Evaluating Explainability Faithfulness and Shortcut Learning in AI-Based Mental Health Text Screening: A User-Level Evaluation Framework

Anuradha Bhoge#1, Madhavi Shamkuwar#2
#
Student (1) and Guide (2), ZIBACAR, Affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India
ORCID ID: 0009-0006-4243-2049 (Anuradha Bhoge); 0000-0002-9277-0912 (Madhavi Shamkuwar)
anuradha5555bh@gmail.com; madhavi.shamkuwar@zealeducation.com

Abstract—Text-based machine learning is widely proposed for the early screening of depression and anxiety, and published models often report high classification scores. Accuracy alone, however, does not show whether a model relies on clinically meaningful language, whether its explanations describe what the model actually does, or whether reported performance survives evaluation on people the model has never seen. This paper proposes an evaluation framework that examines four dimensions together: (i) prediction performance under user-grouped evaluation, (ii) faithfulness and stability of SHAP and LIME explanations, (iii) sensitivity to non-clinical shortcut features, and (iv) the usefulness of explanations for human reviewers. The framework compares a TF-IDF logistic regression baseline, a gradient-boosted tree model and a fine-tuned transformer, and specifies deletion/insertion faithfulness tests, shortcut-ablation experiments, nested user-grouped validation, user-level statistical inference and pre-specified decision rules. This manuscript is a proposed framework and evaluation protocol. No experiments have been run, and no empirical results are reported. The contribution is a reproducible, falsifiable design that can be applied to any suitable labelled corpus, together with an explicit separation of established limitations from claims that a systematic review must still verify.

Keywords— explainable AI; mental health screening; depression detection; SHAP; LIME; explanation faithfulness; shortcut learning; user-grouped evaluation; natural language processing

I. INTRODUCTION

Mental health conditions are common, and many affected people are never assessed. Screening typically relies on self-report questionnaires such as the PHQ-9 [1] and GAD-7 [2] or on clinical interview, both of which depend on a person being willing and able to report their state. Language that people produce in social media posts, forum threads and messages is an additional, passively generated signal, and a sizeable literature applies natural language processing to it [3], [4], [5]. Pretrained transformers [6] have raised benchmark scores, and domain-specific variants have been released [7].

Strong benchmark scores are not sufficient grounds for use in counselling or clinical settings. A prediction such as “elevated risk” is unhelpful to a reviewer who cannot tell what produced it, and it is dangerous if the model has learned something other than the construct of interest. This paper therefore treats the “black-box problem” in AI-based mental health text analysis as a set of measurable questions about performance, explanation quality, shortcut reliance and human usefulness, rather than as a general concern.

The paper is a proposed framework and evaluation protocol. The original project draft left the dataset and all numerical results unspecified; those facts could not be verified, so no outcome is reported here and none is implied. Section II gives the problem statement, Section III reviews related work, Section IV states the research gap, Section V gives objectives, contributions, questions and hypotheses, Section VI the methodology, Section VII the experimental design, and Section VIII the reporting protocol. Sections IX–XII cover ethics, validity, future work and conclusions.

II. PROBLEM STATEMENT

Accuracy is not enough. Aggregate accuracy, F1 or AUC summarise how often predictions match labels in a test set. They do not reveal whether correct predictions rest on clinically relevant language, on artefacts of the data source, or on identity cues. For screening, the costs of errors are also asymmetric and context-dependent: a missed case and a false alarm have different human consequences, so threshold choice and error structure matter as much as the headline score [8], [9].

Explanations may not be faithful. SHAP [10] and LIME [11] are the most common post-hoc explanation methods for text classifiers, but an explanation is a model of the model. Prior work shows that attention weights are not reliable explanations [12], that saliency methods can fail basic sanity checks [13], that explanations can be unstable under small input changes [14], that Shapley-value attributions have conceptual limits as feature-importance measures [15], and that LIME and SHAP can be manipulated by adversarial models [16]. Faithfulness, meaning that an explanation reflects the reasoning actually used by the model, must be tested rather than assumed [17], [18].

Leakage and shortcuts can inflate performance. Social-media corpora contain many posts per person. If posts from one user appear in both training and test data, a model can exploit user-specific vocabulary and style instead of the target condition; leakage of this kind is a recognised source of over-optimistic results in ML-based science [19]. Beyond leakage, models trained on proxy-labelled mental health data have been shown to depend on confounding factors and to lose substantial performance when transferred across platforms [20]. This is one instance of the broader phenomenon of shortcut learning [21], documented in language tasks as annotation artefacts [22] and heuristics that are

right for the wrong reasons [23].

These issues interact. A model may score well because of a shortcut; its explanation may faithfully highlight that shortcut, or may unfaithfully highlight plausible clinical words; and a reviewer shown a plausible-looking explanation may trust the output more without being better informed [24]. Evaluating prediction, faithfulness, shortcut sensitivity and human usefulness separately can therefore give a misleading picture.

Problem addressed. The problem is the absence of a single, pre-specified evaluation design that examines user-level predictive performance, explanation faithfulness and stability, shortcut sensitivity and reviewer usefulness on the same models and data, and that reports them in a way that allows each claim to be falsified. The proposed framework (Section VI) goes beyond a conventional classification pipeline by (a) making user-grouped splitting and nested tuning mandatory, (b) quantifying leakage inflation explicitly, (c) testing explanations against random and reversed rankings, (d) ablating pre-declared shortcut feature families, and (e) fixing hypotheses and decision rules before test data are examined.

III. RELATED WORK

A. Text-based detection of depression and anxiety Research datasets for this task include Twitter users with self-reported diagnoses [25] and Reddit collections such as RSDD [26] and SMHD [27], which infer labels from self-disclosure statements. Reviews describe a rapid growth of this literature [4], [5] and discuss the data, methods and clinical grounding of social-media prediction [3], [28]. A recurring concern is that proxy labels, platform effects and data availability limit what the reported results can support [20], [28], [29].

B. Neural and transformer models Fine-tuned encoders such as BERT [6] and RoBERTa [30] typically outperform bag-of-words baselines on text classification, and MentalBERT [7] adapts pretraining to mental health text. Their predictions are difficult to inspect, which motivates post-hoc explanation. Domain-adapted models also raise a contamination question, because pretraining corpora may overlap with evaluation data; this must be checked per dataset.

C. Post-hoc explanation and its evaluation LIME fits a local surrogate to perturbed inputs [11]; SHAP assigns additive Shapley-value attributions [10], with efficient algorithms for tree ensembles [31]. Faithfulness is commonly evaluated by removing or inserting the features an explanation ranks highest and measuring the change in model output [32]; for text, comprehensiveness and sufficiency measures serve a similar purpose [33]. Removal-based tests can themselves be confounded when masked inputs fall outside the training distribution, which has motivated retraining-based benchmarks [34]. Conceptual work stresses that faithfulness should be defined and tested explicitly [17], and that interpretability claims need functionally grounded and human-grounded evaluation [35].

D. Explainability in health care and human usefulness

Multidisciplinary analyses argue that explainability in health care has technical, legal, ethical and patient-centred dimensions [8], and that current post-hoc approaches may give unwarranted assurance about individual predictions [9]. An alternative view is that inherently interpretable models should be preferred for high-stakes decisions [36]. Controlled studies suggest that making a model more transparent does not automatically improve human decisions [24]. For mental health text, whether explanations help reviewers has, to our knowledge, received little controlled study, but this statement is itself subject to the verification described in Section IV.

E. Shortcut learning and evaluation hygiene Shortcut learning describes models that satisfy a benchmark through unintended cues [21]. In mental health text, candidate cues include subreddit names, hashtags, boilerplate, posting time and the self-disclosure phrases used to construct labels [20]. User-grouped splitting is a basic safeguard against identity leakage, and is part of the broader guidance on avoiding leakage [19].

IV. RESEARCH GAP

We separate two kinds of statement.

Established limitations (supported by the sources above): (a) classification metrics alone do not establish that a model is trustworthy for screening [8], [9]; (b) SHAP, LIME and related explanation methods can be unstable, conceptually limited or manipulable [14], [15], [16]; (c) mental health models trained on proxy-labelled social-media data can rely on confounds and generalise poorly [20]; and (d) record-level evaluation with repeated users risks leakage [19].

Candidate gap requiring verification. We hypothesise that few or no published studies jointly evaluate, for the same mental health text classifiers, user-grouped performance, SHAP/LIME faithfulness and stability, shortcut-ablation sensitivity and reviewer usefulness. We have not established this through a systematic search, so the manuscript does not claim that the combination is novel. Verification will follow a PRISMA 2020-style protocol [37]: databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, PubMed, ACL Anthology), query blocks for mental health condition, text or NLP, and explainability or interpretability, inclusion limited to studies that apply SHAP, LIME or comparable post-hoc methods, and a coding sheet recording which of the four dimensions each study evaluates. If the review finds that the combination has been studied, the contribution of this work will be reframed accordingly.

V. OBJECTIVES, CONTRIBUTIONS, QUESTIONS AND HYPOTHESES

A. Objectives

• O1. Specify a reproducible pipeline that trains three classifiers of increasing complexity under user-grouped evaluation. • O2. Specify faithfulness and stability tests for SHAP and LIME explanations, with random and reversed controls.

• O3. Specify shortcut-ablation and leakage-quantification experiments. • O4. Specify a human-review study for explanation usefulness, to be run only with ethical approval. • O5. Specify statistics, decision rules and a reporting template that make negative results reportable.

B. Contributions The contributions are methodological and are not empirical findings: (1) a framing of the black-box problem as four jointly evaluated, measurable dimensions; (2) a nested user-grouped protocol that also quantifies leakage inflation by contrasting grouped with random splits as a diagnostic only; (3) a faithfulness test suite combining deletion and insertion curves with random and reversed rankings and a stability measure; (4) a taxonomy of pre-declared shortcut feature families with an ablation design; and (5) pre-specified hypotheses, decision rules and a reproducibility checklist (Appendix A). A reference implementation is planned but has not been released.

C. Research questions and hypotheses RQ1. How well do the three classifiers distinguish screened-positive from screened-negative text on users unseen in training? RQ2. Are SHAP and LIME explanations faithful and stable? RQ3. How sensitive are the models to non-clinical shortcut features? RQ4. Do explanations improve the quality of human reviewers' judgements relative to predictions alone? Hypotheses are one-sided only where a directional expectation is justified, and each has a decision rule fixed in advance (Table I).

TABLE I HYPOTHESES AND PRE-SPECIFIED DECISION RULES

ID Hypothesis (null = no advantage) Decision rule

H1 Under user-grouped evaluation the transformer (C) has higher F1 than the TF-IDF baseline (A).

Paired user-level bootstrap 95% CI of F1(C) − F1(A) excludes 0 after Holm correction.

H2 SHAP yields larger faithfulness gains (Eqs. 1–2) than LIME on the same model and instances.

Paired user-clustered CI of gain difference excludes 0; reported per model.

H3 Removing declared shortcut features reduces F1 by less than a margin of 0.03.

Non-inferiority: upper 95% bound of the F1 drop is below 0.03.

H4 Reviewers shown explanations agree more with the reference label than reviewers shown predictions only.

Mixed-effects model with reviewer and item effects; conditional on an approved study.

VI. PROPOSED FRAMEWORK AND METHODOLOGY

A. Overview Fig. 1 shows the six stages. Each stage has defined inputs and outputs, so that the framework can be re-run on any dataset that meets the criteria in Section VI-B.

1 Data audit provenance, labels, inclusion criteria, user IDs

2 User-grouped splitting outer grouped CV; inner split for tuning, threshold

3 Preprocessing and feature sets

full set versus shortcut-ablated set

4 Models A, B, C TF-IDF + LR; XGBoost; fine-tuned transformer

5 Explanations LR coefficients; SHAP; LIME

6 Evaluation and reporting performance, faithfulness, stability, ablation, review

Fig. 1. Proposed evaluation framework. Stages 1–6 are specified; none has been executed.

B. Dataset provenance and inclusion criteria No dataset was identified in the original draft, and none is asserted here. The framework is dataset-agnostic, but a corpus is eligible only if it meets all of the following: (i) an anonymised user identifier supports grouping; (ii) the label source is documented (self-disclosed diagnosis, screening-instrument score, clinician annotation or other) and is recorded as a covariate; (iii) licence or data-use terms permit this analysis and ethical approval has been obtained where required; (iv) metadata needed for the shortcut analysis (timestamps, community or hashtag fields) are available; and (v) there are enough users per class for grouped cross-validation. Candidate corpora described in the literature include CLPsych 2015 [25], RSDD [26] and SMHD [27]; several require data-use agreements, and their eligibility must be confirmed before use. Label definitions, class balance, user counts and exclusions must be reported (Appendix A). Because labels may be proxies for diagnosis, all findings concern screening-label prediction, not clinical diagnosis [20], [28].

C. Preprocessing and features Text is Unicode-normalised, URLs and user mentions are replaced by placeholder tokens, and emojis are retained as tokens. Posts with fewer than 10 tokens are excluded, and at most 50 posts per user are retained (random, seeded) so that very active users do not dominate training or evaluation. For TF-IDF, terms must occur in at least 5 distinct users, which discourages user-specific vocabulary. For Model B, TF-IDF features are supplemented by interpretable descriptors (first-person pronoun rate, negation rate, lexicon-based sentiment, mean sentence length). Descriptors are computed from text only; posting hour and length are kept in the separate shortcut set.

D. Models Model A is an L2-regularised logistic regression on TF-IDF features, whose coefficients are readable and provide an
intrinsic-interpretability reference [36], [38]. Model B is a gradient-boosted tree model (XGBoost [39]) on TF-IDF plus descriptor features, with class imbalance handled by instance weights. Model C is a pretrained transformer encoder (RoBERTa [30] as the general-domain default, optionally MentalBERT [7] after a contamination check) fine-tuned for binary classification with early stopping on an inner grouped validation set, using the Transformers library [40]. Inputs longer than the model limit are truncated, and the truncation rate is reported. Transformer training uses three random seeds. Probabilities are calibrated on inner validation data, and calibration quality is reported [41].

E. Leakage control, tuning and thresholding All splits are by user. An outer repeated stratified grouped cross-validation (5 folds × 3 repeats) estimates generalisation; within each outer training set, an inner grouped split (20% of users) is used for hyperparameter search, calibration and early stopping. The decision threshold is chosen on inner validation data as the highest threshold that reaches a sensitivity of at least 0.80, then frozen before the outer test fold is scored. TF-IDF vocabularies, scalers and calibrators are fitted only on training data. As a leakage diagnostic, the same models are also evaluated under a random post-level split; the gap between random and grouped results is reported as estimated leakage inflation and is never used as a headline result. A user-level label-permutation control should return chance-level performance; if it does not, the pipeline is considered faulty.

F. Explanation generation For Models B and C, token-level SHAP attributions are computed (TreeSHAP for B [31]; a text-masker-based SHAP explainer for C [10]) and LIME explanations are computed with 5,000 perturbation samples per instance [11]. For Model A the coefficients serve as the explanation. Explanations are generated for the predicted class on a random sample of test instances stratified by outcome (true/false positive/negative) and by user, with the sample size fixed in the registered protocol. TreeSHAP values are exact for the tree model's chosen value function, which is not the same as the effect of deleting a word from text [15], [31]; the faithfulness tests below therefore measure the explanation against actual input perturbation.

G. Faithfulness and stability tests Let $f\hat{c}(x)$ be the model probability for the predicted class $\hat{c}$, π an ordering of tokens by attribution (most supportive first), and $x_{\pi,q}$ the input after masking the top fraction q of tokens (grid Q = 0, 0.05, …, 0.5, summed over q in Q). The deletion and insertion areas are

$$\mathrm{DAUC}(x, \pi) = \Sigma q\, f\hat{c}(x_{\pi,q})\, \Delta q / q\max \quad (1)$$

$$\mathrm{IAUC}(x, \pi) = \Sigma q\, f\hat{c}(x \text{ restricted to top } q)\, \Delta q / q\max \quad (2)$$

(trapezoidal sums, normalised by qmax = 0.5). Because lower DAUC and higher IAUC indicate better rankings, the faithfulness gains are $G_{del} = E[\mathrm{DAUC}(x, \pi_{rand})] - \mathrm{DAUC}(x, \pi_{expl})$ and $G_{ins} = \mathrm{IAUC}(x, \pi_{expl}) - E[\mathrm{IAUC}(x, \pi_{rand})]$, with expectations over 100 random orderings per instance. A reversed ordering (least supportive first) is also evaluated and should give the opposite

effect. Because masking can create out-of-distribution inputs [34], two masking operators (token deletion and a neutral placeholder token) are used, and conclusions are reported only if they hold under both. Stability is the mean Jaccard overlap of top-k (k = 10) tokens between an input and its perturbed variants (whitespace and punctuation normalisation, and random removal of 5% of tokens outside the top-k set), restricted to variants for which the predicted class is unchanged [14]. Explanation methods whose gains do not exceed zero are reported as not faithful for that model, and parameter-randomisation sanity checks [13] are applied to Model C where feasible.

H. Shortcut-feature ablation

TABLE II DECLARED SHORTCUT FEATURE FAMILIES

Family Examples Ablation

Metadata Posting hour, post length, post count per user

Removed from Model B; excluded by design for A and C

Platform tokens Subreddit names, hashtags, platform boilerplate

Masked in text

Label-proximal phrases

Self-disclosure statements used to derive labels (e.g. “I was diagnosed with …”)

Masked in text; verify that the dataset's own cleaning removed them

Template text Repeated signatures, auto-generated text

Detected by near-duplicate matching and removed

Each model is evaluated with all features and with the shortcut families in Table II ablated. A shortcut-only model (trained on the ablated families alone) is compared with a majority-class baseline. A small drop after ablation does not prove absence of shortcuts, since cues can be redundant or correlated with content; family-by-family ablation and inspection of top-ranked explanation tokens are used to qualify the conclusion.

I. Human-review study (proposed) The human-review study is a proposed component and is not reported as completed. Participants would be reviewers with relevant background (for example counsellors or psychology students), each rating a held-out set of posts under two counterbalanced conditions: prediction only, and prediction with highlighted explanation. Outcomes are agreement with the reference label, confidence, decision time, and agreement with the model when the model is wrong, which measures over-reliance. Sample size will be set by a power analysis using a pilot, and analysis uses a mixed-effects model with random effects for reviewer and item. Because the reference label may itself be a proxy, the study measures usefulness for label-consistent screening judgement and does not establish clinical utility. It requires approval by an appropriate ethics committee, informed consent and exposure safeguards for potentially distressing content before any recruitment.

VII. EXPERIMENTAL DESIGN AND EVALUATION METRICS

Four experiments are specified. E1 compares the three models under outer grouped cross-validation. E2 evaluates SHAP and LIME faithfulness and stability. E3 performs the shortcut and leakage analyses. E4 is the human-review study. Table III lists measures.

TABLE III EVALUATION DIMENSIONS AND MEASURES

Dimension Measures

Prediction (RQ1) Precision, recall, F1 for the positive class; ROC-AUC; average precision; false-negative rate at the frozen threshold; calibration error. User-level aggregation (mean probability per user) as a secondary analysis.

Faithfulness (RQ2) Gdel, Gins against random and reversed orderings; results by outcome stratum.

Stability (RQ2) Top-k Jaccard overlap under perturbation; cross-seed agreement for Model C.

Shortcut sensitivity (RQ3)

F1 change after ablation, by family; shortcut-only model versus majority baseline; random-versus-grouped gap.

Usefulness (RQ4) Reviewer agreement with reference label, confidence, time, over-reliance rate.

TABLE IV DESIGN PARAMETERS FIXED BEFORE TEST DATA ARE EXAMINED

Parameter Value

Minimum post length 10 tokens

Minimum term support 5 distinct users

Posts retained per user At most 50 (seeded sample)

Outer / inner validation 5×3 repeated grouped CV / 20% inner user split

Sensitivity target for threshold ≥ 0.80 on inner validation

LIME samples per instance 5,000

Deletion grid q = 0 to 0.5, step 0.05

Random orderings per instance 100

Shortcut non-inferiority margin 0.03 F1

Bootstrap 2,000 user-level resamples

Significance level 0.05 with Holm correction

The values in Table IV are design choices proposed here, not quantities estimated from data. They may be revised only in the registered protocol before test folds are scored, and any later change must be reported as a deviation.

A. Statistical analysis Posts from one user are not independent, so uncertainty is estimated with user-level (cluster) bootstrap resampling of pooled out-of-fold predictions, and model comparisons use paired bootstrap on the same users. Fold-wise comparisons with the Wilcoxon signed-rank test [42], [43] are secondary, because five folds cannot yield a two-sided exact p-value below 0.0625, and repeated cross-validation folds overlap, which makes naive tests optimistic [44]. Multiple comparisons are corrected with Holm's procedure [45]. Effect sizes and intervals are reported in preference to p-values alone.

VIII. EVALUATION AND REPORTING PROTOCOL

No results are reported. This section specifies how results will be reported once experiments are run. Every number in a future version must come from the registered pipeline, with software versions, random seeds and user-to-fold assignments archived. Table V states how combinations of outcomes will be interpreted, so that negative and mixed findings are reported rather than suppressed.

TABLE V PRE-SPECIFIED INTERPRETATION OF OUTCOME PATTERNS

Outcome pattern Interpretation

Good grouped performance; positive Gdel and Gins; small shortcut drop

Supports use as a screening aid under the tested conditions; no clinical claim.

Good performance with a large shortcut drop

Performance is partly driven by non-clinical cues; report as inflated.

Good performance; gains near zero Explanations are not faithful for that method and model; do not present them to reviewers as reasons.

Large random-versus-grouped gap Leakage inflation; prior random-split results on this corpus should be treated cautiously.

Explanations raise reviewer confidence but not accuracy

Risk of over-trust; report as a safety concern.

Control checks fail (permutation, reversed ordering)

Treat the pipeline as faulty; do not interpret other results.

Reports will follow the reproducibility checklist in Appendix A and a model-card style summary [46], including intended use, label source, subgroup limits and known failure cases. Error analysis will examine performance by label source, post length, and explicit versus implicit language, and will inspect whether explanations help or mislead on sarcasm and mixed emotion. Individual posts will be paraphrased or summarised, not quoted verbatim, in any publication.

IX. ETHICAL CONSIDERATIONS

Mental health text is highly sensitive, even when public [47]. The framework requires use of data only under their licence or data-use agreement, secure storage, no attempt to re-identify or contact individuals, no redistribution of raw text, and ethics review or documented exemption before data analysis and before any human participant is recruited. Outputs are screening aids for human follow-up and are not diagnoses; a positive flag should always lead to review by a qualified person, and an absent flag must never be read as assurance. False negatives and false positives both carry harm, and thresholds should be set with a clinical or counselling partner. Explanations carry their own risk, since a plausible but unfaithful highlight can mislead [9]. The framework has not been validated for deployment, is not a medical device, and any real-world use needs governance, consent design and compliance with applicable data-protection law.

X. THREATS TO VALIDITY AND LIMITATIONS

• Construct validity. Self-disclosure or platform-derived labels may not correspond to clinical diagnosis [20], [28]. • Internal validity. Grouped splits remove identity leakage but not topic or style confounds; shortcut ablation can miss

correlated cues. • External validity. Results on one platform, language or period may not transfer [20]; the framework itself does not remove this limit. • Conclusion validity. Statistical power depends on the number of independent users, not posts; the bootstrap and corrections address, but do not eliminate, this limit. • Explanation validity. Deletion tests measure agreement with the model, not clinical truth, and removal can create out-of-distribution inputs [34]. • Scope. The current manuscript contains no experiments; its design choices (Table IV) have not been piloted, and its feasibility is argued, not demonstrated.

XI. FUTURE RESEARCH DIRECTIONS

Immediate steps are to complete the systematic review of Section IV, select an eligible corpus, register the protocol, and run E1–E3. Later work could extend to clinical or counselling text with approval, to Marathi and Hindi to test faithfulness across languages, to longitudinal modelling of post sequences, to conformal prediction for calibrated uncertainty, and to co-design of explanation views with counsellors.

XII. CONCLUSION

This paper proposes, but does not yet test, an evaluation framework for AI-based mental health text screening that examines user-grouped performance, SHAP/LIME faithfulness and stability, shortcut sensitivity and reviewer usefulness together under pre-specified decision rules. Its value is a falsifiable design in which leakage, unfaithful explanations and shortcut reliance can be detected and reported, whatever the outcome. Whether the framework yields reliable explanations, and whether this combination of evaluations is new, are open questions that the planned experiments and systematic review must answer.

ACKNOWLEDGMENT

The first author thanks Dr. Madhavi Shamkuwar, project guide, and the Department of MCA, ZIBACAR, for guidance during the research project course.

APPENDIX A: REPRODUCIBILITY CHECKLIST

• Dataset name, version, access terms, label source and definition, inclusion and exclusion counts, users per class. • Ethics approval or exemption reference; data-use agreement. • Preprocessing code, shortcut-family definitions, random seeds, and user-to-fold assignments. • Pinned versions of Python and all libraries; model checkpoints; transformer pretraining source and contamination check. • Tuning grids, selected hyperparameters, calibration method, frozen threshold. • Explanation settings: attribution method, masker, samples, masking operators, orderings, k, Δq.

• Control results: label permutation, random-versus-grouped split, reversed ordering. • Full results with user-level confidence intervals, including negative findings; deviations from the registered protocol.

REFERENCES

[1] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9: Validity of a brief depression severity measure,” J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x. [2] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, “A brief measure for assessing generalized anxiety disorder: The GAD-7,” Arch. Intern. Med., vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092. [3] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, “Detecting depression and mental illness on social media: An integrative review,” Curr. Opin. Behav. Sci., vol. 18, pp. 43–49, 2017, doi: 10.1016/j.cobeha.2017.07.005. [4] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, “Machine learning in mental health: A scoping review of methods and applications,” Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151. [5] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: A narrative review,” npj Digit. Med., vol. 5, Art. no. 46, 2022, doi: 10.1038/s41746-022-00589-7. [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423. [7] S. Ji et al., “MentalBERT: Publicly available pretrained language models for mental healthcare,” in Proc. 13th Lang. Resources Eval. Conf., 2022, pp. 7184–7190. [8] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, “Explainability for artificial intelligence in healthcare: A multidisciplinary perspective,” BMC Med. Inform. Decis. Mak., vol. 20, Art. no. 310, 2020, doi: 10.1186/s12911-020-01332-6. [9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” Lancet Digit. Health, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9. [10] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 4765–4774. [11] 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, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778. [12] S. Jain and B. C. Wallace, “Attention is not explanation,” in Proc. NAACL-HLT, 2019, pp. 3543–3556, doi: 10.18653/v1/N19-1357. [13] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” in Adv. Neural Inf. Process. Syst. 31, 2018, pp. 9505–9515. [14] D. Alvarez-Melis and T. S. Jaakkola, “On the robustness of interpretability methods,” 2018, arXiv:1806.08049. [15] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, “Problems with Shapley-value-based explanations as feature importance measures,” in Proc. 37th Int. Conf. Mach. Learn., PMLR vol. 119, 2020, pp. 5491–5500. [16] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods,” in Proc. AAAI/ACM Conf. AI, Ethics, Soc., 2020, pp. 180–186, doi: 10.1145/3375627.3375830. [17] A. Jacovi and Y. Goldberg, “Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4198–4205. [18] Z. C. Lipton, “The mythos of model interpretability,” Commun. ACM, vol. 61, no. 10, pp. 36–43, 2018. [19] S. Kapoor and A. Narayanan, “Leakage and the

reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: 10.1016/j.patter.2023.100804. [20] K. Harrigian, C. Aguirre, and M. Dredze, "Do models of mental health based on social media data generalize?," in Findings Assoc. Comput. Linguistics: EMNLP 2020, 2020, pp. 3774–3788, doi: 10.18653/v1/2020.findings-emnlp.337. [21] R. Geirhos et al., "Shortcut learning in deep neural networks," Nature Mach. Intell., vol. 2, no. 11, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z. [22] S. Gururangan et al., "Annotation artifacts in natural language inference data," in Proc. NAACL-HLT, 2018, pp. 107–112, doi: 10.18653/v1/N18-2017. [23] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 3428–3448. [24] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in Proc. CHI Conf. Human Factors Comput. Syst., 2021, Art. no. 237, doi: 10.1145/3411764.3445315. [25] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., 2015, pp. 31–39. [26] A. Yates, A. Cohan, and N. Goharian, "Depression and self-harm risk assessment in online forums," in Proc. Conf. Empirical Methods Natural Lang. Process., 2017, pp. 2968–2978. [27] A. Cohan et al., "SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions," in Proc. 27th Int. Conf. Comput. Linguistics, 2018, pp. 1485–1497. [28] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: A critical review," npj Digit. Med., vol. 3, Art. no. 43, 2020, doi: 10.1038/s41746-020-0233-7. [29] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research," in Proc. 7th Workshop Comput. Linguistics Clin. Psychol., 2021, pp. 15–24 (ACL Anthology ID 2021.clpsych-1.2). [30] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, arXiv:1907.11692. [31] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Mach. Intell., vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9. [32] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in Proc. British Mach. Vision Conf., 2018, arXiv:1806.07421. [33] J. DeYoung et al., "ERASER: A benchmark to evaluate rationalized NLP models," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4443–4458. [34] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in Adv. Neural Inf. Process. Syst. 32, 2019, pp. 9737–9748. [35] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, arXiv:1702.08608. [36] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x. [37] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71. [38] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785. [40] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations, 2020, pp. 38–45. [41] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Mach. Learn., 2005, pp. 625–632, doi: 10.1145/1102351.1102430. [42] F. Wilcoxon, "Individual comparisons by ranking methods," Biometrics Bull., vol. 1, no. 6, pp. 80–83, 1945. [43] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.

[44] C. Nadeau and Y. Bengio, "Inference for the generalization error," Mach. Learn., vol. 52, no. 3, pp. 239–281, 2003. [45] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Statist., vol. 6, no. 2, pp. 65–70, 1979. [46] M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, Transparency, 2019, pp. 220–229, doi: 10.1145/3287560.3287596. [47] A. Benton, G. Coppersmith, and M. Dredze, "Ethical research protocols for social media health research," in Proc. 1st ACL Workshop Ethics Natural Lang. Process., 2017, pp. 94–102.