International Journal of Advance Research in Science, Management, Engineering and Technology
Volume 1, Issue 4 | September - October 2026 | www.ijarsmet.com
ISSN: 3139-1141

Explainable AI for Suicide-Risk Detection in Online Text: A Review of Methods, Evaluation Practices and the Black-Box Gap

Monika Borude
Student, Department of MCA, ZIBACAR, affiliated to Savitribai Phule Pune University (SPPU), Pune, Maharashtra, India
ORCID: 0009-0003-4692-3214 (M.Borude)
monikaborude108@gmail.com

Abstract

Automated classifiers are increasingly studied as tools for identifying online posts that may indicate suicide risk, so that trained human reviewers can prioritise them. Such systems can report high accuracy, but a flag that arrives without an account of what produced it is difficult to trust in a safety-critical setting. This paper is a structured narrative review of text-based suicide-risk detection and of the role of explainability in it. It is organised around three dimensions: modelling approach, data and labelling practice, and evaluation design, including whether explanations are tested for faithfulness and whether author-level leakage and missed cases are addressed. The review is purposive rather than exhaustive: it synthesises two prior surveys and a set of five primary studies whose bibliographic details and key characteristics were verified against their published records, and it combines them with the established literature on explanation methods and evaluation pitfalls. No systematic database search, screening count or pooled estimate is claimed. Within the studies examined, accuracy and F1 are the dominant reported measures; explanation, where present, takes the form of attention weights validated against proxies rather than tested for faithfulness; and fine-grained risk prediction remains weak. The paper contributes a critical synthesis, a comparison of explanation methods, and a proposed reporting checklist, and argues that shared evaluation standards are needed before reported performance can be compared fairly or used to justify deployment.

Keywords—suicide risk; suicidal ideation; text classification; explainable AI; interpretability; faithfulness; narrative review; evaluation practices; shortcut learning

I. Introduction

Suicide is a major public-health concern, and the pervasiveness of social media offers new types of data for understanding people at risk and for screening at scale [1]. Many people express distress in writing, often anonymously, before they seek help. Researchers have therefore proposed machine learning as a means of surfacing posts that may need urgent attention, so that moderators, peer-support teams or clinicians can respond sooner. Surveys of this area describe a steady shift from hand-engineered features to deep neural models, and note that many neural models lack interpretability [2]. A critical review of social-media-based mental-health prediction similarly found no standardised process for evaluating the validity of this research and identified concerning trends around construct validity [3]. Public resources have enabled much of the progress: expert-annotated Reddit data [4], a shared task built on it [5], a large microblog corpus [6] and the systems submitted to the shared task [7].

The promise of these tools depends on trust. A missed post can mean that a person in danger receives no response, while a wrongly flagged post can cause distress, stigma and unnecessary workload. In both cases a reviewer needs to understand why a decision was made. Explainable AI (XAI) offers methods for this, but explanation is often presented as a feature of a system rather than tested as a property of it. We refer to this shortfall, together with weak evaluation practice, as the black-box gap: the distance between a reported score and the evidence that the score reflects safe, understandable and generalisable behaviour.

Research gap and objectives

Existing surveys cover methods and applications broadly [2] or evaluate methodological rigour across mental-health conditions [3]. Less attention has been paid to the intersection examined here: whether suicide-risk classifiers on text are explained, whether those explanations are evaluated, and how evaluation design (splits, labels, missed-case reporting) bears on the credibility of both. The objectives are (i) to describe the

ISSN: 3139-1141

modelling approaches and how they have changed, (ii) to examine data, labelling and evaluation practices, (iii) to explain, accurately and in the context of this task, the main explanation methods and their known limits, and (iv) to propose a minimum reporting standard. The review does not rank models, because datasets, labels and tasks differ too much for direct comparison.

Research questions

RQ1. Which modelling approaches have been used for suicide-risk detection in online text, and how have they changed over time? RQ2. What data sources and labelling methods do the studies use, and how are they described? RQ3. How are models evaluated, including control of author-level leakage and reporting of missed cases? RQ4. To what extent are explanation methods applied, and are the explanations tested for faithfulness or usefulness to reviewers?

II. Background

A. Suicide-risk assessment and the role of text classification

Clinical risk assessment relies on structured interviews, validated instruments and professional judgement. Text classification does not replace these. Its realistic role is to prioritise content for human review, and Coppersmith et al. frame such technology as screening that could extend reach beyond the health-care system while raising an explicit trade-off between privacy and prevention [1]. Because the aim is prioritisation, the operating threshold, the cost of missed cases and the form of the explanation must all suit a reviewer's workflow.

B. From bag-of-words to transformers

Early studies represented text with word counts or term frequency-inverse document frequency (TF-IDF) features and trained linear models or tree ensembles; these remain strong baselines, as the shared-task systems in Table I illustrate [7]. Neural approaches followed, including recurrent networks with attention. Transformer models are built on self-attention, in which each token's representation is computed as a learned weighted combination of all other tokens in the sequence [8]. Pre-trained transformers such as BERT are trained on large unlabelled corpora with a masked-language-modelling objective and then fine-tuned on a labelled task [9]. Each step generally improved benchmark accuracy but made the internal reasoning harder to inspect: a fine-tuned transformer has many millions of parameters and no human-readable decision rule.

C. Explainability and its evaluation

Doshi-Velez and Kim argue that interpretability needs a rigorous, evaluable definition rather than intuition [10], and Lipton observes that the term bundles distinct and sometimes conflicting goals [11]. A common division is between intrinsically interpretable models, whose structure can be inspected directly, and post-hoc methods, which explain a trained model from outside. Rudin argues that for high-stakes decisions interpretable models should be preferred to explanations of black boxes [12]. The principal post-hoc and model-internal techniques are as follows.

LIME explains a single prediction by perturbing the input (for text, removing words), observing the black-box output for each perturbed copy, and fitting a sparse, interpretable linear model weighted by proximity to the original instance; the coefficients indicate which words pushed the prediction up or down locally [13]. It is model-agnostic but depends on the perturbation scheme and sampling, so explanations can vary between runs. SHAP assigns each feature a contribution based on Shapley values from cooperative game theory, so that contributions sum to the difference between the prediction and a baseline expectation [14]. For feature i in a feature set F, the value is the weighted average marginal contribution of i across all subsets of the other features:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \left[\frac{|S|!\,(|F|-|S|-1)!}{|F|!} \right] \times \left[f(S \cup \{i\}) - f(S) \right]$$

Exact computation grows exponentially with the number of features, so practical implementations approximate it. Both LIME and SHAP produce word-level attributions, which are easy to display but describe the model's behaviour on perturbed inputs; they are not necessarily the model's actual reasoning, and adversarial classifiers can be constructed that hide biased behaviour from both [20].

Attention-based explanations present the attention weights of a model as a map of which words mattered. Jain and Wallace showed that attention weights often correlate only weakly with gradient-based importance and that very different attention distributions can yield the same prediction, so attention is not by itself a dependable explanation [15]. Serrano and Smith reached a related

ISSN: 3139-1141

conclusion by intervening on attention [16]. Wiegreffe and Pinter countered that the answer depends on what is meant by explanation and proposed tests under which attention can be informative [17]. The defensible position is that attention may be a useful lead, but a claim that it explains a decision requires evidence.

Faithfulness and plausibility. A faithful explanation accurately reflects the process that produced the model's output; a plausible one merely appears convincing to a person. The two can diverge, and Jacovi and Goldberg argue that faithfulness should be evaluated as a graded property rather than assumed [18]. Practical tests include deletion and insertion measures, formalised in the ERASER benchmark as comprehensiveness (how much the prediction falls when the highlighted evidence is removed) and sufficiency (how well the evidence alone supports the prediction) [19]. Stability under small input changes is a further check. For suicide-risk flags, a plausible but unfaithful highlight is hazardous because it can create false confidence in a reviewer.

D. Evaluation pitfalls

Data leakage. Leakage occurs when information unavailable at deployment influences training or evaluation; Kapoor and Narayanan catalogue it as a pervasive cause of over-optimistic and irreproducible results in machine-learning-based science [21]. In this field the characteristic form is author-level leakage: when several posts by one user fall on both sides of a random split, the model can recognise the person's idiosyncratic vocabulary and style rather than risk-relevant language. Splits should therefore be grouped by author (and ideally by time). **Dataset bias and shortcut learning.** Classifiers exploit any feature correlated with the label, so they may learn decision rules that succeed on a benchmark but fail in practice [22]. Candidate shortcuts include post length, community-specific vocabulary and posting patterns. **Label validity.** Labels may come from community membership, self-disclosure, annotation by crowds or clinicians, or later outcomes; each measures a different construct. Evidence that models trained on proxy-diagnosed social-media data transfer imperfectly to clinically grounded signals is a standing caution [23], consistent with the construct-validity concerns in [3]. **False negatives.** The false-negative rate, FN/(TP+FN), is one minus recall on the at-risk class. Accuracy and macro-F1 can conceal it, especially with class imbalance. Because a missed case may leave a person without help, studies should report recall (and precision) for the at-risk class at the stated operating threshold.

III. Review Method and Scope

Design. This paper is a structured narrative review. The submitted draft of this work was framed as a systematic review, but a systematic review requires a documented, reproducible search, screening counts and a full extraction of every eligible study. Those procedures have not been carried out, and this revision does not claim them. Accordingly, no search strings, record counts, inclusion totals or pooled figures are reported, and the paper makes no statement about how many studies exist in the field.

Selection of sources. Sources were selected purposively to answer the research questions: two prior surveys [2], [3] that give broad coverage, five primary studies that represent public benchmark resources, shared-task evaluation and an attention-based model [4]-[7], [1], and the foundational literature on transformers and explanation methods [8]-[20]. Bibliographic details of the suicide-specific sources were checked against their published records (journal, proceedings or anthology entries). The characteristics in Table I were taken from abstracts and published records and should be confirmed against full texts before any claim is relied on.

Appraisal. Each primary study was described on six dimensions: data and labels, approach, evaluation reported, explanation, and limitations. These are descriptive, not a numerical quality score, because the studies differ too much for a single rating to be meaningful. Where a characteristic could not be established from the record consulted, it is marked as not established rather than inferred. A full systematic extension, with databases, dates, strings, a flow diagram and a search log, is identified as future work in Section VII.

ISSN: 3139-1141

IV. Overview of Reviewed Studies

TABLE I. Primary studies examined (details as reported in the records consulted)

Study	Data and labels	Approach	Evaluation reported	Explanation	Limitations / observations
Coppersmith et al. 2018 [1]	Social-media language data; signals around suicide attempts	NLP with machine learning, including deep learning	Feasibility of screening; design of an automated risk-estimation system	Not established from record; not a stated focus	Screening framed against a privacy-versus-prevention trade-off; opt-in setting
Shing et al. 2018 [4]	Reddit r/SuicideWatch; post sequences for 934 users; risk annotated by experts and crowd using a new rubric	Baseline predictive models	Inter-rater reliability of clinicians; expert vs crowd comparison; baseline classification and screening	Not established from record	Dataset access is controlled; baselines only; ratings rely on annotators' reading of text
Zirikly et al. 2019 [5]	Reddit; users labelled no, low, moderate or severe risk; two tasks on SuicideWatch posters, one on non-SuicideWatch posts	Shared task; 15 participating teams	Task-level comparison of submitted systems	Not established from record	Four-level risk scale is hard to predict; shared-task conditions may not match deployment
Ruiz et al. 2019 [7]	CLPsych 2019 data: 496 training users (31,553 posts); 125 test users (9,610 posts)	SVM, Naive Bayes and an ensemble	Macro-F1 0.379 (ensemble, 4-way); F1 0.836 (any risk vs none) and 0.736 (urgent vs not), Naive Bayes	None described in the abstract	Small test set; simple models competitive on binary tasks; four-way performance low
Cao et al. 2019 [6]	Chinese-language microblog posts; corpus reported as 252,901 posts by 3,652 suicide users; tree-hole accounts used to build embeddings	Suicide-oriented word embeddings and two-layer attention network	Accuracy 91.33%, F1 90.92% reported for the proposed model	Attention weights; checked by correlation with tree-hole posts, a proxy rather than a faithfulness test	Labels tied to accounts of users who died; construct validity and generalisation need scrutiny; split protocol not established

Note. Entries are drawn from the abstracts and published records consulted; figures are those reported by the original authors and are not comparable across rows because data, labels and tasks differ. 'Not established' means the record consulted did not state it.

TABLE II. Prior surveys relevant to this review

Survey	Scope	Relevance to this review
Ji et al. 2021 [2]	First comprehensive survey of suicidal-ideation detection: clinical approaches, feature-engineering and deep learning methods, applications by data source (questionnaires, health records, suicide notes, online content)	Establishes the method taxonomy; notes that many neural models lack interpretability, motivating the present focus
Chancellor and De Choudhury 2020 [3]	Systematic review of 75 studies (2013 to 2018) predicting mental-health status from social media; covers annotation, data quality, features and model verification	Identifies construct-validity concerns and proposes reporting standards, which the checklist in Section VII extends to explanation and missed cases

V. Critical Synthesis

A. Modelling approaches (RQ1)

The examined work spans linear and probabilistic models on lexical features [7], deep recurrent models with word embeddings and attention [6], and, in the wider literature, fine-tuned transformers [9]. Surveys describe the same trajectory from feature engineering to deep learning [2]. Two observations follow. First, more complex models do not automatically win: in the CLPsych 2019 system reported in [7], Naive Bayes was the strongest of three models on both binary tasks while an ensemble was strongest on the four-way task, and all four-way scores were low. Second, performance gains from richer models arrive together with a loss of inspectability, which is precisely where explanation is needed. Because Table I is a purposive sample, no frequency or time trend is asserted.

B. Data sources and labelling (RQ2)

The sources examined use Reddit posting behaviour and annotated risk levels [4], [5], [7], or microblog accounts linked to people who later died [6]. These label types measure different things. Posting in a support forum indicates help-seeking, not necessarily risk; annotated severity depends on annotator judgement and the rubric; outcome-linked labels capture a narrow and possibly atypical group. Chancellor and De Choudhury identify exactly this kind of operationalisation as a recurring weakness [3]. Shing et al. are notable for evaluating annotator reliability and comparing experts with crowd workers [4], a practice that should be standard. Data sharing is further constrained by ethics: the Reddit resource is distributed through a controlled application process [4].

C. Evaluation design (RQ3)

Reported measures in Table I include accuracy, F1 and macro-F1 [6], [7]. In the records consulted, author-grouped splitting and false-negative reporting were not established for most studies, which prevents a reader from judging whether scores are inflated by leakage or conceal missed cases. The CLPsych shared task is instructive in the other direction: by fixing training and test users, it avoids user overlap between partitions and permits comparison across 15 teams [5]. Even so, fine-grained risk prediction proved difficult, as the low macro-F1 in [7] shows, whereas coarser flagging was far more tractable. Aggregate accuracy of about 91% on a curated corpus [6] should not be read as expected field performance, because class balance, label construction and the absence of a prospective test all bear on it.

ISSN: 3139-1141

D. Explanation and interpretability (RQ4)

Among the primary studies examined, the only one with an explicit explanatory mechanism is the attention-based model [6], where attention weights were examined against a proxy (tree-hole posts). Agreement with a proxy is evidence that the weights track something related to risk, but it does not show that the weights are faithful to the model's decision, nor that they help a reviewer; given the cautions in [15]-[17], it is a starting point rather than a validation. None of the examined records describes a deletion or insertion test, a stability analysis, or a study of reviewers using the explanations. This is consistent with the observation in [2] that interpretability is a limitation of neural detectors, and it is the core of the black-box gap. These statements are limited to the sources examined; a systematic search could change the proportion.

VI. Discussion

A. Reporting an explanation versus testing it

The distinction that organises this review is between displaying an explanation and testing whether it is faithful and useful. Highlighted words are persuasive, and persuasion is not evidence. In a screening workflow this matters in both directions. A faithful, well-calibrated explanation lets a moderator see that a flag rests on, say, a statement of intent rather than an incidental keyword, and so allocate attention sensibly. An unfaithful one can create automation bias, where reviewers defer to a confident-looking system. Human oversight is therefore not a substitute for explanation testing but a reason for it: oversight is only as good as the information the overseer is given.

B. Why evaluation practice limits interpretation

Three weaknesses compound. Random splits can leak authorship, inflating scores [21]. Labels that proxy for risk may not transfer to clinically grounded settings [3], [23]. And aggregate metrics can hide missed cases. Each makes the reported number optimistic in a way that matters most for the people the system is meant to protect. The more fundamental point is that explanations inherit these defects: an explanation of a model that has learned an authorship or platform shortcut will faithfully describe the shortcut, and a reviewer may mistake it for clinical insight.

C. Shortcut learning

Text classifiers can rely on surface features that correlate with the label, such as post length, community-specific vocabulary or posting patterns [22]. Faithful explanations are one way to surface such shortcuts, since a highlighted token that is topically irrelevant is a warning sign; grouped splits, cross-community tests and controls for length are others. None of the examined records reports such a test, so whether the published scores partly reflect shortcuts cannot be determined from them.

D. Limitations of this review

This is a purposive narrative review, not a systematic one. The primary studies examined are few and were chosen as representative public resources, so the review cannot establish how common any practice is across the field. Characteristics were taken from abstracts and records and require confirmation against full texts, and data that a paper does not report may be reported elsewhere. The studies differ in language, platform, label and metric, so no pooled estimate or ranking was attempted. Findings should be read as a critical description of the examined literature, not a definitive account.

VII. Open Problems and Future Directions

Systematic extension. A reproducible search with stated databases, dates and strings, duplicate removal, screening counts and a flow diagram would turn this narrative review into a systematic one. **Shared reporting standards.** We propose the minimum checklist in Table III. **Author-grouped benchmarks.** Public datasets with author identifiers and timestamps would allow leakage to be controlled routinely, within ethical limits. **Faithfulness testing.** Explanations should be tested with deletion and insertion measures and checked for stability [18], [19], not only displayed. **Reviewer-centred evaluation.** Few studies test whether explanations help the people who review flagged posts; such studies should measure agreement, confidence, time and workload. **Multilingual and cross-platform work.** Examined resources concentrate on a few platforms and languages, limiting generalisation. **Clinical validation.** Studies linking text-based flags to clinical outcomes are needed before any claim of clinical utility.

TABLE III. Proposed minimum reporting checklist (authors' proposal, informed by [3], [18], [19], [21])

Item	What to report
Label source	How the at-risk label was produced (self-disclosure, community membership, annotation, outcome); annotator qualifications and agreement
Split strategy	Whether train, validation and test sets are disjoint by author and ordered in time; how duplicates and reposts were handled
Missed cases	Recall and false-negative rate for the at-risk class at the stated operating threshold, with precision and class prevalence
Shortcut checks	Length, platform and vocabulary controls; out-of-community or out-of-time evaluation
Explanation method	Method, settings and baseline; whether the model is intrinsically interpretable or explained post hoc

ISSN: 3139-1141

Item	What to report
Explanation evaluation	Faithfulness (deletion or insertion, comprehensiveness and sufficiency), stability across runs, and any reviewer study
Ethics and release	Ethics approval, anonymisation, data access terms, and whether examples are quoted

VIII. Ethical Considerations

Suicide-risk detection concerns sensitive personal information and high-stakes decisions. Research on social-media health data raises issues of consent, privacy and the risk of harm from publication, and ethical protocols for such research have been proposed [24]. Reviewed studies should be assessed for how they handled consent, anonymisation and the publication of examples; several resources are distributed only under controlled access [4]. A flagged post may cause distress to its author, and a missed post may leave a person without help. Any tool based on this work should be presented as a screening aid that requires trained human review, and any deployment requires ethics approval, clinical governance and a defined response pathway. This review reproduces no posts from any dataset and performed no new data collection.

IX. Conclusion

This structured narrative review examined text-based suicide-risk detection through the lenses of modelling, data and labels, evaluation and explanation. Within the sources examined, progress is real but evidence of trustworthiness lags behind reported scores: high accuracy on a curated corpus coexists with low macro-F1 on fine-grained risk, labels of uncertain construct validity, little evidence of author-grouped evaluation or missed-case reporting, and explanations, where present, that are not tested for faithfulness. The most important gap is therefore methodological rather than architectural. Before such systems can be compared fairly or considered for deployment, studies should adopt shared reporting standards, test explanations rather than display them, and evaluate with the reviewers who would use them. A systematic extension of this review, with a documented search, is the natural next step.

Acknowledgment

The authors thank Dr. Dinesh Banswal, project guide, and the Department of MCA, ZIBACAR, for guidance during the research project course.

References

[1] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," Biomed. Inform. Insights, vol. 10, 2018, Art. no. 1178222618792860, doi: 10.1177/1178222618792860.

[2] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Trans. Comput. Social Syst., vol. 8, no. 1, pp. 214-226, Feb. 2021, doi: 10.1109/TCSS.2020.3021467.

[3] 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.

[4] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daume III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proc. 5th Workshop Comput. Linguistics Clin. Psychol. (CLPsych), New Orleans, LA, USA, 2018, pp. 25-36, doi: 10.18653/v1/W18-0603.

[5] A. Zirikly, P. Resnik, O. 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, 2019, pp. 24-33, doi: 10.18653/v1/W19-3003.

[6] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, and X. He, "Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention," in Proc. EMNLP-IJCNLP, Hong Kong, China, 2019, pp. 1718-1728, doi: 10.18653/v1/D19-1181.

[7] V. Ruiz, L. Shi, W. Quan, N. Ryan, C. Biernesser, D. Brent, and R. Tsui, "CLPsych2019 shared task: Predicting suicide risk level from Reddit posts on multiple forums," in Proc. 6th Workshop Comput. Linguistics Clin. Psychol., Minneapolis, MN, USA, 2019.

[8] A. Vaswani et al., "Attention is all you need," in Adv. Neural Inf. Process. Syst. 30, 2017, pp. 5998-6008.

[9] 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.

[10] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

ISSN: 3139-1141

[11] Z. C. Lipton, "The mythos of model interpretability," Commun. ACM, vol. 61, no. 10, pp. 36-43, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. NAACL-HLT, 2019, pp. 3543-3556.

[16] S. Serrano and N. A. Smith, "Is attention interpretable?" in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 2931-2951.

[17] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in Proc. EMNLP-IJCNLP, 2019, pp. 11-20.

[18] 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.

[19] J. DeYoung et al., "ERASER: A benchmark to evaluate rationalized NLP models," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4443-4458.

[20] 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.

[21] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023.

[22] R. Geirhos et al., "Shortcut learning in deep neural networks," Nature Mach. Intell., vol. 2, pp. 665-673, 2020.

[23] S. K. Ernala et al., "Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals," in Proc. CHI Conf. Human Factors Comput. Syst., 2019, Paper 134.

[24] A. Benton, G. Coppersmith, and M. Dredze, "Ethical research protocols for social media health research," in Proc. 1st ACL Workshop Ethics NLP, 2017, pp. 94-102.