Journal of Literature and Art Studies, October 2026, Vol. 16, No. 10, 849-860
doi: 10.17265/2159-5836/2026.10.009

Selecting Computational Measures of Affective Language in Mental Health News: A Comparison of Four Methods

Yuze Li
Columbia University, New York, USA

This study compares category-to-dimension mapping, a word lexicon, TF-IDF Ridge regression, and contextual RoBERTa regression as computational measures of affective language across valence, arousal, and dominance. Validation against EmoBank and SemEval precedes analysis of within-category variation and application to HuffPost headlines and short descriptions. RoBERTa achieves the strongest benchmark alignment with stable EmoBank rankings across seeds. Emotion categories leave 52.7%–94.4% of model-score variation within labels, and overlapping SemEval annotations reveal human valence variation unexplained by categories. Seven of twelve method-specific trends in the gap between mental-health-related and other health news pass three uncertainty checks; direct contrasts reveal robust method differences only for arousal. Four archival headlines illustrate contrast, negation, recovery-oriented wording, and modality without additional validation. These findings support selecting and validating measures for the affective distinctions a linguistic question requires, focusing on published language rather than audience emotion or population mental health.

Keywords: affective language, computational measures, mental health news, valence, arousal, dominance, contextual models

Introduction

This study examines whether computational representations yield the same account of affective language in English news headlines and short descriptions. It compares category labels, word norms, linear prediction, and contextual prediction, asking whether their differences alter estimated change in mental-health-related versus other health news. The issue is both linguistic and methodological: sadness identifies an emotion type, whereas continuous scores describe pleasantness, activation, or perceived control. Shared scaling does not make these representations equivalent.

Affective meaning depends on textual relations. Appraisal theory distinguishes attitude, engagement, and graduation in interpersonal evaluation (Martin & White, 2005). This framework motivates contextual interpretation, but this study measures only valence, arousal, and dominance (VAD), not the full Appraisal framework. Thus, these dimensions cannot substitute for appraisal categories, grammatical agency, or social power.

Yuze Li, MA, Graduate School of Arts and Sciences, Columbia University, NY, USA

Large corpora support comparisons across topics and periods, yet require interpretive choices. Corpus-assisted discourse research relates aggregate patterns to contextual reading (Baker et al., 2008), and automated headline analysis can track affective language over time (Rozado et al., 2022). Predictive accuracy alone does not establish equivalent affective distinctions or historical accounts: measurement choices condition textual interpretation (Grimmer & Stewart, 2013; Jacobs & Wallach, 2021).

Mental-health news combines references to distress with reassurance, treatment, stigma criticism, and recovery. Media research distinguishes stigma, agency, and recovery in portrayals of mental illness (McGinty et al., 2016; Stuart, 2006; Whitley & Campbell, 2014). Negative topic vocabulary need not determine a headline's evaluation, nor do positive scores reduce stigma. Social interpretation therefore requires a defined textual target.

The design links representation, human-rating validation, within-category variation, and direct news-trend comparisons to show how measurement conditions account for textual change. Four archival headlines serve as interpretive illustrations of wording, not causal explanations of statistical differences.

Affective Meaning and Computational Representation

Discrete systems assign labels such as anger, joy, fear, and sadness. Dimensional systems locate affect continuously: valence from unpleasant to pleasant, arousal from low to high activation, and dominance from low to high perceived control (Russell & Mehrabian, 1977). VAD distinguishes properties beyond positive–negative polarity.

Category mapping assigns predicted labels their fixed VAD prototypes; a lexicon averages matched word norms. Term frequency–inverse document frequency (TF-IDF) Ridge regression learns word and word-pair associations with human continuous ratings, while RoBERTa VAD learns those targets contextually. Both deriving dimensions from categorical supervision and transparent lexicon scoring have established uses (Park et al., 2021; van der Veen & Bleich, 2025). Suitability depends on the affective distinctions and inferences required.

Two gaps motivate the comparison: prototype assignment may compress within-label differences, which must be distinguished from human-rating variation; and benchmark agreement or performance need not imply matching application trends. Model development is separated from news analysis. The study neither tests understanding of specific constructions nor treats VAD as a complete account of discourse meaning.

Research Questions

RQ1. How do the four approaches differ in their alignment with human continuous-affect ratings on English texts, and are neural-model rankings stable across training seeds?

RQ2. How much continuous variation remains within human emotion categories, and how do text length, lexicon coverage, and multilabel status relate to disagreement between measures?

RQ3. Do the approaches yield different annual trends in the affective-language gap between mental-health-related and other health or wellness news headlines and short descriptions?

Data and Methods

Study Design and Corpora

The design links benchmark validation, representation analysis, and news application. (Table 1; Buechel & Hahn, 2017; Demszky et al., 2020; Misra, 2022; Mohammad et al., 2018). News texts informed none of the model tuning, thresholds, construct prototypes, or candidate-period selection.

Table 1
Corpora and Analytical Roles

Dataset	Domain	Sample	Annotation	Analytical role
EmoBank	Multi-genre English sentences	10,062: 8,062 train; 1,000 dev; 1,000 test	Human VAD, 1–5	Direct VAD training/tuning; held-out within-domain validation
GoEmotions	English Reddit comments	43,410 train; 5,426 dev; 5,427 test	27 emotions + neutral; multilabel	Classifier training/evaluation; category compression and disagreement
SemEval-2018 V-reg	English tweets	937 test; Lexicon: 933	Human valence intensity, rescaled 1–5	Held-out cross-domain valence validation
HuffPost News Category Dataset	WELLNESS/HEALTHY LIVING headlines + descriptions	17,248 unique documents; 32,702 analysis rows	Category/date metadata; no new ratings	Mental-health–control news comparison

Note. VAD = valence, arousal, and dominance. Topic definitions and text variants generate multiple analysis rows per prepared news document.

EmoBank training and development data fitted and tuned the direct regressors; its test set remained held out. GoEmotions used the simplified Hugging Face configuration with multilabel annotations retained. Single-label test analyses included neutral: 4,590 valid records per supervised continuous model and 4,565 for Lexicon.

SemEval scores were linearly rescaled to 1–5; the test set informed no model or threshold selection. The 761 tweets shared across its emotion-classification and valence-regression tasks also permit direct comparison of human category and valence annotations.

News inputs concatenate headline and short description with a period, excluding author names and links; dates and categories provide metadata. We fixed the candidate period, 2012-Q2–2018-Q1, based on joint WELLNESS and HEALTHY LIVING coverage. Only quarters meeting group-size rules entered inference; later records were descriptively separate because archive categories changed.

The narrow topic rule covers mental health, mental illness or disorder, depression, anxiety, bipolar disorder, schizophrenia, post-traumatic stress disorder, suicide, self-harm, eating disorders, obsessive-compulsive disorder, panic disorder, and psychiatry. The broad sensitivity rule adds trauma, stress, burnout, addiction, substance use, alcoholism, loneliness, grief, therapy, psychology, and counseling. A shared control group matches neither rule. These identify topic language, not clinically or manually validated classes.

For each topic definition, group, and quarter, deterministic document hashes selected at most 1,000 articles, with a minimum of 20 per group. The primary narrow analysis retains 20 quarters: 2012-Q2–2015-Q4 and 2016-Q4–2017-Q4. The 2016-Q1–Q3 gap and the excluded 2018-Q1 were left uninterpolated. Four-method common support contains 13,378 control and 1,087 mental-health article-quarter observations.

Four Measures of Affective Language

Outputs were expressed on a 1–5 VAD scale and clipped to its bounds. Supervision differs: Category mapping learns GoEmotions labels, direct regressors learn EmoBank ratings, and Lexicon uses word norms. Shared scaling does not equate these targets.

Category mapping fine-tuned roberta-base for 28 GoEmotions labels (Liu et al., 2019): 128 tokens, three epochs, learning rate 2×10^{-5}, train/evaluation batches 16/32, weight decay 0.01, and warm-up 0.10. The development macro-F1 selected the checkpoint. A global threshold was scanned from 0.20 to 0.60 in 0.05 steps, with micro-F1 breaking ties; the selected 0.25 yielded test micro-/macro-F1 of 0.607/0.469.

NRC Valence, Arousal, and Dominance Lexicon prototypes (NRC-VAD; Mohammad, 2018) uniformly average across labels above the threshold, defaulting to the highest-probability label. Soft mapping instead weights prototypes by predicted probabilities. This fixed-prototype baseline differs from Park et al.'s (2021) training objective. Shared NRC-VAD resources render Category mapping–Lexicon agreement nonindependent.

Lexicon averages NRC-VAD version 2.1 unigram matches per dimension, linearly converting released −1 to 1 norms to 1–5. Unmatched texts remain missing, without neutral imputation. Matched-token counts and proportions diagnose coverage; word order and negation scope are not explicitly modeled.

TF-IDF Ridge lowercases text, strips Unicode accents, and uses unigrams/bigrams, sublinear term frequency, a 100,000-feature cap, and document frequencies from two records to 98% of training records. Multi-output estimation fits one Ridge model per VAD target. The mean development root mean squared error (RMSE) selected a penalty of 1.0 from 0.1, 1, 10, and 100.

RoBERTa VAD fine-tuned roberta-base on three EmoBank targets, converting 1–5 scores to 0–1 for training and back for evaluation. Settings matched the classifier except for five epochs; the checkpoint with the lowest development RMSE was selected.

Frozen instruments scored concatenated news in up to four chunks of at most 192 tokens. Chunk averages used token counts for Category mapping, Ridge, and RoBERTa, and matched-token counts for Lexicon. No text was truncated. Document and scoring-configuration checksums keyed cached outputs.

Validation and Representation Analysis

RQ1 uses Pearson and Spearman correlations, mean absolute error (MAE), RMSE, R^2, and linear calibration on method-specific and common support. Association complements error: mean-concentrated predictions can have modest errors yet distinguish texts poorly. Bootstrap intervals use 1,000 replicates; paired MAE comparisons employ shared valid records. Sign-flip tests use 10,000 permutations and Benjamini–Hochberg correction separately across 18 EmoBank and six SemEval contrasts (Benjamini & Hochberg, 1995). Neural training used seeds 17, 42, and 73; Lexicon and Ridge remained fixed. Seed-42 news instruments were retained; benchmark rankings were checked for initialization sensitivity. SemEval assesses cross-domain valence only.

RQ2 decomposes GoEmotions single-label model-score variance into between- and within-category components:

$$\eta^2_{category} = 1 - \frac{SS_{within}}{SS_{total}}.$$

SS denotes sums of squares; the complementary share remains within categories. Variance shares use all eligible single-label records, including neutral. Within-category standard deviations (SDs) and interquartile ranges summarize only categories with at least 30 valid scores (23 categories, including neutral). GoEmotions lacks human VAD ratings: these are model-score quantities, not direct estimates of human emotion lost through categorization.

For overlapping SemEval annotations, human valence dispersion was summarized for labels with at least 20 cases. Multilabel indicators predicted human valence in ten repetitions of shuffled ten-fold cross-validation; cross-validated R^2 and its complement quantify recovered and unexplained variation.

Disagreement averages absolute pairwise distances across all methods and three dimensions after method–dimension standardization. HC3 regressions use log word count, lexicon coverage, multilabel status, and multilabel-by-length interaction (MacKinnon & White, 1985). Four-method analysis uses 5,400 common-support comments; a three-method analysis excludes Lexicon and uses all 5,427 to check length associations without coverage effects.

Longitudinal Comparison and Uncertainty

RQ3 standardizes document scores by each method–dimension's primary-period reference mean and SD:

$$z_{imd} = \frac{y_{imd} - \mu_{md}}{\sigma_{md}}.$$

Subscripts identify document, method, and dimension. Scaling enables effect-size comparison without equating constructs. Each quarter's gap subtracts the control mean from the mental-health-related mean:

$$G_{tmd} = \underline{z}_{tmd,MH} - \underline{z}_{tmd,control}.$$

Positive gaps indicate higher mental-health-related scores. Change concerns this between-group gap, not either group alone. Twelve method–dimension series are fitted with calendar-quarter trends and seasonal indicators:

$$G_t = \beta_0 + \beta_1 q_t + \gamma_2 I(Q2_t) + \gamma_3 I(Q3_t) + \gamma_4 I(Q4_t) + \varepsilon_t,$$

Time follows the actual calendar, retaining missing-quarter gaps; Q1 is the seasonal reference. Annual change is four times the quarterly slope. Bartlett-kernel heteroskedasticity- and autocorrelation-consistent (HAC) standard errors use calendar lags 1–4 and HC1 correction (MacKinnon & White, 1985; Newey & West, 1987). False discovery rate (FDR) correction covers twelve primary trends.

Robust results require asymptotic HAC $q < .05$, small-sample t-based $q < .05$ (15 degrees of freedom), and a 95% residual calendar-block bootstrap interval excluding zero. Bootstrap estimation uses four-quarter blocks, 1,000 replicates, and no blocks across gaps (Künsch, 1989). Direct contrasts fit the same model to paired gap-series differences, with eighteen contrasts as a separate FDR family and the same robustness criteria. Differing individual significance does not establish differing trends.

Sensitivity checks vary support, mask topic terms, broaden the topic rule, set minimum group counts to 10, 30, or 50, and omit each of eight narrow term groups. Scores are also residualized on original category and log word count; within-category trends require twelve quarters. Diagnostics comprise calendar-segment

Durbin–Watson statistics, exact-lag residual correlations, leave-one-quarter-out estimates, seasonally adjusted Theil–Sen slopes, and quadratic trends; they are not additional confirmatory tests.

Authentic Language Illustrations

Four headlines were purposively selected after quantitative analysis to contrast recovery wording, negation scope, and modality, without scores or construction frequencies. Verification used the public HuffPost mirror (Heegyu, n.d.). All carry *WELLNESS/HEALTHY LIVING* metadata, anxiety/depression wording, and primary-interval dates; hash-sample membership remains unestablished. Archive wording/dates and referenced article links document provenance. Readings are author interpretations, not new rating validation or explanations of model disagreement.

Software and Data Governance

Analyses used Python with Accelerate 1.10.1, Datasets 4.1.1, NumPy 2.1.3, pandas 2.2.3, PyArrow 23.0.1, scikit-learn 1.6.1, SciPy 1.16.3, statsmodels 0.14.5, PyTorch 2.11.0+cpu, and Transformers 4.56.2.

The study uses released benchmarks and public media, with no recruitment, user contact, new ratings, private-account access, or reidentification; no institutional ethics review was sought. Contextual readings add no participant data. EmoBank and GoEmotions licenses are CC BY-SA 4.0 and CC BY 4.0; NRC-VAD permits noncommercial use without redistribution. SemEval and HuffPost retain distributor conditions and source rights. Only brief attributed illustrative excerpts appear here.

Results

Alignment with Human Continuous Ratings

RoBERTa VAD had the highest Pearson r and lowest MAE/RMSE across EmoBank dimensions (Table 2). Valence was most recoverable; arousal and dominance retained substantial unexplained variation. Ridge substantially reduced errors versus Category mapping and Lexicon despite low arousal/dominance correlations.

Table 2
Alignment with Human Continuous Affect Ratings

Dataset	Method	Dimension	n	Pearson r	MAE	RMSE	R^2
EmoBank	Category mapping	Valence	1,000	0.494	0.401	0.567	-1.699
EmoBank	Category mapping	Arousal	1,000	0.264	1.044	1.132	-20.197
EmoBank	CM	V	1,000	0.211	0.602	0.651	-8.570
EmoBank	CM	A	973	0.391	0.395	0.498	-1.099
EmoBank	CM	D	973	0.204	0.293	0.384	-1.551
EmoBank	LX	V	973	0.154	0.252	0.337	-1.573
EmoBank	LX	A	1,000	0.514	0.213	0.296	0.264
EmoBank	LX	D	1,000	0.260	0.189	0.245	0.008
EmoBank	RG	V	1,000	0.230	0.156	0.209	0.013
EmoBank	RG	A	1,000	0.809	0.153	0.208	0.635
EmoBank	RG	D	1,000	0.561	0.181	0.230	0.125
EmoBank	RB	V	1,000	0.501	0.149	0.192	0.166
SemEval	RB	A	937	0.646	0.794	0.994	-0.310
SemEval	RB	D	933	0.618	0.588	0.720	0.315
SemEval	CM	V	937	0.554	0.627	0.759	0.238
SemEval	LX	V	937	0.797	0.463	0.574	0.563

Note. MAE = mean absolute error; RMSE = root mean squared error. Scores/errors use 1–5; r and R^2 are dimensionless. Negative R^2 means worse squared error than predicting the human-score mean. Lexicon n excludes unmatched texts; paired MAE tests use each pair's shared valid records. CM = Category mapping; LX = Lexicon; RG = TF-IDF Ridge; RB = RoBERTa VAD; V/A/D = valence/arousal/dominance (Tables 2–5).

Negative EmoBank R^2 for Category mapping and Lexicon indicates poor error performance/calibration, not absent association. Seventeen of eighteen paired MAE contrasts passed FDR; Category mapping–Lexicon valence did not (q = .4871). Ridge–RoBERTa arousal differed by .0086 (q = .0479), but its bootstrap interval included zero, weakening this particular error advantage.

RoBERTa also led SemEval valence. Category mapping ranked second in correlation but worst in error, with negative R^2; Lexicon had weaker correlation but better calibration and error. On 933 shared records, Lexicon beat Ridge MAE by .0395 (q = .0017); all six paired contrasts passed FDR. Rankings thus depend on domain and association versus error.

Soft mapping improved the hard baseline without closing the validation gap: EmoBank r = .628/.330/.259 and R^2 = −.509/−10.588/−3.859 (V/A/D); SemEval valence r = .730, R^2 = .289. Across seeds 17/42/73, RoBERTa retained the best EmoBank r and MAE in every dimension; r ranges were .805–.809/.558–.574/.467–.522 (V/A/D). Classifier micro-/macro-F1 ranges were .599–.607/.463–.477.

Graded Variation within Categories and Method Disagreement

GoEmotions categories explained 5.6%–47.3% of continuous-score variance, leaving 52.7%–94.4% within labels (Table 3). Valence showed the strongest category structure for every method. Gold-category prototypes have zero within-category variance by construction; continuous outputs vary among texts sharing a human label. This concerns model outputs, not directly lost human emotion.

Table 3
Category Structure and Cross-method Disagreement

Panel	Quantity	Valence	Arousal	Dominance
A	LX category variance share	0.234	0.056	0.101
A	RG category variance share	0.355	0.153	0.072
A	RB category variance share	0.473	0.177	0.272
A	LX median within-category SD	0.368	0.257	0.250
A	RG median within-category SD	0.193	0.120	0.097
A	RB median within-category SD	0.283	0.190	0.092
B	Pairwise Pearson r range	0.438 to 0.628	0.148 to 0.381	0.138 to 0.363
C	4-method log length b	-0.030, p < .001		
C	4-method coverage b	-0.085, p < .001		
C	4-method multilabel b	-0.080, p = .002		
C	3-method log length b	-0.005, p = .167		
C	3-method multilabel b	-0.088, p = .005		
D	SemEval overlap n	761	—	—
D	Labels with n ≥ 20	11	—	—
D	Median human within-label SD	0.616	—	—
D	Human within-label SD range	0.501 to 0.719	—	—
D	Mean cross-validated R^2	0.668	—	—
D	Unexplained human variance share	0.332	—	—
D	Mean cross-validated MAE	0.436	—	—

Note. A: variance shares use all eligible single-label GoEmotions scores, including neutral (RG/RB n = 4,590; LX n = 4,565); median SDs use 23 categories, including neutral, with ≥30 scores. B: pairwise Pearson r ranges. C: coefficients for disagreement averaged across dimensions, not separate V/A/D models. D: overlapping human SemEval annotations; arousal/dominance were not assessed. SD = standard deviation; MAE = mean absolute error.

SemEval supplies complementary human evidence: categories explained 66.8% of valence variation, leaving 33.2% unexplained. Within-label dispersion (Table 3D) shows that categories organize human valence without determining its continuous range.

Agreement was strongest for valence (Table 3B). Four-method disagreement decreased with log length (b = −.03045, HC3 SE = .00336, p < .001), coverage (b = −.08510, SE = .01221, p < .001), and multilabel status (b = −.08005, SE = .02603, p = .0021); the multilabel-by-length interaction was negligible (p = .902).

Without Lexicon, length was nonsignificant (b = −.00528, SE = .00382, p = .167), while multilabel status remained negative (b = −.08818, SE = .03134, p = .0049). Length thus primarily reflects lexicon support. Negative multilabel coefficients need not imply easier measurement of mixed emotions: averaging standardized pairwise distances may favor more central aggregate positions.

Change in the Relative Affective-language Gap

In 20 primary quarters, common support covered 99.96% of controls (13,378/13,384) and 100% of mental-health articles (1,087/1,087). Quarterly-path Spearman ranges were .271–.713/.035–.764/.102–.586 (V/A/D). Eight of twelve annual trends passed asymptotic HAC FDR; seven passed all three criteria (Table 4).

Table 4
Annual Trends in the Standardized Mental-health-related Minus Control Language Gap

Method	Dimension	Trend	SE	HAC q	t q	Bootstrap 95% CI	Robust
CM	A	0.0162	0.0150	.3055	.3241	[-0.0148, 0.0476]	No
CM	D	0.0547	0.0200	.0108	.0265	[0.0032, 0.1090]	Yes
CM	V	0.0378	0.0238	.1338	.1588	[-0.0190, 0.0999]	No
LX	A	-0.0188	0.0317	.5526	.5615	[-0.0717, 0.0325]	No
LX	D	0.0260	0.0162	.1338	.1588	[-0.0235, 0.0744]	No
LX	V	0.0725	0.0335	.0458	.0706	[-0.0046, 0.1475]	No
RG	A	0.0567	0.0150	.0006	.0061	[0.0176, 0.1022]	Yes
RG	D	0.0431	0.0126	.0015	.0093	[0.0221, 0.0659]	Yes
RG	V	0.0722	0.0194	.0006	.0061	[0.0281, 0.1153]	Yes
RB	A	0.0956	0.0158	< .001	.0003	[0.0740, 0.1189]	Yes
RB	D	0.0846	0.0270	.0034	.0136	[0.0213, 0.1459]	Yes
RB	V	0.0856	0.0214	.0004	.0061	[0.0193, 0.1512]	Yes

Note. Trend and SE units are SD/year. HAC q adjusts across twelve trends; t q uses the small-sample t reference. Robust requires both q < .05 and a bootstrap 95% CI excluding zero. HAC = heteroskedasticity- and autocorrelation-consistent; CI = confidence interval.

Ridge and RoBERTa showed robust positive trends across all dimensions, with the mental-health–control gap becoming more positive, activated, and dominant (Table 4). These reflect published language, not readers, clinical states, or stigma. Category mapping yielded robust dominance only. Lexicon arousal was slightly negative and imprecise; its valence passed asymptotic FDR but failed both small-sample and bootstrap checks. Such differences in individual support require direct trend contrasts. Only two of eighteen direct contrasts were robust, both for arousal (Table 5): RoBERTa exceeded Category mapping by .0793 SD/year and Lexicon

by .1144. No valence or dominance contrast passed all three checks; measurement-dependent change is established for arousal only.

Table 5
Direct Differences between Method-specific Annual Trends

Dim.	A	B	Δ	SE	HAC q	t q	Bootstrap 95% CI
A	CM	LX	0.0350	0.0413	.5088	.5259	[-0.0425, 0.1119]
A	CM	RG	-0.0405	0.0148	.0382	.0933	[-0.0903, 0.0043]
A	CM	RB	-0.0793	0.0210	.0014	.0162	[-0.1190, -0.0407]
A	LX	RG	-0.0756	0.0389	.1334	.1773	[-0.1382, -0.0049]
A	LX	RB	-0.1144	0.0272	.0005	.0136	[-0.1537, -0.0757]
A	RG	RB	-0.0388	0.0206	.1334	.1773	[-0.0853, 0.0031]
D	CM	LX	0.0287	0.0223	.3950	.4341	[-0.0300, 0.0889]
D	CM	RG	0.0116	0.0262	.7075	.7140	[-0.0335, 0.0567]
D	CM	RB	-0.0299	0.0321	.4873	.5077	[-0.0868, 0.0241]
D	LX	RG	-0.0171	0.0160	.4690	.4966	[-0.0626, 0.0290]
D	LX	RB	-0.0586	0.0251	.0880	.1524	[-0.0966, -0.0217]
D	RG	RB	-0.0415	0.0209	.1334	.1773	[-0.0942, 0.0190]
V	CM	LX	-0.0347	0.0488	.5729	.5860	[-0.1330, 0.0626]
V	CM	RG	-0.0344	0.0320	.4690	.4966	[-0.0856, 0.0135]
V	CM	RB	-0.0478	0.0243	.1334	.1773	[-0.0939, -0.0015]
V	LX	RG	0.0003	0.0196	.9877	.9879	[-0.0542, 0.0507]
V	LX	RB	-0.0131	0.0306	.7075	.7140	[-0.0777, 0.0504]
V	RG	RB	-0.0134	0.0141	.4873	.5077	[-0.0505, 0.0256]

Note. Δ = Method A − Method B in SD/year; both q columns adjust across eighteen contrasts. Only arousal CM−RB and LX−RB meet all three robustness criteria. Differences use unrounded estimates and may differ from subtracting displayed Table 4 values.

Sensitivity and Diagnostics

Method-specific support changed slopes by ≤.000058 SD/year. Topic masking reversed no signs (maximum change .0210), although four unadjusted HAC decisions changed. The broad rule retained 23 quarters, changed slopes by ≤.0759, and reversed Lexicon valence. Broader keywords define a different linguistic population.

Minimum group counts of 10/20/30 retained 23/20/15 quarters, preserving all twelve slope directions. A minimum of 50 retained ten quarters, below the estimation rule. Omitting each of eight narrow term groups preserved directions; category/length residualization changed slopes by ≤.00322. Within-category analysis was estimable for HEALTHY LIVING (twelve quarters), not WELLNESS (eight).

Leaving out individual quarters preserved eleven of twelve directions; Lexicon arousal ranged −.0364–.0076 (maximum change .0264). Theil–Sen intervals were positive for all Ridge/RoBERTa dimensions. Segmented Durbin–Watson ranged .696–2.474; three series had nominal quadratic terms without multiplicity correction. Annual coefficients summarize a short, discontinuous series, not constant quarterly change.

Contextual Reading of Four Archival Headlines

These contextual readings distinguish topic vocabulary from affective interpretation; they imply no model scores, construction prevalence, or explanation of aggregate slope differences.

Contrast and scope. Yogis's (2012) July 31 *WELLNESS* headline contrasts "Anxiety But Not Fear." Negation restricts a reduction claim to one of the two related emotions rather than negating all affective content. Registering both nouns leaves this relation unspecified; the example does not establish which method captures it.

Recovery transition. Pervez's (2013) June 4 *WELLNESS* headline links "Overcame Depression" with "Happy Nomad." A first-person overcoming event redirects a negative condition toward a positively characterized identity. Coexisting positive/negative vocabulary establishes neither two discrete expressed emotions nor specific numerical scores.

Negated speech. Holmes's (2014) January 29 *WELLNESS* headline uses "Not To Say" concerning someone with depression. Negation targets proposed speech, not depression's existence or the speaker's feelings. The diagnostic term marks the topic, and the infinitival construction signals advice to another.

Modality. Kelland's (2017) September 7 *HEALTHY LIVING* headline cites "Can Ease Depression And Paranoia": the modal marks possibility, ease assigns beneficial direction to negatively associated conditions, and study attribution marks a reported research claim. These relations qualify lexical interpretation without validating the health claim or assigning VAD scores.

Together, these readings motivate future construction-specific tests; they do not establish the causes of aggregate arousal divergence or a corpus-wide shift toward recovery framing.

Discussion

What the Representations Preserve

Three findings connect representation with interpretation: direct continuous supervision, especially RoBERTa, best aligned with human benchmark ratings; categories retained structure but left graded variation unresolved; and direct news-trend differences were robust for arousal only.

Labels identify emotion types; VAD locates affective positions. Fixed prototypes compress within-label differences, a concern supported here by human SemEval valence variation. Neither representation is universally preferable: fear/anger queries may require categories, whereas activation queries require dimensional scales.

Lexicon transparency aids auditability and limited computing resources (van der Veen & Bleich, 2025), but coverage shaped disagreement here. Report matched-token support and leave unmatched texts missing. Contextual examples distinguish lexical presence from evaluation without constituting a lexicon error analysis.

Ridge and RoBERTa share continuous supervision; RoBERTa's higher correlations support better benchmark alignment under this training setup. They do not identify its linguistic mechanism: negation, modality, word order, and other constructions were not experimentally isolated.

Implications for Corpus-assisted Language Research

Contextual accounts of evaluation concern attitudes and textual relations (Martin & White, 2005). VAD covers only part of that territory: valence is not full evaluation, arousal is not automatically rhetorical intensity, and dominance is not grammatical agency or social power. Measures should follow explicitly defined affective targets.

RoBERTa's stronger benchmark alignment and larger arousal-gap trend than Category mapping/Lexicon jointly show how measurement can condition historical description. Without news-specific human VAD ratings, however, benchmark suitability cannot establish the true HuffPost trajectory.

Valence/dominance differences in individual significance remain uncertain findings; only direct contrasts test trend differences. Topic rules likewise condition interpretation: adding stress, therapy, grief, or loneliness changes the population compared.

Positive supervised trends may reflect topic composition, editorial style, clinical vocabulary, advocacy, or external events; the design cannot distinguish causes. Framing includes problem definition, causal interpretation, evaluation, and remedies (Entman, 1993), of which VAD captures only an affective layer. Findings concern one outlet's relative wording, not reduced stigma, increased agency, or changed reader responses.

Limitations and Future Research

Benchmark ratings retain annotation uncertainty and genre/protocol specificity. SemEval tests valence transfer only; news arousal/dominance remain unvalidated. Category mapping and Lexicon share resources. Four representative approaches cannot exhaust alternatives: other models, prompts, lexicons, or calibration may differ.

Unvalidated topic rules may yield false positives/negatives; masking removes one lexical pathway, not all topic meaning. One English outlet, two categories, and twenty nonconsecutive quarters limit generalization; later news is outside inference. Dependence-aware checks reduce specific concerns without removing residual composition differences or establishing causality.

Purposive readings supply no construction inventory, independent annotations, or measured model errors. Future news-specific VAD ratings and annotations of negation scope, evaluative target, modality, and discourse function could test construction-specific disagreement. Replication across outlets/languages would assess generalizability.

Conclusion

Computational measures retain different aspects of affective language: continuous supervision best aligns with human ratings, categories leave graded variation unresolved, and robust method-dependent news trends concern arousal only. Authentic headlines connect interpretation to relations among words. Corpus-assisted research should define the affective target, validate its measure, and test direct contrasts before inferring linguistic change.

Data availability

The study used publicly released resources: EmoBank, GoEmotions, SemEval-2018 Task 1, the HuffPost News Category Dataset, and NRC-VAD version 2.1. The archival headlines were verified in the linked News Category Dataset revision. Access and reuse are subject to the respective distributors' terms.

References

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society, 19*(3), 273-306.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.

Buechel, S., & Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 578-585). Association for Computational Linguistics.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040-4054). Association for Computational Linguistics.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51-58.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297.

Heegyu. (n.d.). News category dataset [Data set]. Hugging Face. Retrieved from https://huggingface.co/datasets/heegyu/news-category-dataset/tree/304a05a55bc6abc0446d8fae0d0771716b6a271a

Holmes, L. (2014, January 29). 6 things not to say to someone with depression. HuffPost. Retrieved from https://www.huffingtonpost.com/entry/what-not-to-say-to-someon_us_5b9dd1cae4b03a1dcc8d49fd

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). Association for Computing Machinery.

Kelland, K. (2017, September 7). Treating insomnia can ease depression and paranoia, study finds. HuffPost. Retrieved from https://www.huffingtonpost.com/entry/treating-insomnia-can-ease-depression-and-paranoia-study-finds_us_59b1d8cfe4b0dfaafcf6ccdc

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics, 17*(3), 1217-1241.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305-325

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan.

McGinty, E. E., Kennedy-Hendricks, A., Choksy, S., & Barry, C. L. (2016). Trends in news media coverage of mental illness in the United States: 1995-2014. *Health Affairs, 35*(6), 1121-1129.

Misra, R. (2022). News category dataset. arXiv: 2209.11429

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174-184). Association for Computational Linguistics.

Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 1-17). Association for Computational Linguistics.

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703-708.

Park, S., Kim, J., Ye, S., Jeon, J., Park, H. Y., & Oh, A. (2021). Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4367-4380). Association for Computational Linguistics.

Pervez, A. (2013, June 4). How I overcame depression and became the happy nomad. HuffPost. Retrieved from https://www.huffingtonpost.com/entry/happiness-tips_us_5b9d31d2e4b03a1dcc85633c

Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE, 17*(10), Article e0276367.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*(3), 273-294.

Stuart, H. (2006). Media portrayal of mental illness and its treatments: What effect does it have on people with mental illness? *CNS Drugs, 20*(2), 99-106.

van der Veen, A. M., & Bleich, E. (2025). The advantages of lexicon-based sentiment analysis in an age of machine learning. *PLOS ONE, 20*(1), Article e0313092.

Whitley, R., & Campbell, R. D. (2014). Stigma, agency and recovery amongst people with severe mental illness. *Social Science & Medicine, 107*, 1-8.

Yogis, J. (2012, July 31). Alcohol reduces anxiety but not fear. HuffPost. Retrieved from https://www.huffingtonpost.com/entry/alcohol-effects_us_5b9c0ebee4b03a1dcc7c0dba