Drift-Aware Stability of Longitudinal NLP Measurements

Anonymous ACL submission

Abstract

Social NLP systems are often used not only to classify texts, but also to measure aggregate social phenomena such as toxicity, misinformation, sentiment, stance, and political framing. In longitudinal settings, researchers may interpret changes in average model scores as evidence of social change. However, text distributions shift over time as topics, events, entities, communities, and linguistic conventions change. We formalize longitudinal NLP measurement as estimating $\mu_t(f) = \mathbb{E}_{X \sim P_t}[f(X)]$, where P_t is the text distribution at time t and f is a measurement model. We show that empirical changes in $\widehat{\mu}_t(f)$ are controlled by distributional drift and finite-sample uncertainty, yielding a scale against which raw trends can be interpreted. This motivates a Trend Stability Ratio comparing observed measurement changes with estimated drift and sampling noise. Across two time-indexed social-media settings, TSR confirms a dominant topic shift in one dataset while substantially re-ranking candidate trend changes in another. These results suggest that longitudinal NLP measurements should be interpreted alongside drift-aware stability diagnostics, rather than from raw changes in aggregate model scores alone.

1 Introduction

Social NLP systems are increasingly used as measurement instruments. Classifiers and language models estimate aggregate quantities such as the prevalence of toxic speech, misinformation, sentiment, stance, ideology, and political framing. In these settings, the object of interest is often not the label assigned to a single text, but a social quantity inferred from many model outputs: whether toxicity increased after an event, whether one community expresses more support for a policy than another, or whether a frame becomes more common over time.

This measurement use of NLP creates a gap between standard evaluation and measurement reliability. Predictive metrics such as accuracy, F1, and AUROC evaluate individual-level classification, but they do not fully determine whether aggregate conclusions are reliable. A model can have acceptable held-out performance while misestimating prevalence, reversing a trend, or distorting comparisons across groups or time periods. This concern is especially acute in longitudinal settings because language is not stationary. Topics, named entities, events, platform composition, and discourse norms change over time, and these shifts can change model scores even when the intended social construct has not changed in the same way.

We study the question of when an observed trend in an NLP measurement isstable enough to interpret. Suppose a toxicity classifier's average score increases from one month to the next. The increase may reflect a genuine rise in toxic speech, but it may also reflect a shift in topics or entities that the model associates with toxicity. A raw trend plot does not distinguish these possibilities. We therefore argue that temporal NLP measurements should be interpreted relative to the distributional drift and sampling uncertainty of the text stream.

Prior work establishes the pieces of this problem. Automated content analysis treats classifiers as tools for estimating aggregate social quantities rather than only document labels (Hopkins and King, 2010; Grimmer and Stewart, 2013; Barberá et al., 2021). Prevalence estimation and quantification study how class frequencies can be estimated from classifier outputs (Forman, 2008; González et al., 2017; Wu and Resnick, 2024), while related label-shift work studies estimation and correction under changing class priors (Lipton et al., 2018). Work on misclassification bias shows that classifier-generated variables can distort downstream social-science analyses even when predictive accuracy is high (TeBlunthuis et al., 2024). Separately, tem-

poral generalization and dataset-shift work show that NLP systems can degrade as language, topics, and entities change over time (Lazaridou et al., 2021; Ushio and Camacho-Collados, 2024; Rabanser et al., 2019). We connect these strands by focusing on the stability of temporal measurement trends. Our question is that when an aggregate model score changes over time, is that change large relative to the drift and noise of the measurement process?

Our contributions are:

1. We formalize longitudinal NLP measurement as estimating time-indexed aggregate model scores $\mu_t(f) = \mathbb{E}_{X\sim P_t}[f(X)]$ over shifting text distributions.

2. We derive simple drift-aware stability bounds showing that observed measurement changes should be interpreted relative to distributional movement and sampling uncertainty.

3. We propose an empirical Trend Stability Ratio for comparing raw temporal changes with estimated drift and finite-sample noise.

4. We demonstrate on two time-indexed social-media settings that drift-adjusted stability can either confirm dominant raw shifts or substantially re-rank candidate trend changes.

2 Longitudinal NLP Measurement

Let $P_1, \ldots, P_T$ denote a sequence of text distributions, where P_t is the distribution of texts at time t. At each time t, we observe a batch $D_t = \{X_{t,1}, \ldots, X_{t,n_t}\}$ with $X_{t,i} \sim P_t$. We do not assume stationarity: the shift from P_t to P_{t+1} may reflect changes in topics, entities, platform composition, community participation, or linguistic conventions.

Let $f : X \to [0,1]$ be a fixed measurement model. For example, $f(x)$ may be a toxicity score, sentiment score, stance probability, or misinformation probability. The model induces the population-level measurement

$$\mu_t(f) = \mathbb{E}_{X\sim P_t}[f(X)], \tag{1}$$

estimated empirically by

$$\widehat{\mu}_t(f) = \frac{1}{n_t}\sum_{i=1}^{n_t} f(X_{t,i}). \tag{2}$$

We emphasize that $\mu_t(f)$ is a model-induced measurement, not automatically the true prevalence of a social construct. If f is invalid or biased, then $\mu_t(f)$ may itself be a poor social measure. Our point is more basic: even before resolving construct validity, longitudinal studies should ask whether changes in $\widehat{\mu}_t(f)$ are stable relative to drift and sampling noise.

The adjacent empirical trend and its magnitude are

$$\widehat{\Delta}_t(f) = \widehat{\mu}_{t+1}(f) - \widehat{\mu}_t(f), \tag{3}$$

$$A_t(f) = \left|\widehat{\Delta}_t(f)\right|. \tag{4}$$

A common analysis would plot $\widehat{\mu}_1(f), \ldots, \widehat{\mu}_T(f)$ and interpret large values of $A_t(f)$ as meaningful temporal changes. We instead ask whether $A_t(f)$ is large relative to the distributional movement from P_t to P_{t+1} and the finite-sample uncertainty in estimating the two means.

3 Drift-Aware Stability

Let $d(P_t, P_{t+1})$ measure distributional drift between adjacent time periods. Distributional distances and divergences are central in domain-adaptation and dataset-shift analyses (Ben-David et al., 2010; Rabanser et al., 2019). For theory, total variation and Wasserstein distances yield clean bounds on changes in expectations. For empirical work, exact distances over text distributions are difficult to estimate, so one can use embedding-based approximations such as maximum mean discrepancy (MMD) (Gretton et al., 2012), sliced Wasserstein distance, centroid shift, or classifier-based two-sample accuracy.

Bounded-score stability. Let $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$. If $f : X \to [0,1]$, then

$$\begin{aligned}&\left|\mu_{t+1}(f) - \mu_t(f)\right| \\ &\quad = \left|\mathbb{E}_{P_{t+1}} f(X) - \mathbb{E}_{P_t} f(X)\right| \\ &\quad \le d_{\mathrm{TV}}(P_t, P_{t+1}).\end{aligned} \tag{5}$$

This follows from the variational characterization of total variation. Thus, if the text distribution barely changes, no bounded measurement score can change much in expectation. Conversely, if the text distribution changes substantially, a raw change in the average score is less directly interpretable as a stable measurement trend.

Embedding-space stability. Total variation is rarely estimated directly for text. Suppose texts are embedded by $h : X \to \mathbb{R}^d$, and let P_t^h be the distribution of $h(X)$ for $X \sim P_t$. If $f(x) =$

$g(h(x))$ and g is L-Lipschitz, then by Kantorovich–Rubinstein duality,

$$\left|\mu_{t+1}(f) - \mu_t(f)\right| \leq L W_1(P_t^h, P_{t+1}^h), \quad (6)$$

where W_1 is the 1-Wasserstein distance in embedding space. This motivates estimating distributional drift in a representation space, as if the embedded text distribution changes little and the measurement function is smooth, the aggregate measurement should not change much.

Finite-sample uncertainty. The previous inequalities concern population quantities. In data, we estimate $\mu_t(f)$ from finite samples. Since $f(X) \in [0,1]$, Hoeffding's inequality gives, with probability at least $1-\delta$,

$$\left|\widehat{\mu}_t(f) - \mu_t(f)\right| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n_t}}. \quad (7)$$

Applying this to adjacent batches and using a union bound, with probability at least $1-\delta$,

$$\left|\widehat{\mu}_{t+1}(f) - \widehat{\mu}_t(f)\right| \leq d_{\mathrm{TV}}(P_t, P_{t+1}) + \sqrt{\frac{\log \frac{4}{\delta}}{2n_t}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n_{t+1}}}. \quad (8)$$

Under the embedding-space assumption, the drift term can be replaced by $LW_1(P_t^h, P_{t+1}^h)$. The result gives the diagnostic principle that a raw temporal change should be interpreted relative to both distributional drift and finite-sample noise. A change that is small relative to this drift-and-noise scale is difficult to justify as a stable measurement change; a change that is large relative to this scale may warrant further validation, though it still does not prove construct validity.

4 Trend Stability Ratio

The stability bound suggests a practical diagnostic. Let $\widehat{D}_t = \widehat{d}(\widehat{P}_t, \widehat{P}_{t+1})$ be an empirical estimate of drift between adjacent batches, computed in a text representation space, and let

$$C_t = \sqrt{\frac{\log \frac{4}{\delta}}{2n_t}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n_{t+1}}} \quad (9)$$

be the finite-sample uncertainty term. In theory, if $\widehat{D}_t$ is calibrated in the same units as A_t, one could compare A_t directly to $\widehat{D}_t + C_t$. In practice, embedding drift proxies such as MMD or centroid shift are not in model-score units. We therefore compute a normalized comparative diagnostic:

$$\widetilde{A}_t = \frac{A_t}{\mathrm{med}(A)}, \quad \widetilde{D}_t = \frac{\widehat{D}_t}{\mathrm{med}(\widehat{D})}, \quad \widetilde{C}_t = \frac{C_t}{\mathrm{med}(C)}, \quad (10)$$

where medians are taken over adjacent intervals within the same corpus. The empirical Trend Stability Ratio is

$$\mathrm{TSR}_t = \frac{\widetilde{A}_t}{\widetilde{D}_t + \widetilde{C}_t + \epsilon}, \quad (11)$$

where $\epsilon > 0$ is a small constant for numerical stability.

TSR is a comparative diagnostic, not a calibrated hypothesis test. Higher values indicate that an observed change is large relative to the empirical drift-and-noise scale of that corpus, whereas lower values indicate that the raw trend is more easily explained by instability in the measurement process. Thus, rather than ranking temporal intervals only by A_t, we compare raw trend magnitude with drift-adjusted stability.

In our experiments, $\widehat{D}_t$ is estimated using RBF-kernel MMD in MiniLM sentence-embedding space (Reimers and Gurevych, 2019). We also compute centroid shift as a robustness check. MMD is a standard two-sample discrepancy, while centroid shift provides a simple first-moment baseline. We use the median heuristic for the RBF bandwidth.

5 Empirical Demonstration

We evaluate TSR on two time-indexed social-media classification settings. The goal is not to introduce a new predictive model, but to test whether raw temporal changes in aggregate model scores agree with drift-adjusted trend stability.

All experiments were run with fixed random seeds. The empirical study uses frozen MiniLM embeddings and logistic regression, with no neural fine-tuning; on a standard workstation, the full pipeline runs in under a few hours, with MMD computation capped at 800 examples per adjacent period pair.

Datasets. For topic measurement, we use the CardiffNLP TweetTopic dataset (Antypas et al., 2022) and focus on the `news_&_social_concern` topic. After preprocessing, the dataset contains 9,103 examples across 24 monthly periods from

Data	N	T	ρ	Overlap
Cardiff	9.1k	24	0.956	8/8
TShift	53.6k	45	0.798	0/8

Data	Max A_t	Max TSR
Cardiff	2020-02→03	2020-02→03
TShift	2018-02→03	2014-08→09

Table 1: Comparison of raw trend magnitude and drift-adjusted stability. TShift denotes TweetTempoShift sentiment. ρ is Spearman correlation between A_t and $\mathrm{TSR}_{\mathrm{MMD}}$. Overlap is the intersection size between the eight largest-A_t intervals and the eight largest-TSR intervals.

2019–09 to 2021–08, with 27.9% positive labels. For sentiment measurement, we use the `tweet_sentiment` task from TweetTempoShift (Ushio and Camacho-Collados, 2024), containing 53,604 examples across 45 monthly periods after filtering. Both settings use monthly bins. We use only public benchmark data and create no new annotations.

Setup. For each dataset, we encode texts with MiniLM sentence embeddings (Reimers and Gurevych, 2019) and train a balanced logistic regression classifier on the first 30% of periods using scikit-learn (Pedregosa et al., 2011). The measurement model $f(x)$ is the predicted probability of the target label. We compute $\widehat{\mu}_t(f)$, A_t, C_t, MMD drift, and $\mathrm{TSR}_{\mathrm{MMD}}$ for each adjacent pair of months. For MMD, we cap each adjacent period at 800 randomly sampled examples for pairwise kernel computation. All model scores, aggregate measurements, raw trend magnitudes, sampling terms, centroid drift estimates, and period counts are computed on the full period-filtered data.

Results. Table 1 summarizes the settings. In the Cardiff topic data, raw trend magnitude and TSR largely agree: the February–March 2020 interval is both the largest raw change and the highest-TSR interval, and the top-eight rankings coincide, which is expected for a dominant topical shock. In this case, TSR confirms rather than overturns the raw trend.

In contrast, TweetTempoShift sentiment shows substantial re-ranking. Although A_t and $\mathrm{TSR}_{\mathrm{MMD}}$ remain positively associated ($\rho = 0.798$), none of the top eight intervals by raw change appear among the top eight intervals by TSR. The largest raw change occurs in February–March 2018, while the highest-TSR interval occurs in August–September 2014. Thus, raw temporal volatility and drift-adjusted trend stability can select different candidate changes for substantive interpretation.

These results illustrate two complementary uses of TSR. In some settings, a raw temporal shift is large enough that the diagnostic confirms it as a dominant movement relative to drift and sampling noise. In other settings, TSR changes which intervals appear most stable, even when raw trend magnitude and TSR are positively correlated overall. The diagnostic is therefore not intended to replace raw trend plots, but to add a scale for interpreting them.

6 Discussion and Conclusion

The empirical results support the central claim that temporal NLP measurements should not be interpreted from raw trend magnitude alone. TSR is most useful as a diagnostic for prioritizing interpretation. A high raw change may coincide with high distributional drift, in which case the observed movement is less clearly separable from instability in the text stream. Conversely, a smaller raw change during a relatively stable period may be more notable after drift adjustment. This does not mean that TSR proves a trend reflects genuine social change, but it does identify when the measured change is large relative to the estimated instability of the measurement process.

The diagnostic also clarifies that standard metrics such as F1 and AUROC remain important, but they answer a different question. Longitudinal measurement focuses on whether aggregate changes in model scores are stable relative to distributional drift and sampling uncertainty, rather than individual ones. Intervals with high raw change but low drift-adjusted stability may warrant manual inspection, recalibration, or subgroup analysis before being interpreted substantively.

We studied longitudinal NLP measurement under distribution shift, formalized temporal measurement as estimating $\mu_t(f) = \mathbb{E}_{X \sim P_t}[f(X)]$, derived a drift-and-noise scale for empirical changes, and proposed TSR as a simple diagnostic. Across two social-media settings, TSR confirms a dominant topic shift in one dataset but substantially re-ranks temporal intervals in a sentiment dataset. Temporal NLP measurement should report not only what changed, but whether the measured change is large relative to how much the language distribution changed.

7 Limitations

The bounds in this paper are intentionally simple and can be loose. They provide a cautionary scale for interpreting temporal changes, not a complete theory of construct validity. Empirically, TSR depends on the choice of text representation and drift estimator. MMD and centroid shift capture distributional movement in embedding space, but may miss shifts that are semantically important for a particular construct. Because empirical drift proxies are not naturally in model-score units, we normalize A_t, $\widehat{D}_t$, and C_t within each corpus. Therefore, TSR should be interpreted as a comparative diagnostic across intervals. Our experiments use two social-media datasets and monthly bins, so results may differ with other domains, binning choices, annotation schemes, or measurement models. Finally, a high-TSR interval does not prove that the intended social construct change, but it does indicate that the model-induced measurement changed by a large amount relative to estimated drift and sampling noise.

8 Ethics Statement

This work concerns automated measurement of social constructs from text. Such measurements can affect how groups, communities, platforms, or time periods are characterized. Errors in sentiment, topic, toxicity, misinformation, stance, or related classifiers may reinforce stereotypes, overstate harms in some groups, or understate harms in others. Our goal is to make measurement instability more visible, not to encourage uncritical deployment of automated monitoring systems. We use existing public benchmark datasets and do not collect new data or infer demographic attributes. We report dataset and preprocessing details in the paper and supplementary material, following the broader motivation of data documentation practices in NLP and machine learning (Bender and Friedman, 2018; Gebru et al., 2021). We do not redistribute raw text beyond the official benchmark sources; released artifacts consist of code, configuration files, and derived interval-level metrics. The benchmark texts may include user-generated or offensive content, especially in social-media tasks, and our analysis is intended for research evaluation rather than deployment. Drift-adjusted diagnostics should supplement careful construct definition, transparent annotation practices, subgroup evaluation, privacy protections, and human expertise.

References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pablo Barberá, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José del Coz. 2017. A review on quantification learning. *ACM Computing Surveys*, 50(5):74:1–74:40.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

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

Daniel J. Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. In *American Journal of Political Science*, volume 54, pages 229–247.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. 2018. Detecting and correcting for label shift

with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. 2019. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nathan TeBlunthuis, Valerie Hase, and Chung-hong Chan. 2024. Misclassification in automated content analysis causes bias in regression. can we fix it? yes we can! *Communication Methods and Measures*.

Asahi Ushio and Jose Camacho-Collados. 2024. A systematic analysis on the temporal generalization of language models in social media. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 52–62, Bangkok, Thailand. Association for Computational Linguistics.

Siqi Wu and Paul Resnick. 2024. Calibrate-extrapolate: Rethinking prevalence estimation with black box classifiers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1634–1647.