informatics

MDPI

Systematic Review

Artificial Intelligence for Detecting and Predicting Suicide-Related Outcomes on Social Media: A Systematic Review

Félix Díaz [1,*], Joseph H. Moscol-Rodriguez [2], Kosseth Otero-Krugger [2], Nhell Cerna [3] and Anibal Alviz-Meza [4]

[1] Escuela Profesional de Ingeniería de Sistemas, Facultad de Ingeniería y Arquitectura, Universidad Autónoma del Perú, Lima 15842, Peru
[2] Escuela Académico Profesional de Medicina Humana, Universidad Norbert Wiener, Lima 15306, Peru; a2017102030@uwiener.edu.pe (J.H.M.-R.); a2021101800@uwiener.edu.pe (K.O.-K.)
[3] Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima 15306, Peru; e14167@utp.edu.pe
[4] Escuela de Posgrado, Universidad Continental, Lima 15113, Peru; aalviz@continental.edu.pe
* Correspondence: felix.diaz@autonoma.pe

Abstract

Artificial intelligence (AI) is increasingly used to detect and predict suicide-related outcomes from social-media data, but the methodological basis and generalizability of this research remain heterogeneous. We conducted a systematic review combining bibliometric analysis with structured document-level synthesis of journal articles indexed in Scopus and Web of Science through 22 May 2026, following PRISMA 2020. A total of 178 studies were included using a 50-field extraction framework. Reddit (51.7%) and Twitter/X (39.3%) were the predominant platforms, while data geography was not reported in 70.8% of studies. Traditional machine learning remained the most frequent model family (69.1%), alongside deep learning (53.9%), Transformer-based models (42.7%), and large language models (9.0%). Suicide-related outcomes and ground-truth sources were heterogeneous. Among studies reporting usable metrics, median F1 was 0.8990 and median AUROC was 0.9480, but external validation was identified in only 26 studies (14.6%). Explainability or interpretability was reported in 58 studies (32.6%); source code was publicly or directly available in 22 (12.4%), and only 14 (7.9%) reported both reusable data and source code. These findings highlight heterogeneous ground-truth practices, limited documented external validation, and uneven reporting of interpretability, ethical procedures, and reusable data and code, constraining assessments of generalizability and reproducibility.

Keywords: artificial intelligence; natural language processing; suicide; suicidal ideation; social media; large language models; explainable artificial intelligence; external validation

Check for updates

Received: 14 September 2026
Revised: 5 November 2026
Accepted: 9 November 2026
Published: 11 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Social-media platforms provide posts, comments, conversations, user histories, and metadata from which linguistic, behavioral, and temporal indicators relevant to suicide-related research can be examined [1,2]. Such research addresses distinct constructs: suicidal ideation (thoughts about suicide), suicidal communication (verbal or nonverbal expressions related to suicide, including threats or plans), and suicide attempts (nonfatal, potentially injurious self-directed acts involving some intent to die) [3]. Suicide risk, in turn, refers to an inferred likelihood or level of vulnerability to suicidal behavior rather than a directly

observed event [4–6]. These constructs should not be conflated with the associated computational tasks: detecting suicide-related content or classifying contemporaneous risk labels does not, by itself, establish the ability to predict subsequent suicidal behavior. Artificial intelligence (AI), machine learning (ML), natural language processing (NLP), and related methods have been applied to detection, classification, screening, risk stratification, and prediction across posts, users, longitudinal histories, and population-level signals [2,4,6–8].

The computational methods used for these tasks have changed considerably over time. In this context, NLP refers to computational methods for processing and analyzing human language rather than a specific model family. Earlier studies relied largely on rule-based methods, manually engineered linguistic features, and traditional ML classifiers, while subsequent work increasingly incorporated deep-learning architectures, including Transformer-based language models that use self-attention to capture contextual relationships in text. More recently, large language models (LLMs), typically pretrained on extensive text corpora and often built on Transformer architectures, have been evaluated for classification, representation, reasoning, and generative applications [9]. Previous reviews have documented different stages of this methodological development. Reviews focused on ML identified the predominance of classification approaches and the growing use of more complex predictive models for suicidal-ideation detection [10,11]. Subsequent syntheses examined deep-learning methods applied to online social-media data [12], NLP-based approaches [13], and the use of ML and NLP specifically within Reddit-based suicide research [14]. More recent reviews have expanded the scope to broader applications of AI in suicide prevention and risk assessment [15,16] and to the emerging use of LLMs [9]. This progression reflects increasing methodological diversity rather than a single, stable modeling framework.

The literature has also been reviewed from the perspective of social-media data sources and their use for suicide-related assessment. Earlier work examined the role of social media in technology-based epidemiology and risk assessment [17], while other reviews focused on ML-based suicide-risk assessment using social networks [18] or on suicidal-ideation detection across social-media platforms [19]. More recent syntheses have continued to identify Reddit and Twitter/X as major sources of data and have raised concerns about the predominance of English-language content, limited cross-cultural representation, and the transferability of models across populations and platforms [12,16,20]. These concerns remain relevant as newer architectures are increasingly evaluated on datasets and platforms that have already been used extensively in earlier research.

Another methodological issue concerns the definition of the suicide-related outcome itself. Previous reviews have shown that social-media studies frequently operationalize suicidal ideation or suicide risk as classification problems, often using binary labels [10,18,19]. However, the labels used to train and evaluate these models can originate from different sources, including human annotation, clinical assessment, self-report measures, community membership, rule-based procedures, or labels inherited from existing datasets. Reviews focused on Reddit and digital suicidal-ideation detection have noted that suicide-related expressions identified in online content cannot generally be assumed to represent subsequent suicide attempts or deaths when clinical outcomes are unknown [14,20]. More recent work has similarly emphasized variation between clinically informed labels and heuristic or keyword-based approaches and the limited use of validated clinical frameworks [16]. Consequently, models described using similar terms may differ substantially in the construct they are trained to detect and in the relationship between their ground truth and clinically defined suicide-related outcomes.

Evaluation practices introduce another source of uncertainty. Earlier reviews reported generally high predictive performance for several ML and deep-learning approaches but

also identified methodological heterogeneity, possible overfitting, limited replication, and insufficient validation in independent samples or settings [10,21]. Concerns about external and real-world validation have persisted in more recent syntheses despite the adoption of more complex architectures [15,16,20]. Class imbalance is also recurrent in suicide-related social-media datasets, while precision–recall-oriented measures have been reported less consistently than measures such as accuracy, F1, or AUROC [11,12,21]. These differences in validation design, outcome definition, class distribution, and performance reporting complicate comparisons across studies and limit the extent to which strong results obtained within individual datasets can be interpreted as evidence of generalizability.

Explainability, ethical reporting, and reproducibility add further dimensions to the assessment of this literature. Previous reviews have identified interpretability as a persistent challenge, particularly as modeling has shifted toward deep neural and Transformer-based architectures [10,11,14]. Recent syntheses have discussed interpretability methods such as SHAP (SHapley Additive exPlanations), which estimates the contribution of input features to individual predictions, and LIME (Local Interpretable Model-agnostic Explanations), which explains individual predictions through simpler local approximations of model behavior [16]. At the same time, reviews of social-media and LLM-based suicide research have reported inconsistent documentation of ethical considerations, privacy, consent, and governance procedures [9,17,18,20]. Reproducibility has also been discussed in relation to incomplete methodological reporting, benchmark availability, and access to datasets or computational resources [9,15,18]. These issues are methodologically distinct but jointly affect the extent to which published findings can be interpreted, independently evaluated, and transferred to other research or application settings.

Taken together, existing reviews have established important findings about particular model families, data sources, platforms, and stages of technological development. Their scopes, however, differ substantially. Some focus primarily on ML or deep learning [10–12], others on specific platforms or methodological domains such as Reddit or NLP [13,14], and recent reviews have examined broader AI applications or LLMs [9,15,16]. This creates a need for an updated synthesis that examines the development of the field under a common study-level framework while jointly considering the provenance of social-media data, outcome and ground-truth construction, analytical units, AI tasks and model families, validation strategies, reported performance, class imbalance, explainability, ethical and privacy reporting, and the availability of data and source code.

Accordingly, this systematic review aimed to characterize the development, methodological structure, evaluation practices, and reporting characteristics of AI-based research for detecting and predicting suicide-related outcomes using social-media data. The review combined bibliometric analysis of temporal scientific production, author-country contributions, and international collaboration with structured document-level analysis of the included studies. The synthesis examined social-media platforms and datasets; geographic and linguistic characteristics; input modalities and analytical units; suicide-related outcomes and their operationalization; ground-truth and labeling procedures; AI tasks and model families; validation and external generalizability; reported performance and class imbalance; explainability; ethics and privacy; and the availability of data and source code.

2. Methodology

2.1. Review Design

This review combined a bibliometric analysis with a structured document-level analysis of the included studies. Study identification, screening, eligibility assessment, and inclusion followed the PRISMA 2020 framework [22]. The bibliometric component examined the temporal evolution of scientific production, author-country contributions, and

international collaboration networks. The document-level component used a structured data-extraction matrix to systematically collect methodological, computational, evaluation, and reporting information from the full texts of eligible studies. This systematic review was not prospectively registered, and a separate review protocol was not prepared.

2.2. Information Sources and Search Strategy

The literature search was conducted on 22 May 2026, using Scopus and Web of Science. No lower publication-date limit was applied, and records published up to 22 May 2026, were considered. The search strategy was structured into four conceptual groups addressing artificial intelligence and related computational methods, suicidal outcomes, social media and related online platforms, and detection or prediction tasks. Within each group, terms were combined using the Boolean operator OR, and the four conceptual groups were combined using AND. In Scopus, the search was applied to title, abstract, and keyword fields using TITLE-ABS-KEY; in Web of Science, it was applied to the Topic field using TS=.

The first group comprised terms related to artificial intelligence, machine learning, deep learning, natural language processing, neural networks, transformer-based models, large language models, and explainable artificial intelligence:

("artificial intelligence" OR "machine learning" OR "deep learning" OR "natural language process*" OR NLP OR "text mining" OR "sentiment analysis" OR "neural network*" OR "convolutional neural network*" OR CNN OR "recurrent neural network*" OR RNN OR LSTM OR transformer* OR BERT OR RoBERTa OR "language model*" OR "large language model*" OR LLM OR "explainable AI" OR XAI)

The second group addressed suicide risk and related suicidal outcomes:

("suicide risk" OR "suicidal risk" OR "suicidal ideation" OR "suicide ideation" OR "suicidal thought*" OR "suicidal behavior*" OR "suicidal behaviour*" OR "suicide attempt*" OR suicidality OR suicid* OR "self-harm" OR "self-injur*")

The third group identified social media, social networking platforms, online forums, microblogs, and specific platforms:

("social media" OR "social network*" OR "social networking site*" OR "online social network*" OR Twitter OR Reddit OR Instagram OR Facebook OR Tumblr OR TikTok OR YouTube OR Weibo OR "Sina Weibo" OR "online forum*" OR "discussion forum*" OR microblog*)

The fourth group contained terms corresponding to detection, prediction, classification, identification, screening, assessment, and risk stratification:

(detect* OR predict* OR classif* OR identif* OR screen* OR assess* OR "risk assessment" OR "risk prediction" OR "early detection" OR "automated detection" OR "risk classification" OR "risk stratification")

The search strategy was designed to capture variations in both suicide-related terminology and online platform types. The truncated term "suicid*" encompassed lexical variants such as “suicide” and “suicidal”, while “suicide attempt*” was explicitly included. To broaden platform coverage beyond individual services, the search combined platform-specific names with general descriptors, including “social media”, “social network*”, “online

forum*", "discussion forum*", and 'microblog*'. This approach reduced dependence on specific platform names, although retrieval remained contingent on the terminology used in indexed records.

The search retrieved 1264 records from Scopus and 583 from Web of Science, corresponding to 1847 records before deduplication.

2.3. Record Management and Study Selection

Records retrieved from Scopus and Web of Science were merged and deduplicated using the Bibliometrix package in R, version 4.6.1. After removing 478 duplicates, 1369 unique records remained (Figure 1). Before abstract screening, 16 non-English records and 857 records not indexed as Article or Article Early Access were excluded, leaving 496 records. Two reviewers screened the abstracts, excluding 286 records, and 210 reports were sought for full-text assessment. Seven reports could not be retrieved. The remaining 203 reports were assessed by two reviewers against the eligibility criteria described in Section 2.4, with disagreements resolved by a third reviewer at both screening stages.

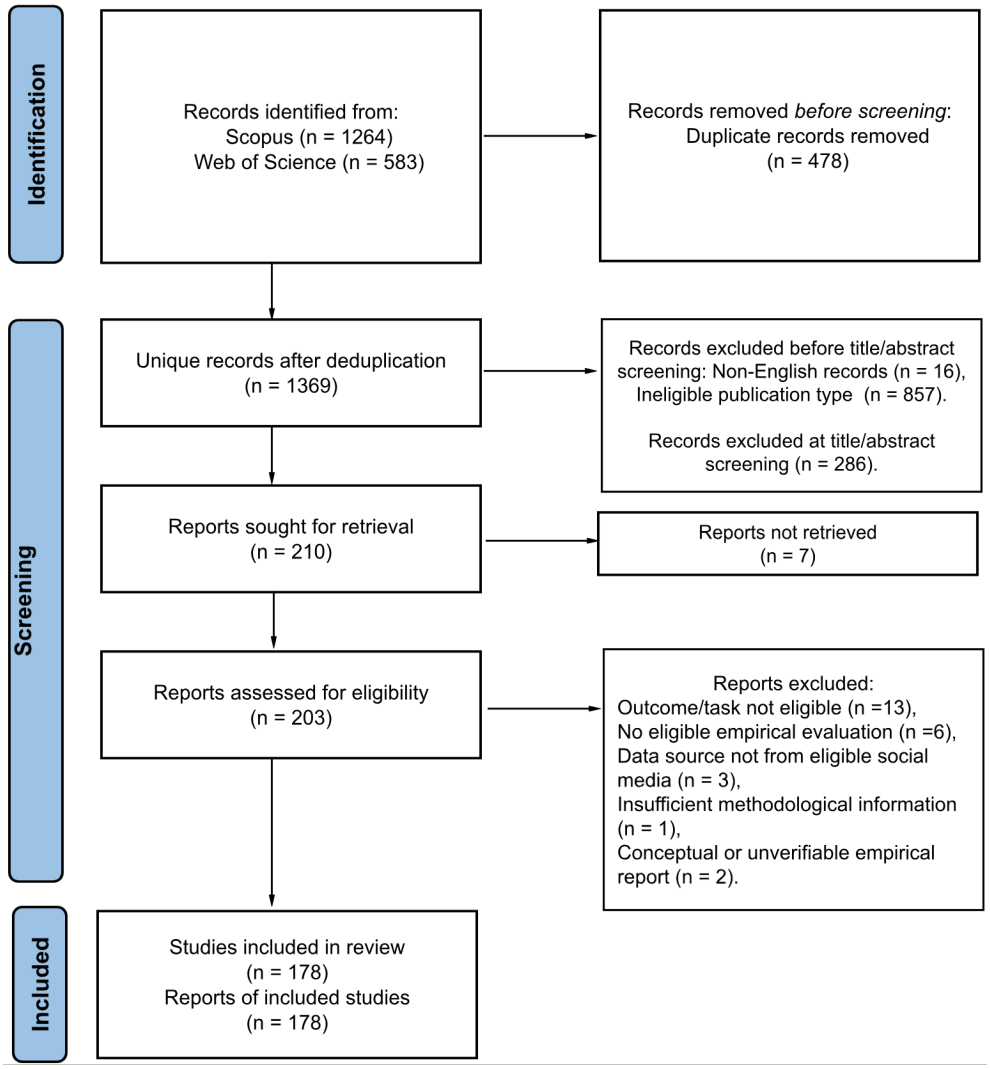


Figure 1. PRISMA flow diagram of the study selection process.

Twenty-five reports were excluded after full-text assessment because of ineligible suicide-related outcomes or computational tasks ($n = 13$), absence of an eligible empirical model evaluation ($n = 6$), ineligible social-media data sources ($n = 3$), insufficient methodological information ($n = 1$), or conceptual or unverifiable empirical reporting ($n = 2$).

The final synthesis included 178 studies. Study-level exclusion reasons are provided in Appendix A.

2.4. Eligibility Criteria

Studies were eligible when they met the following criteria:

- Publication characteristics.
  Original empirical studies indexed as Article or Article Early Access, written in English, available in full text, and published by 22 May 2026, were eligible. No lower publication-date limit was imposed; other document types were excluded.
- Social-media data.
  At least one principal input source had to comprise user-generated content, profiles, interactions, or metadata from social networks, online communities, forums, microblogs, or comparable interactive platforms. Eligible content included text, images, audio, video, and other user-generated information.
- Suicide-related outcome.
  At least one outcome had to explicitly concern suicidal ideation, intention, plans, attempts, behavior, risk, severity, suicide death, or suicidal communication relevant to risk identification. Community-membership proxies and rule-based indicators were not excluded solely because of limitations in ground-truth quality, provided that they were explicitly operationalized as suicide-related outcomes.
- Application of artificial intelligence.
  The study had to apply an AI or equivalent automated computational method, including machine learning, deep learning, NLP, Transformers, LLMs, or explainable AI, to empirical data. Studies merely proposing potential applications without empirical implementation were excluded.
- Suicide-related computational task.
  The computational approach had to detect, identify, classify, screen, stratify, assess, estimate, or predict a suicide-related outcome. AI methods used exclusively for unrelated auxiliary tasks were ineligible unless their outputs contributed to a computational model targeting suicidality.
- Identifiable unit of analysis.
  The analytical unit associated with the suicide-related model output had to be identifiable, including posts, conversations, users, longitudinal histories, communities, populations, or clearly specified multilevel structures.
- Empirical model evaluation.
  At least one model targeting a suicide-related outcome had to undergo task-appropriate empirical evaluation, such as predictive or discriminative performance assessment, comparison with reference approaches, internal or external validation, calibration or error analysis, time-to-event evaluation, or expert assessment.
  Association-only analyses reporting exclusively hazard ratios, odds ratios, regression coefficients, significance tests, in-sample R^2, topic associations, or similar explanatory statistics were insufficient unless the model's performance on the suicide-related task was also evaluated.
- Sufficient methodological reporting.
  The publication had to identify the principal data source, suicide-related outcome, computational approach, analytical unit, and evaluation procedure with sufficient detail to reconstruct the model-evaluation design.

No restrictions were imposed on participant age, sex, country, study population, or language of the analyzed social-media content. During full-text assessment, ambiguities concerning eligible empirical evaluations and minimum methodological reporting were

clarified through general decision rules. These clarifications were applied consistently to all 203 full-text reports before final eligibility decisions, without study-specific exceptions.

2.5. Data Extraction and Analysis

2.5.1. Bibliometric Analysis

Bibliometric data were processed in RStudio (version 2026.04.0) using the Bibliometrix package and its Biblioshiny interface [23]. The analysis examined the temporal evolution of scientific production, author-country contributions, and international collaboration networks. Author-country production was calculated at the author level, with each author contributing one count to the country associated with that author.

For the international collaboration analysis, the association method was used for network normalization, and communities were identified using the Louvain clustering algorithm.

2.5.2. Document-Level Data Extraction and Analysis

The 178 included reports were examined using a structured data-extraction matrix comprising 50 fields. A field-specific extraction instruction was defined for each variable to promote consistent interpretation and coding across studies.

Before full data extraction, the extraction framework was pilot-tested on a subset of eligible studies by two reviewers working independently. Their extraction decisions were compared to identify ambiguities, inconsistencies, and differences in the interpretation of individual fields. Based on this calibration process, operational definitions, allowable response categories, and field-specific decision rules were refined and standardized. The finalized extraction framework was subsequently applied to the full set of included studies. Studies used during the pilot phase were re-evaluated using the finalized extraction rules to ensure consistent coding across the entire corpus.

Data from all included studies were extracted independently by two reviewers. The resulting extraction records were compared, and discrepancies were resolved through discussion and consensus. When consensus could not be reached, a third reviewer adjudicated the decision. A formal pre-consensus inter-rater agreement coefficient was not calculated for the data-extraction stage; instead, extraction consistency was supported by the pilot calibration, independent duplicate extraction, field-specific coding rules, and consensus or third-reviewer adjudication procedures described above.

The extracted information covered study identification; social-media platform and dataset characteristics; data-collection period and geographical and linguistic context; input modality; sample size and unit of analysis; population characteristics; suicide-related outcome and its operational definition; ground truth, labeling procedures, annotators, clinical instruments, and inter-rater agreement; artificial intelligence task, models evaluated, model family, and best-performing model; validation strategy and external validation; performance metrics, including accuracy, precision, recall or sensitivity, specificity, F1-score, AUROC, and AUPRC; class imbalance and balancing methods; explainability and interpretability approaches; ethical and privacy considerations; methodological limitations; availability of data and source code; and the main empirical finding. A dedicated extraction-notes field was used to document ambiguous, incomplete, discrepant, or otherwise non-standard information requiring clarification during extraction or synthesis.

The complete study-level data-extraction matrix, including article identification information and all 50 extracted variables, is provided as Supplementary File S1.

2.5.3. Data Harmonization and Descriptive Synthesis

Extracted variables were standardized using an explicit mapping dictionary and harmonized at the study level according to outcome definitions, the principal evaluation context,

and extraction notes. Mutually exclusive categories were applied to dataset provenance, data geography, language, input modality, population, and principal outcome operationalization. The latter was classified as binary, nominal multiclass, ordinal, continuous, temporal/survival, or other/mixed, based on the principal suicide-related evaluation. Studies integrating distinct input modalities were classified as multimodal. Multilevel, non-standard, and unreported analytical units were grouped under "Other" in the overall study-characteristics summary, while multilevel designs remained distinguishable in outcome-by-unit analyses. Multiple categories per study were permitted for platforms, outcome types, ground-truth sources, AI tasks, modeling approaches, validation strategies, imbalance mitigation, and explainability.

AI tasks were harmonized into binary, multiclass, and ordinal classification; temporal prediction or early detection; detection, ranking, or information extraction; topic modeling; regression; generative or LLM-based tasks; and clustering. Model families were coded using all explicitly evaluated models, rather than only the principal or best-performing model. Non-mutually exclusive categories included traditional machine learning (e.g., support vector machines and random forests), deep learning (e.g., CNNs and recurrent neural networks), Transformer-based models (e.g., BERT and RoBERTa), LLMs (e.g., GPT and LLaMA), hybrid architectures, ensembles, and other approaches. Transformers and LLMs were treated as distinct categories rather than automatically grouped under deep learning. A family was counted only when its models were explicitly evaluated within the predictive architecture; use solely for feature or embedding generation did not qualify. Transfer learning and multi-task learning were treated as training paradigms rather than model families, whereas hybrid and ensemble categories could coexist with their constituent model families.

Validation strategies were classified as hold-out or data-split evaluation, cross-validation, repeated resampling or bootstrap, temporal or out-of-time validation, other procedures, or insufficiently specified validation. External validation required evaluation on data independent of model development and tuning. Its non-mutually exclusive categories distinguished differences in dataset, platform, population, language, geography, collection period, institution or cohort, and clinical setting.

Ground-truth sources were classified as inherited dataset labels, non-clinical human annotation, clinical expert annotation or assessment, community- or platform-based proxies, instrument- or self-report-based measures, rule-, keyword-, or pseudo-label-based approaches, documented events or administrative records, other sources, or not reported. Inherited labels were retained as a separate provenance category even when the original labeling mechanism was identifiable. Direct administration of clinical or psychometric instruments was distinguished from their use as annotation frameworks, severity taxonomies, or classification criteria.

Class imbalance was recorded when identified in the original dataset, collection process, or analytical configuration, including cases subsequently balanced for modeling. Mitigation strategies comprised oversampling or SMOTE, data augmentation or synthetic generation, undersampling, class weighting or loss modification, balanced sampling or batches, and other approaches. Explainability was grouped into post-hoc XAI, feature-based or descriptive, attention-based, intrinsic or knowledge-based approaches, LLM-generated explanations, and other methods. Ethics and privacy reporting covered ethics review, consent, public-data status, anonymization, privacy, terms of service, and access restrictions. Data and code availability were classified as publicly or directly available, partially or conditionally available, unavailable, or not reported.

Following pilot calibration, a predefined decision hierarchy was applied to identify one principal model–evaluation context per study whenever unambiguous. External validation of the same suicide-related outcome and task was prioritized, followed by an independent

internal test set and then cross-validation or the principal internal evaluation. Within the selected context, the authors' principal or best-performing model was retained when this designation was consistent with the reported results. When competing models performed best on different metrics without a defined primary metric, or multiple evaluation contexts were equally eligible, no selection was made solely on numerical superiority; the ambiguous item was coded as NR and documented in the extraction notes.

Available performance metrics were extracted from the same selected model–evaluation context, without combining results from different models, datasets, or evaluation settings or using training-set performance. For binary outcomes, precision, recall or sensitivity, and F1 were preferentially extracted for the explicitly identified suicide-related positive class; for multiclass outcomes, the principal global or averaged values reported by the authors were retained. Averaging conventions were recorded when available.

NR denoted applicable information that was unreported or indeterminate, whereas NA indicated non-applicability. Percentages were converted to proportions when necessary, and NR/NA values were excluded from metric-specific distributions. Categorical findings were summarized using study counts and percentages, with non-mutually exclusive categories potentially exceeding 100% in total. Performance values were described using medians, interquartile ranges (25th–75th percentile), and observed ranges. No pooled performance estimates were calculated because of heterogeneity in outcomes, ground truth, datasets, analytical units, metric definitions, and validation settings.

No formal study-level risk-of-bias, methodological-quality, reporting-bias, or certainty-of-evidence assessment was conducted. Extracted methodological characteristics were synthesized descriptively, without formal sensitivity analyses, statistical heterogeneity testing, subgroup analyses, or meta-regression.

3. Results

3.1. Bibliometric Characteristics

Publication activity expanded considerably from 2020 onward, reaching its highest annual output in 2024 with 43 articles, followed by 28 in 2025, as displayed in Figure 2. An additional 19 articles were identified in 2026, although this count covers only publications available through 22 May. Overall, 178 studies were included. This growth reflects increasing research activity but does not, by itself, indicate improvements in methodological rigor or the generalizability of the reported models.

Scientific production involved 43 countries and 645 author–country contributions, as shown in Figure 3. China ($n = 132$), the United States ($n = 122$), and India ($n = 63$) collectively accounted for 317 contributions (49.1%). These figures represent author-level country contributions rather than distinct publications or the geographic origin of the analyzed social-media data. Consequently, the geographic distribution of research output should not be interpreted as evidence of equivalent geographic representation among the populations or datasets investigated.

The international coauthorship network revealed several interconnected collaboration groups, with the United States, China, and India among its most prominent nodes, as displayed in Figure 4. Connections extended beyond these highly productive countries to smaller national research communities, indicating an internationally distributed collaboration structure. Nevertheless, international coauthorship does not necessarily imply multinational data collection or cross-population model validation. The distinction is relevant when assessing whether increasing scientific collaboration has been accompanied by broader empirical coverage.

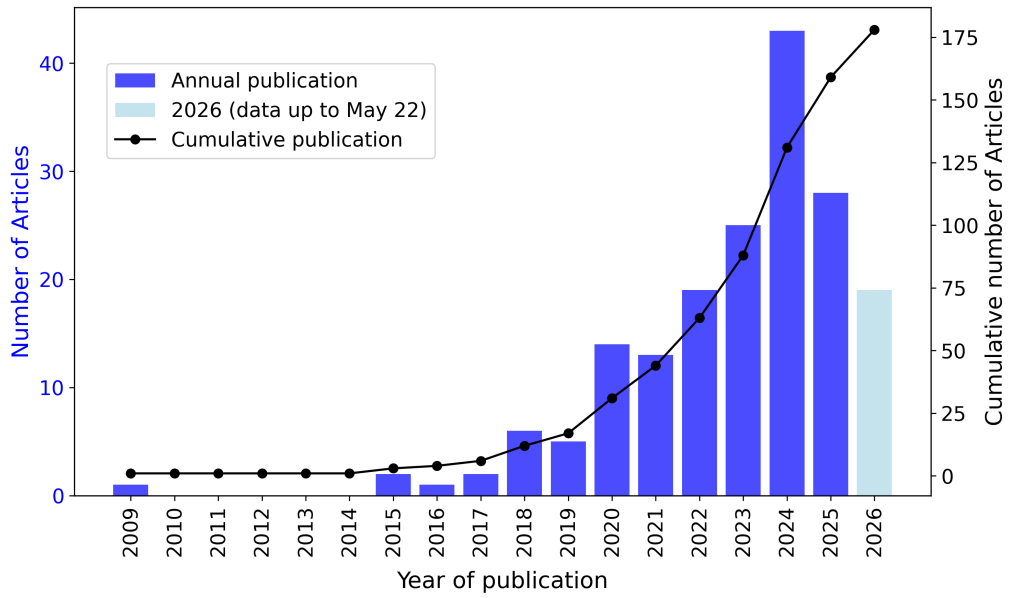


Figure 2. Annual and cumulative scientific production from 2009 to 2026. The 2026 value includes publications available up to 22 May 2026.

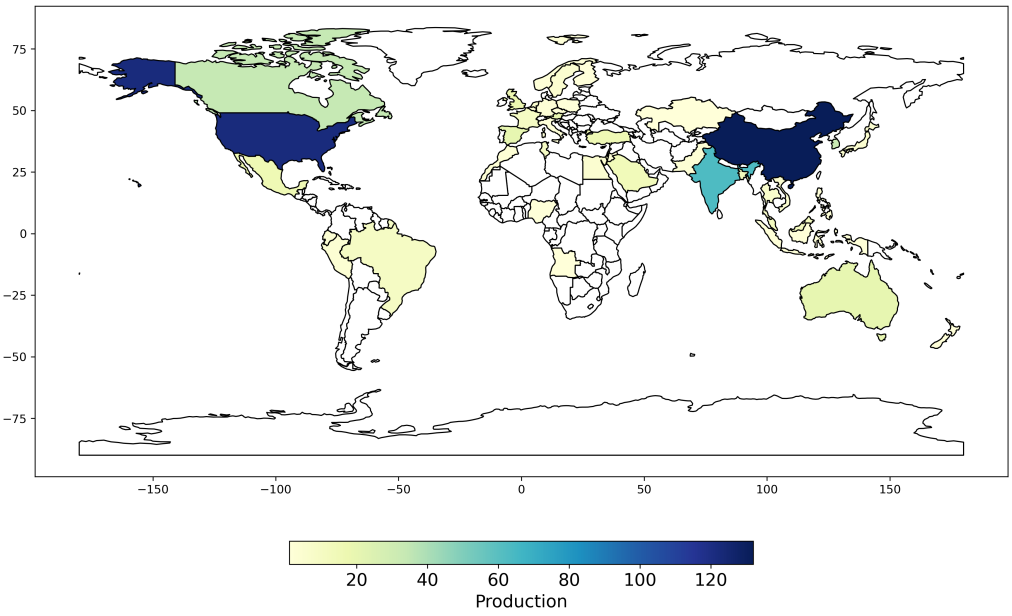


Figure 3. Geographic distribution of scientific production by country. Production represents author-level country contributions rather than the number of articles; each author contributes one count to the corresponding country.

3.2. Data Sources and Study Characteristics

The characteristics of the data sources and study populations are summarized in Table 1. Reddit (n = 92, 51.7%) and Twitter/X (n = 70, 39.3%) were the predominant platforms, followed by Sina Weibo (n = 23, 12.9%), Facebook (n = 12, 6.7%), and Instagram (n = 9, 5.1%). Other platforms included less frequently represented online communities, forums, and social-media services. Platform categories were non-mutually exclusive because studies using multiple platforms contributed to each applicable category.

Dataset provenance was almost evenly divided between original datasets (n = 82, 46.1%) and existing or reused datasets (n = 83, 46.6%), while 11 studies (6.2%) combined both sources. Geographic provenance was substantially less documented: 126 studies (70.8%) did not report the geographic origin of the analyzed data, whereas 50 (28.1%) reported a single country or region and only two (1.1%) involved multiple countries or a global context. This distinction concerns the origin of the analyzed data rather than the institutional affiliations of the authors.

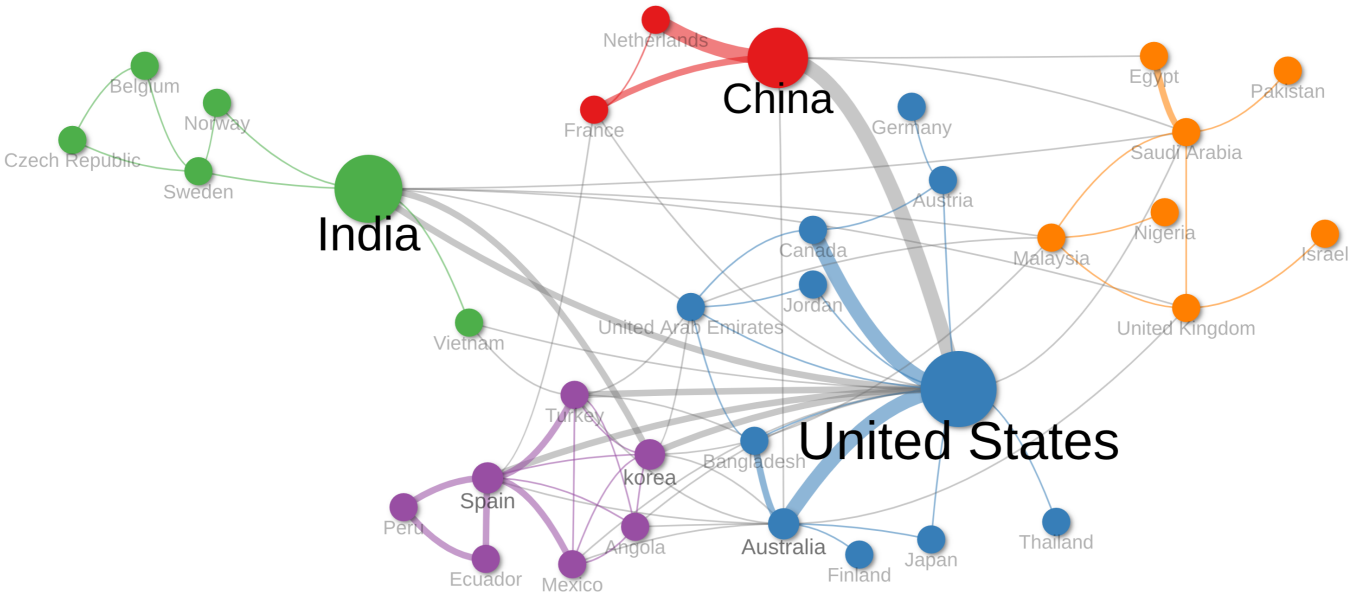


Figure 4. International collaboration network based on country co-authorship. Node size reflects the relative weight of each country in the network visualization, links represent co-authorship connections between countries, and node colors indicate the collaboration groups identified in the network.

Table 1. Data sources and characteristics of the included studies.

Domain	Category	*n*	%	Studies
Platform	Reddit	92	51.7	[4,5,8,24–112]
	Twitter/X	70	39.3	[1,2,4,6,7,26,27,29,38,42,52,59,60,62,64,65,67,70,71,83,85,89,90,94,95,100,101,104,107,113–153]
	Sina Weibo	23	12.9	[63,86,154–174]
	Facebook	12	6.7	[67,83,103,104,129,135,150,175–179]
	Instagram	9	5.1	[4,129,132,135,150,180–183]
	Other specified platforms	18	10.1	[8,27,35,57,58,89,103,119,129,163,184–191]
	Unspecified (NR)	3	1.7	[192–194]
Dataset provenance	Original/own dataset	82	46.1	[1,2,4–7,37,44,53,56,61,62,70,81,82,89,103,109,113,115–119,121–125,127,130–139,143–153, 155–157,159–162,164–179,183,184,186–191]
	Existing/reused dataset	83	46.6	[8,24–36,38,41–43,45–52,54,55,57–60,64–68,71–76,78–80,83–85,87,88,90–102,105,107,108, 110–112,114,120,126,128,129,141,154,163,180–182,192,193]
	Original and existing data	11	6.2	[39,40,63,69,77,86,104,106,142,158,185]
	NR	2	1.1	[140,194]
Data geography	Single country/region reported	50	28.1	[2,4,6–8,52,57–59,63,86,115,125,131–133,135,138,143,147,154–172,174,176,177,180–187]
	Multiple countries/global	2	1.1	[39,190]
	NR	126	70.8	[1,5,24–38,40–51,53–56,60–62,64–85,87–114,116–124,126–130,134,136,137,139–142,144–146, 148–153,173,175,178,179,188,189,191–194]
Language	English	67	37.6	[1,7,26,28,29,32,36,39,41,46,49,51,56–60,64,69,73,74,78,81–85,87–102,105–108,110–112,121, 122,130,134,138–142,144–147,176,181,183,184]
	Chinese	22	12.4	[154–174,185]
	Spanish	4	2.2	[4,27,65,148]
	Multilingual	7	3.9	[40,52,63,86,104,115,129]
	Other	17	9.6	[2,103,118,123,125,132,133,135–137,143,180,182,186,187,189,194]
	NR/NA	61	34.3	[5,6,8,24,25,30,31,33–35,37,38,42–45,47,48,50,53–55,61,62,66–68,70–72,75–77,79,80,109,113, 114,116,117,119,120,124,126–128,131,149–153,175,177–179,188,190–193]
Modality	Text	167	93.8	[1,2,5–8,24–62,64–127,129,130,133,134,136–149,151–177,180–182,184–187,189–194]
	Image	1	0.6	[178]
	Multimodal	5	2.8	[4,63,150,179,183]
	Other	5	2.8	[128,131,132,135,188]

Table 1. *Cont.*

Domain	Category	*n*	%	Studies
Unit of analysis	Post/document	113	63.5	[1,2,5,7,8,24–27,29,31,34–44,46,48–51,53,56–58,61,62,64,68,70–73,75,76,79–83,87–94,96–98,102,103,105–107,109,110,112–118,121–126,129,130,133,134,136,137,139–146,148,149,151–154,159–161,164,165,167,168,172,184–186,189,191–194]
	User	34	19.1	[33,45,55,59,60,63,66,74,99,108,111,119,128,132,135,147,150,156–158,162,166,169,173–176,178–180,182,183,188,190]
	User longitudinal	13	7.3	[4,6,30,47,78,84,120,127,155,163,170,171,177]
	Sentence/segment	3	1.7	[28,65,67]
	Population/time series	1	0.6	[131]
	Other	14	7.9	[32,52,54,69,77,85,86,95,100,101,104,138,181,187]
Population	General social-media users	82	46.1	[1,2,7,24,26,27,29,31,34,38,42,50,52,53,59,60,62–64,68,71,76,77,79,80,82,85,86,91,93,97,99,100,102–104,107,112–114,116–118,120–128,130,133,134,136–141,143–146,148,149,151–153,157–159,163,165,175,179,185,186,189,192,193]
	Suicide-related communities	54	30.3	[5,8,25,30,33,36,37,40,41,43–49,51,54–56,61,66,70,72–75,78,81,84,87,88,92,96,98,101,105,106,108–111,154,156,160,161,164,166–168,171–173,188]
	Clinical/identified population	20	11.2	[4,6,39,57,58,115,119,147,150,155,162,169,170,174,176–178,180,182,183]
	Specific demographic group	6	3.4	[132,135,181,184,190,191]
	Other/NR	16	9.0	[28,32,35,65,67,69,83,89,90,94,95,129,131,142,187,194]

Note. Percentages were calculated using the total number of included studies (N = 178) as the denominator. The "Studies" column identifies the studies contributing to each category. Categories are mutually exclusive within each domain except for platforms, where studies could contribute to multiple categories. "Other specified platforms" groups additional online communities and interactive sources, with each study counted once; Sina Weibo includes Sina Microblog. Dataset provenance distinguishes exclusively original, exclusively reused, and mixed data sources. Data geography refers to the origin of the analyzed data rather than author affiliations. "Multilingual" denotes more than one language, while "Other" groups less frequent languages. Language NR/NA combines unreported and non-applicable cases. Modality reflects the primary model input, and "Other" analytical units include multilevel, nonstandard, and unreported cases. NR, not reported; NA, not applicable.

Dataset reuse was concentrated in a limited number of recurrent resources, as displayed in Figure 5. The Suicide and Depression Detection dataset appeared in 14 studies [36,48,68,72,87,88,90,93,94,97,98,102,105,110], followed by the UMD/CLPsych Reddit Suicidality Dataset in 13 [8,33,45,54,57,59,60,73,74,86,99,108,111]. The C-SSRS/Gaur Reddit suicidality dataset and the Twitter suicidal-intention/ideation dataset were each identified in seven studies. These patterns indicate substantial reuse of established resources alongside the development of original datasets.

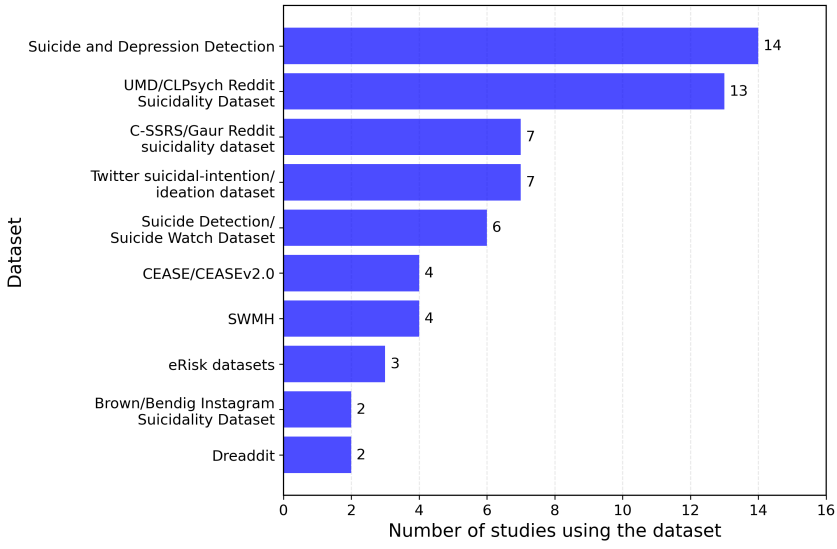


Figure 5. Ten most frequently reused datasets and corpora across the included studies. Counts represent the number of studies using each normalized resource; studies using multiple datasets could contribute to more than one category.

Most studies analyzed individual posts or documents (n = 113, 63.5%), followed by user-level (n = 34, 19.1%) and longitudinal user-level analyses (n = 13, 7.3%). Sample-size reporting varied according to the analytical unit: user counts were not reported in

119 studies, while post or document counts were not reported in 38. These quantities were retained separately because users, posts, and derived analytical instances do not represent interchangeable sampling units. Regarding study populations, general social-media users predominated ($n = 82$, 46.1%), followed by suicide-related communities ($n = 54$, 30.3%) and clinical or otherwise identified populations ($n = 20$, 11.2%).

English was the most frequently reported corpus language ($n = 67$, 37.6%), followed by Chinese ($n = 22$, 12.4%). Only seven studies (3.9%) analyzed multilingual data, while language was not reported or was not applicable in 61 (34.3%). Text was the predominant input modality ($n = 167$, 93.8%), whereas multimodal inputs were examined in only five studies (2.8%). Other linguistic and modality categories are detailed in Table 1.

3.3. Suicide-Related Outcomes, Operationalization, and Ground Truth

Suicidal ideation was the most frequently modeled outcome ($n = 87$), followed by suicide risk ($n = 49$), suicidal communication or content ($n = 22$), and composite or other suicide-related constructs ($n = 15$). Suicide attempts and suicide deaths were each examined in six studies [39,44,52,119,131,132,135,138,160,170,187,188], while suicidal behavior and suicide plans were less frequently represented. These outcome categories were not mutually exclusive.

The analytical unit varied across constructs, as shown in Figure 6a. Suicidal ideation was predominantly examined at the post or document level ($n = 59$), whereas suicide risk was distributed across post-level ($n = 22$), user-level ($n = 16$), longitudinal, and multilevel analyses. All 22 studies addressing suicidal communication or content operated at the post or document level. Suicide-death studies involved several analytical units, including population-level time series. Thus, studies addressing apparently similar suicide-related outcomes did not necessarily represent equivalent computational or inferential tasks.

Ground-truth sources were similarly heterogeneous, as displayed in Figure 6b. Inherited dataset labels were identified in 70 studies, non-clinical human annotation in 61, community- or platform-based proxies in 26, and clinical expert annotation or assessment in 25. Instrument- or self-report-based measures and rule-, keyword-, or pseudo-label-based approaches were each identified in 18 studies, while documented events or administrative records were used in 14. These categories were not mutually exclusive. Notably, all six studies examining suicide death used documented events or administrative records as ground truth [39,119,131,138,170,187], although one also incorporated non-clinical human annotation [39].

Outcome operationalization and labeling characteristics are summarized in Table 2. Binary formulations predominated ($n = 126$, 70.8%), followed by nominal multiclass ($n = 22$, 12.4%) and ordinal outcomes ($n = 21$, 11.8%). Continuous ($n = 5$), temporal or survival-based ($n = 2$), and other or mixed operationalizations ($n = 2$) were less frequent. Importantly, similar binary formulations could represent distinct constructs, ranging from suicidal content identification to user-level risk classification. The number of output classes therefore does not, by itself, establish conceptual equivalence between studies.

Table 2. Outcome operationalization and ground-truth reporting characteristics of the included studies.

Domain	Category	*n*	%	Studies
Outcome operationalization	Binary	126	70.8	[2,4,5,24,25,27–29,31,33,34,36–44,46,48–52,55–57,59–61,63,64,68–72,74–77,80–82,85–88,90–95,97–103,105–107,110–112,116–127,129,132,134–137,140,141,144,147–154,156–158,160,163,165,167–170,172,174,176–186,188,189,192–194]
	Multiclass nominal	22	12.4	[1,7,26,35,53,58,65,67,83,89,96,104,114,115,130,133,139,142,145,146,175,191]
	Ordinal	21	11.8	[8,30,32,45,47,54,62,66,73,78,79,84,108,109,113,128,143,155,161,164,171]
	Continuous	5	2.8	[6,131,162,166,173]
	Temporal/survival	2	1.1	[138,187]
	Other/mixed	2	1.1	[159,190]

Table 2. *Cont.*

Domain	Category	*n*	%	Studies
Clinical instrument use	Administered to users/participants	11	6.2	[6,132,147,150,162,169,174,176–178,183]
	Used only as annotation framework/taxonomy	15	8.4	[28,30,47,49,66,78,79,84,85,104,115,142,143,155,171]
	NR	1	0.6	[135]
	NA	151	84.8	[1,2,4,5,7,8,24–27,29,31–46,48,50–65,67–77,80–83,86–103,105–114,116–131,133,134,136–141,144–146,148,149,151–154,156–161,163–168,170,172,173,175,179–182,184–194]
Annotator profile	Profile reported	73	41.0	[1,2,4,5,7,8,28,30,33,37,39,40,44,45,47,49,51,52,54,57,59,60,66,74,77,78,81,82,84,86,91,103,104,108,109,113–118,120,123,125,127,130,134,136,138,143,149,152,154–161,163–165,167,168,170–172,181,184–186,191]
	NR	50	28.1	[24,26,27,31,32,34,43,55,58,62,65,69,70,73,79,80,85,87–90,92–95,98–100,107,110–112,121,122,124,129,133,137,139–141,144,145,148,151,153,179,188,192,194]
	NA	55	30.9	[6,25,29,35,36,38,41,42,46,48,50,53,56,61,63,64,67,68,71,72,75,76,83,96,97,101,102,105,106,119,126,128,131,132,135,142,146,147,150,162,166,169,173–178,180,182,183,187,189,190,193]
Number of annotators	Reported	53	29.8	[1,2,4,5,7,8,28,30,33,37,40,44,47,51,52,66,78,81,82,84,86,91,103,104,109,113–115,117,118,123,125,130,134,136–138,143,152,154–158,160,165,167,168,170,171,185,186,191]
	NR	70	39.3	[24,26,27,31,32,34,39,43,45,49,54,55,57–60,62,65,69,70,73,74,77,79,80,85,87–90,92–95,98–100,107,108,110–112,116,120–122,124,127,129,133,139–141,144,145,148,149,151,153,159,161,163,164,172,179,181,184,188,192,194]
	NA	55	30.9	[6,25,29,35,36,38,41,42,46,48,50,53,56,61,63,64,67,68,71,72,75,76,83,96,97,101,102,105,106,119,126,128,131,132,135,142,146,147,150,162,166,169,173–178,180,182,183,187,189,190,193]
Inter-rater agreement	Agreement information reported	44	24.7	[1,4,5,7,8,33,40,44,52,59,62,66,74,78,81,82,86,91,99,103,109,113–115,120,123,125,129,130,136,138,143,149,152,154,155,165,168,171,181,184–186,191]
	NR	73	41.0	[2,24,26–28,30–32,34,39,43,45,47,49,51,54,55,57,58,60,65,69,70,73,77,79,80,84,85,87,89,92,95,98,100,104,107,108,110–112,116–118,121,122,124,127,133,134,137,139–141,144,145,148,151,153,156–161,163,164,167,170,172,179,188,192]
	NA	61	34.3	[6,25,29,35–38,41,42,46,48,50,53,56,61,63,64,67,68,71,72,75,76,83,88,90,93,94,96,97,101,102,105,106,119,126,128,131,132,135,142,146,147,150,162,166,169,173–178,180,182,183,187,189,190,193,194]

Note. Percentages use the total corpus (N = 178) as the denominator. Categories are mutually exclusive within each domain. Outcome operationalization reflects the principal suicide-related evaluation context retained for each study. Direct administration of clinical instruments to participants was distinguished from their use solely as annotation frameworks or taxonomies. Annotator profiles were considered reported when professional background or role was specified; participants providing self-reported outcomes were not treated as annotators. For annotation-related fields, NR indicates applicable information that was not reported, whereas NA indicates that human annotation was not applicable. Inter-rater agreement refers to statistics or explicit agreement criteria reported for annotation within the included studies. NR, not reported; NA, not applicable.

Clinical and psychometric instruments played distinct roles in ground-truth construction. They were administered directly to users or participants in only 11 studies (6.2%), whereas 15 (8.4%) used an instrument solely as an annotation framework, severity taxonomy, diagnostic reference, or classification criterion. The C-SSRS (n = 14) illustrates this distinction: some studies administered it directly to participants [132,176,178], while others used its categories to annotate social-media content without direct participant assessment [30,47,49,66,78,79,84,85,143,155,171]. Consequently, the use of a named clinical instrument should not automatically be interpreted as evidence of clinically assessed ground truth.

Reporting of human annotation procedures was also incomplete. Annotator background or role was reported in 73 studies (41.0%), the number of annotators in 53 (29.8%), and inter-rater agreement information in 44 (24.7%). When applicable, these characteristics were not reported in 50, 70, and 73 studies, respectively. Reported agreement procedures included kappa statistics [81,82], Krippendorff's alpha [78,81,168], intraclass correlation coefficients [165], percentage agreement [33,82,168], and explicit agreement thresholds [7,149,152]. Incomplete reporting limits the ability to assess the consistency and comparability of labeling procedures, but does not establish that those procedures were inadequately implemented.

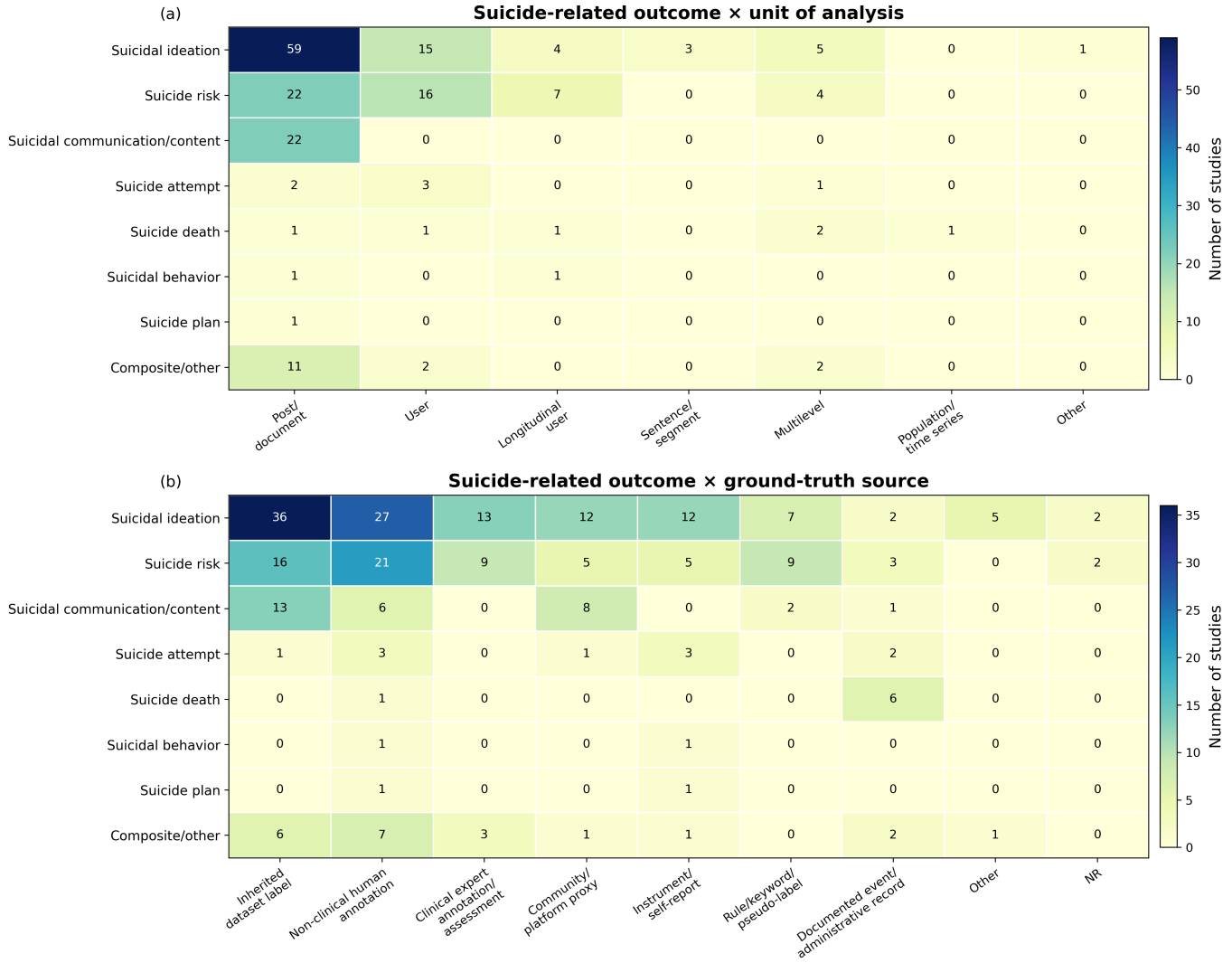


Figure 6. Suicide-related outcomes by unit of analysis and ground-truth source across the included studies. Panel (**a**) shows the number of studies for each combination of suicide-related outcome and unit of analysis. Panel (**b**) shows the corresponding distribution by ground-truth source. Studies reporting multiple suicide-related outcomes or more than one ground-truth source contribute to each applicable category; therefore, counts are not mutually exclusive across rows or ground-truth columns. For panel (**a**), multilevel designs were displayed as a separate analytical category, whereas they were grouped under "Other" in the overall study-characteristics summary in Table 1.

3.4. AI Tasks and Modeling Approaches

Classification dominated the computational tasks evaluated across the included studies, as presented in Figure 7a. Binary classification was the most frequent (n = 127, 71.3%), followed by multiclass classification (n = 27, 15.2%) and ordinal classification or stratification (n = 22, 12.4%). Temporal prediction or early detection was identified in 18 studies (10.1%), detection, ranking, or information extraction in 14 (7.9%), topic modeling in 7 (3.9%), and regression in 6 (3.4%). Generative or LLM-based tasks (n = 4, 2.2%) [32,33,99,108] and clustering (n = 3, 1.7%) [61,132,134] were less frequent. Task categories were not mutually exclusive because individual studies could address multiple computational objectives.

The evaluated model families also showed substantial diversity, as shown in Figure 7b. Traditional machine learning was represented in 123 studies (69.1%), deep learning in 96 (53.9%), and Transformer-based models in 76 (42.7%). Hybrid architectures (n = 58, 32.6%) and ensembles (n = 40, 22.5%) were also common, while LLMs were evaluated in 16 studies (9.0%). These model-family categories were non-mutually exclusive, so the

percentages reflect the proportion of included studies evaluating each family and are not expected to sum to 100%. Importantly, LLM use was not limited to generative applications: although 16 studies evaluated LLM-family models, only four were classified under the generative or LLM-based task category [32,33,99,108]. Other studies incorporated LLMs into classification, representation, or hybrid workflows [24,26–28,30,31,65,71,85,154,155].

The temporal distribution indicates methodological diversification rather than replacement of earlier approaches, as displayed in Figure 7c. Deep learning first appeared among the included studies in 2018 [70,117], Transformer-based models in 2020 [8], and LLMs in 2024 [24,27,65,99,108]. By 2025, Transformer-based and deep-learning models were each evaluated in 75.0% of the 28 included studies published that year, while traditional machine learning remained represented in 53.6%. These modeling families continued to coexist among studies published through 22 May 2026. The observed progression therefore reflects an expanding methodological repertoire rather than convergence toward a single model family.

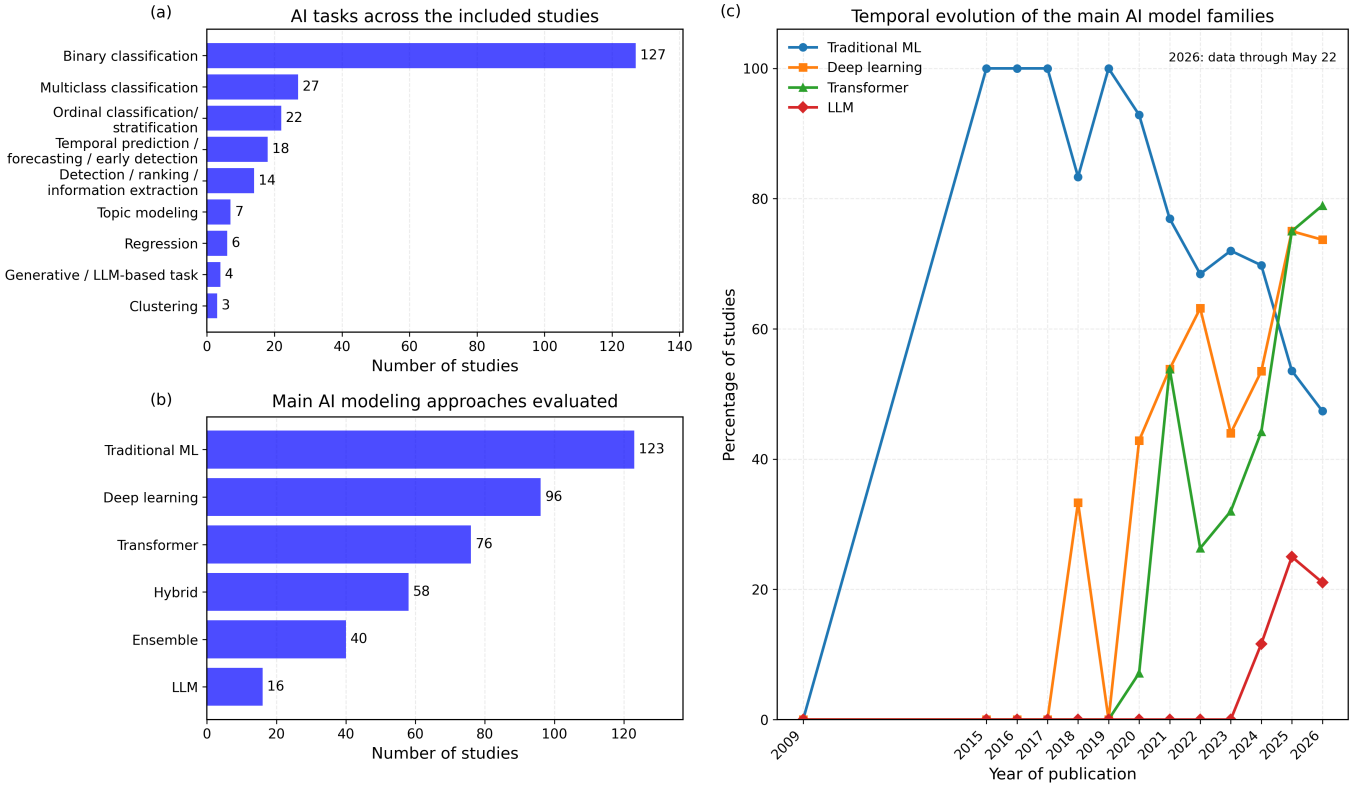


Figure 7. AI tasks and evolution of modeling approaches across the included studies. Panel (**a**) presents normalized AI task categories, panel (**b**) the evaluated model families, and panel (**c**) their annual distribution. Categories in panels (**a**,**b**) are non-mutually exclusive. Annual percentages in panel (**c**) use the number of included studies published in each year as the denominator. Other or non-standard modeling approaches are not displayed separately in panel (**b**). Data for 2026 include publications available through May 22.

3.5. Validation, Generalizability, and Reported Performance

Evaluation predominantly relied on conventional data-splitting and resampling procedures (Figure 8a). Hold-out or train–test-based evaluation was identified in 121 studies (68.0%), cross-validation in 77 (43.3%), and repeated resampling or bootstrap procedures in 19 (10.7%). Temporal or out-of-time validation strategies were identified in six studies (3.4%) [7,131,138,155,171,187], while validation procedures were insufficiently specified in 23 (12.9%). These categories were not mutually exclusive.

Independent external validation was identified in only 26 of the 178 studies (14.6%) [2,5–8,29,42,51,52,56,57,59,60,63,79,85,86,89,100,104,120,145,158,160,163,187]. Among these studies, cross-dataset validation was most frequent (n = 19, 73.1%), followed by cross-platform validation (n = 7, 26.9%) and temporal external validation (n = 6, 23.1%) [7,56,63,145,160,187]. Cross-population validation was identified in four studies [2,5,52,57], whereas cross-language [52,86] and independent clinical validation [6,57] were each reported in two. Cross-geographic validation [52] and independent institutional or cohort validation [6] were each identified in only one study. These external-validation types were not mutually exclusive, as independent evaluation datasets could differ along multiple dimensions (Figure 8b).

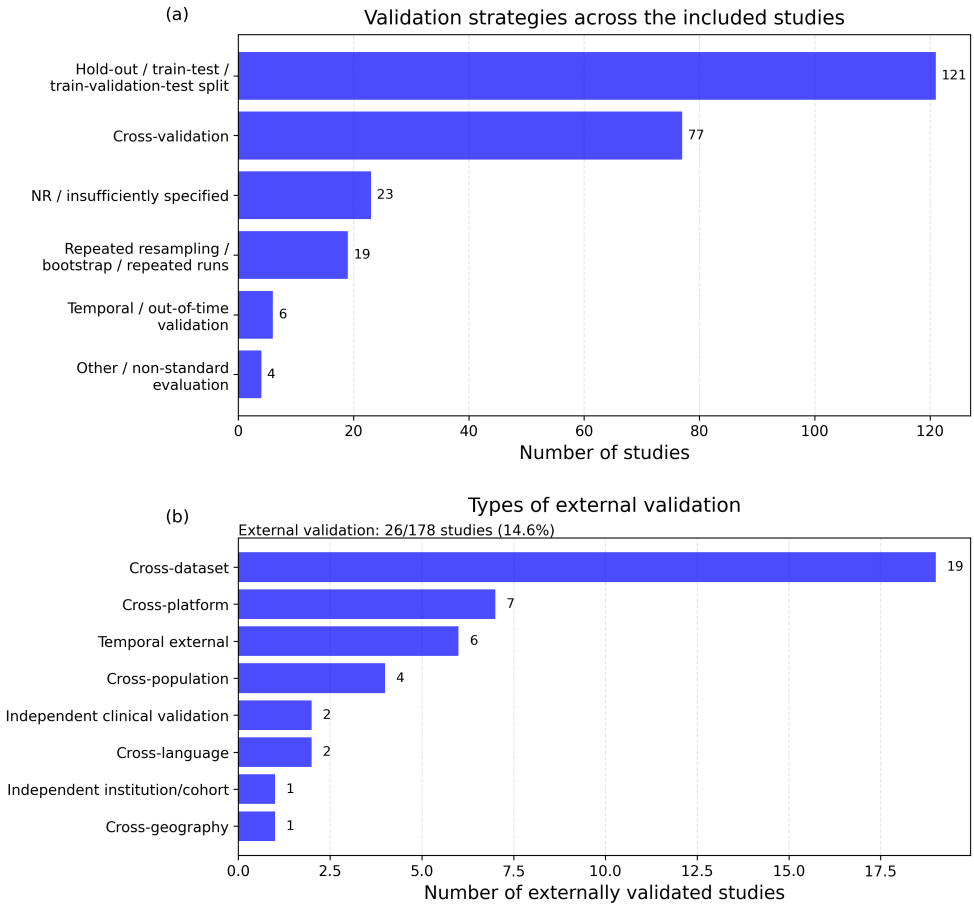


Figure 8. Validation strategies and external generalizability across the included studies. Panel (**a**) summarizes reported validation strategies, while panel (**b**) presents the types of external validation among the 26 studies (14.6% of N = 178) reporting independent external evaluation. Categories are non-mutually exclusive in both panels.

Reported predictive performance was generally high among studies providing usable numerical estimates (Table 3). F1 was the most frequently available metric (n = 136, 76.4%), followed by recall or sensitivity (n = 119, 66.9%), precision (n = 116, 65.2%), and accuracy (n = 114, 64.0%). AUROC was available in 39 studies (21.9%), whereas AUPRC was reported in only five (2.8%) [57,92,122,152,184]. Median F1 was 0.8990, median accuracy was 0.9249, median recall or sensitivity was 0.8949, and median AUROC was 0.9480. These values represent descriptive distributions of study-reported performance rather than pooled estimates, given the heterogeneity in outcomes, ground truth, analytical units, datasets, metric definitions, and evaluation settings.

Table 3. Reported performance and class-imbalance characteristics of the included studies.

Domain	Metric/Category	*n* (%)	Median	IQR	Range
Reported performance	Accuracy	114 (64.0)	0.9249	0.8600–0.9600	0.3833–0.9999
	Precision	116 (65.2)	0.9200	0.8175–0.9524	0.2200–1.0000
	Recall/Sensitivity	119 (66.9)	0.8949	0.7975–0.9495	0.1217–0.9999
	Specificity	18 (10.1)	0.8145	0.7195–0.9468	0.5656–0.9880
	F1	136 (76.4)	0.8990	0.7928–0.9493	0.3184–0.9995
	AUROC	39 (21.9)	0.9480	0.8700–0.9800	0.6100–0.9980
	AUPRC	5 (2.8)	0.9300	0.8750–0.9390	0.8400–0.9970
Class imbalance	Yes	115 (64.6)			
	No	35 (19.7)			
	NR	21 (11.8)			
	NA	7 (3.9)			
Imbalance management	No explicit mitigation	61 (53.0)			
	Explicit mitigation strategy	48 (41.7)			
	NR	6 (5.2)			
Management methods	Oversampling/SMOTE	12 (10.4)			
	Data augmentation/synthetic generation	12 (10.4)			
	Undersampling	12 (10.4)			
	Class weighting/loss	8 (7.0)			
	Balanced sampling/batches	4 (3.5)			
	Other	3 (2.6)			

Note. Performance-metric percentages use the full corpus ($N = 178$) as the denominator and include only usable numerical values. Median, interquartile range (IQR; 25th–75th percentile), and range describe study-reported results and are not pooled performance estimates. Class-imbalance percentages also use $N = 178$; studies were classified as imbalanced when imbalance was identified in the relevant data or analytical configuration, even if subsequently mitigated. Imbalance-management and management-method percentages use the 115 studies with identified imbalance as the denominator. Management methods are non-mutually exclusive. NR, not reported; NA, not applicable; AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision–recall curve.

Class imbalance was identified in 115 studies (64.6%). Among these, 61 (53.0%) reported no explicit mitigation, 48 (41.7%) reported or evaluated a mitigation strategy, and six (5.2%) provided insufficient information. The most frequent approaches were oversampling or SMOTE, data augmentation or synthetic generation, and undersampling (12 studies each), followed by class weighting or loss modification ($n = 8$) and balanced sampling ($n = 4$). The combination of frequent class imbalance and limited AUPRC reporting highlights an important limitation in the available information for assessing performance under imbalanced class distributions.

3.6. Explainability, Ethics, Privacy, and Reproducibility

Explainability or interpretability was identified in 58 of the 178 studies (32.6%), while 120 (67.4%) did not report an explicit approach (Figure 9a). Post-hoc XAI methods were the most frequent ($n = 23$), followed by feature-based or descriptive interpretability ($n = 19$), attention-based approaches ($n = 15$), and intrinsic, rule-, or knowledge-based interpretability ($n = 8$). Four studies used other or non-standard approaches, and two incorporated LLM-generated rationales or explanations [30,85]. These categories were not mutually exclusive. Representative techniques included SHAP and LIME [25,34,41,47,48], permutation importance [8,74], Integrated Gradients [85], attention visualizations [31,35,54], and interpretable rules or knowledge-informed representations [63,120,128,159]. These findings characterize the approaches reported by the studies rather than establishing the validity or clinical utility of their explanations.

Ethics and privacy reporting varied across the reviewed studies (Figure 9b). Ethics-review information was not reported in 118 studies (66.3%), whereas 31 (17.4%) explicitly reported approval or review, 10 (5.6%) reported an exemption, and 5 (2.8%) stated that no ethics review was conducted. Consent information was not reported in 117 studies (65.7%). Anonymization was explicitly documented in 84 studies (47.2%), privacy considerations in 102 (57.3%), and platform terms-of-service considerations in 16 (9.0%). Ethics approval and consent were documented in studies linking social-media data to participant-level clinical or psychometric assessments [6,132,162,169,177,178], while formal exemptions were reported for some secondary-data investigations [78,127,138,154,188]. These results describe the extent of reporting and should not be interpreted as evidence that unreported ethical procedures were absent.

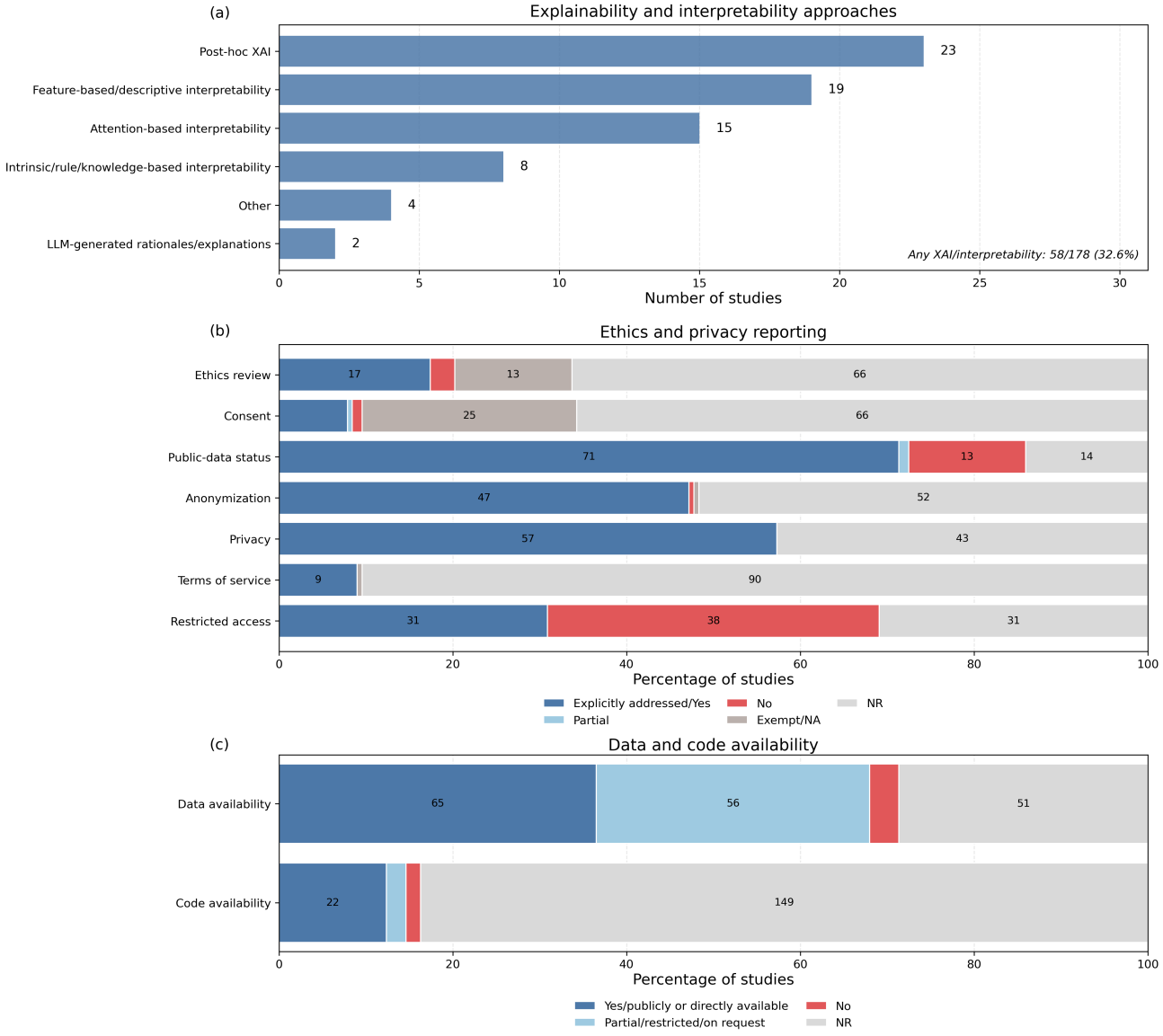


Figure 9. Explainability, ethics, privacy, and reproducibility across the included studies. Panel (**a**) summarizes interpretability approaches, which were not mutually exclusive; panel (**b**) presents ethics- and privacy-related reporting; and panel (**c**) shows data and source-code availability. Public-data status in panel (**b**) refers to the visibility of the original source data, whereas data availability in panel (**c**) concerns reusable study resources. Panels (**b**,**c**) use $N = 178$ as the denominator; labels represent rounded percentages in panel (**b**) and study counts in panel (**c**).

Public visibility of source data did not necessarily correspond to availability for reuse. Although 127 studies (71.3%) analyzed source data classified as public, reusable study data were publicly or directly available in only 65 (36.5%). Another 56 studies (31.5%) reported partial, restricted, or request-based access, 6 (3.4%) explicitly indicated that data were unavailable, and availability was not reported or could not be verified in 51 (28.7%) (Figure 9c). Resources available only upon request or under specific access conditions were classified as partially rather than directly available [28,74,109,162,166,172].

Source-code availability was more limited: code was publicly or directly available in 22 studies (12.4%), partially or conditionally available in 4 (2.2%), explicitly unavailable in 3 (1.7%), and not reported in 149 (83.7%). Only 14 studies (7.9%) had both reusable data and source code classified as publicly or directly available, as detailed in Table 4. These availability classifications describe access to reported research resources rather than demonstrating that the corresponding analyses can be fully reproduced.

Table 4. Studies reporting publicly or directly available data and source code.

Study	Year	Platform/Data Resource	Data Availability	Code Availability
[56]	2020	Reddit; Reddit Mental Health Dataset	Dataset available through the Open Science Framework	Public GitHub repository
[148]	2020	Twitter/X; Spanish suicidal-intent Twitter dataset	Dataset available through the public study GitHub repository	Code available through the same GitHub repository
[84]	2021	Reddit; C-SSRS-aligned Reddit suicide-severity dataset	Dataset/resource publicly available through GitHub	ASHA implementation publicly available through GitHub
[130]	2022	Twitter/X; suicide-related content annotation dataset	Tweet IDs and derived data available through GitHub	Analysis and training code available through GitHub
[137]	2022	Twitter/X; Arabic suicidal-content dataset	Dataset available through the public project repository	Model implementation available through the same repository
[114]	2023	Twitter/X; TWISCO	Tweet IDs and labels distributed through GitHub	Code available through the same public repository
[38]	2024	Twitter/X; Reddit; TwitterDS1, TwitterDS2, and Reddit_SNS (RedditDS1)	Public datasets referenced by the study	Code available through the TCN-based-Model-for-Suicidal-Ideation-Detection repository
[65]	2024	Twitter/X; Reddit; combined Spanish emotion and mental-health corpus	Corpus/raw data available through GitHub and Supplementary Files	Code available through Supplementary Files and/or the project repository
[72]	2024	Reddit; Suicide and Depression Detection	Dataset publicly available through Kaggle	Analytical files/code available as supplementary resources
[188]	2024	Sanctioned Suicide forum; Suicide Discussion subforum	Data and study materials reported in the Supplementary Materials	Analytical scripts/code reported in the Supplementary Materials
[154]	2025	Sina Weibo; SOS-HL-1K	Dataset available through the study GitHub repository	Code available through the study GitHub repository
[67]	2025	Reddit; Twitter/X; Facebook; Sentiment Analysis for Mental Health	Dataset publicly available through Kaggle	Code publicly available through GitHub
[167]	2025	Sina Weibo; Customized Weibo Super Topic suicidal-ideation dataset	Dataset published through GitHub	Model implementation available through GitHub
[50]	2026	Reddit; Suicide Watch Dataset	Dataset publicly available through Kaggle	Associated code/resources reported by the study

Note. The table includes the 14 studies (7.9%; $N = 178$) with both data and code classified as publicly or directly available. Availability refers to reusable research resources rather than the public visibility of source content. Reuse may depend on platform policies and continued content accessibility, particularly when only post identifiers or derived labels are shared. Availability does not by itself establish full computational reproducibility.

4. Discussion

4.1. Principal Findings

This review of 178 studies identifies a central tension between the increasing diversity of AI methods and the limited comparability of the reported evidence. Research remains

concentrated in recurrent social-media platforms and datasets, while suicide-related outcomes and ground-truth procedures vary substantially. Importantly, apparently similar classification tasks may represent distinct suicide-related outcomes, analytical units, labeling procedures, and inferential objectives; detecting suicide-related content does not, by itself, establish the ability to predict subsequent suicidal behavior. Although reported performance was often high among studies providing usable metrics, external validation was identified in only 26 studies (14.6%), and reporting of explainability, ethical procedures, and reusable data and code was uneven. These findings extend concerns raised in earlier reviews [10–12,14,21] by showing that model performance must be interpreted in relation to the target construct, analytical unit, label provenance, independent evaluation, and reporting transparency.

4.2. Methodological Diversification and Empirical Concentration

The findings indicate diversification rather than replacement of modeling approaches: traditional machine learning, deep learning, Transformer-based models, and LLMs coexist in the literature, consistent with earlier syntheses [10,12,15,16]. LLMs have been used not only for generative applications but also for classification, representation, and hybrid workflows [24,27,65,71,85]. Comparisons across reviews nevertheless depend on model taxonomy. Holmes et al. [9] included models such as BERT and T5 within their broader LLM category, whereas the present review classified BERT-type encoders as Transformer-based models. Differences in the apparent timing of LLM adoption therefore partly reflect coding definitions rather than contradictory findings.

This methodological diversification has occurred within a relatively concentrated empirical base. Reddit and Twitter/X remained the predominant platforms, while resources such as the Suicide and Depression Detection dataset and the UMD/CLPsych Reddit Suicidality Dataset were repeatedly reused. Established datasets can facilitate comparisons between modeling approaches, but repeated evaluation on the same resources does not establish generalizability to other populations, platforms, languages, or settings. Similar concentrations of social-media data sources have been identified in previous reviews [9,12,20].

The limited reporting of data provenance further constrains interpretation. The geographic origin of the analyzed social-media data was not reported in 126 of the 178 studies (70.8%). Moreover, the countries associated with author affiliations do not necessarily represent the geographic origin of the users or datasets analyzed. These reporting limitations, together with the concentration of data sources, restrict the ability to assess geographical and cross-cultural generalizability, consistent with concerns raised in earlier syntheses [12,16,20]. Consequently, growth in publication activity and model complexity should not be interpreted as direct evidence of stronger external validity or clinical applicability.

4.3. Outcome and Ground-Truth Heterogeneity as a Central Challenge

Outcome heterogeneity complicates comparisons even among studies using similar classification approaches. Although binary operationalization predominated in 126 studies (70.8%), the positive class represented different constructs, including suicidal ideation, suicidal communication, suicide risk, suicide attempts, suicide death, and composite outcomes. Analytical units also varied, ranging from individual posts to users, longitudinal histories, and population-level signals. Consequently, identifying suicide-related language in a post is not equivalent to estimating an individual's suicide risk or predicting subsequent suicidal behavior. This distinction reinforces concerns raised in earlier reviews about reducing multidimensional suicidal phenomena to binary labels and treating online expressions as proxies for clinically verified outcomes [10,14,18–20].

The 18 studies (10.1%) assigned to the broad temporal-prediction, forecasting, or early-detection category encompassed different objectives, from identifying existing risk indicators in sequential content to forecasting subsequent population-level suicide counts [187]. They therefore cannot be equated with studies prospectively predicting individual suicidal behavior. Likewise, temporal external validation in six studies could involve evaluating classifiers on later-collected posts [7] without assessing future clinical outcomes. Because prediction horizons and predictor–outcome temporal ordering were not systematically extracted, the frequency of genuinely prospective individual-level prediction could not be reliably determined.

Ground-truth sources were similarly heterogeneous, encompassing inherited dataset labels, clinical and non-clinical annotation, community-based proxies, self-report or instrument-based measures, rule-based labels, and documented events. These approaches differ in their relationship to the underlying suicide-related construct and should not be treated as interchangeable [16]. A particularly important distinction concerns clinical and psychometric instruments: they were administered directly to participants to establish the outcome in only 11 studies, whereas 15 used such instruments as annotation frameworks, taxonomies, or classification references. The C-SSRS illustrates this difference, having been administered directly in participant-linked studies [132,176,178] but also used to categorize social-media content without direct participant assessment [30,47,49,66,78,79,84,85,143,155,171]. By contrast, all six studies examining suicide death relied on documented events or administrative records [39,119,131,138,170,187]. These differences mean that favorable performance against a particular operational label does not necessarily establish equivalent predictive or clinical validity across studies.

Human annotation introduced additional variability. Annotator profiles were reported in 73 studies, not reported in 50, and not applicable in 55. When annotation was relevant, information on annotator qualifications, adjudication procedures, and inter-rater agreement was frequently incomplete. Reported practices included clinical professionals, non-clinical annotators, and crowdworkers [5,66,78,155], with agreement assessed through different measures when reported. Earlier reviews similarly highlighted concerns regarding annotation quality, expertise, and potential labeling bias [11,14,18]. Importantly, incomplete reporting does not establish that labeling procedures were inadequate, but it limits assessment of their consistency and methodological comparability. Accordingly, reported model performance should be interpreted in relation to the specific outcome definition, analytical unit, ground-truth provenance, and labeling procedure rather than treating "suicidality" as a homogeneous prediction target.

4.4. High Reported Performance but Limited Evidence of Generalizability

Although reported performance was generally high among studies providing usable metrics (median F1 = 0.8990; median AUROC = 0.9480), these values represent descriptive distributions rather than pooled estimates of predictive performance. Differences in suicide-related outcomes, ground-truth sources, datasets, analytical units, positive-class definitions, averaging procedures, and validation settings limit comparability across studies. In particular, a median F1 of 0.8990 does not imply approximately 90% classification accuracy or equivalent performance in independent populations. The standardized selection of a principal model–evaluation context reduced within-study extraction ambiguity but could not eliminate between-study heterogeneity. Similar interpretive limitations have been highlighted in previous reviews reporting favorable yet methodologically heterogeneous results [10,21].

An additional limitation concerns the distinction between discrimination and calibration. The extracted performance metrics primarily characterize classification or discrimina-

tion, but do not establish whether predicted probabilities correspond to observed outcome frequencies. Calibration was not a predefined variable in the extraction framework and was therefore not systematically assessed in this review. Consequently, we cannot determine how frequently calibration was evaluated or adequately reported across the included studies. This limitation is particularly relevant to models intended for individual risk prediction or clinical decision support, rather than tasks limited to classifying existing social media content. For clinically oriented prediction models, calibration should be evaluated and transparently reported alongside discrimination, consistent with TRIPOD+AI [195].

Independent external validation was identified in only 26 studies (14.6%), predominantly through cross-dataset evaluations. Other dimensions of transportability were less frequently examined. Temporal external validation was reported in six studies [7,56,63,145,160,187], while cross-language [52,86] and independent clinical validation [6,57] were each identified in two studies. Cross-geographic validation [52] and independent institutional or cohort validation [6] were each identified in only one study. Consequently, successful internal testing or evaluation on another dataset does not necessarily establish generalizability across populations, languages, geographic settings, or clinical contexts. Earlier reviews have similarly emphasized the need for independent, prospective, and real-world validation [15,16,20,21].

Evaluation independence presents a related source of uncertainty. In social media datasets, overlapping users or duplicated content across training and test sets, as well as resampling or data augmentation performed before splitting, can introduce data leakage and inflate apparent performance. Although validation strategies were extracted, user-level partitioning, duplicate-control safeguards, and the sequencing of resampling or augmentation were not systematically assessed. Consequently, this review cannot determine whether data leakage affected the reported performance estimates. This limitation of the methodological synthesis should not be interpreted as evidence that leakage occurred in the included studies.

Class imbalance was identified in 115 studies (64.6%), yet AUPRC was available in only five (2.8%) [57,92,122,152,184]. Although this does not invalidate studies reporting accuracy, F1, or AUROC, it limits the interpretation of performance under class distributions that may differ substantially from those used during model development. Greater transparency regarding positive-class prevalence, imbalance mitigation, and complementary precision–recall-based evaluation would therefore improve the interpretation of reported results, consistent with concerns raised in earlier reviews [11,12,21]. Taken together, high performance within recurrent benchmark settings and limited independent validation leave the transportability of many reported models uncertain.

4.5. Uneven Reporting of Explainability, Ethics, and Reproducibility

Explainability or interpretability was identified in 58 studies (32.6%), encompassing heterogeneous approaches such as post-hoc XAI, feature-based analyses, attention visualizations, interpretable rules, and LLM-generated rationales. These methods are not interchangeable, and their presence does not establish that the resulting explanations are stable, faithful to model behavior, or clinically meaningful. Previous reviews have similarly identified interpretability as a methodological challenge in suicide-related modeling [10,11,14,16]. Because the present review examined the documented use of interpretability approaches rather than independently evaluating their validity or clinical utility, these findings characterize methodological reporting rather than the effectiveness of explanations for supporting clinical decisions.

Ethical reporting showed comparable limitations. Ethics-review information was not reported in 118 studies (66.3%), and consent information was not reported in 117 (65.7%).

However, absence of reporting does not establish that ethical review, consent assessment, or other governance procedures were not performed. The findings therefore concern reporting transparency rather than ethical compliance. Earlier syntheses have likewise identified inconsistent documentation of ethical considerations, privacy, and consent [9,14,17,18,20], although differences in their assessment criteria prevent direct numerical comparisons.

Reproducibility-related findings require distinguishing publicly accessible source content from reusable analytical datasets and computational resources. Although 127 studies (71.3%) analyzed source data classified as public, study data were publicly or directly available for reuse in only 65 (36.5%). Source code was publicly or directly available in 22 studies (12.4%), while only 14 (7.9%) had both data and code classified as publicly or directly available. Nevertheless, resource availability does not by itself establish computational reproducibility, particularly when shared materials consist of identifiers, incomplete analytical workflows, or resources subject to platform restrictions and changing access conditions. Consistent with concerns raised in earlier reviews [9,15,18], these findings should be interpreted as indicators of reporting transparency and potential reusability rather than evidence that published analyses can be independently reproduced in full.

4.6. Implications for Future Research and Clinical Translation

Future research should prioritize clearly defined suicide-related outcomes, transparent ground-truth construction, and independent evaluation. Studies should specify the target construct, analytical unit, and label provenance, distinguishing clinical instruments administered directly to participants from those used only as annotation frameworks. When human annotation is involved, annotator expertise, adjudication procedures, and inter-rater agreement should be reported where applicable. External validation should extend beyond testing on an additional dataset to differences in platforms, languages, geographic settings, populations, collection periods, and independent clinical or institutional cohorts. Prospective or temporally independent evaluation is particularly important when predicting subsequent suicidal behavior. These priorities are consistent with recommendations from earlier reviews [9,15,16,21]; however, only six studies in the present corpus reported temporal external validation.

Performance reporting should specify the positive class, averaging procedures for precision, recall, and F1, and any methods used to address class imbalance. Imbalance was identified in 115 studies, whereas AUPRC was reported in only five, indicating the value of incorporating precision–recall-based measures alongside other task-appropriate metrics, particularly when suicide-related outcomes are rare. Explainability should likewise be evaluated in relation to the prediction and its intended use rather than inferred solely from the presence of an interpretability technique. For LLM-based approaches, generated rationales may be linguistically plausible without faithfully representing the underlying model decisions, making explicit evaluation of their reliability particularly important.

Ethical and reproducibility reporting should clearly document ethics-review or exemption status, consent procedures where applicable, privacy safeguards, anonymization, and relevant platform-access conditions. Publicly accessible social-media content should be distinguished from reusable analytical datasets and restricted resources. Sharing sufficiently detailed code, preprocessing and evaluation procedures, and computational resources would facilitate independent verification while respecting legitimate restrictions on sensitive data. Ultimately, clinical translation requires evidence of generalizability beyond development datasets, prospective evaluation where appropriate, and human oversight, with AI supporting rather than replacing clinically grounded assessment [9,13,16,17,21]. Given the limited evidence from independent clinical, cross-language, cross-geographic,

and cross-population validation, caution remains necessary when extrapolating reported performance to untested settings.

4.7. Strengths and Limitations

A principal strength of this review is its structured and traceable synthesis of 178 studies selected from Scopus and Web of Science following PRISMA 2020. Abstract and full-text screening involved two reviewers, with disagreements resolved by a third author. Following pilot calibration, the 50-field extraction framework was independently applied by two reviewers, with discrepancies resolved through consensus or third-reviewer adjudication. Explicit harmonization rules and a predefined model–evaluation hierarchy reduced coding inconsistencies and prevented performance metrics from different models or evaluation contexts from being combined within individual studies. The supplementary extraction matrix and study-level notes further supported methodological transparency and traceability.

The review was restricted to English-language publications indexed as Article or Article Early Access in Scopus and Web of Science, excluding conference papers and other publication types. Seven reports sought for full-text assessment could not be retrieved. These restrictions may have excluded relevant evidence. In addition, substantial heterogeneity in suicide-related outcomes, ground-truth sources, analytical units, performance metrics, and validation settings precluded meaningful statistical pooling. The reported medians should therefore be interpreted as descriptive distributions of study-reported values rather than estimates of overall diagnostic or predictive performance. The search was completed on 22 May 2026; consequently, publication trends and the representation of emerging approaches, particularly LLM-based methods, reflect only the evidence available at that date.

No formal study-level risk-of-bias or methodological-quality assessment was performed, as this review aimed to characterize methodological practices across heterogeneous AI tasks rather than formally appraise individual prediction models. Consequently, the reported frequencies of validation strategies, ground-truth procedures, and other methodological characteristics describe their occurrence or reporting, not their adequacy or the reliability of the resulting evidence. In particular, high reported performance metrics and the presence of external validation should not be interpreted as establishing unbiased or reliable predictive performance. Future systematic reviews focused on clinically oriented prediction models could apply PROBAST+AI [196] to assess model-development quality, risk of bias in performance evaluation, and applicability. In addition, the absence of a quantitative pre-consensus inter-rater agreement measure limits the assessment of the reproducibility of initial data-extraction decisions. Furthermore, the synthesis depended on information available in published reports and associated resources. Information coded as not reported does not establish that the corresponding procedure was absent or inadequately implemented, particularly regarding ethics, privacy, validation, and reproducibility. Data and code availability were assessed according to resources reported or accessible at the time of extraction and may change with repository status, platform policies, and access restrictions.

5. Conclusions

This review of 178 studies shows that increasing methodological diversity in social-media-based AI research has not been accompanied by equally broad evidence of comparability or generalizability. Although traditional machine learning, deep learning, Transformer-based models, and LLMs coexist, empirical evaluations remain concentrated in recurrent Reddit and Twitter/X datasets, and the geographic origin of the analyzed data was not reported in 70.8% of studies. More importantly, models target distinct suicide-related

constructs and analytical units using heterogeneous ground-truth sources. Consequently, high study-reported performance should not be interpreted as evidence that detecting suicide-related content necessarily predicts subsequent suicidal behavior.

External validation was identified in only 26 studies (14.6%), with cross-language, cross-geographic, and independent clinical evaluations remaining uncommon. Explainability was identified in 58 studies (32.6%), code was publicly or directly available in 22 (12.4%), and both reusable data and source code were publicly or directly available in only 14 (7.9%). Ethics-review and consent information were also frequently unreported. Taken together, these findings identify limitations in the published evidence available to assess independent validation, ground-truth validity, interpretability, ethical oversight, and reproducibility. They do not establish that unreported procedures were absent or inadequately implemented. Ultimately, the credibility and potential clinical applicability of these systems depend on clearly defined outcomes, transparent label provenance, independent evaluation, and reproducible reporting, rather than model complexity or internal predictive performance alone.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/informatics13110191/s1, File S1: Study-level data-extraction matrix for the 178 studies included in the systematic review, including article title, publication year, DOI/URL, and the 50 extracted variables (XLSX); PRISMA 2020 Checklist.

Author Contributions: Conceptualization, F.D.; methodology, F.D. and N.C.; software, F.D. and N.C.; validation, J.H.M.-R., K.O.-K., N.C. and A.A.-M.; formal analysis, F.D., N.C. and A.A.-M.; investigation, F.D., N.C., J.H.M.-R., K.O.-K. and A.A.-M.; resources, N.C., J.H.M.-R. and K.O.-K.; data curation, F.D., N.C. and A.A.-M.; writing—original draft preparation, F.D. and N.C.; writing—review and editing, F.D. and N.C.; visualization, N.C. and A.A.-M.; supervision, F.D.; project administration, F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this systematic review are available in Supplementary File S1 (XLSX), which contains the study-level data-extraction matrix for all 178 included studies, including article title, publication year, DOI/URL, and the 50 extracted variables.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AUPRC	Area under the precision–recall curve
AUROC	Area under the receiver operating characteristic curve
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional long short-term memory
CNN	Convolutional neural network
C-SSRS	Columbia-Suicide Severity Rating Scale
DL	Deep learning
IQR	Interquartile range
LIME	Local Interpretable Model-agnostic Explanations
LIWC	Linguistic Inquiry and Word Count
LLM	Large language model
LSTM	Long short-term memory
ML	Machine learning

NA	Not applicable
NLP	Natural language processing
NR	Not reported
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNN	Recurrent neural network
RoBERTa	Robustly Optimized BERT Pretraining Approach
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support vector machine
XAI	Explainable artificial intelligence

Appendix A. Full-Text Reports Excluded After Eligibility Assessment

A total of 203 reports were retrieved and assessed in full text against the eligibility criteria. Following full-text assessment, 25 reports were excluded because they did not satisfy one or more criteria required for inclusion in the systematic review. Tables A1–A3 identify these reports, the principal eligibility criterion not met, and the corresponding reason for exclusion. When more than one criterion was not satisfied, the criterion or combination of criteria most directly responsible for the final exclusion decision is reported.

Table A1. Reports excluded after full-text eligibility assessment and primary reasons for exclusion (Part 1).

Study	Eligibility Criterion Not Met	Reason for Exclusion
[197]	Empirical evaluation of the suicide-related model	Suicide-related themes were examined using hazard models, but no predictive or discriminative evaluation of the suicide-related outcome was performed.
[198]	Suicide-related outcome and computational task	The evaluated classification task distinguished depressed from normal content rather than explicitly assessing suicidality.
[199]	Empirical evaluation of the suicide-related model	AI-derived Facebook topics were associated with C-SSRS scores, but no out-of-sample predictive evaluation of suicidality was performed.
[200]	Empirical evaluation of the suicide-related model	Topic and regression analyses examined associations with suicide attempts, but predictive or discriminative performance for the suicide-related outcome was not evaluated.
[201]	Outcome explicitly related to suicidality	The predicted outcome was self-harm irrespective of suicidal intent and therefore was not sufficiently specific to suicidality.
[202]	Suicide-related computational task	The Transformer classified the intended meaning of geographical mentions rather than detecting, assessing, stratifying, or predicting suicidality.
[203]	Empirical evaluation of the suicide-related model	Topic models and regression analyses examined factors associated with suicidal tendencies, but no predictive evaluation of the suicide-related outcome was performed.
[204]	Suicide-related computational task	The models classified Health Belief Model components and emotional expressions; suicidal ideation was discussed but was not the target of the AI task.
[205]	Suicide-related computational task	The models predicted emotions and sentiment from healthcare experiences rather than an explicit suicide-related outcome.

Table A2. Reports excluded after full-text eligibility assessment and primary reasons for exclusion (Part 2).

Study	Eligibility Criterion Not Met	Reason for Exclusion
[206]	Empirical evaluation of the suicide-related model	The study analyzed messages previously identified as high risk, but did not provide a new empirical evaluation of the suicide-related model.
[207]	Data obtained from social media and user-generated content	Models were developed using US Army administrative records and the National Death Index rather than social-media or user-generated data.
[208]	Outcome explicitly related to suicidality	The system classified emotional distress in blogs rather than an explicit suicide-related outcome.

Table A2. *Cont.*

Study	Eligibility Criterion Not Met	Reason for Exclusion
[209]	Data obtained from social media	The CEASE suicide-note corpus was used, but the study did not establish that the experimental data originated from an eligible social-media source.
[210]	Suicide-related outcome and computational task	The experimental task classified depression versus non-depression without independently labeling or evaluating a suicide-related outcome.
[211]	Suicide-related computational task and empirical evaluation	Machine learning classified illegal drug selling and diversion; suicide-related content was subsequently identified without an evaluated AI model targeting suicidality.
[212]	Empirical evaluation of the suicide-related model	A previously developed suicide-signal model was applied, but its predictive or discriminative performance was not evaluated in the present study.
[213]	Suicide-related outcome and computational task	Suicide-related classifiers generated social-media indicators, but the forecasting model predicted mental-health consultations rather than a suicide-related outcome.

Table A3. Reports excluded after full-text eligibility assessment and primary reasons for exclusion (Part 3).

Study	Eligibility Criterion Not Met	Reason for Exclusion
[214]	Data obtained from social media	The principal behavioral input consisted of interactions with articles on The New York Times website, which was not considered an eligible social-media source.
[215]	Identifiable unit of analysis and sufficient methodological information	Insufficient information was provided on sample composition, prediction unit, ground truth, labeling, and evaluation design to assess the study reproducibly.
[216]	Suicide-related computational task	The evaluated models classified emotions within suicide-related texts rather than detecting, assessing, stratifying, or predicting suicidality itself.
[217]	Suicide-related outcome and computational task	The models classified sentiment in microblogs; suicide was discussed as a potential application but was not an empirically evaluated outcome.
[218]	Suicide-related outcome and computational task	The evaluated labels reflected sentiment and mental-health content but did not establish an explicit suicide-related computational outcome.
[219]	Original research, empirical evaluation, and sufficient methodological information	The report inconsistently described its numerical results as both real-world and simulated, preventing verification of original empirical evaluation.
[220]	Suicide-related outcome and computational task	Neural models reproduced VADER sentiment labels interpreted as suicide propensity; the ground truth therefore represented sentiment rather than suicidality.
[221]	Original research, empirical evaluation, and sufficient methodological information	A multimodal architecture was proposed, but no implemented dataset, training procedure, validation design, or original performance results were reported.

Note. Reports were excluded after full-text assessment. When multiple eligibility criteria were not satisfied, the criterion or combination of criteria most directly responsible for the final exclusion decision is shown.

References

1. O'dea, B.; Wan, S.; Batterham, P.J.; Calear, A.L.; Paris, C.; Christensen, H. Detecting suicidality on twitter. *Internet Interv.* **2015**, *2*, 183–188. [CrossRef]
2. Wang, S.; Ning, H.; Huang, X.; Xiao, Y.; Zhang, M.; Yang, E.F.; Sadahiro, Y.; Liu, Y.; Li, Z.; Hu, T.; et al. Public Surveillance of Social Media for Suicide Using Advanced Deep Learning Models in Japan: Time Series Study From 2012 to 2022. *J. Med. Internet Res.* **2023**, *25*, e47225. [CrossRef]
3. Silverman, M.M.; Berman, A.L.; Sanddal, N.D.; O'Carroll, P.W.; Joiner, T.E., Jr. Rebuilding the Tower of Babel: A Revised Nomenclature for the Study of Suicide and Suicidal Behaviors Part 2: Suicide-Related Ideations, Communications, and Behaviors. *Suicide Life-Threat. Behav.* **2007**, *37*, 264–277. [CrossRef]
4. Ramirez-Cifuentes, D.; Freire, A.; Baeza-Yates, R.; Punti, J.; Medina-Bravo, P.; Alejandro Velazquez, D.; Maria Gonfaus, J.; Gonzalez, J. Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis. *J. Med. Internet Res.* **2020**, *22*, e17758. [CrossRef]
5. Yao, H.; Rashidian, S.; Dong, X.; Duanmu, H.; Rosenthal, R.N.; Wang, F. Detection of Suicidality Among Opioid Users on Reddit: Machine Learning-Based Approach. *J. Med. Internet Res.* **2020**, *22*, e15293. [CrossRef]
6. Kaminsky, Z.; McQuaid, R.J.; Hellemans, K.G.; Patterson, Z.R.; Saad, M.; Gabrys, R.L.; Kendzerska, T.; Abizaid, A.; Robillard, R. Machine Learning-Based Suicide Risk Prediction Model for Suicidal Trajectory on Social Media Following Suicidal Mentions: Independent Algorithm Validation. *J. Med. Internet Res.* **2024**, *26*, e49927. [CrossRef]
7. Burnap, P.; Colombo, G.; Amery, R.; Hodorog, A.; Scourfield, J. Multi-class machine classification of suicide-related communication on Twitter. *Online Soc. Netw. Media* **2017**, *2*, 32–44. [CrossRef]

8. Howard, D.; Maslej, M.M.; Lee, J.; Ritchie, J.; Woollard, G.; French, L. Transfer Learning for Risk Classification of Social Media Posts: Model Evaluation Study. *J. Med. Internet Res.* **2020**, *22*, e15371. [CrossRef]
9. Holmes, G.; Tang, B.; Gupta, S.; Venkatesh, S.; Christensen, H.; Whitton, A. Applications of Large Language Models in the Field of Suicide Prevention: Scoping Review. *J. Med. Internet Res.* **2025**, *27*, e63126. [CrossRef]
10. Heckler, W.F.; de Carvalho, J.V.; Barbosa, J.L.V. Machine learning for suicidal ideation identification: A systematic literature review. *Comput. Hum. Behav.* **2022**, *128*, 107095. [CrossRef]
11. Ji, S.; Pan, S.; Li, X.; Cambria, E.; Long, G.; Huang, Z. Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *IEEE Trans. Comput. Soc. Syst.* **2021**, *8*, 214–226. [CrossRef]
12. Malhotra, A.; Jindal, R. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Appl. Soft Comput.* **2022**, *130*, 109713. [CrossRef]
13. Arowosegbe, A.; Oyelade, T. Application of Natural Language Processing (NLP) in Detecting and Preventing Suicide Ideation: A Systematic Review. *Int. J. Environ. Res. Public Health* **2023**, *20*, 1514. [CrossRef]
14. Yeskuatov, E.; Chua, S.L.; Foo, L.K. Leveraging Reddit for Suicidal Ideation Detection: A Review of Machine Learning and Natural Language Processing Techniques. *Int. J. Environ. Res. Public Health* **2022**, *19*, 10347. [CrossRef]
15. Atmakuru, A.; Shahini, A.; Chakraborty, S.; Seoni, S.; Salvi, M.; Hafeez-Baig, A.; Rashid, S.; Tan, R.S.; Barua, P.D.; Molinari, F.; et al. Artificial intelligence-based suicide prevention and prediction: A systematic review (2019–2023). *Inf. Fusion* **2025**, *114*, 102673. [CrossRef]
16. Mokheleli, T.; Makaba, T.; Ndayizigamiye, P.; Ndlovu, N.; Twinomurinzi, H. Artificial intelligence in suicide risk assessment: A systematic literature review. *Discov. Artif. Intell.* **2026**, *6*, 296. [CrossRef]
17. Pourmand, A.; Roberson, J.; Caggiula, A.; Monsalve, N.; Rahimi, M.; Torres-Llenza, V. Social Media and Suicide: A Review of Technology-Based Epidemiology and Risk Assessment. *Telemed. e-Health* **2019**, *25*, 880–888. [CrossRef]
18. Castillo-Sánchez, G.; Marques, G.; Dorronzoro, E.; Rivera-Romero, O.; Franco-Martín, M.; de la Torre-Díez, I. Suicide Risk Assessment Using Machine Learning and Social Networks: A Scoping Review. *J. Med. Syst.* **2020**, *44*, 205. [CrossRef]
19. Abdulsalam, A.; Alhothali, A. Suicidal ideation detection on social media: A review of machine learning methods. *Soc. Netw. Anal. Min.* **2024**, *14*, 188. [CrossRef]
20. Trentarossi, B.; Romão, M.E.; Barello, S.; Baptista, M.N.; Visonà, S.D.; Belli, G. Digital Detection of Suicidal Ideation: A Scoping Review to Inform Prevention and Psychological Well-Being. *Behav. Sci.* **2025**, *15*, 1601. [CrossRef]
21. Bernert, R.A.; Hilberg, A.M.; Melia, R.; Kim, J.P.; Shah, N.H.; Abnousi, F. Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5929. [CrossRef]
22. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*. [CrossRef]
23. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Inf.* **2017**, *11*, 959–975. [CrossRef]
24. Qorich, M.; El Ouazzani, R. Advanced deep learning and large language models for suicide ideation detection on social media. *Prog. Artif. Intell.* **2024**, *13*, 135–147. [CrossRef]
25. Malhotra, A.; Jindal, R. XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks. *Cogn. Syst. Res.* **2024**, *84*, 101186. [CrossRef]
26. Kallstenius, T.; Capusan, A.J.; Andersson, G.; Williamson, A. Comparing traditional natural language processing and large language models for mental health status classification: A multi-model evaluation. *Sci. Rep.* **2025**, *15*, 24102. [CrossRef]
27. Álvarez López, J.P.; Castro-Sánchez, N.A. Word Embeddings and Machine Learning Classifiers Applications for Automatic Detection of Suicide Tendencies in Social Media. *Program. Comput. Softw.* **2024**, *50*, 612–620. [CrossRef]
28. Aragón, M.E.; López-Monroy, A.P.; Montes-y-Gómez, M.; Losada, D.E. Online expressions, offline struggles: Using social media to identify depression-related symptoms. *Online Soc. Netw. Media* **2025**, *50*, 100338. [CrossRef]
29. Balakrishnan, V.; Loganathamoorthy, D.; Ayyasamy, R.K.; Alhashmi, S.M.; P., C. Online suicide ideation detection (OnSIDe): A context-aware transfer learning approach using BERT-CNN. *PeerJ Comput. Sci.* **2025**, *11*, e3239. [CrossRef]
30. Kim, S.; Lee, J.; Kim, K.; Kim, H. Leveraging timeline integration methods for effective suicidal ideation detection. *Expert Syst. Appl.* **2026**, *300*, 130190. [CrossRef]
31. Kumar, A.; Sharma, A.; Sangwan, S.R. DynaMentA: Dynamic Prompt Engineering and Weighted Transformer Architecture for Mental Health Classification Using Social Media Data. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 4193–4203. [CrossRef]
32. Chen, X.; Li, Q.; Jiang, Y.; Liu, M.; Yang, L.; Jing, X.; Wei, Z.; Cao, L. PsychoDial: A framework for building a suicide risk dialogue dataset in human-AI conversations. *Inf. Process. Manag.* **2026**, *63*, 104787. [CrossRef]
33. Ghanadian, H.; Nejadgholi, I.; Al Osman, H. Improving Suicidal Ideation Detection in Social Media Posts: Topic Modeling and Synthetic Data Augmentation Approach. *JMIR Form. Res.* **2025**, *9*, e63272. [CrossRef]
34. Bouktif, S.; Khanday, A.M.U.D.; Ouni, A. Explainable Predictive Model for Suicidal Ideation During COVID-19: Social Media Discourse Study. *J. Med. Internet Res.* **2025**, *27*, e65434. [CrossRef]

35. Schoene, A.M.; Turner, A.P.; De Mel, G.; Dethlefs, N. Hierarchical Multiscale Recurrent Neural Networks for Detecting Suicide Notes. *IEEE Trans. Affect. Comput.* **2023**, *14*, 153–164. [CrossRef]
36. Baydili, İ.; Tasci, B.; Tasci, G. Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection. *Behav. Sci.* **2025**, *15*, 352. [CrossRef]
37. Chadha, A.; Kaushik, B. A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data. *New Gener. Comput.* **2022**, *40*, 889–914. [CrossRef]
38. Mirtaheri, S.L.; Greco, S.; Shahbazian, R. A self-attention TCN-based model for suicidal ideation detection from social media posts. *Expert Syst. Appl.* **2024**, *255*, 124855. [CrossRef]
39. Hernández-Castañeda, Á.; García-Hernández, R.A.; Ledeneva, Y. Identifying Suicidal Ideation Through Automatic Extraction of Emotional Traces in Suicide Notes. *IEEE Access* **2025**, *13*, 62378–62392. [CrossRef]
40. Acuña Caicedo, R.W.; Gómez Soriano, J.M.; Melgar Sasieta, H.A. Bootstrapping semi-supervised annotation method for potential suicidal messages. *Internet Interv.* **2022**, *28*, 100519. [CrossRef]
41. Banitaan, S.; Alquran, H. Improving Suicide Ideation Detection Through Feature Engineering and Machine Learning. *IEEE Access* **2025**, *13*, 148269–148284. [CrossRef]
42. Anika, S.; Dewanjee, S.; Muntaha, S. Analyzing Multiple Data Sources for Suicide Risk Detection: A Deep Learning Hybrid Approach. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 675–683. [CrossRef]
43. Basyouni, A.; Abdulkader, H.; Elkilani, W.S.; Alharbi, A.; Xiao, Y.; Ali, A.H. A Suicidal Ideation Detection Framework on Social Media Using Machine Learning and Genetic Algorithms. *IEEE Access* **2024**, *12*, 124816–124833. [CrossRef]
44. Deas, N.; Kowalski, R.; Finnell, S.; Radovic, E.; Carroll, H.; Robbins, C.; Cook, A.; Hurley, K.; Cote, N.; Evans, K.; et al. I just want to matter: Examining the role of anti-mattering in online suicide support communities using natural language processing. *Comput. Hum. Behav.* **2023**, *139*, 107499. [CrossRef]
45. Ragheb, W.; Azé, J.; Bringay, S.; Servajean, M. Negatively Correlated Noisy Learners for At-Risk User Detection on Social Networks: A Study on Depression, Anorexia, Self-Harm, and Suicide. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 770–783. [CrossRef]
46. Ezerceli, Ö.; Dehkharghani, R. Mental disorder and suicidal ideation detection from social media using deep neural networks. *J. Comput. Soc. Sci.* **2024**, *7*, 2277–2307. [CrossRef]
47. Naseem, U.; Khushi, M.; Kim, J.; Dunn, A.G. Hybrid Text Representation for Explainable Suicide Risk Identification on Social Media. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 4663–4672. [CrossRef]
48. KINA, E.; Choi, J.G.; Ishaq, A.; Shafique, R.; Villar, M.G.; Alvarado, E.S.; Diez, I.d.l.T.; Ashraf, I. Suicide Ideation Detection Using Social Media Data and Ensemble Machine Learning Model. *Int. J. Comput. Intell. Syst.* **2026**, *19*, 100. [CrossRef]
49. Zhu, J.; Jin, R.; Kenne, D.R.; Phan, N.; Ku, W.S. User Dynamics and Thematic Exploration in r/Depression During the COVID-19 Pandemic: Insights From Overlapping r/SuicideWatch Users. *J. Med. Internet Res.* **2024**, *26*, e53968. [CrossRef]
50. Salame Neto, S.N.; da Silva Pinto, K.; Teles da Silva, S.T.; de Alcântara Neto, M.C. Automatic detection of suicidal ideation on social media using a hybrid CNNBiLSTM architecture. *IEEE Lat. Am. Trans.* **2026**, *24*, 465–475. [CrossRef]
51. Yeskuatov, E.; Chua, S.L.; Foo, L.K. Detecting Suicidal Ideations in Online Forums with Textual and Psycholinguistic Features. *Appl. Sci.* **2024**, *14*, 9911. [CrossRef]
52. Noraset, T.; Chatrinan, K.; Tawichsri, T.; Thaipisutikul, T.; Tuarob, S. Language-agnostic deep learning framework for automatic monitoring of population-level mental health from social networks. *J. Biomed. Inform.* **2022**, *133*, 104145. [CrossRef]
53. Ramírez-Cifuentes, D.; Largeron, C.; Tissier, J.; Baeza-Yates, R.; Freire, A. Enhanced Word Embedding Variations for the Detection of Substance Abuse and Mental Health Issues on Social Media Writings. *IEEE Access* **2021**, *9*, 130449–130471. [CrossRef]
54. Renjith, S.; Abraham, A.; Jyothi, S.B.; Chandran, L.; Thomson, J. An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *J. King Saud Univ.—Comput. Inf. Sci.* **2022**, *34*, 9564–9575. [CrossRef]
55. Ahadi, S.A.; Jazayeri, K.; Tebyani, S. Detecting Suicidality from Reddit Posts Using a Hybrid CNN—LSTM Model. *J. Univers. Comput. Sci.* **2024**, *30*, 1872–1904. [CrossRef]
56. Low, D.M.; Rumker, L.; Talkar, T.; Torous, J.; Cecchi, G.; Ghosh, S.S. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *J. Med. Internet Res.* **2020**, *22*, e22635. [CrossRef]
57. Burkhardt, H.A.; Ding, X.; Kerbrat, A.; Comtois, K.A.; Cohen, T. From benchmark to bedside: Transfer learning from social media to patient-provider text messages for suicide risk prediction. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1068–1078. [CrossRef]
58. Zhang, T.; Schoene, A.M.; Ananiadou, S. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet Interv.* **2021**, *25*, 100422. [CrossRef]
59. Skaik, R.; Inkpen, D. Suicide Ideation Estimators within Canadian Provinces Using Machine Learning Tools on Social Media Text. *J. Adv. Inf. Technol.* **2021**, *12*, 357–362. [CrossRef]
60. Buddhitha, P.; Inkpen, D. Multi-task learning to detect suicide ideation and mental disorders among social media users. *Front. Res. Metrics Anal.* **2023**, *8*, 1152535. [CrossRef]

61. Adarsh, V.; Arun Kumar, P.; Lavanya, V.; Gangadharan, G. Fair and Explainable Depression Detection in Social Media. *Inf. Process. Manag.* **2023**, *60*, 103168. [CrossRef]
62. Rabani, S.T.; Ud Din Khanday, A.M.; Khan, Q.R.; Hajam, U.A.; Imran, A.S.; Kastrati, Z. Detecting suicidality on social media: Machine learning at rescue. *Egypt. Inform. J.* **2023**, *24*, 291–302. [CrossRef]
63. Cao, L.; Zhang, H.; Feng, L. Building and Using Personal Knowledge Graph to Improve Suicidal Ideation Detection on Social Media. *IEEE Trans. Multimed.* **2022**, *24*, 87–102. [CrossRef]
64. Shukla, S.S.P.; Singh, M.P. Stacked Classification Approach using Optimized Hybrid Deep Learning Model for Early Prediction of Behaviour Changes on Social Media. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2024**, *23*, 159. [CrossRef]
65. Salmerón-Ríos, A.; García-Díaz, J.A.; Pan, R.; Valencia-García, R. Fine grain emotion analysis in Spanish using linguistic features and transformers. *PeerJ Comput. Sci.* **2024**, *10*, e1992. [CrossRef]
66. Tsai, Y.S.; Chen, A.L.P. Suicide risk assessment using word-level model with dictionary-based risky posts selection. *Multimed. Tools Appl.* **2024**, *83*, 21435–21454. [CrossRef]
67. Ding, Z.; Wang, Z.; Zhang, Y.; Cao, Y.; Liu, Y.; Shen, X.; Tian, Y.; Dai, J. Trade-offs between machine learning and deep learning for mental illness detection on social media. *Sci. Rep.* **2025**, *15*, 14497. [CrossRef]
68. Aldhyani, T.H.H.; Alsubari, S.N.; Alshebami, A.S.; Alkahtani, H.; Ahmed, Z.A.T. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *Int. J. Environ. Res. Public Health* **2022**, *19*, 12635. [CrossRef]
69. Rentz, D.M.; Heckler, W.F.; Barbosa, J.L.V. A computational model for assisting individuals with suicidal ideation based on context histories. *Univers. Access Inf. Soc.* **2024**, *23*, 1447–1466. [CrossRef]
70. Ji, S.; Yu, C.P.; Fung, S.f.; Pan, S.; Long, G. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity* **2018**, *2018*, 6157249. [CrossRef]
71. Shukla, S.S.P.; Singh, M.P. Enhancing suicidal ideation detection through advanced feature selection and stacked deep learning models. *Appl. Intell.* **2025**, *55*, 303. [CrossRef]
72. Ding, T.; Qu, T.; Zou, Z.; Ding, C. A novel multi-model feature generation technique for suicide detection. *PeerJ Comput. Sci.* **2024**, *10*, e2301. [CrossRef]
73. Li, J.; Zhang, S.; Zhang, Y.; Lin, H.; Wang, J. Multifeature Fusion Attention Network for Suicide Risk Assessment Based on Social Media: Algorithm Development and Validation. *JMIR Med. Inf.* **2021**, *9*, e28227. [CrossRef]
74. Lao, C.; Lane, J.; Suominen, H. Analyzing Suicide Risk From Linguistic Features in Social Media: Evaluation Study. *JMIR Form. Res.* **2022**, *6*, e35563. [CrossRef]
75. Alghazzawi, D.; Ullah, H.; Tabassum, N.; Badri, S.K.; Asghar, M.Z. Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Sci. Rep.* **2025**, *15*, 1111. [CrossRef]
76. Haroon-Sulyman, S.O.; Ahmad, F.K. Diversity-Aware Attention Transformer for Text Anomaly Detection in Suicidal Ideation. *IEEE Access* **2026**, *14*, 36688–36696. [CrossRef]
77. Park, K.; Sakong, W. A Unified Multi-Task Architecture for Suicide Risk Detection via Attention-Based Decision and Cause Sentence Extraction. *IEEE Access* **2026**, *14*, 39154–39170. [CrossRef]
78. Gaur, M.; Aribandi, V.; Alambo, A.; Kursuncu, U.; Thirunarayan, K.; Beich, J.; Pathak, J.; Sheth, A. Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. *PLoS ONE* **2021**, *16*, e0250448. [CrossRef]
79. Mobin, M.I.; Suaib Akhter, A.F.M.; Mridha, M.F.; Hasan Mahmud, S.M.; Aung, Z. Social Media as a Mirror: Reflecting Mental Health Through Computational Linguistics. *IEEE Access* **2024**, *12*, 130143–130164. [CrossRef]
80. Garcia-Galindo, M.D.C.; Hernandez-Castaneda, A.; Garcia-Hernandez, R.A.; Ledeneva, Y. Emotion-Aware LSTM Networks for Detection of Suicide Notes. *Comput. Sist.* **2026**, *30*, 107–118. [CrossRef]
81. Zulfiker, M.S.; Kabir, N.; Biswas, A.A.; Adan, M.M.I.Y.; Uddin, M.S. Identifying suicidal ideations from social media posts using deep learning and explainable AI-driven approach. *Appl. Soft Comput.* **2025**, *184*, 113813. [CrossRef]
82. Emre Aladag, A.; Muderrisoglu, S.; Akbas, N.B.; Zahmacioglu, O.; Bingol, H.O. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *J. Med. Internet Res.* **2018**, *20*, e215. [CrossRef]
83. Mokni, R.; Haoues, M. Hybrid BERT-XGBoost transformer models for predicting mental health from social media. *World Wide-Web-Internet Web Inf. Syst.* **2026**, *29*, 21. [CrossRef]
84. Sawhney, R.; Joshi, H.; Gandhi, S.; Jin, D.; Shah, R.R. Robust suicide risk assessment on social media via deep adversarial learning. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 1497–1506. [CrossRef]
85. Tekale, A.; Masum, M. Clinically Aligned Long-Context Transformers for Cross-Platform Mental Health Risk Detection. *Electronics* **2026**, *15*, 1403. [CrossRef]
86. Zhang, D.; Zhou, L.; Tao, J.; Zhu, T.; Gao, G.G. KETCH: A Knowledge-Enhanced Transformer-Based Approach to Suicidal Ideation Detection from Social Media Content. *Inf. Syst. Res.* **2025**, *36*, 572–599. [CrossRef]
87. Rezapour, M. Contextual evaluation of suicide-related posts. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 895. [CrossRef]
88. Krishnan, L.; Kuppusamy, A. The Prevention and Prediction of Suicide and Depression: A Population-Based Study on Machine Learning Approach. *Int. J. Agric. Stat. Sci.* **2023**, *19*, 743–751. [CrossRef]

89. Kumar, V.S.R.; Kanna, P.R.; Chitra, K.; Gunasundari, C. PEARL-MH: Psycholinguistic emotion-aware reconstruction learning for enhanced multi-class mental health sentiment classification. *Biomed. Signal Process. Control* **2026**, *122*, 110340. [CrossRef]
90. Bokolo, B.G.; Liu, Q. Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts. *Electronics* **2024**, *13*, 3980. [CrossRef]
91. Rahman, A.B.S.; Ta, H.T.; Najjar, L.; Azadmanesh, A.; Gonul, A.S. DepressionEmo: A novel dataset for multilabel classification of depression emotions. *J. Affect. Disord.* **2024**, *366*, 445–458. [CrossRef]
92. Jena, M.; Khan, N.; Lee, M.Y.; Rho, S. Neuro-Symbolic Graph Learning for Causal Inference and Continual Learning in Mental-Health Risk Assessment. *CMES-Comput. Model. Eng. Sci.* **2026**, *146*, 44. [CrossRef]
93. Hidalgo, R.U.C.; Vidal, M.T. Four Supervised Models for Identifying Suicide Indicators in Text Data. *Int. J. Comb. Optim. Probl. Inform.* **2025**, *16*, 505–514. [CrossRef]
94. Priya, S.K.; Karthika, K.P. A Contemporary Multi-Objective Feature Selection Model For Depression Detection Using A Hybrid Pbgsk Optimization Algorithm. *Int. J. Appl. Math. Comput. Sci.* **2023**, *33*, 117–131. [CrossRef]
95. Gorai, J.; Shaw, D.K. A BERT-encoded ensembled CNN model for suicide risk identification in social media posts. *Neural Comput. Appl.* **2024**, *36*, 10955–10970. [CrossRef]
96. Wagay, F.A.; Jahiruddin; Altaf, Y. MentalRoBERTa-Caps: A capsule-enhanced transformer model for mental health classification. *MethodsX* **2025**, *15*, 103483. [CrossRef]
97. Kboubi, F.; Habacha Chaibi, A. Hybrid transformer-based approaches for suicidal posts detection. *Int. J. Comput. Appl.* **2026**, *48*, 269–286. [CrossRef]
98. Srivastava, S.; Arora, N.; Gupta, V. An Efficient Deep Learning Model Using Harris-Hawk Optimizer for Prognostication of Mental Health Disorders. *Int. Res. J. Multidiscip. Technovation* **2024**, *6*, 106–123. [CrossRef]
99. Ghanadian, H.; Nejadgholi, I.; Al Osman, H. Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models. *IEEE Access* **2024**, *12*, 14350–14363. [CrossRef]
100. Sonawane, J.S.; Jain, D. Integrated Multi factor Trajectory Deep Learning Model for Suicide Ideation Detection from Social Media Interactions. *Int. J. Intell. Eng. Syst.* **2025**, *18*, 524–540. [CrossRef]
101. Alshahrani, A.; Liu, X.; Ringer, F.; Alharbi, R.; Alharthi, K.; Alshahrani, S.; Abdelghany, A. Assessing Suicide Risk Through Social Media Posts Using Transformer-Based Models. *J. Theor. Appl. Inf. Technol.* **2025**, *103*, 8562–8570.
102. Esmi, N.; Shahbahrami, A.; Gaydadjiev, G.; de Jonge, P. Suicide ideation detection based on documents dimensionality expansion. *Comput. Biol. Med.* **2025**, *192*, 110266. [CrossRef]
103. Ditta, A.; Iftikhar, N.; Rana, K.G.; Shah, A.A.; Khan, M.A. Suicidal Thoughts Detection on Social Media in Roman Urdu Using an Attention-Based Hybrid Deep Learning Model. *Eng. Rep.* **2026**, *8*, E70810. [CrossRef]
104. Mansoor, M.A.; Ansari, K.H. Early Detection of Mental Health Crises through Artifical-Intelligence-Powered Social Media Analysis: A Prospective Observational Study. *J. Pers. Med.* **2024**, *14*, 958. [CrossRef]
105. Lobo, S.R.; Kamath, S. Leveraging Text Analysis and Machine Learning Models for Suicide Ideation Detection in Social Media Posts. *SN Comput. Sci.* **2025**, *6*, 703. [CrossRef]
106. Kodati, D.; Tene, R. Advancing mental health detection in texts via multi-task learning with soft-parameter sharing transformers. *Neural Comput. Appl.* **2025**, *37*, 3077–3110. [CrossRef]
107. Shukla, S.S.P.; Singh, M.P. Exploring ensemble optimized voting and stacking classifiers through Cross-validation for early detection of suicidal ideation. *J. Intell. Fuzzy Syst.* **2024**, *47*, 335–349. [CrossRef]
108. Lasri, S.; Nfaoui, E.; Mrizik, K. Suicide Ideation and Risk Detection from Social Media Using GPT Models. *J. Comput. Sci.* **2024**, *20*, 1349–1356. [CrossRef]
109. Li, J.; Chen, X.; Lin, Z.; Yang, K.; Leong, H.V.; Yu, N.X.; Li, Q. Suicide risk level prediction and suicide trigger detection: A benchmark dataset. *HKIE Trans. Hong Kong Inst. Eng.* **2022**, *29*, 268–282. [CrossRef]
110. Jere, S.; Patil, A.P. Detection of Suicidal Ideation Based on Relational Graph Attention Network with DNN Classifier. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 321–332.
111. Soumya, K.; Garg, V.K. Named Entity Emotion Intensity Tagging for Suicidal Ideation Detection From Social Media Texts During MT. *Int. J. Intell. Eng. Syst.* **2022**, *15*, 224–236. [CrossRef]
112. Tadesse, M.M.; Lin, H.; Xu, B.; Yang, L. Detection of suicide ideation in social media forums using deep learning. *Algorithms* **2020**, *13*, 7. [CrossRef]
113. Vioulès, M.J.; Moulahi, B.; Azé, J.; Bringay, S. Detection of suicide-related posts in Twitter data streams. *IBM J. Res. Dev.* **2018**, *62*, 7:1–7:12. [CrossRef]
114. Schoene, A.M.; Bojanić, L.; Nghiem, M.Q.; Hunt, I.M.; Ananiadou, S. Classifying Suicide-Related Content and Emotions on Twitter Using Graph Convolutional Neural Networks. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1791–1802. [CrossRef]
115. Tumaliuan, F.B.; Grepo, L.; Jalao, E.R. Development of a two-stage depression symptom detection model: Application of neural networks to twitter data. *Front. Comput. Sci.* **2024**, *6*, 1399395. [CrossRef]

116. Chadha, A.; Kaushik, B. Performance Evaluation of Learning Models for Identification of Suicidal Thoughts. *Comput. J.* **2022**, *65*, 139–154. [CrossRef]
117. Du, J.; Zhang, Y.; Luo, J.; Jia, Y.; Wei, Q.; Tao, C.; Xu, H. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 43. [CrossRef]
118. Diniz, E.J.S.; Fontenele, J.E.; de Oliveira, A.C.; Bastos, V.H.; Teixeira, S.; Rabêlo, R.L.; Calçada, D.B.; dos Santos, R.M.; de Oliveira, A.K.; Teles, A.S. Boamente: A Natural Language Processing-Based Digital Phenotyping Tool for Smart Monitoring of Suicidal Ideation. *Healthcare* **2022**, *10*, 698. [CrossRef]
119. Mbarek, A.; Jamoussi, S.; Hamadou, A.B. An across online social networks profile building approach: Application to suicidal ideation detection. *Future Gener. Comput. Syst.* **2022**, *133*, 171–183. [CrossRef]
120. Lingam, G.; Das, S.K. Fuzzy-Based Deep Reinforcement Learning for Suicidal Ideation Detection in Online Social Networks. *IEEE Trans. Comput. Soc. Syst.* **2026**, *13*, 2661–2677. [CrossRef]
121. Haque, R.; Islam, N.; Islam, M.; Ahsan, M.M. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies* **2022**, *10*, 57. [CrossRef]
122. Allam, H.; Davison, C.; Kalota, F.; Lazaros, E.; Hua, D. AI-Driven Mental Health Surveillance: Identifying Suicidal Ideation Through Machine Learning Techniques. *Big Data Cogn. Comput.* **2025**, *9*, 16. [CrossRef]
123. Abdulsalam, A.; Alhothali, A.; Al-Ghamdi, S. Detecting Suicidality in Arabic Tweets Using Machine Learning and Deep Learning Techniques. *Arab. J. Sci. Eng.* **2024**, *49*, 12729–12742. [CrossRef]
124. Chadha, A.; Kaushik, B. A Survey on Prediction of Suicidal Ideation Using Machine and Ensemble Learning. *Comput. J.* **2021**, *64*, 1617–1632. [CrossRef]
125. Boonyarat, P.; Liew, D.J.; Chang, Y.C. Leveraging enhanced BERT models for detecting suicidal ideation in Thai social media content amidst COVID-19. *Inf. Process. Manag.* **2024**, *61*, 103706. [CrossRef]
126. Priyamvada, B.; Singhal, S.; Nayyar, A.; Jain, R.; Goel, P.; Rani, M.; Srivastava, M. Stacked CNN-LSTM approach for prediction of suicidal ideation on social media. *Multimed. Tools Appl.* **2023**, *82*, 27883–27904. [CrossRef]
127. Roy, A.; Nikolitch, K.; McGinn, R.; Jinah, S.; Klement, W.; Kaminsky, Z.A. A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digit. Med.* **2020**, *3*, 78. [CrossRef]
128. Kumar, E.R.; Venkatram, N. Predicting and analyzing suicidal risk behavior using rule-based approach in Twitter data. *Soft Comput.* **2023**, *28*, 13821–13829. [CrossRef]
129. Acuña Caicedo, R.W.; Gómez Soriano, J.M.; Melgar Sasieta, H.A. Assessment of supervised classifiers for the task of detecting messages with suicidal ideation. *Heliyon* **2020**, *6*, e04412. [CrossRef]
130. Metzler, H.; Baginski, H.; Niederkrotenthaler, T.; Garcia, D. Detecting Potentially Harmful and Protective Suicide-Related Content on Twitter: Machine Learning Approach. *J. Med. Internet Res.* **2022**, *24*, e34705. [CrossRef]
131. Patel, D.; Sumner, S.A.; Bowen, D.; Zwald, M.; Yard, E.; Wang, J.; Law, R.; Holland, K.; Nguyen, T.; Mower, G.; et al. Predicting state level suicide fatalities in the united states with realtime data and machine learning. *npj Ment. Health Res.* **2024**, *3*, 3. [CrossRef]
132. Kim, D.; Jiang, T.; Baek, J.H.; Jang, S.H.; Zhu, Y. Understanding and comparing risk factors and subtypes in South Korean adult and adolescent women's suicidal ideation or suicide attempt using survey and social media data. *Digit. Health* **2024**, *10*, 20552076241255660. [CrossRef]
133. Rostianingsih, S.; Wijaya, S.C.; Budhi, G.S.; Adipranata, R. Machine Learning for Suicide Ideation Detection in Twitter. *TEM J.-Technol. Educ. Manag. Inform.* **2026**, *15*, 485–492. [CrossRef]
134. Sarsam, S.M.; Al-Samarraie, H.; Alzahrani, A.I.; Mon, C.S.; Shibghatullah, A.S. Characterizing Suicide Ideation by Using Mental Disorder Features on Microblogs: A Machine Learning Perspective. *Int. J. Ment. Health Addict.* **2024**, *22*, 1783–1796. [CrossRef]
135. Kim, D.; Quan, L.; Seo, M.; Kim, K.; Kim, J.W.; Zhu, Y. Interpretable machine learning-based approaches for understanding suicide risk and protective factors among South Korean females using survey and social media data. *Suicide Life-Threat. Behav.* **2023**, *53*, 484–498. [CrossRef]
136. Jung, W.; Kim, D.; Nam, S.; Zhu, Y. Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning. *Arch. Suicide Res.* **2023**, *27*, 13–28. [CrossRef]
137. Baghdadi, N.A.; Malki, A.; Balaha, H.M.; AbdulAzeem, Y.; Badawy, M.; Elhosseini, M. An optimized deep learning approach for suicide detection through Arabic tweets. *PeerJ Comput. Sci.* **2022**, *8*, e1070. [CrossRef]
138. Niederkrotenthaler, T.; Tran, U.S.; Baginski, H.; Sinyor, M.; Strauss, M.J.; Sumner, S.A.; Voracek, M.; Till, B.; Murphy, S.; Gonzalez, F.; et al. Association of 7 million+ tweets featuring suicide-related content with daily calls to the Suicide Prevention Lifeline and with suicides, United States, 2016-2018. *Aust. N. Z. J. Psychiatry* **2023**, *57*, 994–1003. [CrossRef]
139. Kancharapu, R.; Ayyagari, S.N. Suicidal ideation prediction based on social media posts using a GAN-infused deep learning framework with genetic optimization and word embedding fusion. *Int. J. Inf. Technol.* **2024**, *16*, 2577–2593. [CrossRef]
140. Xie, G. Suicidal Ideation Detection and Influential Keyword Extraction from Twitter using Deep Learning (SID). *EAI Endorsed Trans. Pervasive Health Technol* **2024**, *10*, 1–7. [CrossRef]

141. Raghu Nath, V.V.N.S.S.; Thind, T.; Kaur, U. AI Against Suicide: Real–Time Risk Detection from Social Media Texts. *Int. J. Drug Deliv. Technol.* **2026**, *16*, 137–144. [CrossRef]
142. Jareanpon, C.; Kiatjindarat, W.; Doenribram, D.; Thanarat, P. The Comparison Of Depressive Symptoms Classification From Twitter Between NaïVe Bayes, Moadsp, and Lstm Based on DSM-5. *ICIC Express Lett. Part B Appl.* **2025**, *16*, 259–267. [CrossRef]
143. Benjachairat, P.; Senivongse, T.; Taephant, N.; Puvapaisankit, J.; Maturosjamnan, C.; Kultananawat, T. Classification of suicidal ideation severity from Twitter messages using machine learning. *Int. J. Inf. Manag. Data Insights* **2024**, *4*, 100280. [CrossRef]
144. Chatterjee, M.; Kumar, P.; Samanta, P.; Sarkar, D. Suicide ideation detection from online social media: A multi-modal feature based technique. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100103. [CrossRef]
145. Lim, Y.Q.; Loo, Y.L. Characteristics of Multiclass Suicide Risks Tweets through Feature Extraction and Machine Learning Techniques. *Int. J. Inform. Vis.* **2023**, *7*, 2297–2305. [CrossRef]
146. Kancharapu, R.; Ayyagari, S.N. Unveiling the Role of Social Media in Mental Health: A GAN-based Deep Learning Framework for Suicide Prevention. *Int. J. Intell. Syst. Appl. Eng.* **2023**, *11*, 489–502.
147. Braithwaite, S.R.; Giraud-Carrier, C.; West, J.; Barnes, M.D.; Hanson, C.L. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Ment. Health* **2016**, *3*, E21. [CrossRef]
148. Valeriano, K.; Condori-Larico, A.; Sulla-Torres, J. Detection of suicidal intent in Spanish language social networks using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 688–698. [CrossRef]
149. Manisha, M.; Kodali, A.; Srilakshmi, V. Machine classification for suicide ideation detection on twitter. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *8*, 4154–4160. [CrossRef]
150. Dogrucu, A.; Perucic, A.; Isaro, A.; Ball, D.; Toto, E.; Rundensteiner, E.A.; Agu, E.; Davis-Martin, R.; Boudreaux, E. Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data. *Smart Health* **2020**, *17*, 100118. [CrossRef]
151. Rajesh Kumar, E.; Rama Rao, K.; Nayak, S.R.; Chandra, R. Suicidal ideation prediction in twitter data using machine learning techniques. *J. Interdiscip. Math.* **2020**, *23*, 117–125. [CrossRef]
152. Rabani, S.T.; Khan, Q.R.; Khanday, A.M.U.D. Detection of suicidal ideation on Twitter using machine learning & ensemble approaches. *Baghdad Sci. J.* **2020**, *17*, 1328–1339. [CrossRef]
153. Rao, K.V. Suicide prediction on social media by implementing sentiment analysis along with machine learning. *Int. J. Recent Technol. Eng.* **2019**, *8*, 4833–4837. [CrossRef]
154. Qi, H.; Fu, G.; Li, J.; Song, C.; Zhai, W.; Luo, D.; Liu, S.; Yu, Y.; Yang, B.; Zhao, Q. Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media. *Bioengineering* **2025**, *12*, 882. [CrossRef]
155. Dai, Y.; Liu, J.; Cao, L.; Xue, Y.; Wang, X.; Ding, Y.; Tian, J.; Feng, L. Leveraging Social Media for Real-Time Interpretable and Amendable Suicide Risk Prediction With Human-in-The-Loop. *IEEE Trans. Affect. Comput.* **2025**, *16*, 1128–1145. [CrossRef]
156. Meng, X.; Wang, C.; Yang, J.; Li, M.; Zhang, Y.; Wang, L. Predicting Users' Latent Suicidal Risk in Social Media: An Ensemble Model Based on Social Network Relationships. *Comput. Mater. Contin.* **2024**, *79*, 4259–4281. [CrossRef]
157. Li, Z.; Cheng, W.; Zhou, J.; An, Z.; Hu, B. Deep learning model with multi-feature fusion and label association for suicide detection. *Multimed. Syst.* **2023**, *29*, 2193–2203. [CrossRef]
158. Li, Z.; Zhou, J.; An, Z.; Cheng, W.; Hu, B. Deep Hierarchical Ensemble Model for Suicide Detection on Imbalanced Social Media Data. *Entropy* **2022**, *24*, 442. [CrossRef]
159. Wang, R.; Yang, B.X.; Ma, Y.; Wang, P.; Yu, Q.; Zong, X.; Huang, Z.; Ma, S.; Hu, L.; Hwang, K.; et al. Medical-Level Suicide Risk Analysis: A Novel Standard and Evaluation Model. *IEEE Internet Things J.* **2021**, *8*, 16825–16834. [CrossRef]
160. Liu, X.; Liu, X.; Sun, J.; Yu, N.X.; Sun, B.; Li, Q.; Zhu, T. Proactive Suicide Prevention Online (PSPO): Machine Identification and Crisis Management for Chinese Social Media Users with Suicidal Thoughts and Behaviors. *J. Med. Internet Res.* **2019**, *21*, e11705. [CrossRef]
161. Huang, Z.; Hu, Q. Tree hole rescue: An AI approach for suicide risk detection and online suicide intervention. *Health Inf. Sci. Syst.* **2024**, *12*, 45. [CrossRef]
162. Gu, Y.; Chen, D.; Liu, X. Suicide Possibility Scale Detection via Sina Weibo Analytics: Preliminary Results. *Int. J. Environ. Res. Public Health* **2023**, *20*, 466. [CrossRef]
163. Fan, H.; Yin, Q.; Kuang, J.; Yan, Z. Suicide risk prediction for users with depression in question answering communities: A design based on deep learning. *Inf. Manag.* **2025**, *62*, 104219. [CrossRef]
164. Yang, B.X.; Xia, L.; Liu, L.; Nie, W.; Liu, Q.; Li, X.Y.; Ao, M.Q.; Wang, X.Q.; Xie, Y.D.; Liu, Z.; et al. A Suicide Monitoring and Crisis Intervention Strategy Based on Knowledge Graph Technology for "Tree Hole" Microblog Users in China. *Front. Psychol.* **2021**, *12*, 674481. [CrossRef]
165. Liu, J.; Shi, M.; Jiang, H. Detecting Suicidal Ideation in Social Media: An Ensemble Method Based on Feature Fusion. *Int. J. Environ. Res. Public Health* **2022**, *19*, 8197. [CrossRef]

166. Li, S.; Pan, W.; Yip, P.S.F.; Wang, J.; Zhou, W.; Zhu, T. Uncovering the heterogeneous effects of depression on suicide risk conditioned by linguistic features: A double machine learning approach. *Comput. Hum. Behav.* **2024**, *152*, 108080. [CrossRef]
167. Meng, X.; Cui, X.; Zhang, Y.; Wang, S.; Wang, C.; Li, M.; Yang, J. Mining Suicidal Ideation in Chinese Social Media: A Dual-Channel Deep Learning Model with Information Gain Optimization. *Entropy* **2025**, *27*, 116. [CrossRef]
168. Yao, X.; Yu, G.; Tang, J.; Zhang, J. Extracting depressive symptoms and their associations from an online depression community. *Comput. Hum. Behav.* **2021**, *120*, 106734. [CrossRef]
169. Cheng, Q.; Li, T.M.H.; Kwok, C.L.; Zhu, T.; Yip, P.S.F. Assessing Suicide Risk and Emotional Distress in Chinese Social Media: A Text Mining and Machine Learning Study. *J. Med. Internet Res.* **2017**, *19*, e243. [CrossRef]
170. Xue, Y.; Li, Q.; Wu, T.; Feng, L.; Zhao, L.; Yu, F. Incorporating stress status in suicide detection through microblog. *Comput. Syst. Sci. Eng.* **2019**, *34*, 65–78. [CrossRef]
171. Cao, L.; Feng, L.; Ding, Y.; Zhang, H.; Wang, X.; Zeng, K.; Dai, Y. Online continuous learning of users suicidal risk on social media. *Artif. Intell. Med.* **2025**, *167*, 103199. [CrossRef]
172. Fu, G.; Song, C.; Li, J.; Ma, Y.; Chen, P.; Wang, R.; Yang, B.X.; Huang, Z. Distant Supervision for Mental Health Management in Social Media: Suicide Risk Classification System Development Study. *J. Med. Internet Res.* **2021**, *23*, e26119. [CrossRef]
173. Pan, W.; Wang, X.; Zhou, W.; Hang, B.; Guo, L. Linguistic Analysis for Identifying Depression and Subsequent Suicidal Ideation on Weibo: Machine Learning Approaches. *Int. J. Environ. Res. Public Health* **2023**, *20*, 2688. [CrossRef]
174. Guan, L.; Hao, B.; Cheng, Q.; Yip, P.S.; Zhu, T. Identifying chinese microblog users with high suicide probability using internet-based profile and linguistic features: Classification model. *JMIR Ment. Health* **2015**, *2*, E17. [CrossRef]
175. Rahman, M.S.; Halder, S.; Uddin, M.A.; Acharjee, U.K. An efficient hybrid system for anomaly detection in social networks. *Cybersecurity* **2021**, *4*, 10. [CrossRef]
176. Ophir, Y.; Tikochinski, R.; Asterhan, C.S.C.; Sisso, I.; Reichart, R. Deep neural networks detect suicide risk from textual facebook posts. *Sci. Rep.* **2020**, *10*, 16685. [CrossRef]
177. Das Swain, V.; Ye, J.; Ramesh, S.K.; Mondal, A.; Abowd, G.; De Choudhury, M. Leveraging Social Media to Predict COVID-19-Induced Disruptionsto Mental Well-Being Among University Students:Modeling Study. *JMIR Form. Res.* **2024**, *8*, e52316. [CrossRef]
178. Badian, Y.; Ophir, Y.; Tikochinski, R.; Calderon, N.; Klomek, A.B.; Fruchter, E.; Reichart, R. Social Media Images Can Predict Suicide Risk Using Interpretable Large Language-Vision Models. *J. Clin. Psychiatry* **2024**, *85*, 23m14962. [CrossRef]
179. Nedunchezhian, P.; Mahalingam, M. The Improved Depression Recovery Motivation Recommendation System (I-DRMRS) in Online Social Networks. *SN Comput. Sci.* **2022**, *3*, 166. [CrossRef]
180. Lekkas, D.; Klein, R.J.; Jacobson, N.C. Predicting acute suicidal ideation on Instagram using ensemble machine learning models. *Internet Interv.* **2021**, *25*, 100424. [CrossRef]
181. Qadir, S.; Alsoubai, A.; Park, J.K.; Ali, N.S.; Choudhury, M.D.; Wisniewski, P. Triaging Casual From Critical-Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media: Algorithm Development and Validation Study. *JMIR Ment. Health* **2026**, *13*, e76051. [CrossRef]
182. Brown, R.C.; Bendig, E.; Fischer, T.; Goldwich, A.D.; Baumeister, H.; Plener, P.L. Can acute suicidality be predicted by Instagram data? Results from qualitative and quantitative language analyses. *PLoS ONE* **2019**, *14*, e0220623. [CrossRef]
183. Vongtangton, P.; Goebert, D. Social Media Content and Suicidality: Implications for Practice. *Psychol. Serv.* **2026**, *23*, 64–72. [CrossRef]
184. Zuromski, K.; Low, D.; Jones, N.; Kuzma, R.; Kessler, D.; Zhou, L.; Kastman, E.; Epstein, J.; Madden, C.; Ghosh, S.; et al. Detecting suicide risk among US servicemembers and veterans: A deep learning approach using social media data. *Psychol. Med.* **2024**, *54*, 3379–3388. [CrossRef]
185. Wu, E.L.; Wu, C.Y.; Lee, M.B.; Chu, K.C.; Huang, M.S. Development of Internet suicide message identification and the Monitoring-Tracking-Rescuing model in Taiwan. *J. Affect. Disord.* **2023**, *320*, 37–41. [CrossRef]
186. Desmet, B.; Hoste, V. Online suicide prevention through optimised text classification. *Inf. Sci.* **2018**, *439*, 61–78. [CrossRef]
187. Lee, K.S.; Lee, H.; Myung, W.; Song, G.Y.; Lee, K.; Kim, H.; Carroll, B.J.; Kim, D.K. Advanced Daily Prediction Model for National Suicide Numbers with Social Media Data. *Psychiatry Investig.* **2018**, *15*, 344–354. [CrossRef]
188. Lekkas, D.; Jacobson, N.C. Breaking the silence: Leveraging social interaction data to identify high-risk suicide users online using network analysis and machine learning. *Sci. Rep.* **2024**, *14*, 19395. [CrossRef]
189. Narynov, S.; Kozhakhmet, K.; Mukhtarkhanuly, D.; Saparkhojayev, N. Applying natural language processing techniques for suicidal content detection in social media. *J. Theor. Appl. Inf. Technol.* **2020**, *98*, 3390–3404.
190. Goh, T.T.; Huang, Y.P. Monitoring youth depression risk in Web 2.0. *Vine* **2009**, *39*, 192–202. [CrossRef]
191. Franz, P.J.; Nook, E.C.; Mair, P.; Nock, M.K. Using Topic Modeling to Detect and Describe Self-Injurious and Related Content on a Large-Scale Digital Platform. *Suicide Life-Threat. Behav.* **2020**, *50*, 5–18. [CrossRef]
192. Nowacki, A.; Sitek, W.; Rybiński, H. LLM-based classifiers for discovering mental disorders. *J. Intell. Inf. Syst.* **2026**, *64*, 161–178. [CrossRef]

193. Saha, S.; Dasgupta, S.; Anam, A.; Saha, R.; Nath, S.; Dutta, S. An Investigation of Suicidal Ideation from Social Media Using Machine Learning Approach. *Baghdad Sci. J.* **2023**, *20*, 1164–1181. [CrossRef]
194. Paul, S.C.; Jahan, B.; Mamun, A.; Hossen, M.J. An Approach to Detect Suicidal Bengali Posts from Social Media Using Machine Learning Algorithms. *Int. J. Eng. Trends Technol.* **2024**, *72*, 43–50. [CrossRef]
195. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* **2024**, *385*, e078378. [CrossRef]
196. Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; van Calster, B.; Celi, L.A.; Denaxas, S.; et al. PROBAST+AI: An Updated Quality, Risk of Bias, and Applicability Assessment Tool for Prediction Models Using Regression or Artificial Intelligence Methods. *BMJ* **2025**, *388*, e082505. [CrossRef]
197. Fennig, U.; Yom-Tov, E.; Savitzky, L.; Nissan, J.; Altman, K.; Loebenstein, R.; Boxer, M.; Weinberg, N.; Gofrit, S.G.; Maggio, N. Bridging the conversational gap in epilepsy: Using large language models to reveal insights into patient behavior and concerns from online discussions. *Epilepsia* **2025**, *66*, 686–699. [CrossRef]
198. Mohamed, R.; Younis, E.M.G.; Ali, A.A. Leveraging large language models and ensemble techniques for real-time mental health monitoring on social media. *Netw. Model. Anal. Health Inform. Bioinform.* **2025**, *14*, 157. [CrossRef]
199. Lissak, S.; Ophir, Y.; Tikochinski, R.; Brunstein Klomek, A.; Sisso, I.; Fruchter, E.; Reichart, R. Bored to death: Artificial Intelligence research reveals the role of boredom in suicide behavior. *Front. Psychiatry* **2024**, *15*, 1328122. [CrossRef]
200. Sobol, I.; Lissak, S.; Tikochinski, R.; Nakash, T.; Brunstein Klomek, A.; Fruchter, E.; Reichart, R. Bridging online behavior and clinical insight: A longitudinal LLM-based study of suicidality on YouTube reveals novel digital markers. *J. Affect. Disord.* **2026**, *400*, 121072. [CrossRef]
201. Tuarob, S.; Chatrinan, K.; Noraset, T.; Tawichsri, T.; Thaipisutikul, T. Forecasting National-Level Self-Harm Trends With Social Networks. *IEEE Access* **2023**, *11*, 64796–64814. [CrossRef]
202. Das, S.; Walker, D.; Rajwal, S.; Lakamana, S.; Sumner, S.A.; Mack, K.A.; Kaczkowski, W.; Sarker, A. Emerging Trends of Self-Harm Using Sodium Nitrite in an Online Suicide Community: Observational Study Using Natural Language Processing Analysis. *JMIR Ment Health* **2024**, *11*, e53730. [CrossRef]
203. Kumar, R.; Mukherjee, S.; Choi, T.M.; Dhamotharan, L. Mining voices from self-expressed messages on social-media: Diagnostics of mental distress during COVID-19. *Decis. Support Syst.* **2022**, *162*, 113792. [CrossRef]
204. Nguyen, K.; Vu, B.; Chandna, S.; Schultz, J.H.; Mayer, G. Between the lines: Investigating health beliefs and emotional expressions in online mental health communities. *Front. Psychol.* **2025**, *16*, 1521623. [CrossRef]
205. Murray, C.; Mitchell, L.; Tuke, J.; Mackay, M. Probabilistic emotion and sentiment modelling of patient-reported experiences. *Artif. Intell. Med.* **2025**, *167*, 103178. [CrossRef]
206. Yang, B.X.; Chen, P.; Li, X.Y.; Yang, F.; Huang, Z.; Fu, G.; Luo, D.; Wang, X.Q.; Li, W.; Wen, L.; et al. Characteristics of High Suicide Risk Messages From Users of a Social Network—Sina Weibo "Tree Hole". *Front. Psychiatry* **2022**, *13*, 789504. [CrossRef]
207. Kennedy, C.J.; Kearns, J.C.; Geraci, J.C.; Gildea, S.M.; Hwang, I.H.; King, A.J.; Liu, H.; Luedtke, A.; Marx, B.P.; Papini, S.; et al. Predicting Suicides Among US Army Soldiers After Leaving Active Service. *JAMA Psychiatry* **2024**, *81*, 1215–1224. [CrossRef]
208. Chau, M.; Li, T.M.H.; Wong, P.W.C.; Xu, J.J.; Yip, P.S.F.; Chen, H. Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-Based Classification. *MIS Q.* **2020**, *44*, 933–955. [CrossRef]
209. Yang, Q.; Zhou, J.; Wei, Z. Time Perspective-Enhanced Suicidal Ideation Detection Using Multi-Task Learning. *Int. J. Netw. Dyn. Intell.* **2024**, *3*, 100011. [CrossRef]
210. Agarwal, D.; Singh, V.; Singh, A.K.; Madan, P. Dwarf Updated Pelican Optimization Algorithm for Depression and Suicide Detection from Social Media. *Psychiatr. Q.* **2025**, *96*, 529–562. [CrossRef]
211. Purushothaman, V.; Li, J.; Mackey, T.K. Detecting Suicide and Self-Harm Discussions Among Opioid Substance Users on Instagram Using Machine Learning. *Front. Psychiatry* **2021**, *12*, 551296. [CrossRef]
212. Mo, L.; Li, H.; Zhu, T. Exploring the Suicide Mechanism Path of High-Suicide-Risk Adolescents—Based on Weibo Text Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 11495. [CrossRef]
213. Saha, K.; Yousuf, A.; Boyd, R.L.; Pennebaker, J.W.; De Choudhury, M. Social Media Discussions Predict Mental Health Consultations on College Campuses. *Sci. Rep.* **2022**, *12*, 123. [CrossRef]
214. Gilgur, A.; Ramirez-Marquez, J.E. Using Deductive Reasoning to Identify Unhappy Communities. *Soc. Indic. Res.* **2020**, *152*, 581–605. [CrossRef]
215. Sadashiva, S.; Manjunath, K.G. Multimodal Deep Learning Framework Using Transformer and LSTM Models for Suicide Risk Detection from Social Media and Voice Data. *Int. J. Saf. Secur. Eng.* **2025**, *15*, 1445–1459. [CrossRef]
216. Kodati, D.; Tene, R. Emotion mining for early suicidal threat detection on both social media and suicide notes using context dynamic masking-based transformer with deep learning. *Multimed. Tools Appl.* **2025**, *84*, 11729–11752. [CrossRef]
217. Kumar, E.R.; Rao, K.V.S.N.R. Sentiment Analysis using Social and Topic Context for Suicide Prediction. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 388–397. [CrossRef]

218. Zuberi, A.H.; Anees, A.; Anjum, N.; Warsi, A.H.; Khan, P.R.; Singh, S.K.; Singh, N.K.; Singh, R.; Abbas, S.H.; Ranjan, R. Machine Learning-Based Sentiment Analysis for Suicide Prevention and Mental Health Monitoring in Educational Institutions. *J. Neonatal Surg.* **2025**, *14*, 396–405. [CrossRef]
219. Shaikh, A.; Jangid, P.; Rana, M.; Sall, S.; Rana, R.; Mhapasekar, D.P.; Jangid, A. Enhancing Suicide Ideation Detection with Multimodal Deep Learning: Ethical, Explainable, and Privacy-Preserving Framework. *Int. J. Appl. Math.* **2025**, *38*, 1034–1053. [CrossRef]
220. Kancharapu, R.; Ayyagari, S.N.A. A comparative study on word embedding techniques for suicide prediction on COVID-19 tweets using deep learning models. *Int. J. Inf. Technol.* **2023**, *15*, 3293–3306. [CrossRef]
221. Malhotra, A.; Jindal, R. Multimodal Deep Learning based Framework for Detecting Depression and Suicidal Behaviour by Affective Analysis of Social Media Posts. *EAI Endorsed Trans. Pervasive Health Technol.* **2020**, *6*, e1. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.