The Divergence Hypothesis: Unmasking Lexical Interference and Label Bias in Mental Health NLP

Moustafa Yehia Hassan
Doha Institute for Graduate Studies
Doha, Qatar
moustafa.hassan@dohainstitute.edu.qa

Abstract

Computational mental health (CMH) classifiers often degrade under distribution shift because human annotators and distant-supervision pipelines reward different linguistic signals. We introduce TSS (Triple-Stream Stress probe), a multi-channel diagnostic framework that decomposes text into (A) lexical character n-grams, (B) a small, mostly content-free morpho-syntactic channel, and (C) a 154-feature psycholinguistic style channel. Across four English datasets ($N = 12{,}906$), TSS reveals a *lexical interference effect*: adding lexical features to the style channel reduces Macro-F1 on human-labeled data (mean drop 0.072, $p < 10^{-4}$) but not on auto-labeled data. We propose Degree of Divergence (DoD), a difference-in-differences statistic adapted from econometrics for label-source auditing, with instance-level bootstrap inference; the headline estimate is $\mathrm{DoD}_{\mathrm{BC-A}} = 0.0374$, 95% CI $[0.0097, 0.0651]$, $p = 0.0032$. A platform-stratified Twitter-only DoD (which removes the Reddit vs. Twitter contrast) reproduces the pattern with bootstrap inference: $\mathrm{DoD}^{\mathrm{Tw}}_{\mathrm{BC-A}} = +0.096$ ($p < 0.001$) and $\mathrm{DoD}^{\mathrm{Tw}}_{\mathrm{AC-A}} = -0.089$ ($p < 0.001$). Interventional masking (`pos_only`) retains ∼95–99% of Channel C's performance after destroying content words on human datasets, indicating that the style channel does not rely primarily on lexical surface form. TSS is positioned as a diagnostic audit framework, not a clinical screening tool: it flags label-source-specific shortcut learning before generalization claims are made.

1 Introduction

Computational Mental Health (CMH) promises scalable screening and monitoring from language, but its progress is constrained by a mismatch between benchmark labels and the construct they are taken to measure. Large corpora are frequently built via distant supervision—keyword filters, self-reported diagnoses, or community-membership proxies—which can systematically reward lexical cues (Coppersmith et al., 2015; Harrigian et al., 2021). Human annotators, in contrast, often rely on *how* something is said: fragmentation, negation scope, function-word usage, and stylistic signatures (Tausczik and Pennebaker, 2010; Pennebaker et al., 2003). When the two label sources are conflated, models can score well by exploiting spurious lexical correlations (*shortcut learning*; Geirhos et al., 2020) rather than learning robust psycholinguistic markers (Ernala et al., 2019; D'Amour et al., 2022).

Task. We study *binary stress detection* from short social-media posts. Each instance receives a label $y \in \{0, 1\}$, where $y{=}1$ indicates that the post expresses current psychological distress or stress-related experience as judged by the dataset's labeling protocol, and $y{=}0$ indicates the absence of such expression. Throughout this paper, "stress" refers to the dataset-level distress/stress label, *not* to a clinical diagnosis; no claim of diagnostic validity is made. The goal is not to build a state-of-the-art classifier, but to audit which linguistic signals different label sources reward.

Research questions. **RQ1 (Label-source divergence)** Does changing the label source (human vs. distant supervision) systematically change which linguistic channels are rewarded? **RQ2 (Lexical interference)** Does adding lexical content to stylistic features help or harm under human annotation? **RQ3 (Quantification)** Can the human-vs-auto gap be quantified as a single, statistically grounded scalar?

Contributions. (i) We release TSS, a channel-separable, length-robust auditing probe that decomposes text into lexical (A), mostly content-free morpho-syntactic (B), and psycholinguistic style (C) channels, with explicit length normalization and conditional scaling for cross-platform trans-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 1–14

fer. (ii) We document a *lexical interference effect*: on human-labeled data, augmenting C with lexical A degrades Macro-F1; an interventional masking suite (`pos_only`, `content_only`, `function_only`) tests whether performance survives deliberate destruction of lexical shortcuts. (iii) We introduce DoD, a difference-in-differences auditing statistic with instance-level bootstrap inference (n=10,000) and Benjamini–Hochberg FDR control, plus a *platform-stratified* variant that removes the Reddit vs. Twitter contrast. (iv) We provide leave-one-domain-out (LODO) evaluation, unsupervised stylistic phenotyping (bootstrap ARI ≈ 0.98), paired baseline comparisons against MentalBERT (Ji et al., 2022) and few-shot LLaMA-3 (Grattafiori et al., 2024; Brown et al., 2020), and a qualitative conflict-zone workbook. The full pipeline, decontamination scripts, and qualitative workbook are publicly available.[1]

2 Related Work

Psycholinguistic style. A long line of work shows that function words and stylistic patterns reflect cognitive style more reliably than consciously controlled topical content (Pennebaker et al., 2003; Tausczik and Pennebaker, 2010; Boyd et al., 2022). Cognitive therapy frameworks predict linguistically measurable correlates of distress, most notably absolutist language (Beck, 1976; Al-Mosaiwi and Johnstone, 2018) and specific cognitive distortions (Shickel et al., 2020).

Distant supervision and shortcut learning in CMH. Distant-supervision pipelines built from self-reported diagnoses (Coppersmith et al., 2015) or community membership have repeatedly been shown to yield models that rely on surface lexical cues and degrade under distribution shift (Harrigian et al., 2021; Aguirre et al., 2021; Ernala et al., 2019). These failures align with the broader finding that neural classifiers tend to latch onto annotation artifacts and spurious correlations (Geirhos et al., 2020; Gururangan et al., 2018; McCoy et al., 2019; D'Amour et al., 2022), and that distributional robustness requires explicit intervention (Sagawa et al., 2020; Kaushik et al., 2020). Our work differs in goal: rather than proposing a more robust classifier, we propose a *diagnostic framework* that quantifies how much a given dataset's labels reward lexical shortcuts, in the spirit of behavioral auditing (Ribeiro et al., 2020).

Stress detection benchmarks. Our experimental backbone uses Dreaddit (Turcan and McKeown, 2019) as the structured human-labeled benchmark and MentalBERT (Ji et al., 2022) as a representative domain-pretrained transformer baseline. We also draw on the Twitter mental-health corpora derived from Coppersmith et al. (2015), treated here as a distant-supervision proxy rather than a stress gold standard (see §3).

3 Data and Task

3.1 Datasets

We evaluate on four English datasets after within-dataset exact-match deduplication. Table 1 summarizes the platform, size, label source, and intended use of each dataset.

3.2 Twitter-auto as a Proxy, Not Gold

Twitter-auto is not used as a gold stress dataset. It is constructed by treating posts whose authors self-reported a mental-health diagnosis (depression, PTSD, anxiety, following Coppersmith et al., 2015) as *auto-positive*, under the working assumption that clinically diagnosed populations exhibit elevated linguistic distress markers. We acknowledge that this introduces a *disorder→stress conflation*: a diagnosed user is not necessarily expressing distress in every post, and the resulting label is a noisy proxy for the post-level stress construct studied here. We treat Twitter-auto purely as a distant-supervision proxy for testing whether such labels reward different linguistic signals than human distress annotations. Twitter-gold is a manually annotated subsample drawn from the same underlying corpus, where annotation targeted *perceived psychological distress* rather than the presence of mental-health keywords; stress-keyword removal was applied to reduce trivial lexical leakage. We do not claim expert-clinical annotation for this set; methodological constraints of the manual-validation protocol are acknowledged in §7. A related, substantially smaller manually annotated Twitter validation subset (120 examples) was constructed by Rastogi et al. (2022) to validate automated annotation strategies, which we cite here as methodological context rather than as the same artifact.

[1] Code and appendix: `https://github.com/MoustafaMohamedMoustafaHassan/TSS-Probe-CMH`

Dataset	Platform	N	Label source	Positive label means	Used for
Dreaddit-test	Reddit	715	Human (expert)	Stress / distress expression	Human-labeled evaluation
Twitter-gold	Twitter	2,863	Human (manual annotation)	Perceived psychological distress	Human-labeled evaluation
Twitter-auto	Twitter	6,218	Distant supervision	Mental-health / distress *proxy*	Auto-label comparison
Reddit-combi	Reddit	3,110	Distant supervision	Community-/keyword-derived proxy	Auto-label comparison

Table 1: Dataset overview. Positive prevalence: Dreaddit-test 0.516, Twitter-gold 0.338, Twitter-auto 0.478, Reddit-combi 0.880 (i.e., the positive label is the *majority* class on Reddit-combi; minority prevalence = 0.120). Dreaddit follows Turcan and McKeown (2019); Twitter-auto and the Twitter-gold subsample derive from Coppersmith et al. (2015); Reddit-combi is a combined Reddit corpus assembled under distant supervision (see §3.3).

3.3 Reddit-combi as a Distant-Supervision Proxy

Reddit-combi is an internally assembled Reddit corpus constructed under distant supervision and released with the artifact bundle as `data/raw/Reddit_Combi.csv` (with title/body fields and binary labels). It is not an externally published benchmark; it is an artifact-level distant-supervision corpus assembled for this study and released with the repository, and we report its provenance through the artifact path. Because its construction is not equivalent to expert annotation, we use Reddit-combi exclusively as a Reddit-side auto-label proxy and pair it with Twitter-auto for the auto-source half of the DoD contrast. Crucially, the positive label is the *majority* class on Reddit-combi (0.880); we therefore favor Macro-F1, PR-AUC, and balanced accuracy over raw accuracy or class-1 F_1 when reporting this dataset, and we never use Reddit-combi as a stand-in for human annotation.

3.4 Decontamination and Imbalance

To prevent leakage in cross-platform CMH evaluation, we perform exact-match cross-dataset decontamination on `cleaned_text`, removing overlapping posts from the training pool while preserving evaluation sets. The datasets exhibit severe class imbalance, especially Reddit-combi, where the positive label is the *majority* class (positive prevalence 0.880; minority prevalence 0.120). Accordingly we report Macro-F1 as the primary metric, complemented by PR-AUC under skewed prevalence (PR-AUC is preferred over ROC-AUC when one class dominates). A naive majority-class baseline yields Macro-F1 of 0.340, 0.343, 0.398, and 0.468 on Dreaddit-test, Twitter-auto, Twitter-gold, and Reddit-combi respectively; all TSS channels substantially exceed these baselines, confirming signal acquisition beyond prevalence effects. For paired comparisons we use McNemar's test (McNemar, 1947) on discordant predictions and paired bootstrap CIs (Efron, 1979) for Macro-F1 differences, controlling false discoveries with Benjamini–Hochberg FDR (Benjamini and Hochberg, 1995).

Dreaddit additionally provides five heterogeneous stressor domains (abuse, anxiety, financial, PTSD, social), enabling leave-one-domain-out (LODO) evaluation that probes cross-stressor generalization beyond platform transfer.

4 The TSS Probe

TSS is a diagnostic decomposition, not a competitive classifier. The three channels are designed so that their contrast is informative: comparing what each rewards under each label source isolates lexical-shortcut behavior.

4.1 Channel A: Lexical Surface Form

Channel A uses character-level TF–IDF n-grams of length 3–5 (Salton and Buckley, 1988), with χ^2 feature selection at $k = 500$ (Yang and Pedersen, 1997). We use character n-grams over word n-grams for two reasons. First, social-media text contains hashtags, contractions, emoji, intentional misspellings, and morphological variation that destabilize word tokenization; character-level features are robust to these (Coppersmith et al., 2015). Second, we want Channel A to provide a strong, sub-word lexical baseline so that any *lexical interference effect* we observe cannot be trivially attributed to vocabulary coverage gaps in a word-based representation. The χ^2 filter at k=500 prevents dimensionality blow-up and is comparable in size to standard sparse lexical baselines for short-text classification.

4.2 Channel B: Mostly Content-Free Morpho-Syntax

Channel B converts POS bigrams and abstract POS-SVO triples (extracted over the first 500 characters per document for efficiency, using the Penn Treebank tagset via spaCy; Honnibal et al., 2020) into six interpretable features: five POS-derived plus one absolutist-ratio (B_{abs}), which is the only Channel B feature that accesses lexical tokens (a closed absolutist-word list, following Al-Mosaiwi and Johnstone, 2018). Channel B uses adaptive length smoothing and log-dampened mass features for cross-platform robustness; the exact formulations are given in Appendix A.

Channel B is a diagnostic, not a performance booster. We do *not* present Channel B as an ingredient that improves accuracy when combined with C. As shown in §5, adding B to C is approximately neutral on most datasets (occasionally negative on Dreaddit-test). This is itself informative: C already absorbs much of the useful structural variation; B is useful as (i) a near-content-free, privacy-oriented diagnostic (5 of 6 features discard lexical tokens), and (ii) a structural stress test that probes whether performance survives a near-complete loss of vocabulary.

4.3 Channel C: Psycholinguistic Style

Channel C is the principal stylistic representation, with 154 features grouped into seven feature families (Table 2). It applies a length-robust squashing transform $c_{\text{out}} = \tanh(2c)$ to convert continuous features to soft binary triggers resistant to cross-platform distributional shift, and uses Yule's I rather than TTR for lexical diversity because TTR mathematically decays with text length (Appendix A).

Beyond sentiment-like lexicons, Channel C explicitly encodes cognitive fragmentation (a hallmark of anxiety) via the coefficient of variation of sentence length, short-burst ratios, and abrupt punctuation patterns, as well as structural negation scope distinguishing helplessness (NEG→VERB) from negative self-evaluation (NEG→ADJ).

4.4 Unified Classifier and Calibration

All channels are trained with linear models for interpretability. Feature vectors are L_2-normalized: $x' = x/\|x\|_2$. Conditional scaling disables `StandardScaler` for any channel containing B, to avoid reintroducing length bias. Class imbalance is handled by `class_weight='balanced'` (Pedregosa et al., 2011) rather than undersampling. Decision thresholds are not fixed at 0.5: we calibrate by maximizing F_1 over a bounded grid $[0.20, 0.80]$ on out-of-fold (OOF) probabilities (no leakage). Dense lexical/style channels use L_2 (ridge); any channel containing B uses Elastic-Net (Zou and Hastie, 2005) with internal 5-fold CV; the selected ℓ_1-ratio converges to 0.95, indicating an algorithmic preference for sparse structural evidence.

Feature block	Examples	#
Base style / readability / POS-ratio	Flesch–Kincaid, sentence-length CV, function-word ratios, pronoun densities	29
Open lexicon categories	affect, cognitive, social process, sensory; LIWC-style counts and ratios	110
Structural rhythm / negation / punctuation	ellipses, "!?", NEG→VERB, NEG→ADJ, scope balance, short-burst ratio	11
Raw intensity + Yule's I	caps ratio, elongated tokens, punctuation density, length-independent diversity	4
Total		**154**

Table 2: Channel C feature blocks with feature counts as implemented in `ChannelC_Extended` of the released codebase. Open lexicon categories follow LIWC-style groupings (Tausczik and Pennebaker, 2010; Boyd et al., 2022); absolutist markers follow Al-Mosaiwi and Johnstone (2018); negation-scope features operationalize the helplessness vs. self-evaluation distinction motivated by Beck (1976).

4.5 Degree of Divergence (DoD)

DoD is a difference-in-differences statistic adapted from causal econometrics to quantify label-source bias in NLP (Appendix A, Eq. 3):

$$\text{DoD} = \Delta_h - \Delta_a \quad (1)$$

where each Δ compares a structural channel (e.g., BC) against lexical A within the same label source. DoD is estimated with an instance-level bootstrap ($n = 10{,}000$) over all 12,906 instances. Because DoD as introduced above mixes label source with platform (human Reddit/Twitter vs. auto Reddit/Twitter), we additionally compute a *platform-stratified Twitter-only* DoD that removes the Reddit vs. Twitter contrast (§5.4).

4.6 Masking Suite

Five masks are evaluated: `none`, `pos_only`, `content_only`, `function_only`, `random_pos` (negative control). Unlike standard ablations that only remove feature blocks, the masking suite intervenes directly on the text: `pos_only` destroys lexical content while preserving syntactic traces; `content_only` preserves topical content while erasing function words and syntactic cues; `function_only` isolates function-word dynamics; `random_pos` acts as a destructive control. This design probes whether a channel's performance survives deliberate destruction of lexical shortcuts, complementing the observational DoD analysis.

5 Results

We first summarize per-channel Macro-F1 across human and auto label sources, then quantify lexical interference, DoD, and a platform-stratified Twitter-only DoD; finally we report interventional masking, baselines, and robustness audits.

5.1 Channel Performance

Table 3 and Table 4 report Macro-F1 per channel on human-labeled and auto-labeled datasets respectively. Two patterns stand out. First, the structural / style combination BC achieves nearly identical Macro-F1 on the two human-labeled datasets (0.690 on Dreaddit-test vs. 0.690 on Twitter-gold; $\Delta \approx 0.0002$), suggesting low degradation under platform shift when relying on structure and style. Second, moving from human-labeled Twitter-gold to auto-labeled Twitter induces a sharp performance drop in C and BC (e.g., C: $0.701 \rightarrow 0.504$; BC: $0.690 \rightarrow 0.455$). From an auditing standpoint, this drop is consistent with *label-source divergence*: auto labels reward lexical proxy cues that structural and style probes are less willing to mimic.

5.2 Lexical Interference Effect

Table 5 reports the instance-bootstrap test for the lexical interference effect, defined as the Macro-F1 change when augmenting C with A. On human-labeled data, the mean drop is 0.072 (95% CI $[0.052, 0.092]$, one-sided $p < 10^{-4}$); on auto-labeled data, the change is -0.020 (95% CI $[-0.033, -0.007]$, one-sided $p = 0.9988$ for "drop > 0"). All instance-level tests are paired (same instances) and corrected via Benjamini–Hochberg FDR (27 tests, 14 rejections).

Channel	Dreaddit (H) [CI]	Tw-G (H) [CI]
A	0.634 [.598,.669]	0.661 [.643,.678]
AB	0.632 [.596,.667]	0.650 [.631,.669]
B	0.561 [.525,.598]	0.586 [.566,.605]
C	**0.739 [.706,.771]**	**0.701 [.682,.719]**
AC	0.679 [.644,.713]	0.617 [.599,.635]
BC	0.690 [.655,.724]	0.690 [.671,.709]
ABC	0.690 [.655,.724]	0.689 [.670,.708]

Table 3: Macro-F1 by channel on human-labeled datasets with 95% bootstrap CIs (Channel C in bold). Tw-G = Twitter-gold; leading-decimal abbreviation in CIs.

Channel	Tw-A (A) [CI]	Red-c (A) [CI]
A	0.522 [.510,.535]	0.611 [.587,.635]
AB	0.460 [.448,.472]	0.589 [.564,.613]
B	0.463 [.451,.475]	0.579 [.554,.604]
C	0.504 [.492,.516]	0.681 [.658,.704]
AC	0.567 [.555,.580]	0.658 [.634,.681]
BC	0.455 [.444,.467]	0.689 [.665,.712]
ABC	0.455 [.444,.467]	0.689 [.665,.712]

Table 4: Macro-F1 by channel on auto-labeled datasets with 95% bootstrap CIs. Tw-A = Twitter-auto, Red-c = Reddit-combi; leading-decimal abbreviation in CIs.

5.3 Degree of Divergence

Table 6 reports the full DoD estimates. $\mathrm{DoD}_{\mathrm{C–A}} = +0.047$ and $\mathrm{DoD}_{\mathrm{AC–A}} = -0.045$ form a symmetric pair: human annotators implicitly weight style, while auto labels reward lexical proxies. The negative $\mathrm{DoD}_{\mathrm{AC–A}}$ is the strongest quantitative signature of the lexical interference effect: lexical augmentation actively penalizes human-source performance.

A note on the permutation diagnostic. A regime-permutation diagnostic (Good, 2005) gives $p = 0.092$ (FDR-adjusted $p = 0.1656$). We treat this as a conservative robustness audit: the bootstrap evaluates stability across instances, whereas the permutation diagnostic probes whether the observed channel ranking could arise under channel-design constraints. We prioritize the bootstrap CI for effect estimation while reporting the permutation result transparently.

5.4 Platform-Stratified Twitter-Only DoD

A potential confound is that human-labeled and auto-labeled datasets differ not only in label source but also in platform composition. The standard DoD above assumes that platform effects are additive and channel-invariant within each label source. We test this directly by computing a *Twitter-only*

Source (C → AC)	Mean drop	95% CI	p (1-s.)
Human	0.0720	$[0.052, 0.092]$	$< 10^{-4}$
Auto	-0.0201	$[-0.033, -0.007]$	0.9988

Table 5: Lexical interference test (instance bootstrap). Augmenting Channel C with Channel A reduces Macro-F1 on human labels and slightly *increases* it on auto labels.

DoD	Est.	95% CI	p	d
BC–A	$+0.0374$	$[0.010, 0.065]$	0.0032	2.63
C–A	$+0.0469$	$[0.020, 0.075]$	0.0001	3.32
AC–A	-0.0453	$[-0.068, -0.022]$	<0.001	-3.89
AB–A	$+0.0360$	$[0.010, 0.062]$	0.0036	2.68

Table 6: DoD estimates (instance bootstrap, n=10,000). For positive DoD the alternative is DoD > 0; for $\mathrm{DoD}_{\text{AC–A}}$ the alternative is DoD < 0.

DoD that holds the platform fixed (Twitter-gold vs. Twitter-auto) and removes the Reddit vs. Twitter contrast:

$$\mathrm{DoD}^{\text{Tw}}_{X-A} = [\, F_1^{\text{Tw-gold}}(X) - F_1^{\text{Tw-gold}}(A) \,] - [\, F_1^{\text{Tw-auto}}(X) - F_1^{\text{Tw-auto}}(A) \,]. \quad (2)$$

We attach uncertainty to this platform-stratified check via a paired bootstrap that resamples each Twitter dataset independently with replacement (N_{boot}=2,000; Table 7). The effect remains stable and highly significant: $\mathrm{DoD}^{\text{Tw}}_{\text{C–A}} = +0.058$ (95% CI $[+0.031, +0.086]$, $p < 0.001$), $\mathrm{DoD}^{\text{Tw}}_{\text{BC–A}} = +0.096$ ($[+0.066, +0.125]$, $p < 0.001$), and $\mathrm{DoD}^{\text{Tw}}_{\text{AC–A}} = -0.089$ ($[-0.113, -0.065]$, $p < 0.001$). The signs and magnitudes match the cross-platform DoD: structural channels gain relative to A under human labels and lose under auto labels, even when both label sources come from the same platform. This does not eliminate every label-construction difference (e.g., annotation protocol, sampling strategy), but it makes a pure platform-shift explanation unlikely. A controlled within-platform dual-annotation experiment would still be required for definitive causal identification (§7).

5.5 Interventional Masking

Table 8 reports masking-suite Macro-F1 on human-labeled datasets, including 95% bootstrap CIs and paired bootstrap CIs for the difference vs. `none` (N_{boot}=2,000). On Dreaddit-test, `C/pos_only` is statistically indistinguishable from `C/none` ($\Delta = -0.006$, 95% CI $[-0.037, +0.023]$, $p = 0.72$), supporting the interpretation that Channel C's signal is not primarily carried by lexical surface form. On Twitter-gold, the same condition shows a small but reliable drop ($\Delta = -0.034$, 95% CI $[-0.052, -0.017]$, $p < 0.001$), consistent with noisier, shorter texts where POS traces carry less recoverable information after lexical deletion. Notably, `BC/pos_only` on Dreaddit *exceeds* unmasked `BC/none` ($\Delta = +0.037$, 95% CI $[+0.008, +0.066]$, $p = 0.013$): destroying content under the combined channel forces the classifier off syntactically irrelevant morpho-lexical patterns and onto genuinely structural ones. `random_pos` acts as a destructive control; on Twitter-gold it produces large significant drops for both C ($\Delta = -0.080$, $p < 0.001$) and BC ($\Delta = -0.062$, $p < 0.001$), confirming that performance does not survive arbitrary syntactic scrambling.

Contrast	DoD^{Tw} [95% CI]	p
C – A	$+0.058$ $[+0.031, +0.086]$	<0.001
BC – A	$+0.096$ $[+0.066, +0.125]$	<0.001
AC – A	-0.089 $[-0.113, -0.065]$	<0.001
AB – A	$+0.051$ $[+0.022, +0.080]$	<0.001

Table 7: Platform-stratified Twitter-only DoD with 95% bootstrap CIs and two-sided p-values (N_{boot}=2,000). DoD^{Tw} is the difference of within-source $\Delta = F_1(\text{channel}) - F_1(A)$ between Twitter-gold (human) and Twitter (auto). Signs and magnitudes match the cross-platform DoD, making a pure platform-shift explanation unlikely.

Channel / mask	Dreaddit F1 [CI]	Tw-G F1 [CI]
C / none	0.739 [.708,.772]	0.701 [.682,.719]
C / pos_only	0.734 [.700,.767]	0.666 [.647,.685]
C / function_only	0.736 [.705,.769]	0.603 [.582,.624]
C / content_only	0.724 [.691,.756]	0.652 [.632,.671]
C / random_pos	0.742 [.709,.774]	0.621 [.599,.641]
BC / none	0.690 [.654,.724]	0.690 [.670,.708]
BC / pos_only	0.728 [.694,.760]	0.594 [.574,.614]
BC / function_only	0.727 [.692,.761]	0.665 [.645,.683]
BC / content_only	0.724 [.692,.754]	0.625 [.606,.644]
BC / random_pos	0.739 [.706,.771]	0.628 [.607,.649]

Table 8: Masking-suite Macro-F1 on human-labeled datasets with 95% bootstrap CIs (cell-level, N_{boot}=2,000). Tw-G = Twitter-gold; CIs use leading-decimal abbreviation (e.g., [.708,.772] ≡ [0.708, 0.772]). Paired CIs for Δ vs. `none` are in Table 9.

5.6 Baselines as Diagnostics

Table 10 reports paired comparisons against MentalBERT (Ji et al., 2022) and 3-shot LLaMA-3-8B (Grattafiori et al., 2024; Brown et al., 2020). LLaMA-3 is evaluated only in a few-shot diag-

Comparison	Dreaddit Δ [95% CI]	p	Tw-gold Δ [95% CI]	p
C: pos_only − none	-0.006 [$-0.037, +0.023$]	0.72	-0.034 [$-0.052, -0.017$]	<0.001
C: function_only − none	-0.003 [$-0.031, +0.025$]	0.88	-0.098 [$-0.118, -0.078$]	<0.001
C: content_only − none	-0.016 [$-0.041, +0.010$]	0.20	-0.049 [$-0.063, -0.033$]	<0.001
C: random_pos − none	$+0.003$ [$-0.018, +0.025$]	0.77	-0.080 [$-0.096, -0.063$]	<0.001
BC: pos_only − none	$+0.037$ [$+0.008, +0.066$]	0.013	-0.096 [$-0.112, -0.081$]	<0.001
BC: function_only − none	$+0.037$ [$+0.002, +0.071$]	0.039	-0.025 [$-0.045, -0.007$]	0.010
BC: content_only − none	$+0.033$ [$+0.005, +0.062$]	0.020	-0.064 [$-0.080, -0.049$]	<0.001
BC: random_pos − none	$+0.049$ [$+0.023, +0.076$]	0.001	-0.062 [$-0.077, -0.047$]	<0.001

Table 9: Paired bootstrap differences for masking conditions vs. `none` (N_{boot}=2,000). On Dreaddit, Channel C is statistically unaffected by lexical destruction; on Twitter-gold, the drops are significant but small in absolute terms (95–96% retention for `pos_only`). The BC pos_only case on Dreaddit shows that combined structure-plus-style representation can *benefit* from lexical destruction.

Dataset	Model	Macro-F1
Dreaddit (H)	TSS-C	0.739
Dreaddit (H)	TSS-BC	0.690
Dreaddit (H)	MentalBERT	0.801
Dreaddit (H)	LLaMA-3 (3s)	0.613
Tw-gold (H)	TSS-C	0.701
Tw-gold (H)	TSS-BC	0.690
Tw-gold (H)	MentalBERT	0.715
Tw-gold (H)	LLaMA-3 (3s)	0.800
Tw-auto (A)	TSS-C	0.504
Tw-auto (A)	TSS-BC	0.455
Tw-auto (A)	MentalBERT	0.418
Tw-auto (A)	LLaMA-3 (3s)	0.522
Reddit-c. (A)	TSS-C	0.681
Reddit-c. (A)	TSS-BC	0.689
Reddit-c. (A)	MentalBERT	0.760
Reddit-c. (A)	LLaMA-3 (3s)	0.794

Table 10: Baseline comparison (Macro-F1). Full paired ΔF_1, bootstrap CIs, and McNemar p-values are in Appendix B.3.

nostic capacity. Its instability is informative (Appendix C.5): on Dreaddit-test recall is 0.984 with precision 0.606 (over-firing); on Twitter-auto recall collapses to 0.203 (under-firing). We interpret this as evidence that few-shot lexical priors track the availability of stress vocabulary rather than the underlying state.

We treat baselines as epistemic diagnostics rather than direct competitors: when MentalBERT outperforms TSS on human labels, it plausibly leverages legitimate contextual signal; when both strong baselines excel on auto-labeled corpora, they may be matching label-source-induced proxy noise; TSS's relative stability across label sources is the central observation. On Twitter-auto, C and BC outperform MentalBERT yet remain below LLaMA-3; on Twitter-gold MentalBERT is approached by TSS-C ($\Delta Macro - F1 = -0.0141$, paired bootstrap 95% CI $[-0.0367, +0.0082]$, $p = 0.108$; Appendix B.3), but LLaMA-3-8B (3-shot) attains a higher Macro-F1 (0.800). We therefore *do not* claim overall competitive superiority on Twitter-gold; we read the result as evidence of representation efficiency relative to one domain-pretrained transformer baseline, not as a SOTA claim.

5.7 LODO and Phenotyping

Under leave-one-domain-out evaluation on Dreaddit's five stressor domains (Appendix C.1), Channel C has the highest mean out-of-domain Macro-F1 among TSS channels (0.685) with notably low variance (std 0.015), supporting a relative cross-stressor invariance claim. MentalBERT attains a higher mean (0.818) but with greater variance (std 0.030), consistent with domain-dependent lexical sensitivity.

K-Means (Lloyd, 1982; MacQueen, 1967) with K=3 on Channel C yields highly stable clusters (bootstrap ARI 0.980, 95% CI $[0.961, 0.991]$, n_{boot}=100). Cluster–dataset association (AMI 0.281, p=0.0005) flags partial platform entanglement, an explicit leakage audit reinforcing the need for label-source-aware evaluation. Channel orthogonality is measured rather than assumed (mean $|\rho| \approx 0.595$); the weakest pair (A vs. B, $|\rho| = 0.296$) supports the intended factorization. Channel C trains in ≈ 27.5 s on a single CPU (Appendix C.4), supporting the diagnostic-density framing.

6 Discussion

TSS is best understood as a *diagnostic tool*: it quantifies when results are label-source-specific and provides a single statistically grounded scalar (DoD) to summarize that gap. Below we consolidate the interpretation, formalize when "strong baselines"

are mostly a lexical illusion, and position structural invariants as the signal that survives label-source shifts.

6.1 Why Lexical Baselines Look Strong

In auto-labeled or closed-community settings, topical stress lexicon co-occurs with structural distress markers, producing a *semantic–structural intersection (SSI)*. A model can succeed by detecting topics that correlate with distress rather than the distress state itself, consistent with shortcut learning under non-causal correlation (Geirhos et al., 2020). This reconciles the high performance of MentalBERT (Macro-F1 0.760) and LLaMA-3 (Macro-F1 0.794) on Reddit-combi with their pronounced degradation on Twitter-auto (MentalBERT Macro-F1 0.418), where SSI is weaker and intent is mixed (news, ads, meta-talk; see qualitative cases in Appendix D).

A direct signal-saturation test on Reddit-combi: augmenting C with morpho-syntax yields no detectable gain ($+0.008$ Macro-F1; $p = 0.259$; Appendix B.1), consistent with SSI-induced overlap. The fact that C still beats A on Reddit-combi by $+0.070$ ($p < 0.0001$; Appendix B.4) is best read as an SSI-saturation effect: under high SSI, all channels benefit; the diagnostic question is the *relative* channel advantage *across* label sources, which is exactly what DoD captures and confirms ($\mathrm{DoD}_{\mathrm{C–A}} = +0.047$, $p < 10^{-4}$).

6.2 Representation Efficiency, Not SOTA

On Twitter-gold, the gap between TSS-C and MentalBERT is small ($\Delta Macro - F1 = -0.0141$, 95% CI $[-0.0367, +0.0082]$, $p = 0.108$; Appendix B.3)—consistent with, but not proving, statistical equivalence relative to *this* baseline. We do *not* claim that TSS-C matches the strongest baseline overall: LLaMA-3-8B (3-shot) reaches Macro-F1 0.800. The reading is *representation efficiency*: a 154-feature, fully interpretable linear channel approaches one domain-pretrained transformer on human-labeled data with radically lower representational and computational cost.

6.3 Implicit Stress vs. Keyword-Driven False Positives

Conflict-zone examples (Appendix D) instantiate two complementary failure modes: *keyword-driven false positives* (stress meta-talk, psychoeducational content, or promotional posts classified as distress because stress words are present even when structural distress is absent), and *implicit stress* (distress expressed without explicit stress lexicon, via hedging, negation, or fragmentation). At the instance level, across 800 audited cases Channel C alone corrects 253 Channel-A errors (195 keyword-driven false positives + 58 implicit-stress false negatives) and BC corrects 241, showing that lexical shortcuts fail in two directions and that structural style features recover both. TSS audits state inference, not topic detection.

6.4 Cross-Stressor Invariance and Clinical Bridge

LODO results (Appendix C.1) show Channel C varies little across held-out stressor domains (std 0.015), supporting the view that structural and stylistic markers track a cognitive–affective state that persists across heterogeneous causes; cognitive fragmentation, absolutist appraisal, and negation scope remain candidate distress signatures (Beck, 1976; Al-Mosaiwi and Johnstone, 2018). We use social-media data only as a stress-test environment for label-source bias, not as a clinical stand-in; analogous shortcut-learning risks arise in clinical NLP (Ernala et al., 2019), and TSS is compatible with privacy-constrained workflows because Channel B is mostly de-lexicalized.

7 Conclusion

We introduced TSS, a diagnostic framework for testing whether different label sources reward different linguistic evidence in computational mental health classification. Across four English datasets, the central result is a *lexical interference effect*: adding lexical surface features to style features reduces Macro-F1 on human-labeled datasets (mean drop 0.072, $p < 10^{-4}$) but not on auto-labeled datasets, with the gap quantified by $\mathrm{DoD}_{\mathrm{BC–A}} = 0.0374$ ($p = 0.0032$). Platform-stratified Twitter-only DoD (Table 7, all contrasts $p < 0.001$) and interventional masking on Dreaddit, where Channel C is statistically unaffected by full deletion of content words ($\Delta = -0.006$, 95% CI $[-0.037, +0.023]$, $p = 0.72$), suggest the effect is not reducible to ordinary platform shift; a controlled within-platform dual-annotation study remains necessary for causal identification. TSS is therefore an audit workflow that flags label-source-specific shortcut learning before generalization claims are made.

Limitations

Data and annotation. Evaluation is restricted to four English datasets and to a binary stress label. For Twitter-gold, inter-annotator agreement (Cohen's κ) was not computed under the manual-annotation protocol available to us; this limits claims about "human annotation" as a general construct. Twitter-auto inherits a disorder→stress conflation from Coppersmith et al. (2015), which is itself a form of distant-supervision noise.

Confounding between label source and platform. The four datasets cross two label sources with two platforms but do not exhaust the design. The platform-stratified Twitter-only DoD (§5.4) removes the most obvious platform confound, but a fully controlled within-platform dual-annotation experiment is required for causal identification of label-source effects.

Feature-level ablation and masking control. Per-cell bootstrap CIs for the masking suite are now reported (Tables 8–9). We do not perform a feature-level within-channel ablation: TSS is presented as a family-level diagnostic decomposition, and a full feature-attribution study is left for future work. We treat feature-family ablation as a separate attribution study rather than a camera-ready addition, because the central claim concerns channel-level label-source divergence rather than ranking individual features within Channel C. This does not affect the channel-level claim tested here, but it limits feature-level interpretability inside Channel C. The `random_pos` negative control does not collapse to chance because POS distributional cues remain after shuffling.

Word- vs. character-level lexical baseline. We did not include a word-unigram/bigram lexical baseline in this camera-ready version; Channel A is intentionally character-based to stress-test lexical surface dependence under noisy social-media spelling. Future work should compare word- and character-level lexical baselines directly to verify that the *lexical interference effect* is not specific to sub-word tokenization.

Clinical translation. Social-media data are not clinical data. We make no claim that TSS is a diagnostic medical tool; any deployment in a clinical setting requires clinician oversight, informed consent, and risk management. DoD may also require calibration for morphologically richer languages (e.g., Arabic, Chinese), where surface-form variation interacts differently with character-level lexical representations.

Few-shot baseline. LLaMA-3-8B is reported only in a 3-shot diagnostic capacity; instability is interpreted as evidence about few-shot lexical priors, not as a comment on fine-tuned LLM performance.

Ethical Considerations

This work is not a diagnostic medical tool. It focuses on auditing label bias and improving methodological rigor in CMH NLP. All evaluations use previously collected research datasets or released artifact-level data; no new user data was collected for this study. We use qualitative examples only after redacting personally identifying details. Any deployment of distress-detection systems built on social-media text requires clinician oversight, informed consent, and careful risk management, particularly given documented disparities in CMH model behavior across demographic groups (Aguirre et al., 2021).

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2932–2949.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542.

Aaron T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press.

Yoav Benjamini and Yosef Hochberg. 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.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and Prafulla Dhariwal. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10. Association for Computational Linguistics.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, and Alex Beutel. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, and Matthias Bethge. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Phillip I. Good. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edition. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. `https://doi.org/10.5281/zenodo.1212303`.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*.

Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, and Olivier Grisel. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Aryan Rastogi, Qian Liu, and Erik Cambria. 2022. Stress detection from social media articles: New dataset benchmark and analytical study. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

Benjamin Shickel, Scott Siegel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. 2020. Automatic detection and classification of cognitive distortions in mental health text. In *Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 275–280.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107. Association for Computational Linguistics.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 412–420.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Appendices

Sections A–F below provide mathematical formulations, full statistical tables, robustness audits, qualitative case analysis, extended metrics, and SHAP audits referenced in the main text.

A Mathematical Formulations

A.1 Degree of Divergence

$$\begin{aligned}\mathrm{DoD} \;=\; &\big[M_h(BC) - M_h(A)\big] \\ &-\; \big[M_a(BC) - M_a(A)\big]\end{aligned} \tag{3}$$

DoD is a difference-in-differences statistic adapted from causal econometrics; M_h and M_a denote Macro-F1 under human and auto label sources respectively. Inference is by instance-level bootstrap (n=10,000).

A.2 Adaptive Length Smoothing (Channel B)

$$\lambda \;=\; \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\,\min(L/L_{\text{ref}},\,1) \tag{4}$$

with $\lambda_{\min}$=3.0, $\lambda_{\max}$=20.0, L_{ref}=50.0. Short texts retain sparse POS signals; long texts are regularized.

A.3 Length-Normalized Log-Odds Mass (Channel B)

$$B_{\text{pos\_mass}} \;=\; \frac{\log(1+M_{\text{pos}})}{\log(1+L_{\text{eff}})}, \qquad L_{\text{eff}} = L + \lambda \tag{5}$$

The remaining Channel B features are $B_{\text{polarity}} = M_{\text{pos}}/(M_{\text{pos}}+M_{\text{neg}}+\varepsilon)$, $B_{\text{load}} = \log_e(1+M_{\text{pos}}+M_{\text{neg}})/\max(\log_e(1+\text{cnt}),\,1)$, and $B_{\text{abs}} = |\{t : t \in \mathcal{A}\}|/L_{\text{eff}}$, where $\mathcal{A}$ is a closed absolutist-word list (Al-Mosaiwi and Johnstone, 2018). Of the six Channel B features, B_{abs} is the only one that accesses lexical tokens.

A.4 Structural Negation Balance (Channel C)

$$\text{NegBalance} \;=\; \frac{R_{\text{Neg}\to\text{Verb}} - R_{\text{Neg}\to\text{Adj}}}{N_{\text{neg\_verb}} + N_{\text{neg\_adj}} + \varepsilon} \tag{6}$$

Operationalizes the helplessness vs. negative self-evaluation distinction motivated by Beck (1976).

A.5 Length-Robust Squashing (Channel C)

$$c_{\text{out}} \;=\; \tanh(2c) \tag{7}$$

A.6 Length-Independent Lexical Diversity

$$I \;=\; \frac{{M_1}^2}{M_2 - M_1} \tag{8}$$

where M_1 is the vocabulary size and $M_2 = \sum_i f_i^2$ is the sum of squared per-type frequencies. This formulation follows the convention used in our pipeline (`features.py`, `_calc_yule_i`); values are capped at 10^3 to prevent extreme outliers from very short texts. Yule's I is preferred over TTR because TTR decays mathematically with text length, biasing cross-platform comparison from long Reddit posts to short tweets.

B Detailed Statistics

B.1 SSI Saturation Test (Reddit-combi)

Setting	ΔF1 (BC−C)	p
Reddit-combi (Auto)	+0.0080	0.259

Table 11: SSI signal-saturation equivalence test (paired instance bootstrap).

B.2 Representation Efficiency Note

The single-row paired-bootstrap result for TSS-C vs. MentalBERT on Twitter-gold ($\Delta Macro-F1 = -0.0141$, 95% CI $[-0.037, +0.008]$, $p = 0.108$) appears in the top row of Table 12. The CI does not exclude zero, but LLaMA-3 (3-shot)

reaches Macro-F1 0.800 on the same dataset, so this is *not* a SOTA claim.

B.3 Paired Baseline Comparisons

Pair (dataset)	ΔF1	95% CI
C / MB (Tw-G)	-0.014	$[-0.037, +0.008]$
C / LLaMA (Tw-G)	-0.099	$[-0.121, -0.078]$
C / MB (Tw-A)	$+0.086$	$[+0.073, +0.099]$
C / LLaMA (Tw-A)	-0.018	$[-0.032, -0.004]$
BC / MB (Tw-G)	-0.025	$[-0.045, -0.004]$
BC / LLaMA (Tw-G)	-0.110	$[-0.132, -0.088]$
BC / MB (Tw-A)	$+0.037$	$[+0.027, +0.048]$
BC / LLaMA (Tw-A)	-0.067	$[-0.080, -0.054]$

Table 12: Paired baseline comparisons (paired bootstrap CI). Pairs are TSS channel vs. baseline. MB = MentalBERT; LLaMA = LLaMA-3-8B (3-shot); Tw-G = Twitter-gold; Tw-A = Twitter-auto.

Pair	p_{McN}	d_z
TSS-C vs. MB (Tw-gold)	4.49×10^{-5}	-1.47
TSS-C vs. LLaMA-3 (Tw-gold)	2.55×10^{-15}	-11.40
TSS-C vs. MB (Tw-auto)	4.49×10^{-13}	$+14.77$
TSS-C vs. LLaMA-3 (Tw-auto)	5.15×10^{-4}	-2.83
TSS-BC vs. MB (Tw-gold)	1.81×10^{-4}	-2.54
TSS-BC vs. LLaMA-3 (Tw-gold)	5.11×10^{-13}	-12.47
TSS-BC vs. MB (Tw-auto)	4.32×10^{-3}	$+6.56$
TSS-BC vs. LLaMA-3 (Tw-auto)	$< 10^{-15}$	-10.64

Table 13: McNemar p and paired effect size d_z. MB = MentalBERT.

B.4 Paired Channel Contrasts

Dataset	Cmp	ΔF1	95% CI	p_{perm}
Dread (H)	C–A	$+0.106$	$[+0.07, +0.15]$	0.0001
Dread (H)	BC–A	$+0.057$	$[+0.02, +0.10]$	0.0032
Tw-gold (H)	C–A	$+0.039$	$[+0.02, +0.06]$	0.0005
Tw-gold (H)	BC–A	$+0.029$	$[+0.00, +0.05]$	0.0131
Tw-auto (A)	C–A	-0.019	$[-0.03, -0.00]$	0.9877
Tw-auto (A)	BC–A	-0.067	$[-0.08, -0.05]$	1.0000
Red-c. (A)	C–A	$+0.070$	$[+0.04, +0.10]$	0.0001
Red-c. (A)	BC–A	$+0.078$	$[+0.05, +0.11]$	0.0001

Table 14: Paired channel contrasts ($\Delta Macro - F1$) by dataset, with permutation p-values.

Channel/Model	Mean	Std	Min	Max
TSS-A	0.662	0.042	0.628	0.723
TSS-B	0.570	0.045	0.531	0.637
TSS-BC	0.679	0.032	0.653	0.734
TSS-C	0.685	0.015	0.666	0.700
MentalBERT	0.818	0.030	0.772	0.851

Table 15: Leave-one-domain-out cross-stressor robustness on Dreaddit (out-of-domain Macro-F1, mean over five domains).

Stressor	TSS-A	TSS-C	TSS-BC	MB
abuse	0.639	0.672	0.653	0.772
anxiety	0.723	0.700	0.734	0.828
financial	0.632	0.690	0.683	0.833
ptsd	0.686	0.695	0.669	0.851
social	0.628	0.666	0.659	0.806
Mean	0.662	0.685	0.679	0.818
Std	0.042	0.015	0.032	0.030

Table 16: Per-stressor LODO breakdown (MentalBERT = MB).

C Robustness Audits

C.1 LODO

C.2 Phenotyping

C.3 Channel Orthogonality

C.4 Efficiency

C.5 LLaMA-3 Few-Shot Instability

D Qualitative Conflict Zone

We extracted 800 instances (200 per dataset) for instance-level qualitative analysis of cases where the lexical baseline (Channel A) and the structural probes (Channels B/C) disagree. The full workbook (`error_analysis_qualitative.xlsx`) is released with the code.

Workbook statistics. The `Shift_Cases_AvsB` sheet documents 682 instances where Channel B corrects Channel A's errors: 415 (60.9%) are A-false-positives (keyword-driven false positives) and 267 (39.1%) are A-false-negatives (implicit stress). The complementary `Reverse_Shift` sheet documents 667 instances where A corrects B; restricting to human-labeled data yields 319 A-corrections (Twitter-gold: 200; Dreaddit-test: 119), comprising 177 B-false-positives and 142 B-false-negatives. The `All_Channels_Compare` sheet shows that Channel C alone corrects 253 of A's errors

Statistic	Value	95% CI / p
Bootstrap ARI	0.980	$[0.961, 0.991]$ (n_{boot}=100)
Cluster–dataset AMI	0.281	p=0.0005 (perm 2000)
Cluster–label AMI	0.084	p=0.0005 (perm 2000)

Table 17: K-Means (K=3) on Channel C (N=12,924, d=154). DoD analysis uses N=12,906; the 18-instance gap reflects zero-length POS sequences excluded from DoD but retained in clustering via Channel C features.

Summary	Value
Mean $\lvert\rho\rvert$ across pairs	0.595
Orthogonality ratio	0.095
A vs. B (key pair)	0.296
Effective N	271,026

Table 18: Instance-level Spearman ρ across channel pairs.

(195 keyword-driven false positives + 58 implicit stress), and BC corrects 241.

Table 21 gives 13 representative cases illustrating the two failure modes; the full characterization is in the released workbook.

E Extended Metric Dashboard

F SHAP Lexical Concentration

Channel	d	Train (s)	Reg.	ℓ_1-ratio
A	500	17.6	L_2	—
B	6	1007.9	ElasticNet	0.95
C	154	27.5	L_2	—
BC	160	335.9	ElasticNet	0.95

Table 19: Efficiency profile. Channel B's training time is an artifact of an exhaustive ElasticNetCV grid; once fitted, inference is near-instantaneous.

Dataset	F1	Prec.	Rec.	Pattern
Dreaddit-test (H)	0.613	0.606	0.984	Over-firing
Twitter-auto (A)	0.522	0.846	0.203	Under-firing
Twitter-gold (H)	0.800	0.780	0.680	Balanced

Table 20: LLaMA-3-8B (3-shot) precision–recall instability across label sources.

Case type	Sample (truncated)	GT	A	TSS
Keyword-driven FP (advert)	"call us at +971 4 324 3244 to consult … #MentalHealth #Stress #DryEye"	0	1	0
Implicit stress (suicidality)	"I am on the edge of committing… a few hours remaining…"	1	0	1
Boundary collapse (HR narrative)	"I man the front desk… HR Customer Service Representative…"	0	1	0
Keyword-driven FP (resilience)	"Resilience is not a trait… #mentalhealth #mentalstrength"	0	1	0
Keyword-driven FP (counseling ad)	"Sometimes love is complicated. Let's talk it out. #counseling #mentalhealth"	0	1	0
Keyword-driven FP (product ad)	"New Product Alert! Silicone Suction Snapper. Use code SOCIAL10 #Stress"	0	1	0
Keyword-driven FP (achievement)	"Squatting 315 lbs for 6 reps… Tonight I just did 305!!!! I'm so excited!!!"	0	1	0
Keyword-driven FP (policy)	"When moving into their tiny house, they would be given a state I.D.…"	0	1	0
Implicit stress (medical trauma)	"…they needed to strap me down… I completely dissociated."	1	0	1
Implicit stress (relational)	"The one person you thought you could talk to blows you off. tired alone hurt"	1	0	1
Implicit stress (masked depression)	"feel sad all the time then happy for a moment… faking it for days"	1	0	1
Implicit stress (burnout)	"stress at work… 60 hours a week (even when I had Covid)…"	1	0	1
Implicit stress (financial crisis)	"…paid 1,000 dollars for hospital bills… now we have nothing left."	1	0	1

Table 21: Representative qualitative conflict-zone cases. GT = ground truth, A = Channel A, TSS = Channel B/C combined verdict. The first eight cases are keyword-driven false positives where A fires on topical markers; the last five are implicit-stress cases where A misses distress expressed without explicit stress vocabulary.

Data	Ch.	F1	PR-AUC	MCC	B.Acc	Prev
Dread (H)	A	0.634	0.809	0.324	0.646	0.516
Dread (H)	C	0.739	0.823	0.495	0.740	0.516
Dread (H)	BC	0.690	0.819	0.421	0.697	0.516
Tw-G (H)	A	0.661	0.540	0.342	0.680	0.338
Tw-G (H)	C	0.701	0.637	0.422	0.689	0.338
Tw-G (H)	BC	0.690	0.708	0.438	0.677	0.338
Tw-A (A)	A	0.522	0.510	0.065	0.531	0.478
Tw-A (A)	C	0.504	0.656	0.209	0.568	0.478
Tw-A (A)	BC	0.455	0.658	0.167	0.545	0.478
Red-c (A)	A	0.611	0.951	0.223	0.608	0.880
Red-c (A)	C	0.681	0.964	0.371	0.710	0.880
Red-c (A)	BC	0.689	0.968	0.380	0.702	0.880

Table 22: Extended metric dashboard. Tw-G = Twitter-gold, Tw-A = Twitter-auto. Macro-F1 is the headline metric; PR-AUC is the imbalance-sensitive complement; MCC and balanced accuracy are further robustness checks under skewed prevalence.

Mode	n	Mean conc.	95% CI
lexical_wins	40	0.333	$[0.308, 0.355]$

Table 23: Top-10 SHAP concentration on Channel A (Lundberg and Lee, 2017) when lexical predictions beat structural ones. The top 10 features (2% of the 500-feature space) capture 33.3% of total absolute SHAP weight (16× the uniform-contribution expectation), consistent with strong shortcut reliance in SSI-rewarded regimes.