Scientia Psychiatrica
Vol. 7, No. 2, 2026, pp. 99–110 · e-ISSN 2715-9736 journal homepage: scientiapsychiatrica.com

ORIGINAL RESEARCH ARTICLE

Temporal Transportability of an Ordinal Suicide-Risk Semantic Axis on Reddit, 2018–2020: An Open-Data Reanalysis

Danila Adi Sanjaya[1], Amir Serikova[2], Arya Ganendra[3*]

[1]*Department of Literature Science, CMHC Research Center, Palembang, Indonesia*
[2]*Department of Cultural and Heritage Preservation, Emeritus Research Center, Skopje, North Macedonia*
[3]*Department of Literacy, Emerald Language Institute, Balikpapan, Indonesia*

ARTICLE INFO

Article history
Received: February 20, 2026
Revised: March 19, 2026
Accepted: April 17, 2026
Available online: April 27, 2026
Published: April 30, 2026

Keywords
Columbia-Suicide Severity Rating Scale
Natural language processing
Reddit
Suicide risk
Temporal transportability

***Corresponding author**
Arya Ganendra
E-mail: arya.ganendra@enigma.or.id

DOI
10.37275/scipsy.v7i2.213

ABSTRACT

Background. Language models applied to online mental-health communities meet shifting vocabulary and social context. Temporal transportability matters for population research but is often conflated with clinical validation.

Objective. To develop and internally validate a Columbia-Suicide Severity Rating Scale-informed semantic axis and examine aggregate transport across matched r/SuicideWatch windows, 2018–2020.

Methods. This secondary open-data study used 500 anonymous Reddit users labeled Supportive, Indicator, Ideation, Behavior or Attempt and 46,203 deduplicated r/SuicideWatch posts from matched January–April windows in 2018–2020. A two-stage model (Supportive versus risk-related content, then three cumulative thresholds) used term frequency-inverse document frequency (TF-IDF) unigrams and bigrams. Nested repeated cross-validation selected regularization; the locked model was applied to the temporal corpus. Drift was measured by cosine distance and Jensen-Shannon divergence. Calendar-matched 2020-minus-2019 differences were bootstrapped by day, and 2020 daily means underwent exploratory interrupted-time-series (ITS) analysis.

Results. Internal validation gave an area under the receiver operating characteristic curve (AUC) of 0.841 (95% confidence interval [CI] 0.799–0.876), Spearman rho 0.568 (0.503–0.624), weighted kappa 0.464 (0.392–0.527) and macro-F1 0.292 (0.257–0.326). Attempt recall was 0%. Mean projected scores were 1.969, 1.923 and 1.928 in 2018, 2019 and 2020; matched 2020-minus-2019 difference was 0.0053 (−0.0020 to 0.0128). Exploratory ITS level change was −0.01976 (−0.04062 to 0.00110). Vocabulary drift was measurable, whereas probability-composition divergence remained small.

Conclusion. The axis captured moderate ordinal structure, yet no Attempt-labeled user had Attempt as the highest-probability class. Stable aggregate outputs are not evidence of individual clinical validity. The model supports aggregate transport research, not screening, diagnosis or intervention.

Copyright: © 2026 The Author(s). Published by HM Publisher in collaboration with CMHC Research Center on behalf of Scientia Psychiatrica. The authors retain copyright of their work.

Access licence: This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Assessing suicidal thoughts and behaviors is a high-stakes clinical task. It rests on direct inquiry, temporal context, prior behavior, current intent, access to means, protective factors and professional judgment. The Columbia-Suicide Severity Rating Scale (C-SSRS) offers a structured nomenclature for suicidal ideation and behavior, and it has shown convergent, divergent and predictive validity in several clinical samples.[1] Even so, a category taken from a clinical interview is not a clinically valid outcome merely because comparable language appears in a social-media post. A post may describe a remote event or another person's experience. It may reflect hypothetical reasoning, supportive communication or an immediate personal crisis. Linguistic evidence and a contemporaneous clinical state are different things.

Recent Reddit research has gone beyond simple keyword counts. Gaur and colleagues annotated users for C-SSRS-informed severity and temporality. Their time-variant and time-invariant models emphasized different parts of the construct.[2] Temporal analyses of

suicide-support forums have revealed meaningful diurnal and longitudinal variation,[3] and growth-trajectory modeling has separated heterogeneous patterns among Reddit users at elevated suicide risk.[4] Large-scale linguistic work has recovered dimensions of suicidality across mental-health communities.[5] Movement of users between r/Depression and r/SuicideWatch alters the thematic context in which apparent risk language appears.[6] These studies show that online language carries structured information, and they also show that class definition, temporal context and transport need explicit evaluation.

Online communication sits inside changing social systems, and this is why the distinction matters. Campus-level social-media discussions have predicted later mental-health consultations.[7] Social-media activity has been linked to suicidal behavior during a youth suicide cluster,[8] and surveys during the pandemic have tied patterns of social-media use to mental-health indicators.[9] At the individual level, positive and negative social-media experiences may relate differently to proximal suicidal ideation.[10] Such work supports the research value of online data, yet none of it makes a single post equivalent to a contemporaneous clinical assessment, because posting behavior, moderation, community composition and outside events can shift independently of the underlying clinical phenomenon.

Modelling studies report wide heterogeneity. Transformer models can recognize suicide-note language.[11] Interpretable linguistic features can separate broad social-media risk strata,[12] and ensemble or deep-learning systems can detect suicidal-ideation content in benchmark corpora.[13,14] Newer work has added explainability,[15] turned to the quality of training data[16] and explored continuous learning under changing online distributions.[17] A model can still keep broad discrimination while failing a rare and serious class; its performance can also shift when vocabulary and user composition change.

What remains open is whether the learned semantic axis stays interpretable on later language distributions, which is a separate matter from classifying suicide-related language inside a development corpus. Models that use metadata,[18] linguistic analyses of another platform (Weibo)[19] and semi-supervised social-media classifiers[20] show that the operational signal depends on text and on context. In the r/SuicideWatch corpus used here, however, no contemporaneous C-SSRS labels exist for 2018–2020. A responsible temporal analysis must therefore evaluate observable properties: vocabulary coverage, centroid movement, projected probability composition and aggregate score contrasts; it can make no claim of external clinical accuracy.

The novelty of the study is an integrated, role-aware transport framework with three linked parts. The first is a two-stage ordinal semantic instrument that keeps supportive communication apart from risk-related content, and the second is repeated out-of-fold evaluation of whether the learned ordering holds within the labeled source corpus. The third is a locked-feature transport analysis. It separates movement in recognized vocabulary from movement in projected class composition. Together these parts keep semantic transport distinct from clinical validation. The primary aim was to develop and internally validate a C-SSRS-informed ordinal semantic axis in an anonymous labeled Reddit corpus. The secondary aims were to quantify aggregate transport across matched 2018–2020 r/SuicideWatch windows through paired-day contrasts and supporting drift measures. The 11 March 2020 interrupted time series (ITS) was exploratory. A final objective was to define a safe intended use for the resulting evidence.

2. Methods

2.1. Study design and reporting

This was a secondary open-data study of model development and temporal transport. Internal validation in a labeled user-level corpus was combined with aggregate application to an unlabeled post-level temporal corpus. The development estimands were discrimination between Supportive and risk-related content, the monotonic association between the probabilistic ordinal score and observed labels, agreement of exact predicted categories, class-specific recall and expected-score error. The primary temporal estimands were paired mean differences between 2020 and 2019 calendar-day aggregates. Two descriptive transport measures were tracked: monthly vocabulary-centroid distance from a 2018 reference, and divergence of projected class composition from that reference. Level and slope changes around 11 March 2020 were exploratory. Reporting was structured with reference to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, extended for artificial intelligence (TRIPOD+AI) statement.[21]

2.2. Data sources and analytic samples

The labeled corpus was the version 2 open release that accompanies the study by Gaur and colleagues.[2] The released dataset (doi:10.5281/zenodo.4543776) held 500 rows, 500 anonymous user identifiers, one aggregated text field per user and one label. Every row had non-empty text. There were no duplicate identifiers or duplicate aggregated text fields. The release holds 108 Supportive, 99 Indicator, 171 Ideation, 77 Behavior and 45 Attempt labels. The repository narrative describes 448 annotated users, but the released CSV contains 500. The present analysis followed the auditable released file and reports the discrepancy as provenance instead of altering the data.

The temporal corpus was the Reddit Mental Health Dataset (doi:10.17605/OSF.IO/7PEYQ). The r/SuicideWatch files that cover the same seasonal window, 1 January through 20 April, in 2018, 2019 and 2020 were selected. The broader “pre” file was excluded. It runs from December 2018 through December 2019 and overlaps the matched 2019 window. The source files held 10,215, 14,592 and 21,410 rows, respectively. Dates were parsed, and posts with missing or empty text were excluded. Exact

duplicates after whitespace normalization were removed in chronological order, with the first occurrence retained: none in 2018, three in 2019 and 11 in 2020; the final temporal sample comprised 46,203 unique posts across 331 observed days and 12 calendar months, as detailed in Table 1.

Table 1. Data sources, sampling windows and analytic flow.

Component	Source and unit	Source records	Exclusions	Analysed	Analytic role
Labeled corpus	Zenodo 4543776; anonymous user	500	0	500	Development and internal validation
r/SuicideWatch 2018	Post; 1 Jan–20 Apr	10,215	0	10,215	Temporal reference
r/SuicideWatch 2019	Post; 1 Jan–20 Apr	14,592	3 duplicates	14,589	Calendar comparator
r/SuicideWatch 2020	Post; 1 Jan–20 Apr	21,410	11 duplicates	21,399	Temporal application

Notes: Duplicate detection used whitespace-normalized text and retained the earliest occurrence. The broader "pre" file was excluded because it overlapped 2019.

2.3. Outcome definitions, text representation and model fitting

Outcome definitions and ordinal structure. The source labels were kept as documented. Supportive covered users whose language mainly offered support. It was not treated as a C-SSRS severity level. Among the risk-related labels, Indicator denoted indirect signals or factors that did not reach explicit ideation. Ideation denoted suicidal thoughts or wishes as annotated in the source. Behavior denoted preparatory or other suicide-related behavior. Attempt denoted a stated or documented attempt in the source annotation. An Attempt label did not necessarily mean imminent current risk, since the text could refer to an earlier event.

The model accordingly used two linked probability structures. Stage 1 estimated the probability that a user text was risk related rather than Supportive. Stage 2 estimated three cumulative probabilities among risk-related texts: greater than Indicator, greater than Ideation and greater than Behavior. A cumulative minimum was applied to constrain these probabilities monotonically. A more severe threshold therefore could not receive a higher probability than a less severe one. Conditional probabilities for Indicator, Ideation, Behavior and Attempt came from differences between adjacent cumulative thresholds. Final five-class probabilities multiplied the conditional risk probabilities by the Stage 1 risk probability. Supportive probability equalled one minus the Stage 1 probability; the continuous semantic-axis score was the expected value across the five ordered classes coded 0–4. Projected Behavior-plus-Attempt probability was the sum of those two output probabilities. These quantities are model outputs. They are not probabilities of a contemporaneous clinical state or of a future event.

Text representation and model fitting. Text was lower-cased within the vectorizer, and white space was normalized. Stemming and stop-word removal were avoided because they can strip negation or function-word signals that carry meaning. The representation used term frequency-inverse document frequency (TF-IDF) unigrams and bigrams with sublinear term frequency. The minimum document frequency was two, the maximum document frequency 0.98 and the feature cap 25,000. Logistic regression used inverse-frequency class weights computed separately within each fitted training partition. The candidate inverse-regularization parameters were 0.25, 1, 4 and 16.

Each training partition received one binary logistic model for Supportive versus risk-related content. Three further logistic models handled the cumulative thresholds among risk-related observations. No individual n-gram coefficients were exported. Rare combinations can expose sensitive details or form an unsafe catalog of language linked to self-harm. Model development and temporal projection ran in Python 3.14 with NumPy 2.4.4, pandas 3.0.2, SciPy 1.17.0 and scikit-learn 1.8.0. The deterministic random seed was 20260831.

2.4. Internal validation and uncertainty

Internal validation used five-fold stratified cross-validation repeated five times. Inside each outer training set, an inner three-fold stratified cross-validation chose the regularization parameter by minimizing the mean absolute error of the expected ordinal score. The vectorizer and all four logistic models were refitted within every inner training fold. After selection, they were fitted on the full outer training set and used to predict its held-out fold. Each user therefore received five outer held-out probability vectors, one per repeat. These were averaged before the final metrics and the confusion matrix were calculated. All 25 outer folds selected C = 16. In the final locked model, deterministic monotonic correction was needed for 110 of 500 labeled probability vectors (22.0%; maximum threshold adjustment 0.114) and for 9 of 46,203 temporal vectors (0.02%; maximum adjustment 0.031). This audit describes an enforced modeling constraint. It does not describe an ordering learned without correction.

The primary internal measures were the area under the receiver operating characteristic curve (AUC) for Supportive versus risk-related content and the Spearman correlation between observed class and expected score. Quadratic weighted kappa, macro-averaged F1, exact accuracy, balanced accuracy, accuracy within one adjacent category and the mean absolute error of the expected score completed the set. The 95% confidence intervals (CIs) came from 1,000 user-level bootstrap samples. Class recall was calculated from the averaged out-of-fold confusion

matrix. Predicted probabilities did not correspond to a target clinical prevalence, and no independent labeled sample existed. Clinical calibration and decision-curve analysis were therefore not estimated.

2.5. Temporal projection, drift and interrupted-time-series analyses

Temporal projection and drift. The final vectorizer and models were fitted on all 500 labeled users with C = 16 and then locked. Every temporal post was transformed in that feature space. Vocabulary coverage was the proportion of posts with at least one recognized feature. Post-level probability vectors and expected scores were aggregated by day, month and year, and only aggregates without author fields or raw text were exported.

Vocabulary-centroid drift for each month was one minus the cosine similarity between its mean locked TF-IDF vector and the mean vector of all 2018 posts. Projected class-composition drift was the Jensen-Shannon divergence, with base-2 logarithms, between each month's mean five-class probability vector and the matching 2018 reference. The two measures answer different questions. Cosine distance tracks movement in recognized vocabulary. Jensen-Shannon divergence tracks movement once that vocabulary is compressed into the model's five output probabilities. Eleven precomputed semantic domains supplied with the temporal dataset were used for triangulation. The 2020-minus-2019 difference in each domain was divided by the pooled post-level standard deviation. These descriptive standardized differences were not treated as independent clinical outcomes.

Calendar-matched and interrupted-time-series analyses. The primary temporal contrast matched 2020 and 2019 by month and day. This gave 110 paired calendar days; 29 February 2020 had no match and was omitted. Differences in daily mean expected score and daily mean projected Behavior-plus-Attempt probability were resampled at the calendar-day level 2,000 times; this yielded percentile confidence intervals and two-sided bootstrap p values. The day, rather than the tens of thousands of individual posts, was thereby kept as the unit of uncertainty.

An exploratory interrupted time series used the 111 daily observations from 1 January to 20 April 2020, of which 70 fell before the interruption and 41 on or after it. The predefined interruption was 11 March 2020, the date on which the World Health Organization stated that COVID-19 could be characterized as a pandemic.[22] Weighted least squares with the daily number of posts as weights estimated a post-volume-weighted daily trajectory; the model included continuous time, an indicator for observations on or after the interruption, time after the interruption and six day-of-week indicators. Newey-West standard errors with seven lags, covering one weekly cycle, were robust to heteroskedasticity and autocorrelation; the analysis was observational and had no external control series. It was not interpreted as a causal effect of the pandemic. No multiplicity adjustment was applied because the two correlated outcomes and the interruption parameters were exploratory. Exact effect estimates with confidence intervals were prioritized. An equal-day-weight sensitivity model tested dependence on volume weighting.

2.6. Ethics, privacy and safety

The CMHC Research Center Ethics Committee Indonesia granted ethical approval (No. 2026/020). The analysis used secondary public datasets and involved no direct contact with participants. Public availability was not taken to remove ethical responsibility. Raw files were kept in a restricted case directory. Usernames, author fields, raw text, text quotations and user hashes were left out of analytic exports, manuscript tables and figures. The separate unlabeled JSON release, which can contain identifiable usernames, was neither downloaded nor analyzed. The model was evaluated as a research instrument. It was never used to contact, label, diagnose, triage or intervene with any individual.

3. Results

3.1. Analytic flow and corpus composition

All four downloaded CSV files matched their repository checksums. The labeled corpus included all 500 anonymous users. The temporal flow retained 10,215 posts from 2018, 14,589 from 2019 and 21,399 from 2020 after removal of 14 duplicate posts; the complete year-specific flow is detailed in Table 1. Dates were complete after parsing and all retained posts contained text. Locked-vocabulary coverage was 99.98% in 2018 and 100% in both 2019 and 2020. Sensitivity analyses using exact-string deduplication (46,205 posts) or retaining all 46,217 posts produced substantively unchanged paired contrasts: the maximum absolute change was 0.00013 for expected score and 0.00003 for projected Behavior-plus-Attempt probability. The alternative deduplication and weighting estimates are detailed in Table 4.

3.2. Internal validation

The binary Supportive-versus-risk AUC was 0.841 (95% CI 0.799–0.876). The expected ordinal score showed a moderate monotonic association with observed class (Spearman rho 0.568, 95% CI 0.503–0.624), and quadratic weighted kappa was 0.464 (0.392–0.527). Exact five-class accuracy was 43.2% (38.8%–47.6%), balanced accuracy 33.1% (30.0%–36.0%) and macro-F1 0.292 (0.257–0.326). Accuracy within one category was 80.0% (76.6%–83.2%), while mean absolute error of the probabilistic score was 0.830 (0.779–0.884). The complete metric set, including class-specific recall, is detailed in Table 2.

Table 2. Internal validation performance of the five-class semantic axis.

Metric	Estimate	95% confidence interval
AUC, Supportive versus risk related	0.841	0.799–0.876
Spearman rho, observed label versus expected score	0.568	0.503–0.624
Quadratic weighted kappa	0.464	0.392–0.527
Accuracy within one category	0.800	0.766–0.832
Exact five-class accuracy	0.432	0.388–0.476
Balanced accuracy	0.331	0.300–0.360
Macro-F1	0.292	0.257–0.326
Mean absolute error, expected score	0.830	0.779–0.884
Recall, Supportive	0.741	Not bootstrapped
Recall, Indicator	0.081	Not bootstrapped
Recall, Ideation	0.678	Not bootstrapped
Recall, Behavior	0.156	Not bootstrapped
Recall, Attempt	0.000	Not bootstrapped

Notes: Confidence intervals used 1,000 anonymous-user bootstrap samples. AUC, area under the receiver operating characteristic curve.

The confusion matrix revealed substantial class asymmetry, as shown in Figure 1A. Recall was 74.1% for Supportive, 8.1% for Indicator, 67.8% for Ideation, 15.6% for Behavior and 0% for Attempt. Of 45 Attempt-labeled users, 27 were predicted Ideation, eight Behavior, six Supportive and four Indicator; none had Attempt as the highest-probability class. The expected-score distributions increased from Supportive through Behavior, whereas Attempt overlapped the Behavior and Ideation distributions rather than separating above them (Figure 1B).

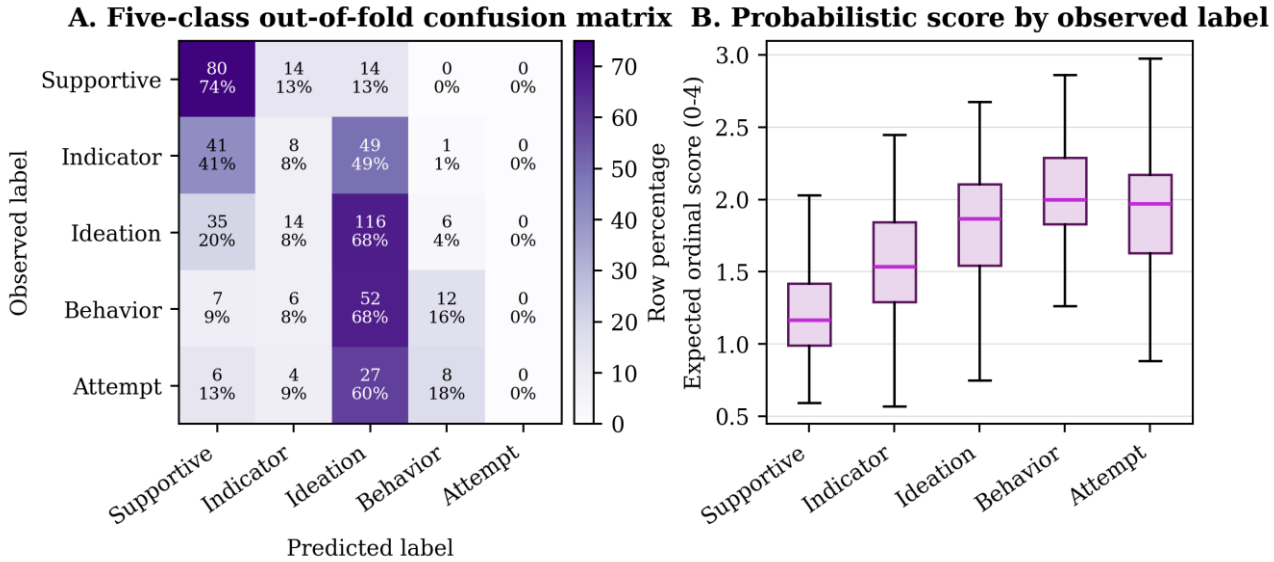


Figure 1. Internal validation of the semantic axis. Panel A shows the averaged repeated out-of-fold confusion matrix; each cell gives count and row percentage. Panel B shows expected ordinal score distributions by observed source label. Exact Attempt recall was 0%, despite a monotonic rise in median score through Behavior.

3.3. Aggregate temporal projections

Mean expected scores were 1.969 in 2018, 1.923 in 2019 and 1.928 in 2020. Corresponding mean projected Behavior-plus-Attempt probabilities were 29.61%, 28.37% and 28.63%. These were averages of model outputs and were not clinical prevalence estimates. The temporal corpus contained no contemporaneous C-SSRS assessment, clinical outcome or prediction horizon. In the 110 matched calendar days, the mean 2020-minus-2019 score difference was 0.00531 (95% bootstrap CI −0.00204 to 0.01280; p = 0.176). The difference in projected Behavior-plus-Attempt probability was 0.00279 (0.00017–0.00551; p = 0.034). Although the latter interval excluded zero, its absolute magnitude was 0.279 percentage points and it could not be mapped to clinical risk without temporal labels. The annual projections and paired-day contrasts are detailed in Table 3.

Table 3. Aggregate temporal projections and paired-day contrasts.

Analysis	Outcome/parameter	Estimate	95% confidence interval	p
Yearly 2018	Mean score / Behavior+Attempt probability	1.969 / 0.2961	Descriptive	—
Yearly 2019	Mean score / Behavior+Attempt probability	1.923 / 0.2837	Descriptive	—
Yearly 2020	Mean score / Behavior+Attempt probability	1.928 / 0.2863	Descriptive	—
Paired 2020–2019	Mean score difference	0.00531	-0.00204 to 0.01280	0.176
Paired 2020–2019	Behavior+Attempt probability difference	0.00279	0.00017 to 0.00551	0.034

Notes: Paired analyses used 110 common month-days and 2,000 bootstrap samples. Behavior+Attempt probability is P(Behavior) + P(Attempt), a model output rather than absolute clinical risk. CI, confidence interval.

Monthly means varied within a narrow range, as shown in Figure 2A. Vocabulary-centroid distance from the combined 2018 reference was measurable and reached 0.0046 in April 2020. Jensen-Shannon divergence of projected class composition remained below 0.0007 across all months (Figure 2B). The combination indicates detectable movement in recognized vocabulary with much smaller movement after projection onto the five-class output space.

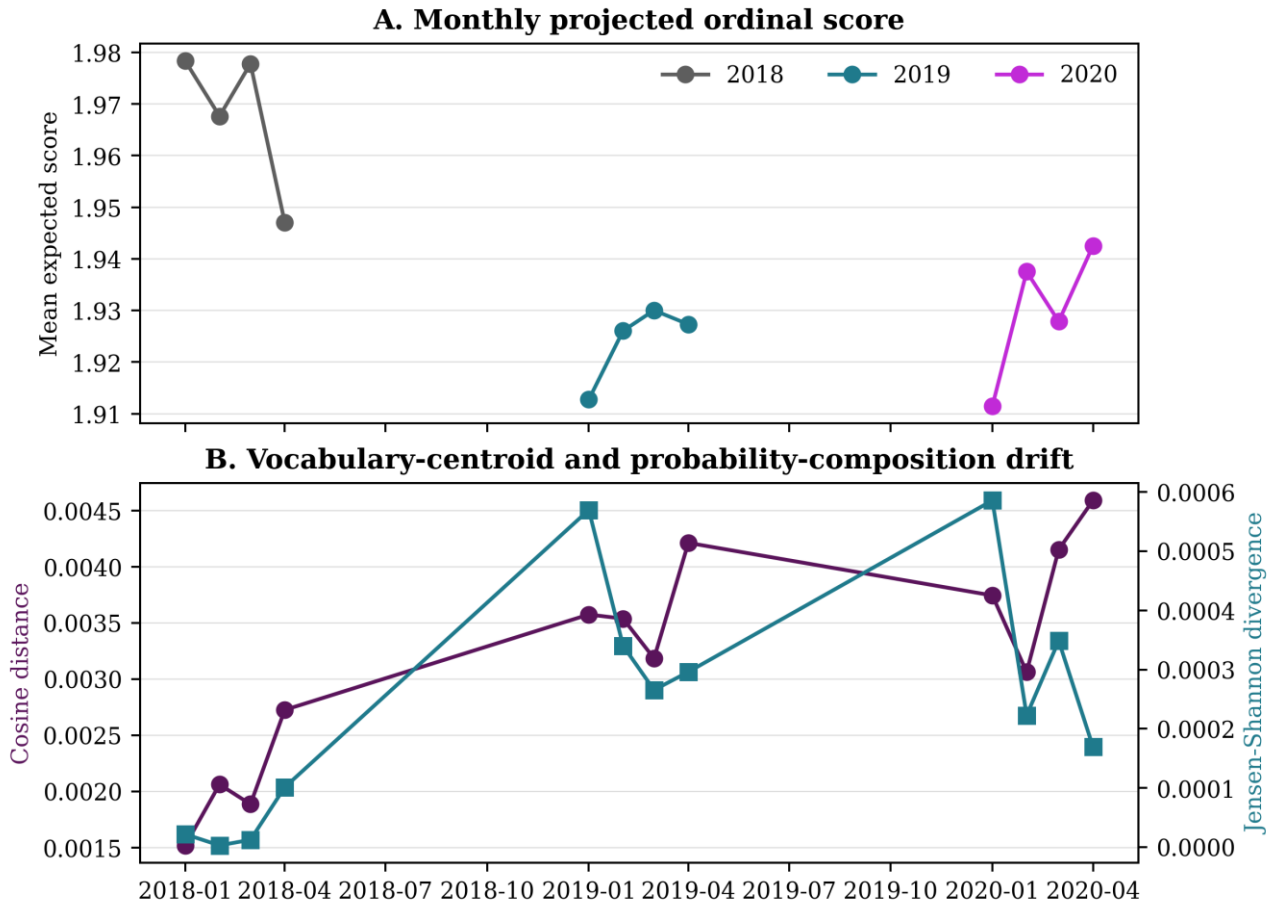


Figure 2. Monthly temporal projections and semantic drift. Panel A shows monthly mean expected scores; gaps between years intentionally represent matched January–April observation windows. Panel B shows cosine distance between each monthly locked-feature centroid and the combined 2018 reference (left axis), with Jensen-Shannon divergence between each monthly five-class probability composition and the 2018 reference (right axis).

3.4. Interrupted time series and semantic-domain triangulation

The observed daily series and fitted pre- and post-interruption trajectories are shown in Figure 3. Before 11 March 2020, mean score increased by 0.000498 per day (95% CI 0.000088–0.000908; p = 0.017). The estimated level change at the interruption was −0.01976 (−0.04062 to 0.00110; p = 0.063), and the change in slope was 0.000240 per day (−0.000583 to 0.001064; p = 0.567). For projected Behavior-plus-Attempt probability, the pre-interruption trend was 0.000152 per day (0.000023–0.000281; p = 0.021), the level change was −0.00814 (−0.01678 to 0.00050; p = 0.065) and the slope change was 0.000169 (−0.000117 to 0.000454; p = 0.247). Equal-day weighting gave similar effect directions and magnitudes: level changes were −0.02037 for score and −0.00861 for projected Behavior-plus-Attempt probability, while slope changes remained small. The equal-day-weight level changes did differ inferentially from the primary estimates: their intervals excluded zero (−0.03928 to −0.00146, p =

0.035, and −0.01636 to −0.00087, p = 0.029), whereas those of the primary weighted models did not (p = 0.063 and 0.065). The complete interrupted-time-series and robustness estimates are detailed in Table 4.

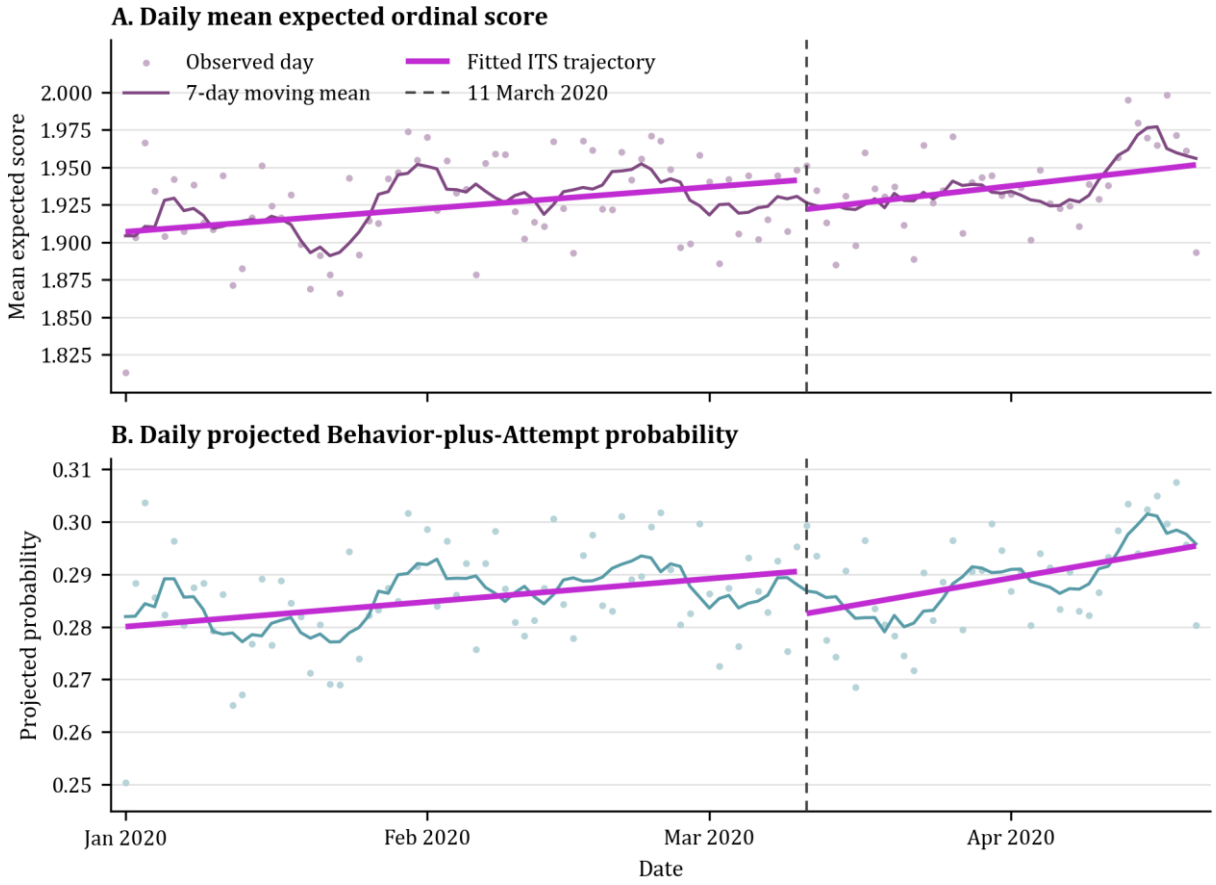


Figure 3. Daily 2020 trajectories around the prespecified interruption. Observed daily mean expected score (Panel A) and projected Behavior-plus-Attempt probability (Panel B) are shown with seven-day moving means and fitted post-volume-weighted interrupted-time-series trajectories. The vertical line marks 11 March 2020. Fitted segments are descriptive and are not causal estimates of a pandemic effect.

Table 4. Interrupted-time-series and robustness analyses.

Analysis/specification	Outcome or diagnostic	Estimate	95% CI or comparison
A. Interrupted time series (primary post-volume-weighted and equal-day-weight models)			
Post-volume-weighted ITS	Mean-score pre-interruption trend/day	0.000498	0.000088 to 0.000908; p = 0.017
Post-volume-weighted ITS	Mean-score level change	-0.01976	-0.04062 to 0.00110; p = 0.063
Post-volume-weighted ITS	Mean-score slope change/day	0.000240	-0.000583 to 0.001064; p = 0.567
Post-volume-weighted ITS	Behavior+Attempt pre-trend/day	0.000152	0.000023 to 0.000281; p = 0.021
Post-volume-weighted ITS	Behavior+Attempt level change	-0.00814	-0.01678 to 0.00050; p = 0.065
Post-volume-weighted ITS	Behavior+Attempt slope change/day	0.000169	-0.000117 to 0.000454; p = 0.247
Equal-day-weight ITS	Mean-score level change	-0.02037	-0.03928 to −0.00146; p = 0.035
Equal-day-weight ITS	Mean-score slope change/day	0.000508	-0.000267 to 0.001282; p = 0.199
Equal-day-weight ITS	Behavior+Attempt level change	-0.00861	-0.01636 to −0.00087; p = 0.029
Equal-day-weight ITS	Behavior+Attempt slope change/day	0.000250	-0.000017 to 0.000518; p = 0.067
B. Deduplication sensitivity			
Exact-string deduplication	n; score difference; Behavior+Attempt difference	46,205; 0.005337; 0.002794	Versus primary: +0.000025; +0.000006
No deduplication	n; score difference; Behavior+Attempt difference	46,217; 0.005183; 0.002761	Versus primary: −0.000129; −0.000028
C. Monotonic-correction diagnostics			
Monotonic correction, labeled	Corrected vectors; maximum adjustment	110/500 (22.0%); 0.114	Diagnostic
Monotonic correction, temporal	Corrected vectors; maximum adjustment	9/46,203 (0.02%); 0.031	Diagnostic

Notes: CI, confidence interval; ITS, interrupted time series. The primary deduplication analysis used whitespace-normalized text and retained 46,203 posts. Differences are 2020 minus 2019 over 110 matched month-days.

The precomputed semantic domains changed heterogeneously, as shown in Figure 4. The largest positive standardized differences in 2020 versus 2019 were negative sentiment (0.050), isolation (0.029), domestic stress (0.016) and anxiety language (0.012). Economic-stress, sadness, positive-emotion, substance-use, guns and suicidality domains were lower, and LIWC negative emotion was essentially unchanged, with all absolute standardized differences below 0.051. This mixed profile did not support a uniform increase across negative-language dimensions.

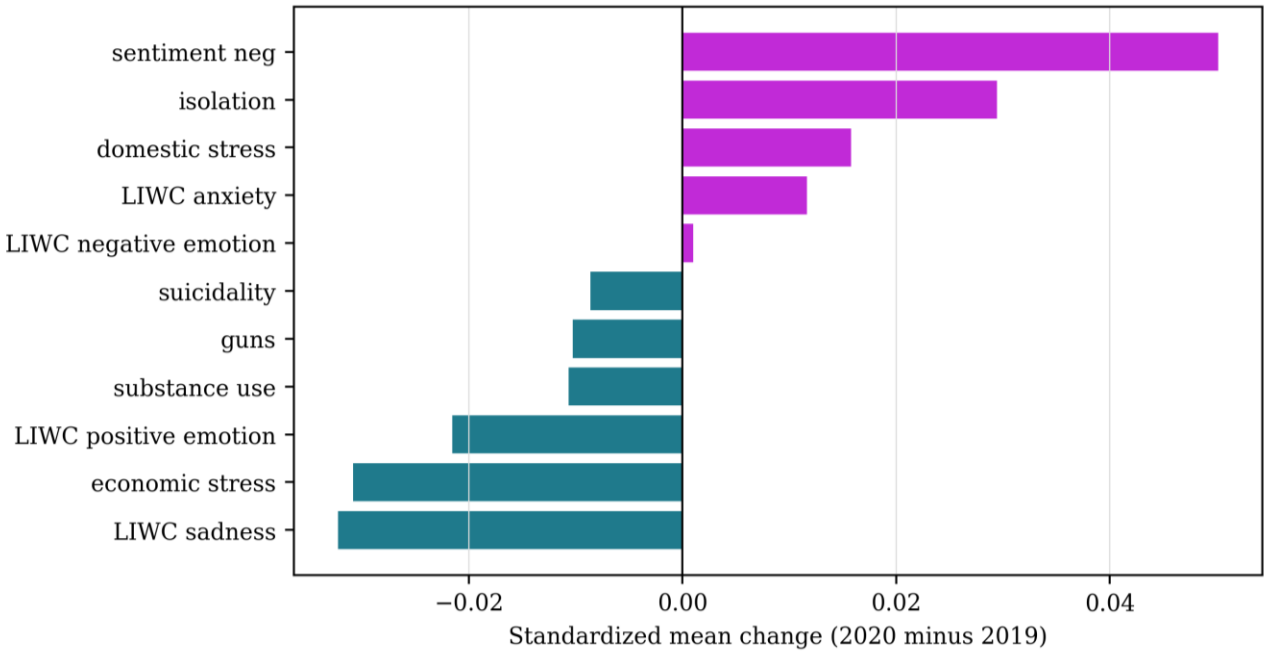


Figure 4. Standardised changes in precomputed semantic domains, 2020 versus 2019. Bars are mean differences divided by the pooled standard deviation. Positive values indicate higher means in 2020. All absolute standardized differences were smaller than 0.051. LIWC, Linguistic Inquiry and Word Count.

4. Discussion

4.1. Principal findings

In this open-data reanalysis, a transparent two-stage text model captured a moderate ordinal semantic axis. It separated supportive from risk-related language well. It did not, however, reliably identify the highest source label. An AUC above 0.84 alongside zero Attempt recall is the result with the most consequence. It shows that broad discrimination can coexist with failure at a clinically important end of the label distribution; applied across the matched 2018–2020 windows, the model showed near-complete vocabulary coverage and small changes in probability composition despite measurable centroid movement. The calendar-matched difference in mean score was close to zero. The difference in projected Behavior-plus-Attempt probability was statistically distinguishable but very small in absolute terms. It did not establish a change in clinical prevalence.

Two different questions are therefore answered. Within the source corpus, the expected score kept meaningful rank information, as the monotonic correlation and weighted kappa show; across time, the locked model kept recognising vocabulary and produced fairly stable aggregate outputs. Neither finding tells us the sensitivity, specificity, calibration or decision value that would result if clinicians assessed the temporal posts or their authors. Temporal semantic transport describes how a measurement pipeline behaves when the distribution changes. Clinical external validation needs an independently labeled population and a defined clinical time horizon.

4.2. Comparison with existing literature

The internal findings agree with recent social-media studies in which suicide-related language carries a learnable but imperfect signal. Gaur and colleagues found that severity and temporality cannot be swapped for each other.[2] Work on interpretable linguistic evaluation,[12] metadata-enhanced modeling[18] and semi-supervised learning[20] reported performance that varied by label and representation. The present two-stage model is deliberately simpler. Its error pattern, however, makes the same point sharper. A post that mentions an attempt may describe a remote history. A post that conveys imminent intent may lack the vocabulary tied to earlier attempts; the expected score can therefore rise monotonically even when exact classification at the highest category is inadequate.

Recent external studies also show why success in one setting does not carry over automatically to another. A deep-learning model applied to social-media data from US servicemembers and veterans found a useful risk signal;[23] an independently validated model tied social-media trajectories after suicidal mentions to later risk classification.[24] A Taiwanese internet-message system paired automated identification with monitoring, tracking and rescue procedures.[25] These studies differed in population, outcome and action pathway. They support the feasibility of language-based research.

They say nothing about the calibration or severe-class sensitivity of the present Reddit model.

Clinical-text studies offer a further benchmark, because their reference standards and denominators are more explicit. Natural-language features from mental-health notes may add predictive value to structured models.[26] Emergency-department triage notes can be used to identify self-harm or suicidal ideation.[27] Large-scale clinical natural language processing (NLP) has improved ascertainment of suicidal ideation and attempt,[28] made population-specific risk models possible[29] and combined structured with unstructured information to predict suicidal behavior after discharge.[30] Parser-based models in text-based mental-health services[31] and zero-shot identification of suicide documentation[32] show that task definition and documentation context shape performance considerably. These systems work with information that a single online text aggregate lacks. The 0% Attempt recall of the present model therefore marks a boundary. On one side lies an aggregate semantic instrument. On the other lies a safety-critical clinical classifier.

Temporal and subgroup studies reinforce the case for transport analysis. NLP has exposed time-varying suicide-risk patterns among high-risk veterans[33] and distinct performance profiles across risk tiers.[34] Models enriched with social determinants have tied documented context to suicide mortality.[35] NLP of death-investigation records has revealed inconsistencies in the recorded circumstances.[36] Explainable modeling has found subgroup-specific mortality associations,[37] and work on racial variation has shown that suicide classification cannot be assumed to behave the same way across groups.[38] Research on multimodal speech and language likewise suggests that the available signal depends on the population and the measurement channel.[39] Taken together, these studies favour monitoring both the input representation and the output distribution. A stable score should not be assumed to have an unchanging meaning.

Population surveillance studies give another point of comparison. Time-series modeling in Japan connected social-media signals with national suicide statistics.[40] Analyses of more than seven million suicide-related tweets related online content to crisis-line calls and suicide counts.[41] Other work has looked at how Twitter events correspond to suicides,[42] at sentiment around suicides at a landmark site[43] and at potentially harmful versus protective suicide-related content.[44] Those designs had external events, outcomes or content standards. The present study had none. Stable probability composition could therefore reflect real stability, model insensitivity or opposing shifts in the people and topics present on Reddit; the small divergence after projection may mean that vocabulary movement was largely orthogonal to the learned C-SSRS-informed axis. Only newly labeled temporal data could separate that explanation from compression by the linear model.

4.3. Interpretation and mechanism

Treating Supportive as a separate role matters conceptually. Mental-health communities include people who offer reassurance, people who disclose personal distress, people who discuss another person and people who move between these roles. A support provider and a distressed author may write about the same topic while the personal meaning of their words differs. Forcing Supportive into the lowest position of a single severity scale would let topic similarity stand as evidence about the provider's own state. The two-stage structure reduced that mismatch and may have contributed to the strong binary AUC. An aggregated text field cannot reliably establish speaker role, event timing or personal history, however. Role separation by itself therefore cannot recover clinical intent or behavior.

The ordinal threshold construction encodes a reasonable monotonic relationship. Evidence that is enough for a more severe threshold should also support the less severe ones; the construction also gives a smooth expected score that is less brittle than exact class assignment. Accuracy within one category reached 80%, but this metric has an asymmetric clinical meaning. Misclassifying Attempt as Ideation is numerically close on an ordinal scale. It could matter greatly if the model allocated care. In this retrospective dataset, on the other hand, an Attempt label may refer to past behavior and not to immediate danger. A clinically deployable system would need explicit prediction windows, validated event timing, calibration to the intended population, sensitivity-oriented thresholds, human review and prospective evaluation of impact.

The gap between centroid drift and probability drift offers a further mechanism. Centroid distance summarizes the average direction of the words that the locked vectorizer recognizes. The output distribution summarizes how those features combine after weighting by the four logistic models. Vocabulary can thus move while its weighted projection stays stable; three explanations remain open: real stability on the learned axis, new language moving largely orthogonally to the axis, or insensitivity produced by the compressed linear representation. Only new labeled temporal samples can tell them apart. Monitoring both levels gives more information than monitoring output prevalence alone.

4.4. Implications for research, practice and policy

The immediate use of this model is methodological research in a controlled setting; it can support studies of aggregate semantic transport, help in designing annotation protocols and show why class-wise evaluation is needed; it should not generate individual alerts, diagnostic labels, clinical referrals or moderation sanctions. No score should be presented to a user or clinician as an estimate of imminent suicide risk.

The next validation program should assemble temporally stratified, independently adjudicated samples with explicit event timing and several

clinicians. Where governance permits, the unit of analysis should be longitudinal. Historical behavior, current state and future outcome should be kept apart. Evaluation should report class-specific sensitivity, precision-recall, calibration, decision curves and the consequences of errors. Severe labels are uncommon. Sampling and loss functions may therefore need to prioritize clinically important false negatives without making positive predictive value unusable. External validation should span different platforms, time periods and English-speaking populations before any cross-cultural extension.

Such a program should also separate three deployment questions that are often merged into one performance claim. The first concerns measurement: whether the model reproduces a clearly defined annotation construct when assessors see the same text. The second concerns prognosis: whether the output predicts a future suicidal thought, behavior, attempt or death within a stated time horizon. The third concerns decision support: whether showing the output to a trained professional improves care without avoidable surveillance, coercion, missed crises or diversion of resources. Each question needs its own reference standard and denominator. A model can perform acceptably for corpus measurement and still be unsuitable for prognosis. A prognostic model can lack clinical utility if alerts are too frequent or action pathways are undefined. The current results address the first question internally and otherwise only the transport of its language representation over time.

Prospective work should start with a silent evaluation, in which predictions are generated without being shown or acted upon; that design allows calibration, alert burden, subgroup performance and temporal degradation to be estimated without changing care. Predefined stopping rules should be triggered by severe-class false negatives, marked calibration drift, inequitable error rates or evidence that the model changes user behavior harmfully. If an assisted-use study is later justified, human factors, escalation pathways and documentation responsibility should be evaluated together with conventional discrimination statistics. Continuous monitoring would have to track input drift, outcome drift, calibration and the consequences of both alerts and non-alerts.

For Indonesian and other non-English settings, direct transfer would be inappropriate. Translating English features alone would miss differences in idiom, indirect expression, help-seeking norms, stigma and platform use. A local corpus would call for culturally and linguistically appropriate constructs, community involvement, bilingual or local-language clinical adjudication and independent ethics oversight. Fairness cannot be inferred from a dataset that lacks the relevant demographic variables. Governance should also name intended and prohibited uses, access controls, audit logs, incident response and human authority; it should include a process for retiring the model when language or platform conditions change.

Policy should keep a firm line between aggregate research and intervention on identifiable people; aggregate analyses may help investigators choose periods for closer qualitative study, check whether a research instrument has become stale or compare broad community-level themes. Intervention demands far more: a lawful basis for processing, evidence of clinical benefit, consent or another defensible authority, rapid access to trained responders, mechanisms for appeal and correction, and explicit protection against punitive use. Platform moderation and clinical triage are different decisions. A score built for one should not move silently to the other. The privacy-minimizing design used here (no quotations, identifiers or exported individual features) should be seen as a minimum. It is not a complete governance framework.

4.5. Strengths and limitations

The study used checksum-verified open releases, a fully nested repeated validation procedure, user-level bootstrap intervals and a locked final model. Rank-based and exact class performance were both reported. The severe-class error pattern was exposed, and two drift measures that answer different questions were used. Calendar matching reduced seasonal imbalance. Day-level bootstrap with autocorrelation-robust inference avoided treating all posts as independent. Every exported analytic product was stripped of identities, quotations and raw text. The transparent linear architecture and deterministic code keep the analysis auditable.

Several constraints shape interpretation. First, the labeled sample was modest, particularly for Attempt, and the source adjudication details did not allow re-review of every label. Second, the temporal corpus had no C-SSRS labels. Projected probabilities therefore cannot be read as clinical prevalence, accuracy or calibration. Third, Reddit users are self-selected, the temporal corpus represents posts and not a defined clinical population, and platform demographics were unavailable. Fairness across age, sex, ethnicity, language or clinical setting could not be estimated. Posting behavior may also change with moderation, access or external events. Fourth, one user may have written several posts. Even so, user identifiers were not retained in aggregate outputs. Day-level inference does not fully model within-user dependence. Fifth, the interrupted time series had no control community and could reflect changes in volume, moderation, composition or calendar conditions instead of a pandemic effect. Sixth, English-language corpora were used, and transport to clinical services was not evaluated. Finally, the model was deliberately simple. More expressive models might capture additional structure, but they would still need external labels, calibration and safety governance.

5. Conclusion

A two-stage, C-SSRS-informed semantic axis captured moderate ordinal structure and separated supportive from risk-related Reddit language well, but it did not

identify Attempt as the highest-probability class; across the matched 2018–2020 windows, vocabulary drift was measurable while aggregate projected severity changed little. The findings support transparent research on population-level semantic transport and do not support individual screening, diagnosis, triage or automated intervention. Before any prospective use, external validation on temporally labeled data with clinical adjudication is required.

Declarations

Use of Generative Artificial Intelligence. No generative-AI tool was used for drafting, editing, coding, or analysis of this manuscript.

Author Contributions. D.A.S.: conceptualization, methodology, software, formal analysis, data curation, visualization, writing – original draft. A.S.: validation, writing – review & editing. A.G.: project administration, writing – review & editing. All authors reviewed and approved the final manuscript and agree to be accountable for all aspects of the work.

Funding. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest. The authors declare no competing financial or non-financial interests in relation to this work.

Ethical Approval. Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2026/020). The study used secondary open datasets and was conducted with privacy and data-minimization safeguards.

Informed Consent. Direct informed consent was not obtained for this secondary analysis of publicly released datasets. No direct participant contact occurred, and no username, quotation, raw post, or other direct identifier is reproduced.

Data Availability. The labeled corpus is openly available at Zenodo, doi:10.5281/zenodo.4543776. The Reddit Mental Health Dataset is openly available at OSF, doi:10.17605/OSF.IO/7PEYQ. Reuse remains subject to repository licenses, applicable Reddit terms, and independent ethics oversight. Reproducible code and privacy-preserving aggregate outputs accompany the submission; raw text is not redistributed.

Acknowledgements. None.

References

1. Posner K, Brown GK, Stanley B, et al. The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 2011;168(12):1266-77. doi:10.1176/appi.ajp.2011.10111704
2. Gaur M, Aribandi V, Alambo A, et al. Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. PLoS One. 2021;16(5):e0250448. doi:10.1371/journal.pone.0250448
3. Dutta R, Gkotsis G, Velupillai S, et al. Temporal and diurnal variation in social media posts to a suicide support forum. BMC Psychiatry. 2021;21(1):259. doi:10.1186/s12888-021-03268-1
4. Yan Y, Li J, Liu X, et al. Identifying Reddit Users at a High Risk of Suicide and Their Linguistic Features During the COVID-19 Pandemic: Growth-Based Trajectory Model. J Med Internet Res. 2024;26:e48907. doi:10.2196/48907
5. Bauer B, Norel R, Leow A, et al. Using Large Language Models to Understand Suicidality in a Social Media-Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts. JMIR Ment Health. 2024;11:e57234. doi:10.2196/57234
6. Zhu J, Jin R, Kenne DR, et al. 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. doi:10.2196/53968
7. Saha K, Yousuf A, Boyd RL, et al. Social Media Discussions Predict Mental Health Consultations on College Campuses. Sci Rep. 2022;12(1):123. doi:10.1038/s41598-021-03423-4
8. Swedo EA, Beauregard JL, de Fijter S, et al. Associations Between Social Media and Suicidal Behaviors During a Youth Suicide Cluster in Ohio. J Adolesc Health. 2021;68(2):308-316. doi:10.1016/j.jadohealth.2020.05.049
9. Bailey E, Boland A, Bell I, et al. The Mental Health and Social Media Use of Young Australians during the COVID-19 Pandemic. Int J Environ Res Public Health. 2022;19(3):1077. doi:10.3390/ijerph19031077
10. Hamilton JL, Dalack M, Boyd SI, et al. Positive and negative social media experiences and proximal risk for suicidal ideation in adolescents. J Child Psychol Psychiatry. 2024;65(12):1580-1589. doi:10.1111/jcpp.13996
11. Zhang T, Schoene AM, Ananiadou S, et al. Automatic identification of suicide notes with a transformer-based deep learning model. Internet Interv. 2021;25:100422. doi:10.1016/j.invent.2021.100422
12. Lao C, Lane J, Suominen H, et al. Analyzing Suicide Risk From Linguistic Features in Social Media: Evaluation Study. JMIR Form Res. 2022;6(8):e35563. doi:10.2196/35563
13. Liu J, Shi M, Jiang H, et al. Detecting Suicidal Ideation in Social Media: An Ensemble Method Based on Feature Fusion. Int J Environ Res Public Health. 2022;19(13):8197. doi:10.3390/ijerph19138197
14. Aldhyani THH, Alsubari SN, Alshebami AS, et al. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. Int J Environ Res Public Health. 2022;19(19):12635. doi:10.3390/ijerph191912635
15. Bouktif S, Khanday AMUD, Ouni A, et al. Explainable Predictive Model for Suicidal Ideation During COVID-19: Social Media Discourse Study. J Med Internet Res. 2025;27:e65434. doi:10.2196/65434
16. Lin E, Sun J, Chen H, et al. Data Quality Matters: Suicide Intention Detection on Social Media Posts Using RoBERTa-CNN. Annu Int Conf IEEE Eng Med Biol Soc. 2024:1-5. doi:10.1109/EMBC53108.2024.10782647
17. Cao L, Feng L, Ding Y, et al. Online continuous learning of users suicidal risk on social media. Artif Intell Med. 2025;167:103199. doi:10.1016/j.artmed.2025.103199
18. Jung W, Kim D, Nam S, et al. Suicidality Detection on Social Media Using Metadata and Text Feature Extraction and Machine Learning. Arch Suicide Res. 2023;27(1):13-28. doi:10.1080/13811118.2021.1955783
19. Pan W, Wang X, Zhou W, et al. Linguistic Analysis for Identifying Depression and Subsequent Suicidal Ideation on Weibo: Machine Learning Approaches. Int J Environ Res Public Health. 2023;20(3):2688. doi:10.3390/ijerph20032688
20. Lovitt M, Ma H, Wang S, et al. Suicide Risk Assessment on Social Media with Semi-Supervised Learning. Proc IEEE Int Conf Big Data. 2024:8541-9. doi:10.1109/bigdata62323.2024.10825422

21. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
22. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Geneva: World Health Organization; 2020 [cited 2026 Oct 22]. Available from: https://www.who.int/news-room/speeches/item/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020
23. Zuromski KL, Low DM, Jones NC, et al. Detecting suicide risk among U.S. servicemembers and veterans: a deep learning approach using social media data. Psychol Med. 2024;54(12):3379-3388. doi:10.1017/S0033291724001557
24. Kaminsky Z, McQuaid RJ, Hellemans KG, et al. 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. doi:10.2196/49927
25. Wu EL, Wu CY, Lee MB, et al. Development of Internet suicide message identification and the Monitoring-Tracking-Rescuing model in Taiwan. J Affect Disord. 2023;320:37-41. doi:10.1016/j.jad.2022.09.090
26. Levis M, Leonard Westgate C, Gui J, et al. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol Med. 2021;51(8):1382-1391. doi:10.1017/S0033291720000173
27. Rozova V, Witt K, Robinson J, et al. Detection of self-harm and suicidal ideation in emergency department triage notes. J Am Med Inform Assoc. 2022;29(3):472-480. doi:10.1093/jamia/ocab261
28. Bejan CA, Ripperger M, Wilimitis D, et al. Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. Sci Rep. 2022;12(1):15146. doi:10.1038/s41598-022-19358-3
29. Levis M, Levy J, Dufort V, et al. Leveraging unstructured electronic medical record notes to derive population-specific suicide risk models. Psychiatry Res. 2022;315:114703. doi:10.1016/j.psychres.2022.114703
30. Carson NJ, Yang X, Mullin B, et al. Predicting adolescent suicidal behavior following inpatient discharge using structured and unstructured data. J Affect Disord. 2024;350:382-387. doi:10.1016/j.jad.2023.12.059
31. Xu Z, Chan CS, Fung J, et al. Developing and validating a parser-based suicidality detection model in text-based mental health services. J Affect Disord. 2023;335:228-232. doi:10.1016/j.jad.2023.04.128
32. Workman TE, Goulet JL, Brandt CA, et al. Identifying suicide documentation in clinical notes through zero-shot learning. Health Sci Rep. 2023;6(9):e1526. doi:10.1002/hsr2.1526
33. Levis M, Levy J, Dimambro M, et al. Using natural language processing to evaluate temporal patterns in suicide risk variation among high-risk Veterans. Psychiatry Res. 2024;339:116097. doi:10.1016/j.psychres.2024.116097
34. Levis M, Dimambro M, Levy J, et al. Using Natural Language Processing to develop risk-tier specific suicide prediction models for Veterans Affairs patients. J Psychiatr Res. 2024;179:322-329. doi:10.1016/j.jpsychires.2024.09.031
35. Mitra A, Pradhan R, Melamed RD, et al. Associations Between Natural Language Processing-Enriched Social Determinants of Health and Suicide Death Among US Veterans. JAMA Netw Open. 2023;6(3):e233079. doi:10.1001/jamanetworkopen.2023.3079
36. Wang S, Zhou Y, Han Z, et al. A natural language processing approach to detect inconsistencies in death investigation notes attributing suicide circumstances. Commun Med (Lond). 2024;4(1):199. doi:10.1038/s43856-024-00631-7
37. Yin Y, Workman TE, Blosnich JR, et al. Sexual and Gender Minority Status and Suicide Mortality: An Explainable Artificial Intelligence Analysis. Int J Public Health. 2024;69:1606855. doi:10.3389/ijph.2024.1606855
38. Rahman N, Mozer R, McHugh RK, et al. Using natural language processing to improve suicide classification requires consideration of race. Suicide Life Threat Behav. 2022;52(4):782-791. doi:10.1111/sltb.12862
39. Belouali A, Gupta S, Sourirajan V, et al. Acoustic and language analysis of speech for suicidal ideation among US veterans. BioData Min. 2021;14(1):11. doi:10.1186/s13040-021-00245-y
40. Wang S, Ning H, Huang X, 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. doi:10.2196/47225
41. Niederkrotenthaler T, Tran US, Baginski H, 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(7):994-1003. doi:10.1177/00048674221126649
42. Sinyor M, Williams M, Zaheer R, et al. The Relationship Between Suicide-Related Twitter Events and Suicides in Ontario From 2015 to 2016. Crisis. 2021;42(1):40-47. doi:10.1027/0227-5910/a000684
43. Yip P, Xiao Y, Xu Y, et al. Social Media Sentiments on Suicides at the New York City Landmark, Vessel: A Twitter Study. Int J Environ Res Public Health. 2022;19(18):11694. doi:10.3390/ijerph191811694
44. Metzler H, Baginski H, Niederkrotenthaler T, et al. Detecting Potentially Harmful and Protective Suicide-Related Content on Twitter: Machine Learning Approach. J Med Internet Res. 2022;24(8):e34705. doi:10.2196/34705