds, find the best threshold (`best_t`), and save it to `artifacts/threshold.txt`.
10. Save model: Save the trained Keras model to `artifacts/mental_health_model.keras`.
11. Save reference corpus: Create a balanced TF-IDF reference corpus and save it to `artifacts/tfidf_reference.csv`.
12. Evaluation and interpretation: Generate metric visualizations, perform SHAP analysis, and save results as artifacts.

The Streamlit interface uses the saved threshold (`artifacts/threshold.txt`) to maintain consistent decision boundaries across runs. High-risk phrase flags are triggered independently of classifier confidence. When probability drops below 0.2, the system retrieves the nearest TF-IDF reference and exposes its label/similarity as contextual evidence; if probability falls below 0.1 and dark-expression matches are present, a manual-review warning is raised. The uncertainty cutoffs (0.2 and 0.1) were selected from validation confidence inspection to balance false reassurance risk against alert sensitivity. Because the study does not include runtime benchmarking, the deployment evidence should be understood as a functional prototype demonstration rather than a fully benchmarked real-time system. Latency, memory footprint, throughput, hardware configuration, and stress testing remain required before making a strong operational real-time claim.

3.5. Implementation and Reproducibility

The complete workflow, from preprocessing to evaluation, is managed through GitHub for versioning and collaborative development. Data artifacts, trained models, and experiment outputs are organized in DagsHub, while MLflow records metrics and run histories. These resources are shared to support reproducibility and transparent reporting; repository and experiment links are listed in the Appendix A.

The proposed research framework is shown in Figure 3.

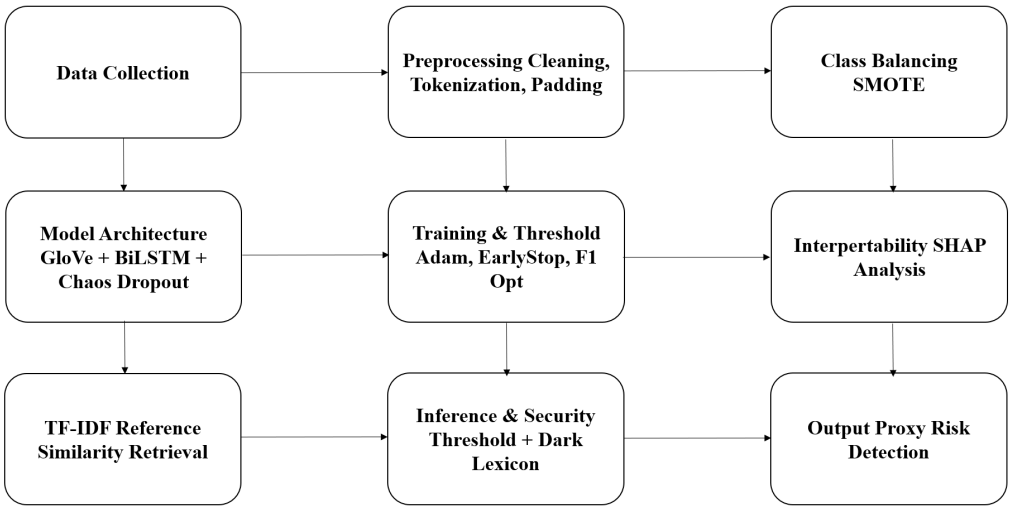


Figure 3. Proposed research framework.

4. Results

After resampling, the label distribution used for model fitting became balanced, which reduced prior bias toward the majority class. The best operating threshold was 0.37 (loaded from `threshold.txt`), yielding an F1-score of 0.86, accuracy 0.79, and AUC 0.93 in the saved classification report summarized in Table 2. These values should be interpreted as validation results from the reported pipeline artifact, not as clinical diagnosis accuracy or as a leakage-audited external test result. To avoid ambiguity, Table 2 is treated as the primary numeric source for the reported metrics; auxiliary visual artifacts are used only to illustrate confusion patterns, threshold behavior, and explanation outputs.

Table 2. Classification results.

Class	Precision	Recall	F1-Score	Support
0 (Non-risk/neutral)	0.88	0.81	0.85	5892
1 (Risk proxy: hate/offensive)	0.83	0.90	0.86	5996
Accuracy	–	–	0.79	11,888
Macro avg	0.86	0.85	0.85	11,888
Weighted avg	0.86	0.85	0.85	11,888
AUC	–	–	0.93	–

Figure 4a shows that correct assignments (TP and TN) dominate error counts, indicating a stable confusion pattern. Figure 4b demonstrates that predictions are linked to lexical contributors rather than completely opaque model behavior. However, the current SHAP visualization remains global and generic; it should be complemented in future work with token-level local explanations for true positives, false positives, and high-risk false negatives. Figure 4c,d further illustrate separability and precision–recall tradeoff across thresholds, in line with the validation metrics reported in Table 2. A low value of a feature is represented by blue, and the baseline is the dash or center line, which is 0 on the x-axis. The projected outcome is decreased by blue dots to the left of the dash and increased by blue dots to the right.

4.1. Qualitative Error Analysis

Without rerunning the model, the available artifact-level inspection suggests three main error risks. First, explicit self-harm phrases can be missed when their lexical pattern is weakly represented in the proxy-label training distribution. Second, hostile or offensive wording can be over-associated with risk even when the post reflects anger, sarcasm, or political conflict rather than psychological distress. Third, short posts with limited context create uncertainty because the model receives too little sequential evidence for reliable discrimination. These patterns reinforce that the system should be used only as a conservative proxy-screening aid with human review, not as an autonomous mental-health decision tool.

4.2. Deployment Results and Safety Behavior

The deployed Streamlit interface combines dark-expression detection with TF-IDF reference retrieval. Under uncertainty, users receive the nearest reference tweet plus label/similarity evidence for contextual interpretation. High-risk expressions are flagged even when classifier confidence is weak, adding a conservative prevention layer [6,13]. This interface demonstrates functional feasibility, but it does not yet quantify latency, memory use, throughput, or stress-test behavior.

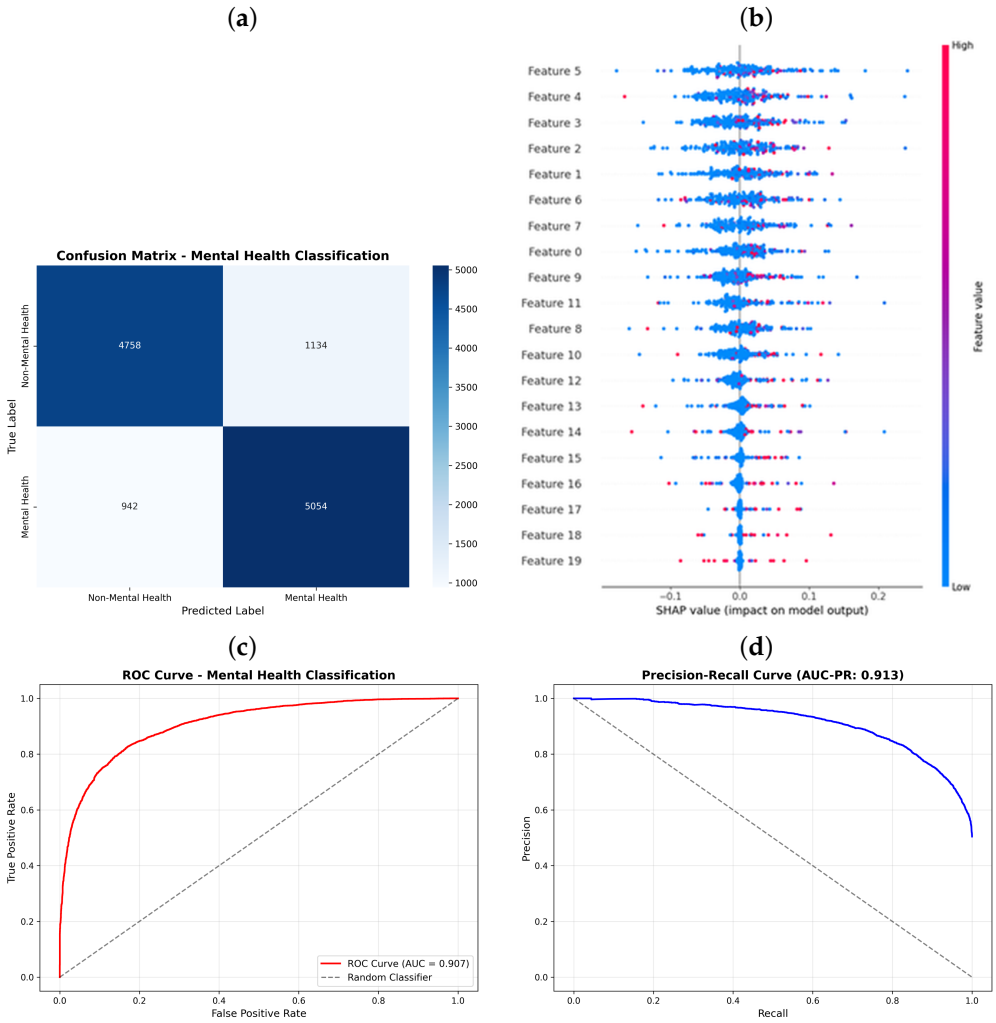


Figure 4. Evaluation and interpretability outputs: (**a**) confusion matrix; (**b**) SHAP summary plot; (**c**) ROC curve; (**d**) precision–recall curve.

5. Discussion

This discussion interprets the empirical results with emphasis on practical strengths and deployment limitations. We focus on four dimensions: predictive performance, interpretability, uncertainty-aware safety behavior, and external validity boundaries that inform responsible future use.

5.1. The Strength of the Study

The framework provides a practical balance between predictive performance, implementation simplicity, and interpretability. With an F1-score of 0.86 and AUC of 0.93, the BiLSTM configuration shows promising behavior for noisy social text. However, because no BERT, RoBERTa, DistilBERT, or lightweight transformer baseline was run under the same dataset split, thresholding rule, and deployment constraints, this study cannot claim empirical superiority or lower operational cost relative to transformer models. Our goal is therefore framed more conservatively: deployable and explainable proxy risk screening under constrained resources [11–13].

Another strength lies in explainable and safety-aware operation. SHAP supports attribution analysis, while TF-IDF fallback injects contextual references when confidence is low. The deployed interface adds runtime safeguards through threshold loading, dark-expression flags, and manual-review escalation for extreme uncertainty. Together, these features improve transparency in ambiguous cases. Because the available files contain

artifact-level outputs rather than validation-level counterfactual predictions, Table 3 reports component-level evidence rather than claiming a full numeric ablation. The operational tradeoff with transformer alternatives is summarized conceptually in Table 4, and deployment snapshots are shown in Figure 5.

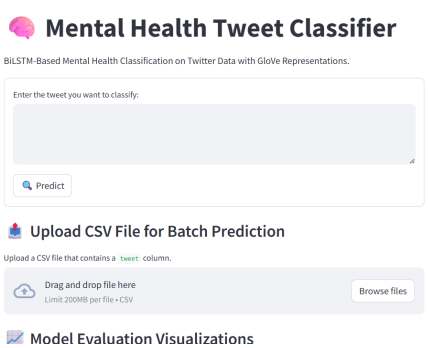
Table 3. Artifact-based component evidence for safety and uncertainty modules.

Component	Artifact Evidence	Contribution Beyond Classifier	Remaining Limitation
BiLSTM classifier	Confusion matrix, ROC curve, precision–recall curve, and metric summary	Provides the base proxy-risk decision boundary and validation performance baseline	Does not explain or contextualize uncertain predictions by itself
BiLSTM + SHAP	SHAP summary plot	Adds attribution evidence showing which lexical features influence model output	Current evidence is global; local case-level explanations remain future work
BiLSTM + uncertainty threshold	Saved decision-threshold description and Streamlit decision behavior	Separates ordinary prediction from low-confidence cases that require additional support	Coverage and rejection-rate statistics were not saved as validation artifacts
BiLSTM + TF-IDF retrieval	Balanced reference-bank design and Streamlit output screenshot	Supplies nearest-reference context for low-confidence inputs, supporting human interpretation	Incremental effect on F1 or false-negative reduction was not isolated
Full safety framework	Combined classifier, SHAP, thresholding, lexicon check, TF-IDF retrieval, and deployment screenshots	Integrates prediction, explanation, uncertainty handling, reference support, and conservative escalation	A full counterfactual ablation requires rerunning or saved validation-level predictions

Table 4. Conceptual operational tradeoff summary of BiLSTM + Chaos + SHAP + TF-IDF vs. transformer models.

Operational Aspect	BiLSTM + Chaos + SHAP + TF-IDF	BERT/RoBERTa/DistilBERT
Predictive score (F1)	Observed validation F1 = 0.86 in this study	Often strong in prior studies, but not evaluated here
Inference efficiency	Expected to be lightweight, not benchmarked here	Often higher computational cost, not benchmarked here
Explainability readiness	Implemented through SHAP and TF-IDF reference support	Typically requires task-specific post hoc analysis
Safety fallback integration	Implemented through threshold, lexicon, and reference retrieval	Usually requires an additional pipeline
Deployment readiness	Functional prototype demonstrated	Requires separate implementation and benchmarking

(**a**) Streamlit input interface



(**b**) Streamlit classification result

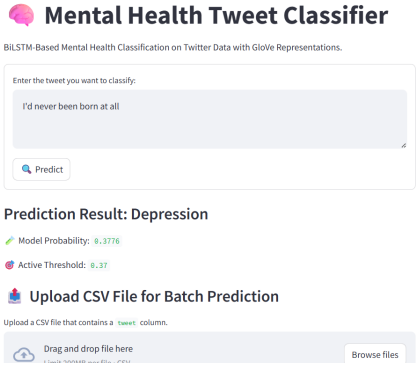


Figure 5. Streamlit deployment views: (**a**) input interface; (**b**) classification result with safety support.

Figure 5a illustrates the input interface used for tweet submission in proxy risk screening. Figure 5b shows the corresponding output view, including class prediction, confidence-aware decision behavior, and supporting safety evidence from rule-based flags or similarity references during uncertain inference.

5.2. The Limitation of the Study

Despite these strengths, testing also exposed limitations. Some explicit self-harm statements (e.g., "I want to die") can remain under-detected, suggesting that the learned boundary is partly constrained by proxy-label semantics. Although critical terms exist in the vocabulary, nuanced contextual sensitivity is still lower than what may be expected from contextual transformer encoders. Additionally, because labels originate from hate/offensive categories, outputs must remain interpreted as proxy risk screening rather than clinical depression diagnosis. This caveat is part of the core research design, not merely a minor limitation.

Several empirical gaps remain. First, the study includes artifact-based component evidence for the uncertainty threshold, TF-IDF retrieval module, lexicon rule layer, and SHAP explanation, but it does not yet provide a rerun ablation that separately removes each module under identical validation conditions. Consequently, the current results demonstrate the complete pipeline and its operational support modules but do not quantify the marginal F1, AUC, false-negative, or rejection-rate contribution of each component. Second, the comparison with transformer approaches is conceptual rather than experimental. Third, if any experiment artifact was produced using SMOTE before train/validation splitting, the resulting validation scores may be optimistic; a leakage-audited rerun should split first and apply resampling only within the training fold. Fourth, the SHAP evidence remains mostly global and should be expanded with local token-level explanations for true positives, false positives, false negatives, and high-risk missed cases.

From an external-validity standpoint, generalization to new slang, multilingual contexts, and cross-platform discourse remains limited. Ethical deployment further requires strict human-in-the-loop governance, explicit escalation protocols, and privacy protection. Future work should prioritize richer mental-health labels, leakage-safe resampling, class weighting or focal loss alternatives to token-index SMOTE, transformer/hybrid baselines under identical deployment constraints, targeted augmentation for high-risk language, latency and memory benchmarking, and active learning around uncertain predictions to improve sensitivity and reliability [3,22,24].

6. Conclusions

This study presents an application-specific BiLSTM-based proxy risk framework that combines chaos dropout, SMOTE balancing, SHAP explanation, and TF-IDF fallback for noisy social text. The results show promising screening performance with reference-supported outputs under uncertainty. However, the task is limited by the use of hate/offensive labels as proxy indicators, so the system should not be interpreted as a depression, self-harm, or clinical mental-health detector. The current evidence provides artifact-based support for the safety and uncertainty modules, but it does not yet replace rerun ablation, transformer baselines, or latency and resource benchmarking. Overall, the work contributes an explainable and deployment-oriented proxy-screening pipeline, while future validation should use clinically grounded labels, leakage-safe resampling, ablation studies, and controlled baseline comparisons.

Author Contributions: A.B.S.: Conceptualization, methodology, software, writing–original draft; H.Y.R.: data curation, writing–review and editing, supervision, correspondence; T.R.: investigation, data curation; H.A.P.: investigation, data curation; R.A.A.: editing, data curation; A.B.F.M.: review and editing; M.A.: review and editing; A.H.B.: review, editing and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the Deanship of Scientific Research (DSR) at King Abdullaziz University, Jeddah, Saudi Arabia under grant no. (IPP:1358-830-2025).

Data Availability Statement: The dataset used in this study is publicly available from Kaggle at https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset, accessed on 15 November 2025. The source code and implementation details are accessible via GitHub at https://github.com/AkasSakti/MentalheaLth, accessed on 7 June 2026. Experiment tracking artifacts and model runs are provided through DagsHub and MLflow at https://dagshub.com/AkasSakti/MentalheaLth, accessed on 7 June 2026 and https://dagshub.com/AkasSakti/MentalheaLth.mlflow/#/experiments/0/runs/eea54d8e183848b1a7c0175d80c70e42, accessed on 7 June 2026. All resources are publicly accessible without restriction. No personally identifiable information is distributed beyond what is already contained in the original public dataset.

Acknowledgments: This project was funded by the Deanship of Scientific Research (DSR) at King Abdullaziz University, Jeddah, Saudi Arabia under grant no. (IPP:1358-830-2025). The authors, therefore, acknowledge with thanks DSR for technical and financial support.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the curve
BiLSTM	Bidirectional long short-term memory
MLflow	Machine Learning Flow
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
TF-IDF	Term frequency–inverse document frequency

Appendix A. Repository and Author Identifiers

The public implementation repository is available at https://github.com/AkasSakti/MentalheaLth, accessed on 5 May 2026. Experiment tracking is available through DagsHub at https://dagshub.com/AkasSakti/MentalheaLth, accessed on 5 May 2026.

References

1. Chancellor, S.; Choudhury, M. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digit. Med.* **2020**, *3*, 43. [CrossRef] [PubMed]
2. Aldkheel, A.; Zhou, L. Depression detection on social media: A classification framework and research challenges and opportunities. *J. Health Inform. Res.* **2023**, *8*, 88–120. [CrossRef] [PubMed]
3. Ramirez-Cifuentes, D.; Freire, A.; Baeza-Yates, R.; Puntí, J.; Medina-Bravo, P.; Velazquez, D.A.; Gonfaus, J.M.; Gonzàlez, J. Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. *J. Med. Internet Res.* **2020**, *22*, e17758. [CrossRef] [PubMed]
4. Jamali, A.A.; Berger, C.; Spiteri, R.J. Momentary depressive feeling detection using X (formerly Twitter) data: Contextual language approach. *JMIR AI* **2023**, *2*, e49531. [CrossRef] [PubMed]
5. Kim, J.; Lee, J.; Park, E.; Han, J. A deep learning model for detecting mental illness from user content on social media. *Sci. Rep.* **2020**, *10*, 11846. [CrossRef] [PubMed]
6. Kabir, M.; Ahmed, T.; Hasan, B.; Laskar, T.R.; Joarder, T.K.; Mahmud, H.; Hasan, K. DEPTWEET: A typology for social media texts to detect depression severities. *Comput. Hum. Behav.* **2023**, *139*, 107503. [CrossRef]
7. Baydili, I.; Tasci, B.; Tasci, G. Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection. *Behav. Sci.* **2025**, *15*, 352. [CrossRef] [PubMed]

8. Wongvorachan, T.; He, S.; Bulut, O. A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information* **2023**, *14*, 54. [CrossRef]
9. Kim, M.; Hwang, K.-B. An empirical evaluation of sampling methods for the classification of imbalanced data. *PLoS ONE* **2022**, *17*, e0271260. [CrossRef] [PubMed]
10. Yang, W.; Pan, C.; Zhang, Y. An oversampling method for imbalanced data based on spatial distribution of minority samples SD-KMSMOTE. *Sci. Rep.* **2022**, *12*, 16820. [CrossRef] [PubMed]
11. Bokolo, B.G.; Liu, Q. Deep learning-based depression detection from social media: Comparative evaluation of ML and transformer techniques. *Electronics* **2023**, *12*, 4396. [CrossRef]
12. Kerasiotis, M.; Ilias, L.; Askounis, D. Depression detection in social media posts using transformer-based models and auxiliary features. *Soc. Netw. Anal. Min.* **2024**, *14*, 196. [CrossRef]
13. Qasim, A.; Mehak, G.; Hussain, N.; Gelbukh, A.; Sidorov, G. Detection of depression severity in social media text using transformer-based models. *Information* **2025**, *16*, 114. [CrossRef]
14. Kokalj, E.; Skrlj, B.; Lavrac, N.; Pollak, S.; Robnik-Sikonja, M. BERT meets Shapley: Extending SHAP explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 16–21.
15. Sadeghi, Z.; Alizadehsani, R.; Cifci, M.A.; Kausar, S.; Rehman, R.; Mahanta, P.; Bora, P.K.; Almasri, A.; Alkhawaldeh, R.S.; Hussain, S.; et al. A review of explainable artificial intelligence in healthcare. *Comput. Electr. Eng.* **2024**, *118*, 109370. [CrossRef]
16. Marx, C.; Park, Y.; Hasson, H.; Wang, Y.; Ermon, S.; Huan, L. But are you sure? An uncertainty-aware perspective on explainable AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*; PMLR: Cambridge, MA, USA, 2023.
17. Thuy, A.; Benoit, D.F. Explainability through uncertainty: Trustworthy decision-making with neural networks. *arXiv* **2024**, arXiv:2403.10168.
18. Sarsam, S.M.; Al-Samarraie, H.; Alzahrani, A.I.; Alnumay, W.; Smith, A.P. A lexicon-based approach to detecting suicide-related messages on Twitter. *Biomed. Signal Process. Control* **2021**, *65*, 102355. [CrossRef]
19. Liang, M.; Niu, T. Research on text classification techniques based on improved TF-IDF algorithm and LSTM inputs. *Procedia Comput. Sci.* **2022**, *208*, 460–470. [CrossRef]
20. Jia, N.; Wang, T. A GCM neural network with piecewise logistic chaotic map. *Symmetry* **2022**, *14*, 506. [CrossRef]
21. Zhao, W.; Joshi, T.; Nair, V.N.; Sudjianto, A. SHAP values for explaining CNN-based text classification models. *arXiv* **2020**, arXiv:2008.11825. [CrossRef]
22. Aldhyani, T.H.H.; Alsubari, S.N.; Alshebami, A.S.; Alkahtani, H.; Ahmed, Z.A.T. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12635. [CrossRef] [PubMed]
23. Vilenchik, D. Simple statistics are sometimes too simple: A case study in social media data. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 402–408.
24. Liu, J.; Shi, M.; Jiang, H. Detecting suicidal ideation in social media: An ensemble method based on feature fusion. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8197. [CrossRef] [PubMed]