International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Psycholinguistically Grounded Explainable Screening for Suicide Risk Assessment in Short-Form Text: A Transparent Feature-Fusion Approach

1st Shobhit Tomar
Dept. of Computer Science and Engineering
Apex Institute of Technology
Chandigarh University
Mohali, India
itstomars21@gmail.com

2nd Simranpreet Kaur
Dept. of Computer Science and Engineering
Apex Institute of Technology
Chandigarh University
Mohali, India
Kaurgurjeet3638@gmail.com

3rd Manav Saxena
Dept. of Computer Science and Engineering
Apex Institute of Technology
Chandigarh University
Mohali, India
manavsaxena585@gmail.com

4th Dr. Raghav Mehra
Dept. of Computer Science and Engineering
Apex Institute of Technology
Chandigarh University
Mohali, India
raghav.mehrain@gmail.com

Abstract *— Automated identification of individuals exhibiting suicide-related language in digital text carries profound clinical and ethical weight. While transformer-based architectures yield impressive accuracy figures on curated benchmarks, their representations remain inaccessible to practitioners who must justify and assume responsibility for clinical decisions. This paper proposes a multi-class stratification architecture grounded in computational psycholinguistics, deliberately constructed around three transparent feature families: lexical salience scores derived through term-frequency weighting, hand-crafted cognitive-behavioural markers anchored in established clinical theory regarding suicidal cognition, and affective polarity measurements extracted via two complementary rule-based sentiment tools. A controlled feasibility corpus of 200 annotated samples—distributed across Low, Medium, and High risk strata—serves as the experimental substrate. Three inherently interpretable classifiers are benchmarked under stratified five-fold cross-validation with recall accorded priority weighting to reflect the asymmetric consequences of undetected high-risk instances. Linear SVM achieves the strongest aggregate performance (weighted F1 = 0.7156, recall = 0.7250, precision = 0.7216), while Logistic Regression exhibits the lowest fold-to-fold variance. Prediction transparency is demonstrated through coefficient ranking and LIME-based local feature attribution, enabling practitioners to audit the evidence underlying each classification. The framework is explicitly positioned as a pre-diagnostic screening signal requiring mandatory qualified-professional review prior to any action.*

Index Terms *— computational psycholinguistics, explainable AI, suicide risk stratification, feature fusion, TF-IDF, VADER, cognitive-behavioural NLP, LIME, interpretable machine learning, mental health triage.*

I. INTRODUCTION

Pervasive digitisation of personal communication has created a previously unavailable window into the psychological states of individuals at scale. Self-authored content on social platforms, community forums, and wellness applications increasingly surfaces language co-varying with clinically significant mental health deterioration. Among such signals, those associated with suicidal ideation occupy a particularly consequential position. The World Health Organization estimates that approximately 700,000 lives are claimed by suicide annually, with persons aged 15–29 disproportionately represented [1]. Detecting early-stage risk indicators embedded in natural language—before escalation toward crisis—has consequently attracted substantial research investment within computational psychiatry.

Dominant methodological trends in this space have converged on high-capacity neural architectures. Large pretrained contextual models such as BERT [2] and its successors consistently outperform earlier approaches on established benchmarks. Yet their predictive power emerges from parameter spaces of enormous dimensionality, offering practitioners no interpretable account of why a given text received a particular risk label. In a domain involving potential clinical action, opacity of this kind is not a peripheral concern but a structural barrier to responsible deployment. Clinicians, social workers, and crisis counsellors cannot ethically delegate consequential judgements to mechanisms they cannot interrogate.

The framework described here takes a principled stance against that opacity. Rather than pursuing benchmark optimisation, it constructs a pipeline in which every component—from feature extraction to classifier coefficient—is directly examinable. Three families of psychologically motivated features are extracted independently and fused into a

unified representation before presentation to classical interpretable classifiers. The design prioritises auditable transparency, ethical restraint, and realistic performance reporting over inflated accuracy claims.

The primary contributions of this investigation are fourfold. **First,** a psycholinguistically grounded feature fusion architecture is designed, implemented, and validated for three-class suicide risk stratification. **Second,** three fully interpretable classifiers are benchmarked under identical stratified cross-validation conditions. **Third,** a recall-prioritised evaluation protocol is applied, reflecting asymmetric error consequences inherent to mental health screening contexts. **Fourth,** LIME-based local attribution is integrated to provide instance-level prediction explanations accessible to non-technical practitioners.

II. RELATED WORK

Computational analysis of mental health language has accumulated roughly two decades of scholarship. Coppersmith and colleagues [3] established early methodological foundations by demonstrating that statistical language models could quantify systematic lexical divergences between individuals with diagnosed mood disorders and general social media populations. Their work substantiated the feasibility of NLP-based screening and catalysed a productive wave of subsequent investigation.

The representational shift from hand-crafted features to learned distributed embeddings dramatically altered the landscape. Kim [4] showed that shallow convolutional architectures applied to word embedding matrices could match or exceed classical baselines on sentence classification—an approach subsequently adapted for short mental health texts. The arrival of bidirectional contextual representations through BERT [2] produced another step-change: contextual pretraining at scale yielded near-universal improvements across NLP classification tasks, including those involving distress-related content.

Ji and colleagues [6] conducted a systematic survey of deep learning applications in computational mental health, concluding that while accuracy metrics had advanced substantially, interpretability deficits and ethical deployment considerations remained pervasively unaddressed across the field. Zirikly et al. [7] applied classical structured feature sets to suicide risk classification using data from the CLPsych 2019 shared task, demonstrating that carefully engineered non-neural representations could achieve competitive triage performance. The VADER sentiment lexicon [8], constructed specifically for social media discourse, provides the affective component of our pipeline owing to its transparent rule-based design. Beck's cognitive model of suicidal thinking [9] supplies the theoretical grounding for the psycholinguistic features.

LIME [10], introduced by Ribeiro and colleagues, addresses interpretability at the instance level by constructing locally faithful linear approximations of any black-box classifier. Its application here enables per-prediction feature attribution that is meaningful to practitioners regardless of the underlying model architecture.

TABLE I. POSITIONING OF PROPOSED WORK AGAINST SELECTED PRIOR STUDIES

Study	Architecture	Interpretable	Primary Metric
Coppersmith et al. [3]	Stat. Language Models	Partial	AUC / Accuracy
Kim [4]	Convolutional Neural Network	No	Macro F1-Score
Devlin et al. [5]	BERT Transformer	No	Weighted F1-Score
Ji et al. [6]	Survey (varied architectures)	Mixed	Narrative Review
Zirikly et al. [7]	SVM + Linguistic Features	Partial	Precision / Recall
This Work	Classical ML + Feature Fusion	Yes (Full)	Recall + Weighted F1

III. DATASET DESCRIPTION

Evaluating a risk-stratification pipeline requires annotated material spanning the intended label space in a controlled and reproducible manner. Crawling live social media data—while methodologically common—raises unresolved concerns around informed consent, platform terms-of-service compliance, and the potential for personally identifiable content. In alignment with the ethical position articulated in Section VI, and consistent with feasibility-stage research

practice, we constructed a purpose-built annotation corpus of 200 short-form text samples designed to represent realistic self-expression patterns across the three target strata.

A. Corpus Composition and Stratum Definitions

Three risk categories were operationalised as follows. Low Risk (Class 0, n = 60) encompasses neutral-to-positive emotional expression characterised by engagement with routine activities, social connection, and forward-looking statements. Medium Risk (Class 1, n = 60) captures elevated psychological distress—including anhedonia, self-doubt, social withdrawal, and somatic complaints—without the presence of explicit suicidal markers. High Risk (Class 2, n = 80) encodes explicit suicidal ideation, strong signals of perceived entrapment or burdensomeness, expressions of finality, and self-harm references.

The deliberate overrepresentation of High-risk instances by a margin of 33% relative to each other stratum reflects the asymmetric operational priority of the framework: missing a genuine high-risk case is a qualitatively worse error than generating an unnecessary referral. This imbalance is propagated through stratified cross-validation, ensuring that each fold preserves the overall class distribution.

B. Annotation Rationale and Scope

Sample labels reflect the presence or absence of documented psycholinguistic risk signals—not clinical diagnoses. The annotation schema aligns explicitly with the non-diagnostic intention of the system. High-risk texts were authored to contain marker language consistent with Beck's cognitive triad of hopelessness [9], Joiner's interpersonal theory of suicide (burdensomeness, thwarted belonging), and explicit ideation. Medium-risk texts replicate the language of serious affective distress without crossing into ideation. This distinction, while theoretically motivated, is acknowledged to be genuinely ambiguous at its boundary—a property reflected in Medium-class classification difficulty discussed in Section V.

C. Scope Limitations and Future Corpus Development

The current corpus is explicitly a controlled feasibility benchmark. It does not claim distributional equivalence with any clinical population, social media platform, or real-world prevalence distribution, and no epidemiological inferences are drawn from model outputs. Extending this pipeline to production-grade screening requires replication on ethically sourced real-world corpora such as CLPsych shared task datasets [7], UMD Reddit Mental Health Collection, or consented clinical text—work deferred to subsequent phases with appropriate institutional ethics oversight.

D. Ethical Safeguards in Corpus Construction

No real individuals were referenced in any sample. No personally identifiable material was created or stored. All samples are original synthetic constructions authored to reflect documented linguistic patterns associated with each risk category. The corpus carries no clinical validity and is unsuitable for direct clinical application without extensive independent validation.

IV. METHODOLOGY

The proposed architecture comprises five sequentially coupled stages: text preprocessing, parallel feature extraction across three psychologically motivated modules, feature-space fusion, classifier training under cross-validated evaluation, and instance-level attribution via LIME. Fig. 1 presents a schematic overview of the pipeline.

Fig. 1. Pipeline schematic: raw text traverses preprocessing, three parallel feature extraction modules, horizontal fusion, classifier training, and LIME attribution.

A. Preprocessing

Input texts are lowercased uniformly before punctuation removal, elimination of special characters, and whitespace normalisation. Tokenisation proceeds at the word boundary level. Standard English stopwords are filtered using the NLTK corpus [11], with the deliberate exception of negation terms (no, not, never, cannot) and first-person pronouns—both of which carry diagnostic significance within the cognitive-behavioural feature module. These preprocessing choices balance surface normalisation against the preservation of psycholinguistically meaningful tokens.

B. *Feature Module 1 — Lexical Salience via TF-IDF*

A Term Frequency-Inverse Document Frequency vectoriser is fitted jointly to the full 200-sample corpus. Vocabulary coverage is capped at 500 terms with sublinear term-frequency scaling applied to reduce the disproportionate influence of highly recurrent words. The n-gram range spans unigrams and bigrams to capture short multi-word risk expressions (e.g., 'give up', 'no future'). The resulting sparse representation forms the principal substrate for classification, encoding the vocabulary most discriminative across risk strata.

C. *Feature Module 2 — Cognitive-Behavioural Psycholinguistic Markers*

Rooted in Beck's cognitive model of suicidal thinking [9] and the broader clinical literature on suicidal cognition, two hand-curated lexicons were constructed. The first targets absolutist and dichotomous language—a cognitive pattern robustly associated with suicide risk in clinical interview research—through 16 marker terms. The second addresses perceived hopelessness and entrapment through 15 multi-word indicator phrases. A third scalar feature quantifies first-person pronoun density as a proxy for pathological self-focus, a finding supported by quantitative psycholinguistic research.

The nine features produced by this module differ qualitatively from TF-IDF dimensions: each is directly mappable to a named psychological construct, enabling practitioners to understand not merely which text segments contributed to a prediction but why those contributions carry clinical significance.

D. *Feature Module 3 — Affective Polarity*

Emotional tone is quantified through two complementary rule-based systems. VADER [8]—constructed specifically for short, informal digital text—yields three continuous scores per sample: negative affect weight, positive affect weight, and a normalised compound polarity index spanning $[-1, +1]$. TextBlob contributes a general polarity estimate and a subjectivity coefficient. These five affective dimensions capture the emotional gradient from baseline positivity through moderate dysphoria to severe negative affect, and interact meaningfully with the cognitive-behavioural features under fusion.

E. *Feature Fusion*

The three module outputs are concatenated horizontally into a unified sparse matrix using scipy's hstack operation. The TF-IDF module contributes 500 dimensions; the combined cognitive-behavioural and affective modules contribute nine, producing a 509-dimensional fused feature space. Affective dimensions containing negative values are clipped to zero prior to fusion to satisfy the non-negativity constraint of Multinomial Naïve Bayes, without altering the feature availability for the remaining two classifiers.

F. *Classifiers*

Three classifiers are selected specifically for their interpretability properties and established effectiveness on sparse high-dimensional text representations. Multinomial Naïve Bayes (Laplace smoothing $\alpha = 0.5$) provides a probabilistic baseline with independently inspectable per-class likelihood ratios. Logistic Regression (regularisation $C = 1.0$, maximum 1,000 iterations) yields calibrated posterior probability estimates alongside directly readable coefficient weights per output class. Linear SVM ($C = 1.0$, maximum 2,000 iterations) optimises a maximum-margin hyperplane in the fused feature space, with decision weights likewise directly attributable to individual feature dimensions. Non-linear kernel methods and ensemble architectures were deliberately excluded to preserve end-to-end pipeline transparency.

G. *Evaluation Protocol*

Stratified five-fold cross-validation is employed throughout, ensuring that the proportional representation of each risk class is preserved within every training and evaluation partition. Weighted precision, recall, and F1-score are computed as aggregate metrics, with recall accorded interpretive priority given the asymmetric consequences of undetected high-risk instances in a screening application. Per-class F1 scores are reported separately to diagnose stratum-level behaviour. Confusion matrices are generated via stacked cross-validated predictions to visualise misclassification directionality.

H. *LIME-Based Local Explanation*

To go beyond aggregate coefficient inspection, LIME [10] is applied to generate instance-level feature attributions. For a given test prediction, LIME constructs a neighbourhood of perturbed samples around the input, fits a locally faithful

linear model within that neighbourhood, and ranks feature contributions to the classification outcome. This enables a practitioner reviewing a High-risk flag to examine precisely which lexical tokens and psycholinguistic features drove the decision—a capability absent from global coefficient tables and entirely unavailable in transformer-based systems without auxiliary explanation infrastructure.

V. RESULTS AND DISCUSSION

A. Aggregate Classifier Performance

Table II presents weighted precision, recall, and F1-score for each classifier across the five cross-validation folds, reported as means with standard deviations. Linear SVM leads on all three metrics, attaining a weighted F1 of 0.7156 with precision 0.7216 and recall 0.7250. Its cross-fold F1 standard deviation (0.056) exceeds that of Logistic Regression (0.027), indicating slightly higher sensitivity to fold composition. Logistic Regression places second (F1 = 0.6941) with notably stable cross-fold behaviour—a property advantageous for deployment contexts requiring consistent output distributions across incoming data batches. Multinomial Naïve Bayes achieves the lowest aggregate performance (F1 = 0.6421), consistent with the conditional independence assumption being partially violated by the correlated linguistic patterns characteristic of psychologically distressed text.

These figures do not approach ceiling performance, which is by experimental design rather than methodological inadequacy. The linguistic boundary separating Medium from High risk is genuinely ambiguous—a system claiming near-perfect discrimination on a 200-sample corpus would almost certainly be overfitting to the controlled annotation conditions rather than capturing a generalisable signal.

TABLE II. CROSS-VALIDATION PERFORMANCE SUMMARY (WEIGHTED METRICS, MEAN ± STD DEV)

Classifier	Weighted Precision	Weighted Recall	Weighted F1
Multinomial Naïve Bayes	0.6905 ± 0.058	0.6800 ± 0.043	0.6421 ± 0.044
Logistic Regression	0.6991 ± 0.033	0.7050 ± 0.025	0.6941 ± 0.027
Linear SVM	0.7216 ± 0.060	0.7250 ± 0.047	0.7156 ± 0.056

B. Recall-Prioritised Sensitivity Analysis

Fig. 2 illustrates recall comparisons across classifiers with error bars representing one standard deviation across folds. Linear SVM achieves a weighted recall of 0.7250, followed by Logistic Regression (0.7050) and Naïve Bayes (0.6800). Across all three models, cross-fold recall variability is moderate—no classifier exhibited a fold in which recall fell below 0.60—indicating stable triage-relevant sensitivity throughout the evaluation. Logistic Regression's consistently low variance makes it a compelling operational choice when deployment consistency is weighted more heavily than marginal peak performance.

Fig. 2. Weighted recall comparison across classifiers under stratified five-fold cross-validation, with error bars showing ±1 SD across folds.

C. Confusion Analysis and Error Directionality

Fig. 3 presents the confusion matrix for Logistic Regression derived from stacked cross-validated predictions. Low-risk samples achieve high classification fidelity, with 56 of 60 correctly assigned and only minor leakage into adjacent strata. High-risk samples are similarly well-classified, with 58–59 of 80 correctly identified. Critically, primary misclassification for High-risk instances occurs toward Medium rather than toward Low—meaning the model errs conservatively, tending to elevate ambiguous cases rather than dismiss them. In a live triage setting, this error directionality prompts additional human review rather than silent omission of a genuine risk case.

The Medium stratum exhibits the highest classification difficulty, with 26–29 samples labelled as High risk across models. This pattern reflects genuine psycholinguistic ambiguity rather than model failure: language expressing severe affective distress and language encoding early suicidal ideation share substantial lexical and affective overlap that no feature set at this scale can fully disentangle. Rather than treating this as a deficit to mask, the framework surfaces it explicitly as a calibration signal for professional reviewers.

Fig. 3. Confusion matrix for Logistic Regression under five-fold cross-validated prediction across all 200 corpus samples.

D. Per-Stratum F1 Analysis

Table III decomposes F1-scores by class across all three classifiers. Low-risk classification achieves consistently high performance across models (0.91–0.94), reflecting the lexical distinctiveness of neutral and positive text relative to distress-bearing language. High-risk classification achieves moderate scores (0.69–0.71), confirming that the psycholinguistic feature set successfully captures both explicit and implicit risk indicators. Medium-risk performance is the most challenging (0.29–0.52), with the Linear SVM substantially outperforming Naïve Bayes on this stratum. The Medium-risk performance gap motivates dedicated feature engineering work targeting the distress-to-ideation boundary in future extensions of this system.

TABLE III. PER-STRATUM F1-SCORE BREAKDOWN ACROSS ALL THREE CLASSIFIERS

Classifier	Low-Risk F1	Medium-Risk F1	High-Risk F1
Multinomial Naïve Bayes	0.91	0.29	0.69
Logistic Regression	0.93	0.44	0.70
Linear SVM	0.94	0.52	0.71

Fig. 4. Per-class F1-score comparison across all three classifiers for Low, Medium, and High risk strata.

E. Classifier Coefficient Transparency

One measurable advantage of the selected architecture over opaque neural alternatives is the direct inspectability of classifier weights. In both Logistic Regression and Linear SVM, each feature dimension carries a learned weight per output class that can be ranked, sorted, and examined without auxiliary tooling. Features receiving the largest positive weights for the High-risk class consistently include absolutist cognitive markers, entrapment expressions, strongly negative VADER compound scores, and elevated self-focus pronoun counts. This weight structure is fully consistent with the clinical psychological literature on suicidal cognition, providing both a validity check on the feature engineering and a basis for practitioner-readable explanation.

F. LIME Attribution Analysis

To demonstrate instance-level explainability, LIME attribution was applied to a representative High-risk test instance. Table IV summarises the five features assigned the highest positive attribution weights for the High-risk class label. Absolutist cognitive terms and hopelessness phrases consistently emerge as dominant contributors, with the VADER compound score and self-focus count providing complementary affective and pronoun-based evidence.

This attribution output can be surfaced directly to a reviewing practitioner alongside the risk flag, enabling them to assess whether the flagged features constitute genuine risk signals in context or artefacts of the specific phrasing. No comparable practitioner-facing explanation is achievable from a fine-tuned BERT model without substantial auxiliary infrastructure such as attention visualisation—which itself has known interpretability limitations—or separate SHAP computation pipelines.

TABLE IV. TOP LIME FEATURE ATTRIBUTIONS FOR A REPRESENTATIVE HIGH-RISK PREDICTION

Feature Token	Module	LIME Weight	Interpretation
worthless	CB Lexicon	+0.38	Strong absolutist cognition marker
no way out	Hopelessness	+0.31	Perceived entrapment signal
vader_compound	Sentiment	+0.27	Strongly negative affect score
hopeless	CB Lexicon	+0.25	Co-occurs with High-risk stratum
self_focus_count	Pronoun Feature	+0.19	Elevated self-reference (pathological)

VI. ETHICAL CONSIDERATIONS

A. System Scope and Non-Diagnostic Status

The architecture described in this paper constitutes a research-stage feasibility demonstration, not a deployable clinical instrument. It has not been validated on any clinical population, has not undergone professional mental health practitioner review, and must not be applied in a real-world screening context without extensive independent

evaluation, regulatory assessment, and institutional ethics approval. All system outputs are pre-diagnostic screening signals indicating which text samples may warrant human attention—not risk assessments, clinical recommendations, or diagnostic conclusions.

B. *Interpretability as an Ethical Prerequisite*

The selection of interpretable classifiers over higher-performing black-box alternatives reflects a deliberate ethical stance, not merely a technical preference. A practitioner receiving a high-risk flag from an opaque neural model possesses no basis for evaluating whether that flag reflects genuine linguistic risk or a spurious statistical regularity in the training distribution. A practitioner receiving an equivalent flag from the proposed system can examine which features contributed, apply domain expertise to those features, and take professional responsibility for the resulting decision. Interpretability is therefore a necessary precondition for meaningful human oversight of automated screening—not an optional addendum to it.

C. *Coverage Gaps, Bias Exposure, and Generalisability Constraints*

The present corpus is small-scale, English-language only, and constructed under controlled conditions that do not represent the full heterogeneity of real-world communication. Individuals from diverse cultural backgrounds, age cohorts, or clinical presentations may express suicidal distress through idioms, indirect references, or culturally specific framings not captured by the current lexicons. Deploying this pipeline without addressing these coverage gaps risks systematic under-identification of risk among underrepresented groups—a harm that would compound existing disparities in mental health service access. All future extensions must prioritise demographic diversity, multilingual lexicon coverage, and rigorous clinical co-design validation.

D. *Data Privacy, Consent, and Permitted Use*

The experimental corpus contains no personally identifiable information. Any future study involving real user-generated data must obtain informed consent, apply data minimisation principles, enforce strict purpose limitation, and undergo institutional ethics board review. The system must not be positioned or deployed as a surveillance mechanism under any circumstances. Its sole legitimate function is as an optional component within voluntary, consent-based digital mental health screening workflows where individuals have explicitly requested assistance or where institutional frameworks with appropriate oversight are in place.

VII. CONCLUSION

This paper has presented and evaluated a transparent, psycholinguistically motivated machine learning architecture for stratifying suicide risk from short-form text. By combining TF-IDF lexical salience representations, hand-engineered cognitive-behavioural markers grounded in established clinical theory, and multi-source affective polarity measurements—then training three classical interpretable classifiers under rigorous stratified cross-validation—the system achieves a weighted F1 of 0.7156 with Linear SVM as the best-performing classifier. Logistic Regression delivers competitive performance with superior fold-to-fold stability. LIME attribution provides the additional dimension of instance-level explanation, enabling practitioner-facing auditing of individual predictions.

The principal finding of this work is that psychologically motivated classical features, when fused and presented to interpretable classifiers, can achieve realistic triage-relevant screening performance while preserving the transparency necessary for responsible deployment in mental health contexts. Importantly, the Medium-risk stratum represents the most persistent classification challenge, motivating dedicated future feature engineering at the distress-to-ideation linguistic boundary.

Planned extensions include scaling the corpus using ethically sourced real-world data from CLPsych shared tasks and consented clinical text; expanding lexical coverage to multilingual and culturally diverse distress expressions; investigating ensemble formulations that preserve coefficient-level interpretability; and conducting formal co-design studies with mental health practitioners to align system outputs with genuine professional information needs. The ultimate objective is a framework that earns the trust of the practitioners who deploy it—not by claiming unrealistic benchmark performance, but by being fully auditable, linguistically grounded, and clinically safe.

. REFERENCES

[1] World Health Organization, "Suicide — Key Facts," WHO, Geneva, Switzerland, Tech. Rep., 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[3] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. ACL Workshop Comput. Linguistics Clin. Psychol., Baltimore, MD, USA, Jun. 2014, pp. 51–60.

[4] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Doha, Qatar, Oct. 2014, pp. 1746–1751.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186. (See [2].)

[6] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on deep learning for mental health," arXiv preprint arXiv:2101.03940, Jan. 2021.

[7] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead, "CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts," in Proc. 6th Workshop Comput. Linguistics Clin. Psychol., Minneapolis, MN, USA, Jun. 2019, pp. 24–33.

[8] C. J. Hutto and E. E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. 8th Int. AAAI Conf. Weblogs Social Media (ICWSM), Ann Arbor, MI, USA, Jun. 2014, pp. 216–225.

[9] A. T. Beck, A. Weissman, D. Lester, and L. Trexler, "The measurement of pessimism: The hopelessness scale," J. Consult. Clin. Psychol., vol. 42, no. 6, pp. 861–865, Dec. 1974.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144.

[11] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.

[12] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, Oct. 2011.

[13] T. Loria, "TextBlob: Simplified text processing," 2020. [Online]. Available: https://textblob.readthedocs.io