What Makes User Reactions Useful for Fake News Detection?

Yassine Toughrai[1,2] **David Langlois**[1,2] **Kamel Smaïli**[1,2]
[1]Université de Lorraine
[2]Laboratoire Lorrain de Recherche en Informatique et ses Applications
{yassine.toughrai, david.langlois, smaili}@loria.fr

Abstract

Fake news detection is often treated as a content-centered task, yet social media posts are embedded in user reactions that express skepticism, endorsement, correction, ridicule, and affective responses. This paper asks what makes user reactions useful for fake news detection. We study performance according to different dimensions: text only, emotions in comments, and number of available comments. We study this question across Fakeddit and Weibo20 under both standard and chronological splits (in a chronological split, most recent posts are included in the testing split). Robustness of results is measured across two encoder families (BERT-base, RoBERTa-base) and two English emotion taxonomies (dair-ai/emotion[1], GoEmotions [2]). Across these configurations, comment-only representations outperform post-only baselines in most evaluation regimes, though this gain depends on dataset, split, and number of available comments. Emotion fine-tuning yields a consistently positive benefit under the chronological evaluations studied here, while its effects under standard evaluation are mixed.

These findings suggest that user responses should be preserved and considered when developing fake-news detection systems and benchmarks.

1 Introduction

Fake news detection is often framed as a content-centered classification problem: given a claim, article, or social media post, a model predicts whether it is fake or genuine. This formulation has led to a wide range of methods based on lexical features, neural text representations, retrieved evidence, propagation patterns, social context, and multimodal signals. However, social media misinformation is rarely encountered as isolated content. Posts are embedded in user reactions: people question them, endorse them, mock them, correct them, amplify them, or respond with anger, fear, surprise, and skepticism.

This paper asks: *What makes user reactions useful for fake news detection?* We argue that user comments, which we treat as reactions, provide useful veracity signal in their own right, and that their affective structure can improve chronological evaluation robustness.

Comments may help simply because they provide additional text about the event. They may also help because they reveal social interpretation, such as collective skepticism or endorsement. A third possibility is that comments are useful because they encode affective responses. Fake news often attempts to attract attention and provoke emotional reactions, and prior work has shown that emotion-aware signals can improve fake news detection. Yet it remains unclear whether the useful signal lies in explicit emotion distributions, generic comment semantics, or **learned affective representations**.

We investigate this question through a staged empirical analysis. First, we ask whether emotion distributions in user comments are statistically associated with veracity. Second, we test whether these shallow emotion-distribution features are sufficient for fake news detection. Third, we compare generic comment embeddings against embeddings extracted from emotion-fine-tuned encoders.

Across two datasets, two encoder families, two emotion taxonomies on the English side, and reconstructed chronological splits on both datasets, emotion fine-tuning of comment encoders produces a consistent and reproducible effect: significant improvement on chronological splits in 6 of 6 paired comparisons. The single configuration with the weakest base encoder and the richest emotion taxonomy (RoBERTa-base with GoEmotions) is the

[1]https://huggingface.co/datasets/dair-ai/emotion
[2]https://huggingface.co/datasets/google-research-datasets/go_emotions/

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4282–4297
November 6-10, 2026 ©2026 Association for Computational Linguistics

only one we test whose macro-F1 does not degrade between standard and chronological Fakeddit splits.

We further find that the chronological evaluation benefit depends on the availability of substantial reaction context: at low engagement thresholds, emotion fine-tuning significantly degrades performance. Geometry analyses suggest that affective fine-tuning consistently increases global fake/real separation, while local neighborhood effects depend on the dataset and neighborhood scale.

Our contributions are as follows:

- **Reactions carry veracity signal.** We show that comment-only representations outperform post-only baselines in most evaluation regimes.
- **Chronological evaluation robustness gap.** We show that most evaluated representations degrade under chronological evaluation, exposing a robustness gap that standard or random splits tend to hide.
- **A consistent benefit in Chronological evaluation.** Across two encoders (BERT-base, RoBERTa-base), two English emotion taxonomies (dair-ai/emotion, GoEmotions), and two languages (English, Chinese), emotion-fine-tuned comment encoders significantly improve macro-F1 on chronological splits (all $p \leq 0.020$, magnitudes 2 to 9 points).
- **Engagement-cohort dependence.** On Fakeddit, the effect of emotion fine-tuning depends on the engagement-defined population. It is negative or neutral in lower-engagement cohorts and positive in the high-engagement cohort, but this should not be interpreted as a universal numerical cutoff at exactly 50 comments.

2 Related Work

2.1 Fake News Detection

Fake news detection has been studied through multiple modeling paradigms, including content-based classification (Farhangian et al., 2025; Pan et al., 2018), evidence-based verification (Thorne et al., 2018; Xu et al., 2022), propagation modeling (Monti et al., 2019; Bian et al., 2020), social-context modeling (Zhang et al., 2021; Mehta et al., 2022), and multimodal detection (Nakamura et al., 2020; Zhou et al., 2020; Wang et al., 2018).

Content-based methods classify claims, posts, or articles using lexical, stylistic, semantic, or neural representations of the source text. Evidence-based approaches instead verify a claim against external sources or retrieved evidence. Other approaches model how information spreads, exploiting chronological propagation patterns, user interactions, or graph structure. Multimodal methods further incorporate images, videos, metadata, or cross-modal inconsistencies.

Despite this diversity, many fake news detection systems remain centered on the source post or article. This is natural in settings where early detection is required, where user reactions are unavailable, or where the task is formulated as claim verification. However, social media misinformation is not only a content-level phenomenon. Posts are embedded in user responses: users question them, endorse them, mock them, correct them, or express anger, fear, surprise, and skepticism. These reactions can reveal how a crowd interprets a post, and may contain information that is not directly present in the source content. Our work focuses on this reaction signal and asks what makes it useful for fake news detection.

2.2 User Reactions and Social Context

Several lines of work have incorporated social context into fake news and rumor detection (Derczynski et al., 2017). In rumor verification, reply threads and user responses are often modeled through stance, support, denial, questioning, certainty, or evidentiality. In fake news detection, social-context methods use comments, reposts, propagation graphs, user engagement, or chronological interaction patterns to complement content-based representations.

However, comments are often treated as generic auxiliary text or as nodes in a propagation structure. While this can improve detection, it does not directly explain which aspects of user reactions carry the useful signal. Reactions may help because they add topical context, because they reveal collective skepticism or endorsement, because they encode affective responses, or because they provide signals that are more stable than the lexical content of emerging fake news. Our paper is positioned around this question: rather than proposing a new comment architecture, we investigate what makes reactions useful by comparing post-only models, comment-only models, shallow emotion-distribution features, and what we call learned af-

fective comment representations.

2.3 Emotion-Aware Fake News Detection

Emotion has been repeatedly linked to misinformation, since fake news often seeks to attract attention, provoke outrage, exploit fear, or trigger surprise. The most directly related work is Zhang et al. (2021), which introduces *dual emotion* for fake news detection. Their framework distinguishes between publisher emotion, expressed in the news content, and social emotion, expressed in user comments, and models the relationship between them. They show that dual emotion is distinctive between fake and real news and that dual-emotion features can be plugged into existing detectors as an enhancement, establishing that social emotion in comments should not be ignored.

Our work differs in focus. Rather than designing publisher–social emotion gap features, we ask when reaction-based affective signals are useful, and specifically whether they help under chronological distribution shift. A related line of work proposes Intermediate-Layer Emotional Representations (ILER) for the source-content side (Toughrai et al., 2025).

While prior emotion-aware methods establish that affective signals help detection in general, our contribution is to characterize **when** and **why** they do.

3 Experimental Setup

We study post-level binary fake news detection with user reactions. Each instance consists of a source post x, a set of associated user comments $C_x = \{c_1, \ldots, c_n\}$, and a binary veracity label $y_x \in \{0, 1\}$, where the two labels correspond to fake and genuine/real news according to the dataset annotation scheme. The objective is to predict y_x from different views of the instance: the post content alone, the comments alone, or combinations of post and comment representations.

Our goal is not to propose a new fake news detection architecture. Instead, we use controlled baselines to investigate what makes user reactions useful for detection. Specifically, we ask whether the useful signal comes from: (i) the semantic content of comments, or (ii) learned affective representations obtained from emotion-fine-tuned encoders.

In post-only settings, the model predicts y_x from x alone:

$$\hat{y}_x = f(x).$$

In comment-only settings, the model predicts y_x from an aggregated representation of C_x:

$$\hat{y}_x = f(g(C_x)),$$

where $g(\cdot)$ is either an emotion-distribution aggregation function or an embedding-pooling function. In fusion settings, we use scalar late fusion over independently trained post-only and comment-only predictors:

$$p_{\mathrm{LF}}(y \mid x, C_x) = \lambda p_{\mathrm{p}}(y \mid x) + (1 - \lambda) p_{\mathrm{c}}(y \mid C_x),$$

where p_{LF} denotes the scalar late-fusion probability, p_{p} the post-only prediction, and p_{c} the comment-only prediction. λ is selected on the validation set.

3.1 Datasets

We evaluate our hypotheses on two fake-news datasets with user reactions: Fakeddit and Weibo20. Table 7 in Appendix B summarizes the post-level splits used in our experiments.

Fakeddit (Nakamura et al., 2020). Fakeddit is a large multimodal fake-news dataset collected from Reddit, containing post-level labels, textual content, metadata, images, and associated comments. We focus primarily on a comment-rich subset, which we call Fakeddit-50, consisting of posts with at least 50 matched comments. This threshold is not intended to define a representative subset of Fakeddit; rather, it provides a controlled setting for studying reaction-based models when substantial user feedback is available.

We also construct Fakeddit-50-Chronological, a chronological version of Fakeddit-50 in which models are trained on older posts and evaluated on the most recent posts.

Weibo20 (Zhang et al., 2021). Weibo20 is a Chinese fake-news dataset containing social media posts, associated user comments, and binary fake/real labels. For Weibo20, we use all available comments associated with each post. We evaluate both the standard split and a reconstructed chronological split of the same underlying collection. The standard split measures in-distribution performance, while the chronological split places earlier posts in training/validation and more recent posts in testing, approximating future-post detection.

This distinction is central to our analysis: comparing standard and chronological evaluation allows us to test whether affective reaction representations merely improve ordinary classification performance or provide more stable signals when lexical, topical, and event-specific overlap between training and test data is reduced. We do not evaluate on Weibo-21 (Tong et al., 2024), as access to the full release was not granted to us.

3.2 Emotion Annotation of Comments

To analyze the affective structure of user reactions, we assign soft emotion probabilities to each comment using emotion classifiers matched to the dataset language. For Fakeddit-50, we use English BERT and RoBERTa emotion classifiers fine-tuned on `dair-ai/emotion`; we also evaluate a GoEmotions-fine-tuned encoder as a robustness check. For Weibo20, we also use Chinese BERT and RoBERTa emotion classifiers fine-tuned on SMP-2020 [3] weibo sentiment dataset.

The English and Chinese label inventories are not strictly parallel (see comparisons in appendix A.1).

For each post x with comment set C_x, we aggregate comment-level probabilities into a post-level emotion distribution:

$$p_e(x) = \frac{1}{|C_x|} \sum_{c_i \in C_x} P(e \mid c_i).$$

This representation captures the average affective profile of the reactions to a post. In addition to mean emotion probabilities, we compute simple distributional statistics over comments, including the standard deviation and maximum probability for each emotion, as well as entropy-based features. These features are used for statistical analysis of fake/genuine emotion differences.

3.3 Reaction Representations

We compare aggregated generic comment embeddings and aggregated emotion-fine-tuned comment embeddings.

Generic comment embeddings. This uses frozen pretrained BERT encoders to embed comments without emotion-specific fine-tuning. For Fakeddit-50, we use `bert-base-uncased`; for Weibo20, we use `bert-base-chinese`.

[3] https://smp2020ewect.github.io/

Emotion-fine-tuned comment embeddings. This uses the same embedding and aggregation pipeline, but replaces the generic encoder with an encoder fine-tuned on emotion classification. These embeddings are intended to represent comments in an **affectively organized space** while preserving richer information than explicit emotion probabilities alone.

For embedding-based representations, we pool comment embeddings at the post level using mean pooling:

$$z_x = \frac{1}{|C_x|} \sum_{c_i \in C_x} h(c_i),$$

where $h(c_i)$ is the comment representation extracted from the corresponding encoder. Mean pooling is not intended to be an optimal aggregation mechanism: it can discard higher-order information about the token or comment embedding distribution (Hara et al., 2026). We use it because it is simple, parameter-free, and keeps the comparison focused on the representation space itself. More expressive aggregation mechanisms, such as attention pooling, may further improve performance, but would introduce additional modeling capacity and make it harder to isolate the effect of affective representation learning.

3.4 Baselines

We compare post-based, reaction-based, affective, and fusion baselines.

Post-only baselines. We first evaluate models that use only the source post, without comments. We include TF-IDF, BERT and RoBERTa as baselines. These baselines measure how much signal is available from the post content alone.

Comment-only baselines. We evaluate comment-only models using pooled comment embeddings. The generic comment baseline uses BERT and RoBERTa encoders without emotion-specific fine-tuning. The affective comment baseline uses the same architecture and pooling strategy, but replaces the encoder with emotion-fine-tuned BERT and RoBERTa variants. This comparison isolates the effect of emotion fine-tuning while keeping the downstream classifier and aggregation procedure fixed.

Post-reaction fusion baselines. Finally, we evaluate whether reactions complement post content by using scalar late fusion.

Dataset	Input	Feature / Encoder	Test Acc.	Test Macro-F1
Fakeddit-50				
	Post only	TF-IDF	0.8247	0.7484
	Post only	BERT	0.8424	0.7738
	Post only	RoBERTa	0.8503	0.7861
	Comments only	BERT	0.8807	0.8288
	Comments only	BERT + dair-ai/emotion	0.8915	0.8482
	Comments only	BERT + GoEmotions	0.8846	0.8320
	Comments only	RoBERTa	0.8670	0.8075
	Comments only	RoBERTa + dair-ai/emotion	0.8724	0.8183
	Comments only	RoBERTa + GoEmotions	0.8527	0.7987
	Post + comments	BERT post + BERT emotion (dair-ai)	0.9221	0.8858
Fakeddit-50-Chronological				
	Post only	TF-IDF	0.7413	0.6819
	Post only	BERT	0.8041	0.7694
	Post only	RoBERTa	0.8007	0.7614
	Comments only	BERT	0.7550	0.7298
	Comments only	BERT + dair-ai/emotion	0.7796	0.7493
	Comments only	BERT + GoEmotions	0.8031	0.7747
	Comments only	RoBERTa	0.7442	0.7146
	Comments only	RoBERTa + dair-ai/emotion	0.7801	0.7490
	Comments only	RoBERTa + GoEmotions	0.8351	0.8085
	Post + comments	BERT post + BERT emotion (dair-ai)	0.8750	0.8615
Weibo20				
	Post only	TF-IDF	0.7747	0.7697
	Post only	BERT	0.8203	0.8202
	Post only	RoBERTa	0.8218	0.8218
	Comments only	BERT	0.9074	0.9074
	Comments only	BERT + SMP2020	0.9105	0.9105
	Comments only	RoBERTa comments	0.9042	0.9042
	Comments only	RoBERTa + SMP2020	0.9019	0.9018
	Post + comments	BERT post + emotion	0.9438	0.9438
(Zhang et al., 2021)	Post + comments	BERT + Dual Emotion Features	0.915	0.915
Weibo20-Chronological				
	Post only	TF-IDF	0.7009	0.6948
	Post only	BERT	0.6735	0.6625
	Post only	RoBERTa	0.6978	0.6864
	Comments only	BERT	0.7370	0.7244
	Comments only	BERT + SMP2020	0.7967	0.7920
	Comments only	RoBERTa	0.7033	0.6844
	Comments only	RoBERTa + SMP2020	0.7449	0.7343
	Post + comments	BERT post + emotion	0.8215	0.8195
(Zhang et al., 2021)	Post + comments	BERT + Dual Emotion Features	0.734	0.734

Table 1: Post-only, comment-only, and post–reaction fusion results across encoders. Macro-F1 is the primary metric. Comment-only rows show base encoders (orange) and emotion-fine-tuned variants (teal). Paired significance tests for the emotion vs base comparisons are reported in Table 11.

3.5 Evaluation

We evaluate all models at the post level. We report accuracy and macro-F1, using macro-F1 as the primary metric because Fakeddit-50 is label-imbalanced and because fake/real performance can differ across datasets. For each test result, we compute 95% confidence intervals using nonparametric bootstrap over test posts.

For comparisons between generic and emotion-fine-tuned comment encoders, we keep the dataset split, comment aggregation method, classifier architecture, and training protocol fixed. Thus, differences between these models primarily reflect the representation induced by emotion fine-tuning.

4 Are User Reactions Useful?

We first compare post-only, comment-only, and post–reaction fusion models across our standard and chronological splits (Table 1).

Reactions are useful across all configurations. Comment-only representations achieve macro-F1 comparable to or better than post-only baselines in most settings, and post–reaction fusion improves over post-only in every configuration we evaluate. Under standard evaluation, the comment-only advantage is large: on Fakeddit-50, for BERT, post-only macro-F1 = 0.7738 against best comment-only macro-F1 = 0.8482; on Fakeddit-50, for RoBERTa, post-only macro-F1 = 0.7861

against best comment-only macro-F1 = 0.8183; on Weibo20, 0.8202 against 0.9105 (for BERT); 0.8218 against 0.9042 (for RoBERTa). Under chronological evaluation, the advantage of generic comment encoders over post-only narrows or disappears, but emotion fine-tuning restores and exceeds the comment-vs-post lead in every configuration.

Emotion-fine-tuned comments are the strongest standalone representation under chronological shift. The best comment-only result on Fakeddit-50-Chronological is RoBERTa + GoEmotions at 0.8085 macro-F1, compared to 0.7614 for post-only RoBERTa and 0.7146 for base RoBERTa comments. On Weibo20-Chronological, BERT + SMP2020 reaches 0.7920, compared to 0.6625 post-only and 0.7244 base comments. The best-performing configuration in every (dataset, split) combination we tested includes emotion-fine-tuned comment representations, either as the standalone encoder or as part of post–reaction fusion. Table 14 in appendix G shows that this benefit is statistically significant across encoders, emotion taxonomies, and languages — significant in all six chronological-split paired comparisons, with magnitudes ranging from $+0.019$ to $+0.094$ macro-F1.

5 Are Reaction Emotions Associated with Veracity?

In this section, we test whether the distributions of comment emotions on fake and real posts differ statistically.

We assign soft emotion distributions to each comment using BERT-based emotion classifiers (dair-ai/emotion for English, SMP2020 for Chinese), aggregate them into post-level mean distributions, and compare fake and real posts sets by testing whether the distribution of dominant post-level emotions differs between them using a global chi-square test of independence. We report Cramér's V as a measure of effect size. Results are reported in Appendix F.

On Fakeddit-50, the association is statistically detectable but small: the chi-square test over dominant post-level emotion is significant ($p = 0.0074$) with Cramér's $V = 0.026$, and the permutation test over post-level mean emotion vectors is also significant ($p = 0.0001$).

The association is much stronger on Weibo20. Fake posts show substantially more angry and surprised reactions and fewer happy reactions than real posts. The chi-square test is highly significant ($p < 10^{-300}$), with a large effect size (Cramér's $V = 0.475$).

The presence of an association is consistent across both datasets. Whether this association is large enough to be exploitable for classification, and under what conditions, is the question we address next.

6 Do Posts and Reactions Provide Complementary Signals?

Table 1 shows that the scalar late-fusion *"Post + comments"* baseline obtains the best performance among the evaluated baselines in all four dataset–split settings, both in accuracy and macro-F1. This suggests that post content and user reactions provide complementary predictive information.

This result should not be interpreted as showing that comments are always more useful than posts. Rather, it supports the more limited claim that user reactions are a useful signal in their own right and can be at least as informative as the source post under the evaluated conditions. The strong performance of scalar late fusion further suggests that the two views capture partly different information, so preserving and modeling reactions is important for fake-news detection.

7 Does just any supervised fine-tuning help?

Before analyzing when affective representations help, we first test a simpler alternative explanation: the gains observed with emotion-fine-tuned encoders may not come from affective supervision, but from supervised fine-tuning in general. If this were the case, then another supervised semantic task should produce similar improvements over the generic encoder.

We therefore introduce a non-affective supervised control based on natural language inference (NLI). For Fakeddit-50, we use a BERT encoder fine-tuned on MNLI; for Weibo20, we use a Chinese BERT encoder fine-tuned on OCNLI.

The NLI source-task models successfully learn their respective objectives, obtaining test macro-F1 scores of 0.7580 on MNLI and 0.7249 on OCNLI. However, Table 2 shows that this supervised semantic fine-tuning does not reproduce the benefit of emotion fine-tuning. The NLI encoder yields lower macro-F1 than the generic encoder in all four dataset–split settings. More importantly, the emotion-fine-tuned encoder outperforms the NLI

Setting	Gen.	NLI	Emo.	Δ_{E-N}	Holm p
Fakeddit-50	0.8288	0.8264	0.8482	+0.0218	.031/.036
F50-Chron.	0.7298	0.7083	0.7493	+0.0410	.0003/.0003
Weibo20	0.9074	0.8760	0.9105	+0.0345	.0003/.0003
W20-Chron.	0.7244	0.7051	0.7920	+0.0869	.0003/.0003

Table 2: Macro-F1 for generic, NLI-fine-tuned, and emotion-fine-tuned BERT comment encoders. Δ_{E-N} denotes Emotion minus NLI. Holm-corrected p-values are reported as paired bootstrap / approximate-randomization. F50 = Fakeddit-50; W20 = Weibo20.

encoder in every setting, with gains ranging from $+0.0218$ to $+0.0869$ macro-F1. All four Emotion–NLI differences remain significant after Holm correction under both paired bootstrap and approximate randomization tests.

These results argue against the explanation that the improvements are merely caused by exposing the encoder to any supervised NLP task. Instead, they suggest that the relevance of the auxiliary task matters: emotion classification provides a supervision signal that is better aligned with user reactions than the NLI control considered here.

8 When Do Affective Representations Help?

Setting	F1 Δ	95% CI	p
Fakeddit, engagement-cohort analysis			
≥ 1, standard	-0.0220	$[-0.0256, -0.0184]$	$< 10^{-4}$
≥ 1, chronological	-0.0142	$[-0.0183, -0.0101]$	$< 10^{-4}$
≥ 10, standard	-0.0099	$[-0.0148, -0.0052]$	0.0002
≥ 10, chronological	-0.0054	$[-0.0132, +0.0025]$	0.19
≥ 50, standard	$+0.0195$	$[+0.0031, +0.0362]$	0.017
≥ 50, chronological	$+0.0194$	$[+0.0032, +0.0358]$	0.020
Weibo20			
standard	$+0.0031$	$[-0.0094, +0.0165]$	0.62
chronological	$+0.0676$	$[+0.0466, +0.0890]$	$< 10^{-4}$

Table 3: Paired bootstrap tests of emotion-fine-tuned vs. generic comment encoders **under BERT**. Cross-encoder and cross-taxonomy replication is reported in Table 11.

Section 4 establishes that emotion-fine-tuned comment encoders are the strongest standalone reaction representations on chronological splits. We now characterize when this benefit emerges.

Emotion fine-tuning shows an engagement-cohort effect on Fakeddit. Under BERT with dair-ai/emotion, we vary the comment inclusion threshold from ≥ 1 to ≥ 50 comments per post (Table 3). At low engagement (≥ 1 on both splits and ≥ 10 on the standard split), emotion fine-tuning significantly degrades macro-F1. At ≥ 10 under chronological evaluation, it is statistically neutral. At ≥ 50, it significantly improves performance on both splits.

These results suggest that the usefulness of emotion fine-tuning depends on the evaluated population and the amount of reaction context. In the Fakeddit cohort analysis, positive effects appear only in the high-engagement cohort.

9 Geometry Analysis

We now examine whether this benefit is reflected in the structure of the comment-representation space itself. We compute two complementary diagnostics: the cosine distance between fake and real post-level centroids in pooled comment-embedding space, which measures global class separation, and veracity-neighborhood purity under k-nearest-neighbors, which measures local class structure. These diagnostics are run on the BERT + dair-ai/emotion configuration for Fakeddit and bert-base-chinese + SMP2020 for Weibo20.

9.1 Fake/Real Centroid Separation

Table 4 reports fake/real centroid cosine distance under base and emotion-fine-tuned comment encoders across all four (dataset, split) cells.

Dataset	Encoder	Centroid Distance
Fakeddit-50		
	Base	0.000539
	Emotion	0.002434
Fakeddit-50-Chronological		
	Base	0.002873
	Emotion	0.006292
Weibo20		
	Base	0.005024
	Emotion	0.074819
Weibo20-Chronological		
	Base	0.003908
	Emotion	0.030671

Table 4: Fake/real centroid cosine distance in pooled comment-embedding space, under base and emotion-fine-tuned encoders. Base values in orange, emotion-fine-tuned values in teal.

The central observation is that emotion fine-tuning increases fake/real centroid separation in every (dataset, split) cell we test, across both languages and both evaluation regimes. The base encoder produces small centroid distances throughout: pretrained comment embeddings without auxiliary fine-tuning do not place fake and real posts in well-separated regions of the representation space.

The emotion-fine-tuned encoder, in contrast, produces representations with consistently larger class separation under all four conditions.

What is robust to this caveat is that emotion fine-tuning never leaves fake/real centroid separation unchanged or decreased in any cell of our matrix. The geometric reorganization induced by emotion fine-tuning produces class separation in every configuration we examine, regardless of dataset, language, or split.

9.2 Veracity-Neighborhood Purity

Table 5 reports the fraction of each post's k-nearest neighbors in pooled comment-embedding space that share its veracity label, averaged across posts, for k ranging from 1 to 500 on the two chronological splits.

Fakeddit-50-Chronological			Weibo20-Chronological		
k	Base	Emotion	k	Base	Emotion
1	0.7796	0.7393 ↓	1	0.6546	0.7182 ↑
3	0.7663	0.7318 ↓	3	0.6484	0.6894 ↑
5	0.7560	0.7291 ↓	5	0.6411	0.6851 ↑
10	0.7382	0.7244 ↓	10	0.6319	0.6666 ↑
20	0.7230	0.7192 ↓	20	0.6220	0.6540 ↑
50	0.7077	0.7101 ↑	50	0.6072	0.6386 ↑
75	0.7025	0.7060 ↑	75	0.6008	0.6288 ↑
100	0.6986	0.7035 ↑	100	0.5967	0.6231 ↑
200	0.6892	0.6960 ↑	200	0.5860	0.6080 ↑
500	0.6671	0.6762 ↑	500	0.5652	0.5861 ↑

Table 5: k-nearest-neighbors veracity-neighborhood purity under chronological evaluation. Base values in orange, emotion-fine-tuned values in teal.

The neighborhood-purity results are scale-dependent. At neighborhood scales ($k \geq 50$), emotion fine-tuning increases veracity-neighborhood purity in every entry of the table, on both chronological splits. On Fakeddit-50-Chronological, however, it decreases purity for small neighborhoods ($k \in \{1, 3, 5, 10, 20\}$) and improves purity only from $k = 50$ onward. Thus, the local-geometry evidence from Table 4 should not be interpreted as uniformly positive. Instead, emotion fine-tuning consistently increases global fake/real centroid separation, while its effect on local neighborhood structure depends on the dataset and neighborhood scale.

Conclusion

We investigated what makes user reactions useful for fake news detection. The central finding is that emotion-fine-tuned comment encoders provide a consistently positive benefit under the chronological evaluations studied here: across two encoder families, two emotion taxonomies, and two languages, this effect produces significant macro-F1 improvements on chronological splits in 6 of 6 paired comparisons. This pattern suggests that affective fine-tuning can make reaction representations more useful when evaluation moves to later posts, although the chronological splits also differ in engagement, topic, and other population properties.

Additional findings complement this central result. First, the chronological evaluation benefit requires sufficient reaction context: at low engagement thresholds, emotion fine-tuning significantly degrades performance on Fakeddit. Second, geometry analysis shows that emotion fine-tuning increases fake/real separation in reaction-representation space.

Our results argue for treating user reactions, and especially their affective structure, as conditional signals whose usefulness depends on the evaluation population and the amount of available reaction context. They should not be treated as a guaranteed general performance lift.

Our findings also have implications for benchmark construction. Random-split evaluation can give an incomplete and potentially misleading view of model robustness, because it masks whether improvements persist when evaluation moves to later posts. Chronological evaluation should therefore become a standard component of fake-news detection benchmarks, alongside conventional random splits.

Limitations

Our study has several limitations. First, our chronological evaluations are limited to reconstructed chronological settings for Fakeddit and Weibo20. Although they are closer to future-post evaluation than random splits, they do not cover the full diversity of platforms and events that might be encountered in deployment. Our Chinese evaluation is restricted to Weibo-20 because access to the more recent Weibo-21 release was not granted to us.

Second, comments are not always available at prediction time. Our goal is to understand what user reactions contribute once they exist, rather than to optimize early fake news detection. In real deployment scenarios, comments may arrive gradually, and early detection systems may need to op-

erate with few or no reactions. We partially control comment availability in Fakeddit-50 by using posts with at least 50 comments, which creates a subset biased toward viral and controversial content. The engagement-cohort analysis in Section 8 shows this is not merely a sampling artifact: reaction-based methods behave qualitatively differently at low engagement, where emotion fine-tuning actively degrades performance.

Third, our claim of chronological evaluation benefit is supported by 6 of 6 paired comparisons across encoders, taxonomies, and languages, but rests on two underlying datasets (Fakeddit, Weibo20). Extending this evaluation to additional datasets with reactions and chronological structure would strengthen the generalization claim and is an obvious direction for follow-up work.

Fourth, our results do not imply that emotions cause fake news or that affective reactions are universally predictive across platforms. We show statistical association and predictive usefulness under specific dataset conditions.

Finally, our experiments are limited to binary fake news detection. Many real-world misinformation settings involve more fine-grained labels, including satire, misleading content, manipulated media, unverifiable claims, and partially true claims. Future work should examine whether affective reaction representations remain useful for finer-grained veracity labels, multilingual settings, and earlier stages of misinformation spread.

Acknowledgments

We acknowledge the AID and ANR for their invaluable financial support, which made this research for TRADEF project endeavor possible. The guidance and insights provided were instrumental in the successful execution of our study.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Faramarz Farhangian, Leandro Augusto Ensina, George D C Cavalcanti, and Rafael M. O. Cruz. 2025. DRES: Fake news detection by dynamic representation and ensemble selection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20023–20041, Suzhou, China. Association for Computational Linguistics.

Tomomasa Hara, Hiroto Kurita, Masaaki Imaizumi, Kentaro Inui, and Sho Yokoi. 2026. Why mean pooling works: Quantifying second-order collapse in text embeddings. *Preprint*, arXiv:2604.27398.

Nikhil Mehta, Maria Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380, Dublin, Ireland. Association for Computational Linguistics.

Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake news detection on social media using geometric deep learning. *Preprint*, arXiv:1902.06673.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web – ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part I*, page 669–683, Berlin, Heidelberg. Springer-Verlag.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 1178–1186, New York, NY, USA. Association for Computing Machinery.

Yassine Toughrai, David Langlois, and Kamel Smaïli. 2025. Fake news detection via intermediate-layer emotional representations. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW

'25, page 2680–2684, New York, NY, USA. Association for Computing Machinery.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA. Association for Computing Machinery.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2501–2510, New York, NY, USA. Association for Computing Machinery.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3465–3476, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, page 354–367, Berlin, Heidelberg. Springer-Verlag.

A Additional Experimental Details

This appendix provides additional implementation details, dataset statistics, full classification results, comment-count ablations, emotion-distribution analyses, and representation-geometry diagnostics. All classification results are computed at the post level. Unless otherwise stated, confidence intervals are 95% percentile bootstrap intervals computed with 10,000 resamples over the fixed test set. These intervals quantify test-sample uncertainty only and do not incorporate training-run randomness across repeated refits.

A.1 Emotion Classifiers and Label Spaces

Available at Table 6.

Component	Details
English encoders:	`google-bert/bert-base-uncased`. `FacebookAI/roberta-base`.
Chinese encoders	`google-bert/bert-base-chinese`. `hfl/chinese-roberta-wwm-ext`.
English label inventory	
`dair-ai/emotion`	sadness, joy, love, anger, fear, surprise.
`GoEmotions`	admiration, approval, annoyance, gratitude, disapproval,amusement, curiosity, love, optimism, disappointment, joy, realization, anger, sadness, confusion, caring, excitement, surprise, disgust, desire, fear, remorse, embarrassment, nervousness, pride, relief, grief
Chinese label inventory	
`SMP-2020`	angry, fear, happy, neutral, sad, surprise.

Table 6: Emotion classifiers and label inventories.

A.2 Embedding Extraction, Pooling, Classifier Training and Scalar Late Fusion

Comment embeddings are extracted from the final hidden layer, `last_hidden_state`, using the CLS token representation. Mean token pooling is not used. The maximum sequence length is 128. Tokenization uses Hugging Face default right-side truncation. The typical extraction batch size is 64 for comment-embedding runs and 128 for emotion-probability inference. Encoders are frozen during embedding extraction and probability inference.

Post-level comment representations are obtained by mean-pooling CLS vectors across comments.

For comment-only embedding models, we use the following classifier head:

$$\begin{gathered}\mathrm{Linear}(768, 256) \\ \downarrow \\ \mathrm{ReLU} \\ \downarrow \\ \mathrm{Dropout}(0.2) \\ \downarrow \\ \mathrm{Linear}(256, 1)\end{gathered}$$

The classifier is optimized with AdamW, learning rate 5×10^{-4}, batch size 64, for 20 epochs. Checkpoints are selected by validation macro-F1. The sigmoid decision threshold is 0.5. Class weights are not used for the comment MLP. Post-only and fusion baselines use logistic regression with standardization unless otherwise noted.

For scalar late fusion, we train post-only and comment-only classifiers independently and combine their predicted probabilities using the validation-selected weight defined in Section 3. We report mean and standard deviation over five seeds.

B Dataset Statistics

Dataset	Split	# Posts	Real	Fake
Fakeddit-50				
	Train	16,292	3,987	12,305
	Val	2,036	498	1,538
	Test	2,037	499	1,538
Fakeddit-50-Chronological				
	Train	16,292	3,416	12,876
	Val	2,036	862	1,174
	Test	2,037	706	1,331
Weibo20				
	Train	3,816	1,920	1,896
	Val	1,272	640	632
	Test	1,274	641	633
Weibo20-Chronological				
	Train	3,816	1,920	1,896
	Val	1,272	640	632
	Test	1,274	641	633

Table 7: Dataset sizes, label distributions, and comment-count summaries.

C Emotion Distribution Statistics

This section reports the full post-level emotion-distribution analyses. Comment-level emotion probabilities are first aggregated into post-level vectors, and fake/real comparisons are then computed over posts.

C.1 Fakeddit-50

For Fakeddit-50, the dominant-emotion chi-square test gives $p = 0.007426$, with Cramér's $V = 0.026181$. A permutation test over post-level mean emotion vectors gives $p = 0.0001$. The association is statistically detectable but small in magnitude.

C.2 Weibo20

For Weibo20, the dominant-emotion chi-square test gives $p = 4.86 \times 10^{-308}$, with Cramér's $V = 0.4749$. A permutation test over post-level mean emotion vectors gives $p = 0.0001$. Since Weibo20 standard and Weibo20-Chronological share the same underlying post collection, the full-dataset emotion-distribution statistics are identical; the difference between the two settings lies in how examples are partitioned and evaluated.

Emotion	Fake Mean	Real Mean	Fake–Real	95% CI	Welch q	Cohen's d	Cliff's δ
anger	0.2292	0.2094	+0.0198	[+0.0172, +0.0224]	3.93e-49	0.2531	0.1353
joy	0.4627	0.4742	-0.0114	[-0.0149, -0.0080]	2.10e-10	-0.1120	-0.0420
sadness	0.2096	0.2154	-0.0059	[-0.0081, -0.0037]	2.28e-07	-0.0846	-0.0702
love	0.0256	0.0275	-0.0020	[-0.0028, -0.0012]	1.26e-06	-0.0858	-0.0350
surprise	0.0178	0.0191	-0.0013	[-0.0019, -0.0007]	5.81e-05	-0.0695	-0.0285
fear	0.0551	0.0543	+0.0008	[-0.0004, +0.0019]	0.1839	0.0231	0.0506

Table 8: Per-emotion fake–real differences for Fakeddit-50.

Emotion	Fake Mean	Real Mean	Fake–Real	95% CI	Welch q	Cohen's d	Cliff's δ
angry	0.3932	0.2779	+0.1153	[+0.1067, +0.1241]	9.63e-144	0.6583	0.4005
happy	0.1569	0.3267	-0.1697	[-0.1777, -0.1618]	0.0	-1.0513	-0.5561
surprise	0.1775	0.1091	+0.0684	[+0.0640, +0.0730]	3.36e-172	0.7344	0.4026
sad	0.0669	0.0988	-0.0319	[-0.0349, -0.0288]	1.45e-89	-0.5114	-0.3500
fear	0.0202	0.0123	+0.0078	[+0.0064, +0.0094]	6.29e-25	0.2620	0.1763
neutral	0.1846	0.1752	+0.0094	[+0.0038, +0.0152]	1.03e-03	0.0825	-0.0104

Table 9: Per-emotion fake–real differences for Weibo20.

D Emotion-Distribution Classification Baselines

Table 10 reports the strongest emotion-distribution baselines. These models use post-level summaries of comment-emotion probabilities only.

E Full Classification Results

Table 11 reports the paired bootstrap test results.

Dataset	Feature Variant	Classifier	Val Macro-F1	Test Acc	Test Macro-F1	Test Macro-F1 95% CI
Fakeddit-50						
	Mean + std + max + entropy	XGBoost	0.5408	0.7599	0.5188	[0.4957, 0.5423]
	Mean + std + max	XGBoost	0.5338	0.7580	0.5121	[0.4899, 0.5352]
	Mean + entropy	XGBoost	0.5113	0.7541	0.4957	[0.4747, 0.5175]
Fakeddit-50-Chronological						
	Mean + entropy	Small MLP	0.5893	0.7197	0.5879	[0.5641, 0.6119]
	Mean + std + max + entropy	XGBoost	0.5485	0.6907	0.5663	[0.5430, 0.5900]
	Mean + std + max	XGBoost	0.5350	0.6897	0.5584	[0.5347, 0.5817]
Weibo20						
	Mean + std + max	XGBoost	0.8050	0.7951	0.7951	[0.7723, 0.8163]
	Mean + std + max + entropy	XGBoost	0.8065	0.7951	0.7951	[0.7724, 0.8168]
	Mean + std + max	Random Forest	0.8017	0.7865	0.7864	[0.7634, 0.8085]
Weibo20-Chronological						
	Mean + std + max	XGBoost	0.7624	0.6303	0.6014	[0.5734, 0.6285]
	Mean + std + max + entropy	Small MLP	0.7092	0.6279	0.6008	[0.5736, 0.6284]
	Mean + std + max + entropy	XGBoost	0.7649	0.6279	0.5979	[0.5700, 0.6254]

Table 10: Best emotion-distribution baselines across datasets. These models use only post-level summaries of comment-emotion probabilities.

Encoder	Emotion data	Acc Δ	Acc 95% CI	Acc p	F1 Δ	F1 95% CI	F1 p
Standard splits							
Fakeddit-50							
BERT-base	dair-ai/emotion	$+0.0108$	$[-0.0005, +0.0226]$	0.072	$+0.0195$	$[+0.0031, +0.0362]$	0.017
BERT-base	GoEmotions	$+0.0039$	$[-0.0103, +0.0177]$	0.60	$+0.0033$	$[-0.0176, +0.0239]$	0.75
RoBERTa-base	dair-ai/emotion	$+0.0054$	$[-0.0093, +0.0196]$	0.47	$+0.0108$	$[-0.0108, +0.0318]$	0.31
RoBERTa-base	GoEmotions	-0.0142	$[-0.0299, +0.0020]$	0.083	-0.0088	$[-0.0314, +0.0143]$	0.46
Weibo20							
BERT-chinese	SMP2020	$+0.0031$	$[-0.0094, +0.0165]$	0.67	$+0.0031$	$[-0.0094, +0.0165]$	0.62
RoBERTa-chinese	SMP2020	-0.0024	$[-0.0173, +0.0126]$	0.80	-0.0024	$[-0.0174, +0.0126]$	0.74
Chronological splits							
Fakeddit-50-Chronological							
BERT-base	dair-ai/emotion	$+0.0245$	$[+0.0098, +0.0393]$	0.002	$+0.0194$	$[+0.0032, +0.0358]$	0.020
BERT-base	GoEmotions	$+0.0481$	$[+0.0314, +0.0648]$	$< 10^{-4}$	$+0.0448$	$[+0.0264, +0.0636]$	$< 10^{-4}$
RoBERTa-base	dair-ai/emotion	$+0.0358$	$[+0.0177, +0.0545]$	$< 10^{-4}$	$+0.0343$	$[+0.0135, +0.0553]$	$< 10^{-4}$
RoBERTa-base	GoEmotions	$+0.0908$	$[+0.0707, +0.1105]$	$< 10^{-4}$	$+0.0938$	$[+0.0710, +0.1164]$	$< 10^{-4}$
Weibo20-Chronological							
BERT-chinese	SMP2020	$+0.0597$	$[+0.0400, +0.0793]$	$< 10^{-4}$	$+0.0676$	$[+0.0466, +0.0890]$	$< 10^{-4}$
RoBERTa-chinese	SMP2020	$+0.0416$	$[+0.0220, +0.0612]$	$< 10^{-4}$	$+0.0499$	$[+0.0278, +0.0720]$	$< 10^{-4}$

Table 11: Paired bootstrap tests of emotion-fine-tuned vs. generic comment encoders across encoders, emotion-fine-tuning data, datasets, and splits. Differences are computed as emotion $-$ base over the same test posts. Green indicates significant positive effects ($p < 0.05$); Red indicates negative point estimates (none significant). Approximate randomization tests give the same significance pattern (Table 14).

F Post-level emotion-distribution association with veracity

Dataset	Global Test	p-value	Effect Size	Strongest Differences
Fakeddit-50	Dominant emotion χ^2	0.0074	$V = 0.026$	anger higher in fake
Fakeddit-50-Chronological	Dominant emotion χ^2	0.0074	$V = 0.026$	same underlying collection
Weibo20	Dominant emotion χ^2	$\leq 10^{-4}$	$V = 0.475$	fake: more angry/surprise, less happy
Weibo20-Chronological	Dominant emotion χ^2	$\leq 10^{-4}$	$V = 0.475$	same underlying collection

Table 12: Post-level emotion-distribution association with veracity.

G Paired Significance Tests

The model-wise bootstrap confidence intervals reported in the main result tables quantify uncertainty for each model separately. To directly compare generic and emotion-fine-tuned comment encoders, we additionally run paired tests over the same test posts. For each dataset, we compare the predictions of the base comment encoder and the emotion-fine-tuned comment encoder. Differences are computed as emotion minus base.

We use two paired procedures. First, we compute paired bootstrap confidence intervals with 10,000 resamples over test instances. For each resample, we compute the difference in accuracy and macro-F1 between the two models. Second, we run approximate randomization tests with 10,000 permutations, swapping paired predictions between the two systems. These tests evaluate whether the observed differences are larger than expected under the null hypothesis that the two systems are exchangeable on the same test instances.

Dataset	Acc. Δ	Acc. 95% CI	Acc. p	F1 Δ	F1 95% CI	F1 p
Fakeddit-50	+0.0108	[-0.0005, 0.0226]	0.0716	+0.0195	[0.0031, 0.0362]	0.0172
Fakeddit-50-Chronological	+0.0245	[0.0098, 0.0393]	0.0020	+0.0194	[0.0032, 0.0358]	0.0198
Weibo20	+0.0031	[-0.0094, 0.0165]	0.6652	+0.0031	[-0.0094, 0.0165]	0.6226
Weibo20-Chronological	+0.0597	[0.0400, 0.0793]	$< 10^{-4}$	+0.0676	[0.0466, 0.0890]	$< 10^{-4}$

Table 13: Paired bootstrap tests comparing emotion-fine-tuned and generic comment encoders. Differences are computed as emotion minus base over the same test posts. Macro-F1 is the primary metric.

Setting	Acc. p	Macro-F1 p
Fakeddit, engagement-cohort analysis		
≥ 1, standard	$< 10^{-4}$	$< 10^{-4}$
≥ 1, chronological	$< 10^{-4}$	$< 10^{-4}$
≥ 10, standard	$< 10^{-4}$	0.0001
≥ 10, chronological	0.11	0.18
≥ 50, standard	0.0834	0.0212
≥ 50, chronological	0.0017	0.0188
Weibo20		
standard	0.7275	0.6048
chronological	$< 10^{-4}$	$< 10^{-4}$

Table 14: Approximate randomization tests comparing emotion-fine-tuned and generic comment encoders, mirroring the structure of Table 3 (the paired bootstrap counterpart).

The paired tests show that emotion fine-tuning significantly improves macro-F1 on Fakeddit-50, Fakeddit-50-Chronological, and Weibo20-Chronological. On Fakeddit-50, the macro-F1 improvement is significant under both paired bootstrap and approximate randomization, although the accuracy gain is only borderline. On Fakeddit-50-Chronological, emotion fine-tuning significantly improves both accuracy and macro-F1. On Weibo20 standard, the difference is negligible and not significant.

Dataset	k	Base Acc	Base Macro-F1	Emotion Acc	Emotion Macro-F1
Fakeddit-50					
	10	0.8748	$0.8174_{[0.7965, 0.8376]}$	0.8601	$0.8062_{[0.7855, 0.8262]}$
	20	0.8910	$0.8513_{[0.8327, 0.8693]}$	0.8910	$0.8492_{[0.8302, 0.8672]}$
	30	0.8925	$0.8507_{[0.8314, 0.8688]}$	0.8881	$0.8440_{[0.8249, 0.8622]}$
	40	0.8856	$0.8411_{[0.8219, 0.8599]}$	0.8895	$0.8422_{[0.8226, 0.8611]}$
	50	0.8876	$0.8406_{[0.8214, 0.8595]}$	0.8925	$0.8498_{[0.8308, 0.8681]}$
Fakeddit-50-Chronological					
	10	0.7619	$0.7246_{[0.7038, 0.7452]}$	0.7329	$0.6886_{[0.6671, 0.7103]}$
	20	0.7953	$0.7679_{[0.7480, 0.7865]}$	0.7874	$0.7493_{[0.7292, 0.7690]}$
	30	0.8090	$0.7805_{[0.7608, 0.7992]}$	0.7904	$0.7472_{[0.7260, 0.7675]}$
	40	0.8135	$0.7824_{[0.7625, 0.8013]}$	0.7904	$0.7585_{[0.7382, 0.7779]}$
	50	0.7752	$0.7568_{[0.7374, 0.7760]}$	0.7855	$0.7514_{[0.7305, 0.7712]}$
Weibo20					
	10	0.8571	$0.8571_{[0.8375, 0.8765]}$	0.8524	$0.8524_{[0.8326, 0.8720]}$
	20	0.8815	$0.8814_{[0.8634, 0.8989]}$	0.8752	$0.8751_{[0.8570, 0.8932]}$
	30	0.8909	$0.8908_{[0.8732, 0.9079]}$	0.8854	$0.8853_{[0.8677, 0.9026]}$
	40	0.8917	$0.8917_{[0.8744, 0.9081]}$	0.9019	$0.9019_{[0.8852, 0.9176]}$
	50	0.8893	$0.8891_{[0.8715, 0.9062]}$	0.8925	$0.8924_{[0.8752, 0.9089]}$
Weibo20-Chronological					
	10	0.6915	$0.6735_{[0.6464, 0.6993]}$	0.6931	$0.6749_{[0.6482, 0.7008]}$
	20	0.7355	$0.7248_{[0.6994, 0.7501]}$	0.7331	$0.7198_{[0.6936, 0.7450]}$
	30	0.7402	$0.7304_{[0.7053, 0.7552]}$	0.7865	$0.7809_{[0.7576, 0.8034]}$
	40	0.7465	$0.7362_{[0.7113, 0.7610]}$	0.7708	$0.7628_{[0.7386, 0.7867]}$
	50	0.7488	$0.7377_{[0.7130, 0.7623]}$	0.7418	$0.7295_{[0.7046, 0.7546]}$

Table 15: Comment-count ablation for comment-only base and emotion-fine-tuned encoders. For each dataset, the first k comments per post are used. Subscripts report 95% bootstrap confidence intervals for macro-F1.

H Comment-Count Ablations

We repeat the comment-only experiments using the first k comments per post, with $k \in \{10, 20, 30, 40, 50\}$. The post set and train/validation/test split are fixed across k, so differences reflect only the amount of reaction context available.

Comments seem to start being insightful at around 20 to 30 comments per post (figures 1 and 2). Note that this is a truncation effect, complementary to the inclusion-threshold analysis: section 8 varies which posts enter the dataset, while this section varies how many of their comments are seen.

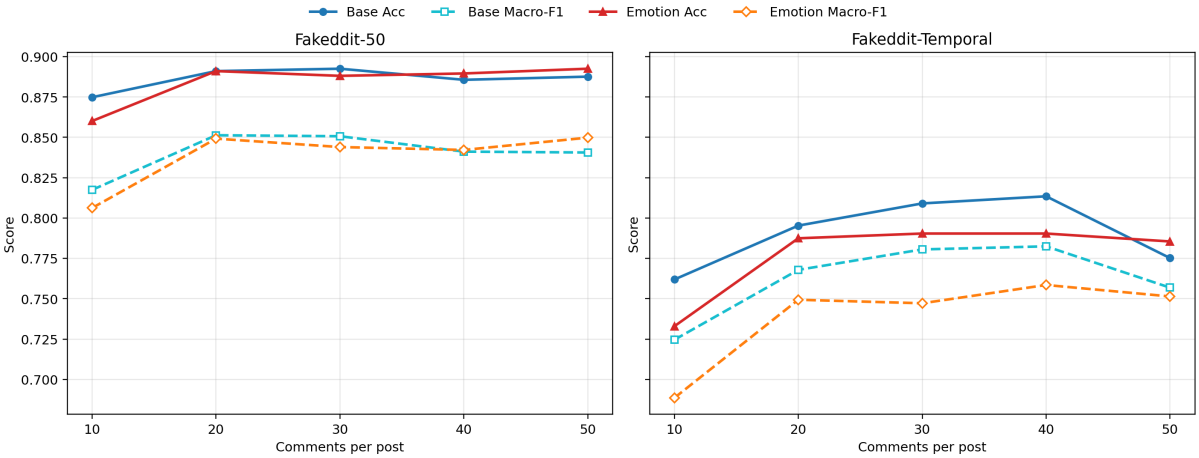


Figure 1: Comment-count ablation for base and emotion-fine-tuned comment encoders.

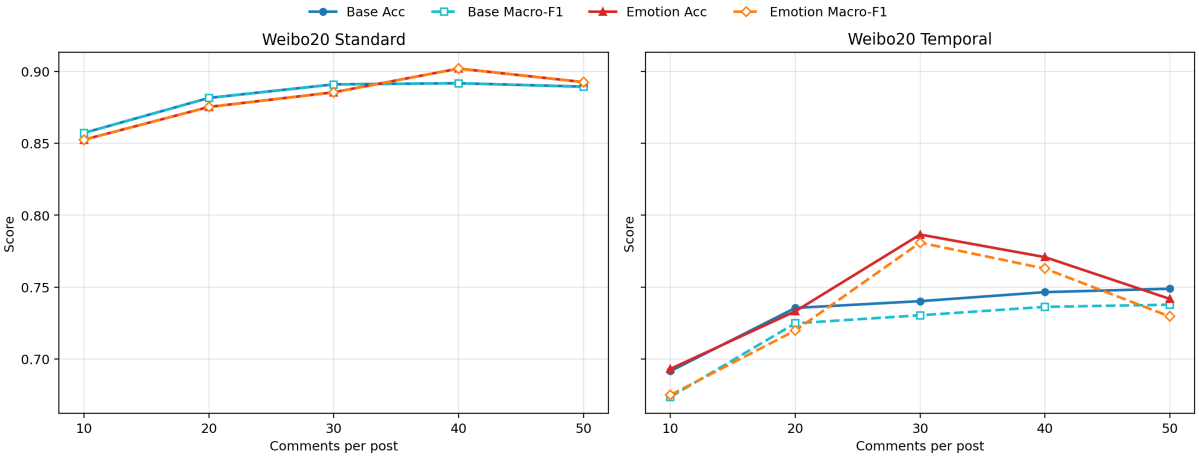


Figure 2: Comment-count ablation for base and emotion-fine-tuned comment encoders.