# Can LLMs Follow the Pulse of a Crisis? Evaluating Crisis Sentiment in Bangladesh's July Uprising

**Md. Samiul Alim**[1,*] **Mahir Shahriar Tamim**[1,*] **Tanvir Ahmed Khan**[1] **Sharjil Khan**[1]
**Rafia Ferdous Duti**[1] **Shahriyar Zaman Ridoy**[1] **Mohammad Ali Moni**[2,†]

[1]North South University, Dhaka, Bangladesh
[2]Charles Sturt University, Australia

samiul.alim01@northsouth.edu, mahir.tamim@northsouth.edu, khan.tanvir01@northsouth.edu,
sharjil.khan@northsouth.edu, rafia.duti@northsouth.edu, shahriyar.zaman01@gmail.com, mmoni@csu.edu.au




Figure 1: Graffiti images collected from publicly available online sources, depicting the July 2024 student-led uprising in Bangladesh and illustrating public expressions of resistance, solidarity, generational identity, and political sentiment captured on urban walls.(July Graffiti, 2024)

## Abstract

Crisis sentiment analysis is especially challenging for low-resource languages such as Bangla, where language, context, and public reaction shift rapidly. We introduce UNRESTSENT200K, a Bangla crisis sentiment dataset with ≈200K Facebook and YouTube comments from the July–August 2024 Bangladesh uprising. The dataset covers five event-aligned phases, from early escalation and internet blackout to regime transition and a later flood crisis. Each comment is linked to its parent post, enabling evaluation with and without discourse context. All comments are annotated through a fully human process involving 14 native Bangla-speaking annotators and senior validation, achieving substantial agreement ($\kappa = 0.73$, $\alpha = 0.71$) and 94.2% blind-audit agreement. We benchmark fine-tuned encoders, prompted LLMs, and LoRA-tuned LLMs. Results show that parent-post context consistently improves performance, while temporal shift across phases causes large performance drops. Strong LLMs perform well, but still struggle with sarcasm, implicit political references, and phase-dependent meaning. UNRESTSENT200K provides a benchmark for studying context-aware and temporally robust sentiment analysis in low-resource crisis discourse. UNRESTSENT200K is available at https://sami0055.github.io/UNRESTSENT200K/

## 1 Introduction

During a political crisis, public language can change almost overnight. A short social-media comment that appears positive at one moment may sound sarcastic, fearful, or critical after a major event. Its meaning may also depend on the news post to which it responds. Yet most sentiment benchmarks treat language as stable over time, evaluate each utterance in isolation, and focus mainly on English (Socher et al., 2013; Maas et al., 2011; Go et al., 2009). These assumptions are especially fragile in crisis settings, where sentiment systems are increasingly used to support situational awareness, journalism, and policy decisions (Tufekci, 2017; STEINERT-THRELKELD, 2017). The central problem is therefore not only a lack of data. It is also a lack of evaluation settings that show whether models can follow a rapidly changing public conversation.

The July–August 2024 Bangladesh uprising (Rana et al., 2026) offers a rare opportunity to study this problem (Figure 1). In only seven weeks, public discussion on Facebook and YouTube moved through five distinct phases: early mo-

[*] These authors contributed equally to this work.
[†] Corresponding author.



*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 1688–1708
November 6-10, 2026 ©2026 Association for Computational Linguistics

bilisation, a nationwide internet blackout, regime collapse, political transition, and an overlapping flood crisis. The language, platforms, and broad user population remained largely the same, but the social and political context changed sharply. This creates a natural stress test: can a model trained during one phase still understand sentiment in the next, and can it interpret a comment without seeing the post that prompted it?

For Bangla, these questions have been difficult to answer. Existing resources such as SentNoB, SentiGOLD, and BnSentMix (Islam et al., 2021, 2023; Alam et al., 2025) have made important progress, but they are mostly static, decontextualised, and focused on consumer or general social-media text. They do not divide a single unfolding event into meaningful time periods, link comments to their parent posts, or support controlled comparisons across phases. As a result, they cannot reveal how quickly model performance changes as an event develops. To fill this gap, we introduce UNRESTSENT200K, the largest Bangla sentiment dataset to date, with approximately 200K comments (2.8× SentiGOLD), as summarised in Table 1. Every comment is timestamped, assigned to one of five event-aligned phases, and paired with its parent post. The labels were produced entirely by 14 native Bangla-speaking undergraduate annotators who had first-hand exposure to the events. Our five-stage **Pilot Label Annotation (PiLA)** framework includes calibration, double annotation, senior adjudication, and an independent blind audit. It achieves substantial agreement (Cohen's $\kappa = 0.73$ and Krippendorff's $\alpha = 0.71$) and 94.2% agreement in the blind audit.

Our experiments tell a consistent story. We evaluate four fine-tuned encoders (BanglaBERT, mBERT, XLM-R, and BanglaElectra), 15 zero- and few-shot LLMs, and two LoRA-tuned LLMs. Giving an encoder the parent post improves performance by 7–11 percentage points, showing that context is often essential rather than optional. However, context alone does not solve the problem: training on earlier phases and testing on later ones causes drops of 19–28 F1 points. Sentiment changes occur in sharp, event-linked phases rather than along a smooth trend. GPT-4o-mini, the strongest prompted model, reaches approximately 76% few-shot accuracy but shares several failure modes with the encoders. LoRA-tuned Gemma-4B performs best overall at 78.4% accuracy, yet even this result remains below 80%. Together, these findings show that larger models and stronger prompting help, but do not remove the need for contextual grounding and temporal adaptation.

**Contributions.**

- **A large Bangla crisis benchmark.** We introduce UNRESTSENT200K, containing approximately 200K comments from the July–August 2024 Bangladesh uprising.
- **A context- and time-aware design.** Each comment is linked to its parent post and an event phase, enabling direct tests of contextual understanding and temporal robustness.
- **Careful human annotation.** PiLA provides a five-stage process with native Bangla-speaking annotators, senior adjudication, and an independent blind audit.
- **Evidence across model families.** Experiments with fine-tuned encoders, prompted LLMs, and LoRA-tuned LLMs show clear gains from parent-post context and large losses under temporal shift.

## 2 Related Work

**Sentiment benchmarks and the low-resource gap.** Sentiment analysis is driven by large static English benchmarks (Socher et al., 2013; Maas et al., 2011; McAuley et al., 2015; Go et al., 2009; Hu and Liu, 2004). In Bangla, SentNoB (Islam et al., 2021) introduced noisy multi-domain social-media comments (15K), SentiGOLD (Islam et al., 2023) scaled to 70K with fine-grained labels, and BnSentMix (Alam et al., 2025) targeted code-mixed text. Pretrained Bangla encoders (Bhattacharjee et al., 2022) and recent task-specific resources (Kundu et al., 2026; Nayeem et al., 2026; Islam and Hossen, 2025; Hasan et al., 2024; Talukder, 2025; Rashid et al., 2024) have raised the ceiling, but all are static, decontextualised, and consumer-oriented; none is anchored to a single event whose internal phases can be used to probe within-episode shift (Hossen et al., 2025; Olsen et al., 2024).

**Temporal and contextual robustness.** Prior work has studied concept drift and temporal generalisation in social-media classification (Tufekci, 2017; STEINERT-THRELKELD, 2017), but often across years, topics, or domains, where temporal change is confounded with demographic and

Table 1: Comparison of Bangla social-media datasets across sentiment, hate-speech, and offensive-language benchmarks.

| Dataset | Size | Task | Domain |
|---|---|---|---|
| SentNoB (Islam et al., 2021) | 15K | Sent. | Multi |
| SentiGOLD (Islam et al., 2023) | 70K | Sent. | Multi |
| BnSentMix (Alam et al., 2025) | 20K | Sent. | Code-mix |
| BD-SHS (Romim et al., 2022) | 50K | HS | Social |
| Bengali Tweets (Das et al., 2022) | 10K | Hate | Code-mix |
| TB-OLID (Raihan et al., 2023) | 5K | Offensive | Transliterated |
| BanTH (Haider et al., 2025) | 37K | HS | Transliterated |
| BIDWESH (Fayaz et al., 2025) | 9K | HS | Dialectal |
| BanglaMultiHate (Hasan et al., 2026) | 50K | HS | YouTube |
| **UNRESTSENT200K (Ours)** | **≈200K** | **Sent.** | **Crisis** |

topical shifts. In contrast, UNRESTSENT200K isolates temporal shift within a single seven-week crisis, while keeping language, platform, and population largely fixed. It also operationalises contextual grounding by pairing each comment with its parent post, enabling direct comment-only versus post+comment evaluation.

**Human annotation for sensitive discourse.** Politically sensitive crisis discourse requires annotators with strong linguistic and cultural grounding. Following best practices in corpus annotation (Pustejovsky, 2012; Sabou et al., 2014; Mukta et al., 2021), we use a fully human five-stage PiLA protocol with guideline piloting, double annotation, senior adjudication, and blind audit. The process involved 14 native-Bangla annotators with first-hand familiarity with the events and achieved substantial reliability: $\kappa = 0.73$, $\alpha = 0.71$, and 94.2% blind-audit agreement. Thus, UNRESTSENT200K contributes not only a benchmark, but also a replicable annotation protocol for low-resource, politically sensitive NLP.

## 3 Dataset

We construct UNRESTSENT200K, a Bangla sentiment dataset based on public Facebook and YouTube discussions about the July–August 2024 Bangladesh uprising (Rana et al., 2026). The dataset contains ≈200K annotated comments. Each comment is linked to its parent post, which allows us to evaluate sentiment classification in both comment-only and post+comment settings. Each comment is assigned one of three sentiment labels: POSITIVE, NEUTRAL, or NEGATIVE. For each instance, we retain the comment text, parent post, timestamp, platform, and event phase.

### 3.1 Data Collection

We collected public posts and comments from Facebook and YouTube. The posts were selected from verified news pages, public-interest groups, and high-engagement discussion threads related to the July–August 2024 uprising. We used the Apify platform[1], the Facebook Graph API, and the YouTube Data API v3 for data collection. The collection period covers July 5 to August 31, 2024. We divide this period into five phases: P1 Pre-Escalation (July 5–15), P2 Crisis and Blackout (July 16–August 4), P3 Post-Blackout (August 5–10), P4 Post-Revolution (August 11–20), and P5 Flood Crisis (August 21–31). These phases are used for temporal analysis and split construction. They are not used as sentiment labels. In total, we collected more than 2,000+ source posts and their comment threads. For each comment, we keep the corresponding parent post so that the dataset can support both comment-only and post+comment sentiment classification.

### 3.2 Preprocessing

We applied several preprocessing steps before annotation. First, we normalized the text using Unicode NFC normalization and removed irregular whitespace. URLs and user mentions were replaced with special tokens. Emojis were kept because they often carry sentiment in social media comments. We removed comments with fewer than three tokens, comments consisting mainly of non-Bangla text, and obvious spam entries. We also removed near-duplicate comments within the same thread using a Jaccard similarity threshold of 0.85. After preprocessing, the final dataset contains ≈200K comments.

### 3.3 Annotation Guidelines

We prepared annotation guidelines for three-way sentiment classification. Annotators were asked to assign one label to each comment.

**Positive.** A comment is labeled positive if it expresses support, approval, hope, praise, relief, or solidarity toward the event, actor, or situation discussed in the parent post.

**Neutral.** A comment is labeled neutral if it is factual, unclear, mixed, question-like, or does not express a clear positive or negative sentiment.

**Negative.** A comment is labeled negative if it expresses criticism, anger, frustration, sadness, fear,

[1] https://apify.com

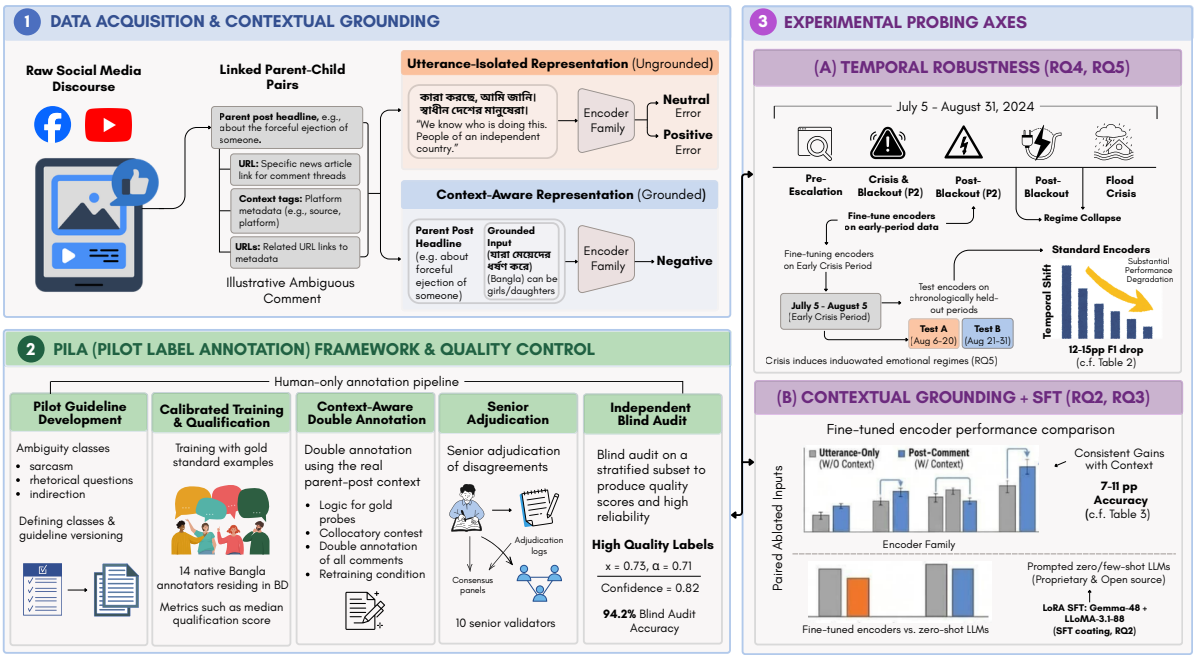



Figure 2: Overview of the UNRESTSENT200K construction and evaluation framework. **(1) Data acquisition and contextual grounding:** public Bangla crisis discourse is collected from Facebook and YouTube, comments are linked to their parent posts and metadata, and utterance-only and context-aware inputs are constructed. **(2) PiLA annotation and quality control:** a five-stage, fully human workflow covers guideline development, annotator calibration, context-aware double annotation, senior adjudication, and an independent blind audit. **(3) Experimental probing axes:** the benchmark evaluates temporal robustness across event phases and the effects of contextual grounding and supervised fine-tuning across encoder and LLM families.

ridicule, sarcasm, blame, or disappointment. Annotators were instructed to use the parent post as context, but the label was assigned only to the comment. The guidelines included examples for difficult cases, including sarcasm, rhetorical questions, religious expressions, political references, code-switched profanity, and emoji-based sentiment. This follows standard practice in corpus annotation, where guidelines are refined through pilot annotation and disagreement analysis (Pustejovsky, 2012; Sabou et al., 2014).

### 3.4 Manual Annotation

The annotation was carried out by 14 native Bangla-speaking undergraduate annotators from Bangladesh. All annotators were familiar with the July–August 2024 events and completed pilot training before the main annotation stage. A separate group of 10 senior validators supervised the process. Each comment was annotated independently by two annotators. If both annotators selected the same label, that label was accepted. If they disagreed, the comment was reviewed by a senior validator. Difficult cases were resolved through adjudication based on the annotation guidelines. The annotation process had five stages: pilot guideline development, annotator training and qualification, double annotation, senior adjudication, and blind audit. The full process took approximately eight months.

### 3.5 Annotation Agreement

We measured annotation reliability using Cohen's $\kappa$ and Krippendorff's $\alpha$. The annotation process achieved a pairwise Cohen's $\kappa$ of 0.73 and Krippendorff's $\alpha$ of 0.71, indicating substantial agreement. We also conducted a blind audit on a stratified subset of the dataset. The audit achieved 94.2% exact-label agreement with the final gold labels. These results show that the annotation process was reliable despite the context-dependent nature of the comments.

### 3.6 Dataset Split

We split the dataset into training, validation, and test sets using a 70/15/15 ratio. The split was stratified by both sentiment label and event phase so that each partition preserves the overall class and temporal distribution. The training set contains 66,382 negative, 29,249 neutral, and 43,950 positive ex-

Table 2: Stratified 70/15/15 train/validation/test split of UNRESTSENT200K by event-aligned phase and sentiment class. Splits are jointly stratified on phase and sentiment to preserve the class and temporal distribution within each partition.

| | | Train (70%) | | | Validation (15%) | | | Test (15%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Phase** | **Period** | **Neg** | **Neu** | **Pos** | **Neg** | **Neu** | **Pos** | **Neg** | **Neu** | **Pos** | **Total** |
| P1 | Pre-Escalation of Revolution (Jul 5–15) | 2,281 | 853 | 1,576 | 489 | 183 | 338 | 488 | 182 | 337 | 6,727 |
| P2 | Crisis & Blackout (Jul 16–Aug 4) | 9,180 | 4,700 | 9,698 | 1,967 | 1,007 | 2,078 | 1,968 | 1,008 | 2,078 | 33,684 |
| P3 | Post-Blackout (Aug 5–10) | 14,153 | 7,482 | 10,704 | 3,033 | 1,603 | 2,294 | 3,032 | 1,603 | 2,293 | 46,197 |
| P4 | Post-Revolution (Aug 11–20) | 23,664 | 10,532 | 14,479 | 5,071 | 2,257 | 3,103 | 5,070 | 2,257 | 3,103 | 69,536 |
| P5 | Flood Crisis (Aug 21–31) | 17,104 | 5,682 | 7,493 | 3,665 | 1,218 | 1,606 | 3,665 | 1,217 | 1,605 | 43,255 |
| **Overall** | | **66,382** | **29,249** | **43,950** | **14,225** | **6,268** | **9,419** | **14,223** | **6,267** | **9,416** | **199,399** |

amples. The validation set contains 14,225 negative, 6,268 neutral, and 9,419 positive examples. The test set contains 14,223 negative, 6,267 neutral, and 9,416 positive examples. The full phase-wise distribution is shown in Table 2. Overall, the dataset contains 94,830 negative (47.6%), 62,785 positive (31.5%), and 41,784 neutral (21.0%) comments.

## 4 Experiments and Analysis

The July–August 2024 Bangladesh uprising was not a static event (Rana et al., 2026). Public discussion changed as the crisis moved from early protest activity to internet blackout, regime collapse, political transition, and later the flood crisis. This makes UNRESTSENT200K useful for studying more than standard sentiment classification. It allows us to ask whether sentiment models can handle three problems that often appear in real crisis discourse: context dependence, temporal change, and abrupt shifts in public reaction. We organize the experiments around these problems. First, we evaluate recent LLMs to see how far prompting alone can go on Bangla crisis sentiment. Second, we test whether supervised adaptation with mid-scale LLMs improves over prompting. Third, we measure the effect of adding the parent post as context. Finally, we evaluate temporal robustness and examine whether sentiment changes gradually or through phase-level shifts. We evaluate four fine-tuned encoder models: BanglaBERT (Bhattacharjee et al., 2022), BanglaElectra, mBERT (Devlin et al., 2019), and XLM RoBERTa (Conneau et al., 2020). We also evaluate 15 LLMs under zero-shot and few-shot prompting. Across experiments, we report accuracy, macro-F1, and chance-corrected metrics such as Cohen's $\kappa$ and MCC where applicable.

### 4.1 Prompted LLMs

**Motivation.** We first ask whether recent LLMs can classify sentiment in UNRESTSENT200K without task-specific training. This is an important baseline because LLMs are often used directly on new events, especially when labeled data is limited.

**Models and prompting setup.** We evaluate open-source models from the Gemma, Qwen, and LLaMA families, along with proprietary models including GPT-4o-mini, GPT-4.1-mini, and GPT-4.1-nano. Each model is tested with zero-shot and few-shot prompts written in Bangla. In the few-shot setting, the prompt includes seven examples covering positive, negative, neutral, sarcastic, and politically implicit comments. The full prompt templates are shown in Appendix B.1.

**Input format.** Each test input is given in the same format: `[post] [SEP] [comment]`. This is the same input format used in the post+comment encoder setting, making the LLM and encoder results comparable.

**Main results.** Table 3 shows that the strongest prompted model is GPT-4o-mini in the few-shot setting. It reaches 76.0% accuracy, 75.0 macro-F1, and $\kappa = 0.55$. This is slightly above the strongest post+comment encoder, BanglaBERT, which reaches 74.72% accuracy and 0.7167 macro-F1. However, the improvement is small. This suggests that strong LLMs can benefit from the provided context, but prompting alone does not fully solve the task.

**Effect of few-shot examples.** Few-shot examples help the stronger models more than the smaller ones. For GPT-4o-mini, $\kappa$ improves from 0.47 to 0.55. For GPT-4.1-mini, it improves from 0.34 to 0.40. In contrast, several smaller open-source models show little improvement from few-shot examples. This suggests that examples are useful only when the model already has enough Bangla lan-

Table 3: Zero-shot and few-shot performance (%) of LLMs with expert prompting (Bangla-native) on UnRestSent200K. Best result per section is **bolded and underlined**.

| Model | Zero-Shot | | | | Few-Shot | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc | F1 | MCC | $\kappa$ | Acc | F1 | MCC | $\kappa$ |
| **Proprietary Models** | | | | | | | | |
| GPT-4o-mini | 71.0 | 69.0 | **50.0** | **47.0** | **76.0** | **75.0** | **58.0** | **55.0** |
| gpt-4.1-mini | **73.0** | **71.0** | 36.0 | 34.0 | 73.0 | 71.0 | 45.0 | 40.0 |
| gpt-4.1-nano | 58.0 | 56.0 | 22.0 | 13.0 | 69.0 | 61.0 | 31.0 | 29.0 |
| **Open-Source Models ($\leq$3B)** | | | | | | | | |
| gemma4-2b | **61.0** | **59.0** | **40.0** | **38.0** | **64.0** | **63.0** | **45.0** | **43.0** |
| gemma3-1b | 43.0 | 37.0 | 18.0 | 15.0 | 52.0 | 47.0 | 26.0 | 22.0 |
| qwen2.5-3b | 49.0 | 44.0 | 25.0 | 20.0 | 55.0 | 52.0 | 31.0 | 28.0 |
| qwen2.5-1.5b | 46.0 | 36.0 | 20.0 | 11.0 | 49.0 | 40.0 | 23.0 | 15.0 |
| llama-3.2-3b | 43.0 | 29.0 | 15.0 | 5.0 | 45.0 | 33.0 | 17.0 | 8.0 |
| qwen2.5-0.5b | 38.0 | 22.0 | 0.2 | 0.0 | 39.0 | 23.0 | 2.0 | 0.4 |
| **Open-Source Models ($>$3B)** | | | | | | | | |
| gemma4-4b | 52.0 | 49.0 | 16.0 | 15.0 | **66.2** | **65.0** | **48.7** | **46.3** |
| gemma3-4b | 49.0 | 40.0 | 27.0 | 16.0 | 59.0 | 54.0 | 39.0 | 33.0 |
| qwen3-4b | **56.0** | **52.0** | **35.0** | **30.0** | 58.0 | 56.0 | **38.0** | **34.0** |
| qwen3-4b-instruct | 56.0 | 52.0 | 35.0 | 30.0 | 59.0 | 57.0 | 23.0 | 14.0 |
| qwen2.5-7b | 47.0 | 45.0 | 23.0 | 21.0 | 61.0 | 60.0 | 42.0 | 39.0 |
| llama-3.1-8b | 53.0 | 52.0 | **37.0** | 11.0 | 58.0 | 55.0 | 37.0 | 33.0 |
| | **LoRA SFT** | | | | | | | |
| **Fine-Tuned Models** | | | | | | | | |
| Gemma-4B | **78.4** | **77.6** | **65.3** | **63.1** | – | – | – | – |
| Llama-3.1-8B | 76.8 | 75.9 | 62.4 | 60.2 | – | – | – | – |

guage ability and event-level reasoning capacity to use them.

**Error patterns.** A manual inspection of errors shows three recurring patterns. First, models often miss sarcasm, especially when a comment appears positive on the surface but is negative in context. Second, neutral comments are often pushed toward negative labels. Third, politically charged comments become harder when their interpretation depends on the phase of the uprising. These errors are also seen in the encoder models, suggesting that they are properties of the task rather than failures of one model family.

### 4.2 Supervised Adaptation with Mid-Scale LLMs

Prompting provides a useful starting point, but UNRESTSENT200K also allows us to test whether in-domain supervised training improves performance. We therefore fine-tune two mid-scale LLMs, Gemma-4B and LLaMA-3.1-8B, using LoRA. Both models are trained with $r$=64, $\alpha$=32, and learning rate $3\times10^{-4}$. The fine-tuned models outperform all prompted models in Table 3. Gemma-4B achieves 78.4% accuracy, 77.6 macro-F1, MCC 65.3, and $\kappa$ 63.1. LLaMA-3.1-8B achieves 76.8% accuracy and 75.9 macro-F1. Gemma-4B also improves over GPT-4o-mini by 2.4 percentage points in accuracy and over grounded BanglaBERT by 3.7 points. These results show that supervised adaptation is still important for low-resource crisis sentiment analysis. The smaller Gemma-4B model also performs better than the larger LLaMA-3.1-8B model, suggesting that the choice of backbone and fine-tuning recipe matters more than parameter count alone. At the same time, the best model remains below 80% accuracy, leaving room for future work on

Table 4: **Context ablation on UNRESTSENT200K.** Fine-tuned encoder performance when the model receives only the *comment* versus the *parent post concatenated with the comment* via `[SEP]`. Corpus, splits, optimiser, schedule, and hyperparameters are held fixed; the only variable is the input. Adding the parent post yields a consistent +7–11pp accuracy gain across every architecture. Best per column **bolded and underlined**.

| | Comment-only input `[comment]` | | | | Post + Comment input `[post] [SEP] [comment]` | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Acc. (%)** | **Prec.** | **Rec.** | **F1** | **Acc. (%)** | **Prec.** | **Rec.** | **F1** |
| BanglaBERT | **64.19** | 0.6195 | 0.6158 | 0.6170 | **74.72** | 0.7245 | 0.7116 | 0.7167 |
| mBERT | 61.06 | 0.5869 | 0.5881 | 0.5874 | 69.90 | 0.6705 | 0.6790 | 0.6731 |
| BanglaElectra | 62.02 | 0.6193 | 0.6202 | 0.6159 | 68.98 | 0.6637 | 0.6705 | 0.6649 |
| XLM-RoBERTa-Base | 63.41 | 0.6287 | 0.6312 | 0.6299 | 72.96 | 0.7289 | 0.7296 | 0.7012 |

Table 5: Temporal generalization performance of supervised fine-tuned (SFT) transformer models under distribution shift on UnRestSent200K. Models are trained on comments collected between July 5 and August 5, 2024, and evaluated on two chronologically non-overlapping future test partitions: Test-A (August 6–20) and Test-B (August 21–31). Results highlight the degradation of model performance under evolving crisis discourse and temporal non-stationarity.

| **Model** | **Test** | **Acc. (%)** | **Prec.** | **Rec.** | **F1 Score** |
|---|---|---|---|---|---|
| BanglaBERT | A | 53.70 | 0.5204 | 0.5236 | 0.5217 |
| | B | **54.83** | 0.5027 | 0.5222 | 0.5064 |
| BanglaElectra | A | 40.11 | 0.4195 | 0.4275 | 0.4006 |
| | B | 41.18 | 0.4164 | 0.4343 | 0.3942 |
| mBERT | A | 40.96 | 0.4234 | 0.4380 | 0.4111 |
| | B | 41.24 | 0.4191 | 0.4354 | 0.3948 |
| XLM-RoBERTa | A | 47.35 | 0.4708 | 0.4784 | 0.4695 |
| | B | 47.28 | 0.4480 | 0.4664 | 0.4436 |

crisis-aware and context-aware sentiment models.

### 4.3 Qualitative Effect of Fine-Tuning

Table 6 illustrates the three failure patterns flagged in §4.1 sarcasm, neutral/positive collapse to negative, and politically charged surface cues and shows the same `(post, comment)` pairs being recovered after LoRA-SFT on Gemma-4B. The pattern is qualitative: SFT exposure to the gold PiLA labels lets the model pick up domain-specific pragmatic cues that the zero-shot model misses despite seeing identical input.

### 4.4 Effect of Parent-Post Context

Many comments in UNRESTSENT200K are difficult to interpret without their parent post, as shown in Table 6. This is common in social media discussions during political events. A short comment may look factual or neutral by itself, but become clearly positive or negative once the surrounding news post is known. For example, the comment *"We know who is doing this."* is hard to label in isolation. When paired with a post about a family being forced from their home, the same comment becomes a negative reaction. This shows why the parent post is not extra information; in many cases, it is part of the meaning of the comment. We test this directly using a paired ablation. In the comment-only setting, the encoder receives only the comment. In the post+comment setting, the encoder receives the parent post and comment separated by `[SEP]`. All other training settings are kept fixed. The results in Table 4 show consistent gains from context. BanglaBERT improves from 64.19% to 74.72% accuracy. XLM-RoBERTa improves by 9.55 points, mBERT by 8.84 points, and BanglaElectra by 6.96 points. The gain appears across all encoder families. This result has a direct implication for future dataset design. In crisis discourse, comment-only sentiment classification can remove information needed for the correct label. For this reason, UNRESTSENT200K supports both comment-only and post+comment evaluation, allowing future work to measure how well models use discourse context.

### 4.5 Temporal Robustness

We next evaluate whether models trained on earlier phases of the uprising generalize to later phases. This setting reflects a realistic use case: during an unfolding crisis, models may be trained on available data and then applied as the event continues to change. We train the encoder models on com-

Table 6: **Qualitative comparison of zero-shot and supervised predictions.** Both settings use Gemma-4B; LoRA SFT denotes the same backbone fine-tuned on UNRESTSENT200K ($r$=64, $\alpha$=32, learning rate $3\times10^{-4}$). ✓ and ✗ indicate agreement and disagreement with the PiLA gold label, respectively. Examples represent the three failure modes discussed in §4.1.

| # | Post–Comment Pair | Gold | Zero-shot | LoRA SFT |
|---|---|---|---|---|
| 1 | **Post** আমার বাসায় কাজ করা পিয়ন এখন ৪০০ কোটি টাকার মালিক: প্রধানমন্ত্রী<br>"The peon who worked at my house now owns 4 billion taka: Prime Minister."<br>**Comment** আহ কি গর্বের কথা! "Ah, what a matter of pride!" | **Negative** | **Positive** ✗ | **Negative** ✓ |
| | ***Sarcasm.*** The ostensibly positive expression "pride" becomes negative under the corruption-related headline. Zero-shot follows the literal wording, whereas LoRA SFT captures the intended sarcasm. | | | |
| 2 | **Post** দেশে ইন্টারনেট সেবায় ধীরগতি, ফেসবুক-মেসেঞ্জারে প্রবেশেও বিঘ্ন<br>"Internet service is slow nationwide; access to Facebook and Messenger is also disrupted."<br>**Comment** এটা হাসিনার কাজ "This is Hasina's doing." | **Negative** | **Neutral** ✗ | **Negative** ✓ |
| | ***Pragmatic negativity.*** Zero-shot treats the comment as a neutral factual attribution, whereas LoRA SFT recognizes its accusatory and negative pragmatic force. | | | |
| 3 | **Post** বন্যায় এখন পর্যন্ত ১৩ জনের মৃত্যু, ক্ষতিগ্রস্ত ৪৫ লাখ মানুষ<br>"13 dead in floods so far; 4.5 million affected."<br>**Comment** আল্লাহ আপনি সবাইকে হেফাজত করেন। "May Allah protect everyone." | **Positive** | **Negative** ✗ | **Positive** ✓ |
| | ***Surface-cue confusion.*** Disaster-related terms in the post mislead the zero-shot model, while LoRA SFT correctly identifies the prayer and expression of solidarity as positive. | | | |

ments from July 5 to August 5, 2024. We then test them on two later windows: Test-A, covering August 6–20, and Test-B, covering August 21–31. These windows do not overlap with the training period. The results in Table 5 show a large performance drop. BanglaBERT remains the best model, reaching 54.83% accuracy on Test-B, but this is far below its in-distribution post+comment accuracy of 74.72%. Other encoders also degrade under the time-based split. Overall, the models lose between 19 and 28 macro-F1 points compared with their in-distribution results. This drop is important because the language, platforms, and broad event remain the same. What changes is the stage of the crisis. After the blackout and regime collapse, public discussion shifts toward new actors, new concerns, and new emotional frames. Appendix A.6 further shows measurable lexical change across phases using Jensen–Shannon divergence. These results show why random splits are not enough for crisis sentiment analysis. A model that performs well on a random test set may still fail when the event enters a new phase. UNRESTSENT200K therefore provides a useful setting for studying temporal robustness in low-resource NLP.

### 4.6 Sentiment Change Across Phases

The temporal experiment shows that performance drops over time. We then ask whether this change is gradual or concentrated around specific moments in the uprising. To measure this, we compute Cohen's $d$ for all pairwise comparisons between the five event phases. The full results are shown in Table 7. Seven of the ten comparisons show negligible effect sizes. The strongest changes are centered around the Post-Blackout phase. The comparison between Crisis and Blackout and Post-Blackout has a medium effect size ($d = -0.556$), and the comparison between Post-Blackout and Flood Crisis also has a medium effect size ($d = 0.567$). This suggests that sentiment did not change smoothly across the full period. Instead, the largest shift occurred around the moment when internet access returned, the regime collapsed, and people began reacting to a new political situation. Sentiment then shifted again as the flood crisis became the dominant concern. For future research, this makes UNRESTSENT200K useful for studying change-point-aware sentiment models. Models for crisis monitoring may need to detect when the discourse has entered a new phase, rather than assuming that sentiment changes slowly over time.

Table 7: Pairwise sentiment-shift comparisons between the five event phases of the 2024 Bangladesh uprising using Cohen's $d$ effect size. Larger absolute values indicate stronger changes in sentiment distribution between phases, with the strongest shifts observed around the immediate post-blackout period.

| Period 1 | Period 2 | $d$ | Effect |
|---|---|---|---|
| Pre-Escalation | Crisis/Blackout | 0.138 | Negligible |
| Pre-Escalation | Post-Blackout | $-0.411$ | Small |
| Pre-Escalation | Post-Revolution | $-0.034$ | Negligible |
| Pre-Escalation | Flood Crisis | 0.144 | Negligible |
| Crisis/Blackout | Post-Blackout | $-0.556$ | Medium |
| Crisis/Blackout | Post-Revolution | $-0.174$ | Negligible |
| Crisis/Blackout | Flood Crisis | 0.003 | Negligible |
| Post-Blackout | Post-Revolution | 0.381 | Small |
| Post-Blackout | Flood Crisis | 0.567 | Medium |
| Post-Revolution | Flood Crisis | 0.178 | Negligible |

### 4.7 Findings

Our experiments yield five main findings.

**Prompting alone does not resolve crisis sentiment.** Few-shot GPT-4o-mini achieves the strongest prompted result at 76.0% accuracy, only slightly exceeding grounded BanglaBERT at 74.72%. Its remaining errors in sarcasm, implicit attribution, and context-dependent language show that strong prompting cannot fully address the task.

**Supervised adaptation produces the best performance.** LoRA-tuned Gemma-4B achieves the highest overall accuracy of 78.4%, outperforming both prompted LLMs and fine-tuned encoders.

**Parent-post context is essential.** Including the parent post improves all encoders by approximately 7–11 accuracy points. This consistent gain shows that the post is often part of the sentiment signal rather than optional background information.

**Temporal shift substantially reduces performance.** Models trained on earlier phases lose approximately 19–28 macro-F1 points when evaluated on later phases. Thus, random splits can overestimate real-world reliability by mixing examples from different stages of an evolving crisis.

**Sentiment changes around major events.** Sentiment shifts appear around event-linked transitions, particularly the Post-Blackout period ($d = -0.556$) and later Flood Crisis ($d = 0.567$), suggesting that crisis sentiment evolves around major events rather than following a smooth temporal trend.

Together, these findings show that reliable crisis sentiment analysis requires both discourse context and temporal adaptation. UNRESTSENT200K supports the study of these challenges through parent-linked comments and event-aligned evaluation phases.

## 5 Discussion

Our results show that crisis sentiment analysis depends strongly on both context and time. Models perform much better when the parent post is included, with consistent gains of 7–11 percentage points across encoder models. At the same time, performance drops substantially when models are trained on earlier phases of the uprising and tested on later phases. This shows that standard random splits can overestimate model performance in an unfolding crisis. The results also show that strong LLMs do not remove these challenges. GPT-4o-mini performs well under few-shot prompting, but it only slightly improves over the best post+comment encoder and remains below the best fine-tuned LLM. Its errors also overlap with those of smaller models, especially on sarcasm, implicit political references, and comments whose meaning depends on the event phase. These findings suggest that future sentiment benchmarks, especially for crisis settings, should include temporal splits and discourse context. For low-resource languages such as Bangla, this is particularly important because models may appear reliable under standard evaluation while failing when the public conversation changes. UNRESTSENT200K provides a setting for studying these issues in a controlled way.

## 6 Conclusion

We introduce UNRESTSENT200K, a Bangla crisis sentiment dataset containing ≈200K comments from Facebook and YouTube discussions around the July–August 2024 Bangladesh uprising. Each comment is linked to its parent post and assigned one of three sentiment labels: POSITIVE, NEUTRAL, or NEGATIVE. The dataset is annotated through a fully human process under the PiLA framework. Experiments with fine-tuned encoders, prompted LLMs, and LoRA-tuned LLMs show three main results. First, parent-post context consistently improves sentiment classification. Second, temporal shift across phases of the uprising leads to large performance drops. Third, supervised adaptation gives the strongest results, but the benchmark remains challenging. Overall, UNRESTSENT200K supports research on context-aware and temporally robust sentiment analysis in low-resource crisis discourse.

**Limitations**

UNRESTSENT200K has several limitations. First, it covers only Facebook and YouTube. These platforms were important sources of public discussion during the uprising, but they do not represent all Bangla speakers or all forms of political communication. Other sources, such as X/Twitter, Telegram, private messaging, and offline discourse, are not included. Second, the dataset uses three sentiment labels. This makes large-scale annotation reliable, but it does not capture finer distinctions such as anger, fear, hope, grief, or stance. Future work can extend the label schema to include emotion and stance labels. Third, the dataset is based on one event in Bangladesh. The results should therefore not be treated as universal claims about all crises or all low-resource languages. Cross-event and cross-language studies are needed to test how well the findings generalize. Finally, the models evaluated in this paper were not specifically designed for temporal adaptation or change-point detection. Developing models that can use context and adapt to new crisis phases is an important direction for future work.

**Reproducibility Statement**

To support reproducibility, the released package has the processed and de-identified UNRESTSENT200K corpus; fixed training, validation, and test splits; parent-post context; timestamps; platform and event-phase metadata; annotation guidelines; prompt templates; preprocessing and evaluation scripts; and the configurations used for the encoder and LoRA experiments. Dataset statistics and phase boundaries are reported in Section 3, while detailed preprocessing, annotation, temporal-analysis, topic-modeling, and prompting procedures are provided in the remainder of this appendix.

**AI Usage Disclosure** Large language models were used as experimental systems to generate benchmark predictions for UNRESTSENT200K under zero-shot and few-shot prompting and supervised fine-tuning. AI-assisted tools were also used for limited non-experimental support, including language editing, figure ideation, and visualization assistance. All AI-assisted material was manually reviewed and verified by the authors, who retain full responsibility for the paper's content, results, and interpretations.

## References


Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2025. BnSentMix: A diverse Bengali-English code-mixed dataset for sentiment analysis. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Azizul Hakim Fayaz, MD Uddin, Rayhan Uddin Bhuiyan, Zakia Sultana, Md Samiul Islam, Bidyarthi Paul, Tashreef Muhammad, and Shahriar Manzoor. 2025. Bidwesh: A bangla regional based hate speech detection dataset. *arXiv preprint arXiv:2507.16183*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. BanTH: A

multi-label hate speech detection dataset for transliterated Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico. Association for Computational Linguistics.

Mahmudul Hasan, Md Rashedul Ghani, and KM Azharul Hasan. 2024. Aspect based sentiment analysis datasets for bangla text. *Data in Brief*, 57:111107.

Md Arid Hasan, Firoj Alam, Md Fahad Hossain, Usman Naseem, and Syed Ishtiaque Ahmed. 2026. LLM-based multi-task Bangla hate speech detection: Type, severity, and target. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, California, United States. Association for Computational Linguistics.

Md. Sabbir Hossen, Md. Saiduzzaman, and Pabon Shaha. 2025. Social media sentiments analysis on the july revolution in bangladesh: A hybrid transformer based machine learning approach. In *2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–8.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

Ariful Islam and Md Rifat Hossen. 2025. Banglasentnet: A hybrid deep learning framework for multi-aspect sentiment analysis in bangla e-commerce reviews. In *Data Science, AI and Applications*. Springer Nature Switzerland.

Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. SentNoB: A dataset for analysing sentiment on noisy Bangla texts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.

Md. Ekramul Islam, Labib Chowdhury, Faisal Ahamed Khan, Shazzad Hossain, Md Sourave Hossain, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023. Sentigold: A large bangla gold standard multi-domain sentiment analysis dataset and its evaluation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4207–4218. Association for Computing Machinery.

July Graffiti. 2024. https://julymemories.org/graffiti-spot/. Accessed: 2026-01-27; cultural collection of graffiti from the July 2024 movement in Dhaka.

Swastika Kundu, Autoshi Ibrahim, Mithila Rahman, and Tanvir Ahmed. 2026. Anubhuti: a comprehensive corpus for sentiment analysis in bangla regional languages. In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, pages 1–6. IEEE.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA. Association for Computational Linguistics.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15. Association for Computing Machinery.

Md. Saddam Hossain Mukta, Md. Adnanul Islam, Faisal Ahamed Khan, Afjal Hossain, Shuvanon Razik, Shazzad Hossain, and Jalal Mahmud. 2021. A comprehensive guideline for bengali sentiment annotation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(2).

Md. Darun Nayeem, Zarin Rafa, Tasnuva Tasnim Nova, Yasin Rahman, Abdul Mumeet Pathan, and Md. Masudul Islam. 2026. Banglamuse: A multimodal bangla sentiment dataset of text–audio pairs for speech and sentiment analysis. *Data in Brief*, 65:112458.

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, St. Julians, Malta. Association for Computational Linguistics.

James Pustejovsky. 2012. The role of linguistic models and language annotation in feature selection for machine learning. In *Proceedings of the Sixth Linguistic Annotation Workshop*, page 1, Jeju, Republic of Korea. Association for Computational Linguistics.

Md Nishat Raihan, Umma Tanmoy, Anika Binte Islam, Kai North, Tharindu Ranasinghe, Antonios Anastasopoulos, and Marcos Zampieri. 2023. Offensive language identification in transliterated and code-mixed Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Md. Sohel Rana, Shadia Sharmin, Mohammad Bin Amin, and Judit Oláh. 2026. From revolt to revolution: exploring motivating factors of bangladesh's july 2024 uprising. *Research in Globalization*, 12:100341.

Mohammad Rifat Ahmmad Rashid, Kazi Ferdous Hasan, Rakibul Hasan, Aritra Das, Mithila Sultana, and Mahamudul Hasan. 2024. A comprehensive dataset for sentiment and emotion classification from bangladesh e-commerce reviews. *Data in Brief*, 53:110052.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. BD-SHS: A benchmark dataset for learning to detect online Bangla hate speech in different social contexts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA. Association for Computational Linguistics.

ZACHARY C. STEINERT-THRELKELD. 2017. Spontaneous collective action: Peripheral mobilization during the arab spring. *American Political Science Review*, 111(2):379–403.

Md. Ashraful Islam Talukder. 2025. Bangla and banglish e-commerce reviews dataset for aspect-based sentiment analysis. *Mendeley Data*.

Zeynep Tufekci. 2017. *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.

# A Appendix

## A.1 Preprocessing Details

This section describes the preprocessing steps used before annotation, as summarized in Section 3.2.

### A.1.1 Normalization

- **Unicode normalization:** Text is normalized to NFC form to reduce inconsistent Bangla character encodings.
- **Whitespace cleanup:** Extra spaces, tabs, and irregular line breaks are removed.
- **Entity masking:** URLs and user mentions are replaced with `[URL]` and `[USER]`.
- **Emoji retention:** Emojis are kept because they often express sentiment in social media comments.

### A.1.2 Filtering

We remove comments with fewer than three tokens after normalization. We also remove comments that consist mainly of non-Bangla text or obvious spam. To keep the released corpus primarily Bangla, comments with substantial English or Bangla–English transliteration are excluded. Figure 3 reports the resulting sentiment distribution across word-length bins of the retained corpus: short comments ($\leq$10 words) skew positive, while neutral expression increases with comment length and negative sentiment stays relatively flat across lengths.

### A.1.3 Deduplication

Near-duplicate comments within the same thread are removed using a Jaccard similarity threshold of 0.85. Duplicates across different posts are retained when they appear as part of separate public discussions.

### A.1.4 Post-Level Framing Labels

In addition to comment-level sentiment labels, each source post is assigned a coarse framing label:

$$\mathcal{Y}^{(p)} = \{\textsc{Hope}, \textsc{Despair}, \textsc{Outrage}\} \quad (1)$$

These labels describe the dominant framing of the parent post. HOPE captures optimistic or change-oriented framing, DESPAIR captures grief or uncertainty, and OUTRAGE captures anger, blame, or calls for accountability. These post-level labels are used as annotation context and metadata; the main prediction task remains comment-level sentiment classification.

## A.2 PiLA: Annotation Protocol Details

This appendix provides additional details about the **Pilot Label Annotation (PiLA)** framework shown in Figure 2. PiLA is a fully human annotation process. No labels are assigned by automated systems. The full annotation effort took approximately eight months and involved 14 primary annotators and 10 senior validators.

### A.2.1 Annotator and Validator Pool

The primary annotation team consists of 14 undergraduate annotators from Bangladesh. All annotators are native Bangla speakers and were familiar with the July–August 2024 events through Bangla news and social media. The validator team consists of 10 senior annotators who did not overlap

Table 8: Table showing topic modeling statistics for each period. $C_v$ denotes the topic coherence score, computed using normalized pointwise mutual information (NPMI) and ranging from 0 to 1, where higher values indicate greater semantic coherence. *Eff.* represents the effective number of topics, calculated as $\exp(H)$, where $H = -\sum_i p_i \log p_i$ is the Shannon entropy of the topic distribution. This quantity reflects the number of equally prominent topics that would yield the same level of topic diversity.

| Period | Docs | Topics | $C_v$ | Eff. |
|---|---|---|---|---|
| P1: Pre-Escalation | 6,727 | 12 | 0.405 | 1.7 |
| P2: Crisis & Blackout | 33,684 | 29 | **0.781** | 1.8 |
| P3: Post-Blackout | 46,197 | 18 | 0.384 | 1.9 |
| P4: Post-Revolution | 69,536 | 25 | 0.420 | 2.0 |
| P5: Flood Crisis | 43,255 | 25 | 0.422 | 1.9 |
| **Global** | 199,399 | 29 | 0.375 | — |

with the primary annotation pool. All participants provided informed consent, and the study protocol was reviewed by the host institution.

#### A.2.2 Guideline Document

The annotation guideline defines the three sentiment labels: POSITIVE, NEUTRAL, and NEGATIVE. It also provides rules and examples for common ambiguous cases, including sarcasm, rhetorical questions, religious expressions, political references, code-switched profanity, and emoji-based sentiment. Annotators were instructed to use the parent post as context, but not to copy the post's sentiment into the comment label. The label was assigned to the comment only. The guideline was refined during the pilot stage and then fixed before the main annotation stage.

#### A.2.3 Qualification and Quality Checks

Before the main annotation stage, annotators completed a qualification round. Only annotators who reached Cohen's $\kappa \geq 0.65$ were admitted to the main task. All 14 annotators met this threshold, with scores ranging from 0.66 to 0.81. During annotation, gold-check examples were inserted at a 2% rate. Annotators whose agreement dropped below the required threshold were paused, retrained on the relevant ambiguity cases, and re-qualified before continuing. Labels from affected batches were reviewed and re-annotated where necessary.

#### A.2.4 Adjudication

Each comment was independently labeled by two annotators. If both annotators agreed, their shared label was accepted. If they disagreed, the comment was reviewed by a senior validator. Difficult cases were escalated to a small validator panel. For adjudicated comments, we record the two initial labels, annotator confidence scores, validator decision, and final gold label. These logs allow later analysis of disagreement patterns and label reliability.

### A.3 Quality Metrics Details

#### A.3.1 Pairwise Inter-Annotator Agreement

We report Cohen's $\kappa$ for agreement between the two primary annotators assigned to each comment. The overall pairwise score is $\kappa = 0.73$, indicating substantial agreement. We compute:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{2}$$

where $p_o$ is observed agreement and $p_e$ is expected agreement under the annotators' marginal label distributions. Per-phase $\kappa$ ranges from 0.69 to 0.78.

#### A.3.2 Multi-Rater Reliability

We also report Krippendorff's $\alpha$ to measure reliability across the full annotator pool. The overall score is $\alpha = 0.71$, showing that the labels are consistent beyond individual annotator pairs.

#### A.3.3 Annotator Confidence

Annotators provided a confidence score in $[0, 1]$ for each label. The mean confidence score is $\bar{c} = 0.82$ with standard deviation 0.14. In total, 87.3% of labels have confidence scores of at least 0.7. These scores are included as metadata so that users can filter examples by confidence if needed.

#### A.3.4 Pre-Adjudication Agreement

The two primary annotators agreed directly on 78.4% of comments. The remaining 21.6% were sent to senior validators for adjudication. Among these, 3.1% required review by a three-validator panel.

#### A.3.5 Independent Blind Audit

As a final quality check, two senior validators who were not involved in adjudication re-labeled a stratified random sample of 2,000 comments, with 400 comments from each phase. They were blind to the final gold labels. The audit achieved 94.2% exact-label agreement and 96.8% agreement within one step on the sentiment scale. Most disagreements involved rhetorical questions and emoji-mediated sentiment, which were also identified as difficult cases during guideline development.

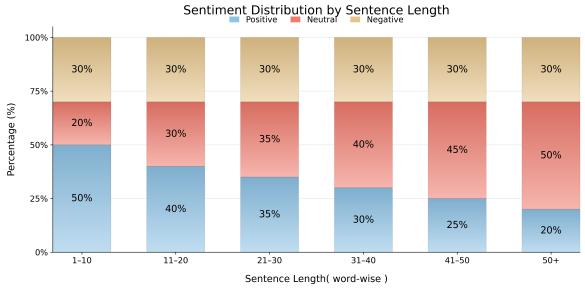



Figure 3: Sentiment distribution across sentence length intervals (word-wise). Short sentences (1–10 words) are dominated by positive sentiment, while the proportion of neutral sentiment increases with sentence length. Negative sentiment remains relatively stable across all length ranges.

## A.4 Effect of Comment Length

We analyze sentiment distribution across word-level comment length intervals to examine whether affective polarity varies with comment verbosity. As shown in Figure 3, shorter comments are more frequently positive, with the positive share decreasing from 50% in 1–10 word comments to 20% in comments longer than 50 words. In contrast, neutral sentiment increases steadily with length, suggesting that longer comments tend to provide explanation, context, or factual discussion rather than direct affective reactions. Negative sentiment remains comparatively stable across all length ranges. This indicates that comment length is associated with sentiment composition, and that models may need to account for verbosity-driven shifts in crisis discourse.

## A.5 Quantifying Temporal Non-Stationarity

We measure lexical change across event phases using three signals: Jensen–Shannon (JS) divergence over unigram distributions, Jaccard similarity of top words, and Shannon entropy. The results are shown in Figure 4.

**JS divergence.** JS divergence values range from 0.3713 to 0.3832 across phase pairs, showing that word distributions change over time.

**Jaccard similarity.** Top-word Jaccard similarity ranges from 0.4235 to 0.4652. This means that less than half of the most frequent terms are shared across phase pairs, indicating vocabulary turnover as the event develops.

**Entropy.** Shannon entropy increases from 7.5989 to 7.7872, suggesting that the discourse becomes more lexically diverse in later phases. Together, these results show that the corpus changes measurably across event phases. This supports the temporal-shift results in Table 5: performance drops are associated with changes in the data distribution, not only with model limitations. The corresponding sentiment-level shift, quantified using Cohen's $d$ on the polarity distributions of all $\binom{5}{2} = 10$ pairwise phase pairs, is reported in Table 7; seven of ten contrasts are negligible, with the strongest effects clustered around the post-blackout transition mirroring the lexical-level drift pattern above.

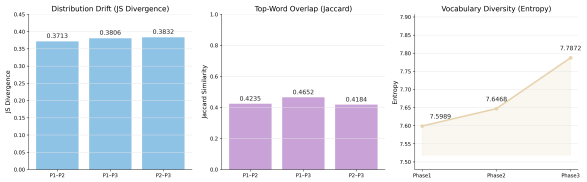



Figure 4: Temporal drift across three chronological phases of Bangla text. Divergence, vocabulary turnover, and rising entropy together reveal substantial non-stationarity and increasing lexical diversity over time.

## A.6 Topic Modeling: Extended Methodology and Results

### Detailed Methodology

To understand how the online conversation changed during the uprising, we follow a simple three-step process. First, a multilingual sentence transformer represents each comment so that comments with similar meanings are placed close together. Second, UMAP reduces these representations to a smaller space. Third, HDBSCAN groups nearby comments into topics. We also use a custom list of 619 Bangla stopwords, including 158 discourse markers, 151 platform-specific artifacts, and 6 special tokens. Removing this noise helps the model focus on meaningful themes. Table 8 reports the topic statistics for each event phase.

We choose the model settings to keep the topics detailed but still easy to interpret. For the global model, UMAP uses `n_components=5`, `n_neighbors=20`, and `min_dist=0.1`. HDBSCAN uses `min_cluster_size=20`, `min_samples=5`, and `cluster_selection_epsilon=0.1`. These lenient settings prevent the crisis discussion from being divided into too many small topics. CountVectorizer uses one- to three-word phrases, with `min_df=10` and `max_df=0.9`, to retain useful Bangla phrases while removing terms that are

too rare or too common.

The five event phases contain different numbers of comments, so one clustering setting would not suit every phase. We therefore scale `min_cluster_size` from 25 in the smallest phase to 100 in the largest. If HDBSCAN marks a comment as an outlier (topic $= -1$), we assign it to the nearest topic centroid using cosine similarity. This gives every comment a topic. As a final check, we vary `min_cluster_size` by $\pm 20\%$. Topic coherence changes by less than 0.05 ($\Delta C_v < 0.05$), showing that the main findings remain stable under reasonable parameter changes.

**Extended Results**

**Attention rises and falls in sharp bursts.** The story begins with an online discussion that does not grow smoothly. Instead, public attention moves in sudden waves as events unfold. Seventy-five percent of the topics have a coefficient of variation above 1.0, meaning that their activity changes greatly relative to their average level. The largest burst appears on August 9, three days after the collapse, when Topic T1 reaches 1,063 posts in a single day. A topic remains active for 44.8 days on average. The dominant topic, T0, changes momentum 15 times and has a volatility value of 4,269.78. Together, these patterns show how quickly attention moved from one concern to another during the crisis.

**The themes change, but they remain connected.** Although attention is highly volatile, the discussion does not repeatedly start from zero. Across 100 topic transformations, the mean semantic similarity is 0.93. In other words, themes usually evolve from earlier themes rather than disappear and reappear in unrelated forms. For example, student-protest discussion (ছাত্রলীগের, আন্দোলন) develops into narratives about state violence (পুলিশ, তোলা) with a similarity of 0.902. Confrontation rhetoric later becomes celebratory discourse (জড়িত, হাসিমাখা) with a similarity of 0.954, while its volume grows by 462% after the collapse.

**Negative sentiment shapes most of the conversation.** The final part of the story concerns emotion. Only Topic T8 (সিদ্ধান্ত, সংস্কার—decisions and reforms) has a positive polarity ($+0.246$). The most negative topics, T15 ($-0.486$) and T6 ($-0.294$), also carry a large share of the discussion. The topics are closely connected: we find 38 significant co-occurrence relationships with $r > 0.5$. The strongest link is between decision-making and public-reaction topics (T3 $\leftrightarrow$ T1, $r = 0.988$), reflecting the rapid cycle of action and reaction common in crisis communication (Tufekci, 2017).

Taken together, the results tell a consistent story. Discussion stays centered on a small set of core grievances; these themes become more focused under stress, attract attention in sudden bursts, and continue to evolve across event phases. At the same time, sentiment remains mostly negative. Crisis discourse is therefore not simply a collection of disconnected conversations. It has a connected but rapidly changing structure, which calls for analysis methods designed specifically for crisis settings.

### A.7 Failure Mode Discovery: Topic-Level Sentiment Instability

Overall performance scores hide an important part of the story: a model may appear stable on average while becoming unreliable for a particular topic. We therefore return to the 25 BERTopic topics and follow each one across the five event phases. For every topic, we measure how much its sentiment polarity changes using the standard deviation ($\sigma$). This topic-level view complements the lexical drift shown in Figure 4 and reveals that sentiment instability is concentrated in specific themes.

**Different topics follow different paths.** Six topics are highly volatile ($\sigma > 30$), 14 have medium volatility ($15 \leq \sigma \leq 30$), and 5 remain stable ($\sigma < 15$). Topic T22 (কোটা—quota reform) provides the clearest example. Its polarity changes from $+57.1\%$ to $-33.3\%$ between P3 and P4, giving $\sigma = 45.76$ and a range of 90.48 (Figure 5). Humanitarian topics, in contrast, remain much steadier, with $\sigma \approx 12$–$15$. The key lesson is that temporal shift does not affect every topic in the same way.

**Two turning points create the greatest risk.** Fifteen topics show a dramatic sentiment flip ($\Delta > 40\%$), but most of the largest changes occur at two moments in the uprising. The first is the P2$\rightarrow$P3 transition, after the internet blackout, when information returns and people begin to interpret what happened. Four topics change by more than 60% at this point. The second is the P3$\rightarrow$P4 transition, during the regime collapse, when celebration and uncertainty appear together. Five topics reverse by more than 45%. For example, Topic T11 (সমস্যা—problems) moves from $-15.0\%$ to $-45.6\%$ during P2$\rightarrow$P3 ($\Delta = 64.2\%$). A classifier trained only on P2 would therefore underestimate P3 negativity for

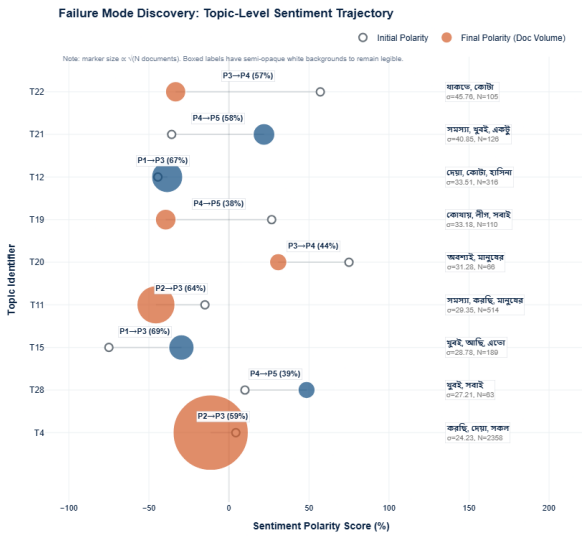



Figure 5: How topic-level sentiment changes across the five phases (P1–P5) of the July Revolution. A hollow circle shows a topic's initial polarity, and a filled circle shows its final polarity; larger filled circles represent more documents. Arrows and percentage labels show the direction and size of each change. T22 (কোটা) is highly volatile, moving from $+57.1\%$ to $-33.3\%$ ($\sigma = 45.76$), while flood-relief topics remain comparatively stable ($\sigma \approx 12$–$15$). The largest reversals occur after the internet blackout (P2→P3) and during the regime collapse (P3→P4), showing that sentiment shift depends strongly on both topic and time.

this topic by more than 30 percentage points.

**This pattern explains where models are likely to fail.** Politically charged themes, such as quota reform and league politics, are three times more volatile than humanitarian concerns. A single fixed model may therefore work well for stable topics but lose reliability when the meaning and sentiment of a political topic change. One practical response is topic-aware temporal recalibration: the model could lower or adjust its confidence when it detects a historically volatile topic. Future work could also prioritize these topics during temporal fine-tuning. Both strategies follow the same principle—adapt the model where the conversation changes most.

### A.8 Failure Mode Discovery through Topic Modeling

An overall accuracy score tells us how often a model is wrong, but it does not tell us whether the errors follow a common pattern. To find that pattern, we collect all 2,789 misclassified comments and apply BERTopic to the error set alone. Using a smaller configuration suited to this subset (`min_cluster_size=15` and one- to two-word phrases), the model finds 33 semantically coherent error topics with $C_v = 0.9499$. We then follow the evidence from class-level errors to topic concentration, confidence, and complete sentiment reversals.

Table 9: The ten topics containing the most classification errors. Topic 0 contains 676 errors (24.2% of all errors), and 57.46% of neutral comments are misclassified across the full error set. The mean confidence on incorrect predictions is 0.977, showing that the model is often highly confident even when it is wrong.

| Topic | Errors | Conf. | Neu→Neg | Neu→Pos | Neg↔Pos |
|---|---|---|---|---|---|
| T0: আন্দোলনের স্পিরিটের | 676 | 0.968 | 261 | 61 | 104 |
| T1: ঘটেছে বৈষম্যবিরোধী | 369 | 0.971 | 186 | 20 | 58 |
| T2: বিদ্যুৎকেন্দ্র | 134 | 0.966 | 29 | 5 | 23 |
| T3: আগের শিক্ষাক্রমে | 110 | 0.968 | 40 | 18 | 29 |
| T4: শিক্ষার্থীদের স্বপ্নকে | 92 | 0.970 | 14 | 13 | 38 |
| T5: আটক মির্জাপুরে | 89 | **1.000** | 0 | 85 | 4 |
| T6: ছাত্রলীগের হামলায় | 75 | 0.980 | 40 | 7 | 10 |
| T7: চিরকালই বিদ্রোহী | 75 | 0.982 | 19 | 9 | 17 |
| T8: শিক্ষার্থীদের অস্বস্তি | 72 | 0.992 | 10 | 16 | 27 |
| T9: দূর আগের | 70 | 0.977 | 10 | 11 | 27 |
| **All Topics (33)** | 2,789 | 0.977 | 895 | 356 | 531 |

**The main problem begins with neutral comments.** The model classifies negative comments with 83.71% accuracy and positive comments with 63.10% accuracy, but neutral accuracy falls to 42.54%. Thus, 57.46% of neutral comments are misclassified. The largest single error type is neutral→negative: it occurs 895 times and accounts for 32.1% of all errors. This is the first sign that the model systematically reads ambiguous language as negative.

**The errors gather around a small set of themes.** They are not spread evenly across the 33 topics. Topic 0 alone contains 676 errors, or 24.2% of the full error set. Among its 322 neutral errors, 261 (81.1%) are predicted as negative. Large error clusters also appear around political discussion (বৈষম্যবিরোধী আন্দোলনের) and education policy (আগের শিক্ষাক্রমে). These examples suggest that BanglaBERT often reacts to words associated with conflict or criticism but misses the neutral pragmatic meaning of the full comment. LoRA-SFT on a mid-scale LLM recovers several such context-dependent cases, as illustrated in Table 6.

**The table confirms a consistent negative bias.** As Table 9 shows, the model misclassifies 1,251 of the 2,177 neutral comments (57.46%). It predicts negative for 895 of them and positive for 356, a ratio of about 2.5:1. The same direction of bias appears in T1, T3, and T6, where neutral→negative errors clearly outnumber neutral→positive errors. The problem therefore extends beyond one topic:

ambiguous Bangla discourse is repeatedly pulled toward the negative class.

**High confidence makes these errors harder to detect.** The model's mean confidence across all incorrect predictions is 0.977. Topic 5 is the clearest warning: the confidence is 1.000 even though 85 neutral comments are predicted as positive. The model is therefore not simply uncertain about difficult examples; it is often confidently wrong. A standard probability threshold may fail to catch these cases, which motivates confidence calibration that also considers the topic.

**Some mistakes reverse the sentiment completely.** In 531 cases, or 19.0% of all errors, the prediction flips directly between negative and positive. Topic 0 contains 104 such flips and Topic 1 contains 58; both focus on বৈষম্যবিরোধী আন্দোলনের, where political framing can change the sentiment of otherwise similar language. Topic 4 (শিক্ষার্থীদের স্বপ্নকে—students' dream) has an especially high flip rate: 38 of its 92 errors, or 41.3%. The full error story is therefore not only about neutral comments. BanglaBERT also struggles to distinguish hopeful and critical framings when they discuss the same political or educational issue.

### A.9 Structural Diagnosis: Where Do Models Fail?

The previous section showed what the model's errors look like. We now ask why those errors gather in particular places. For this purpose, BERTopic (Grootendorst, 2022) gives us two connected views. The first follows the topic structure of the full corpus over time; the second shows where the model's errors sit within that structure. The pipeline—multilingual sentence-transformer embeddings, UMAP, HDBSCAN, phase-specific parameters, and a 619-term Bangla stopword list—is described in Appendix A.6. Table 8 summarizes the results by phase. Two structural patterns explain much of the model behavior.

**First, one political theme dominates the conversation.** The global model finds 29 topics ($C_v = 0.375$, silhouette $= 0.402$), but T0 (লীগ, সরকার, কোটা—league, government, quota) contains 56.3% of all documents ($n = 105{,}226$). It is also highly stable over time (CV $= 0.014$) and has a negative polarity ($-0.224$). The effective number of topics based on Shannon entropy is only 1.7–2.0 across the five phases. Thus, the discussion may look diverse on the surface, but most attention remains concentrated around one central political grievance. We call this pattern *discourse hegemony*.

This dominance creates a risk for sentiment models. Because T0 supplies such a large share of the training data, a model can learn to connect its political vocabulary with negative sentiment. It may then apply that negative tendency to neutral comments or to comments from other topics. This provides a structural explanation for the neutral→negative bias found in Appendix A.8.

**Second, the blackout makes the discussion more focused, not more random.** Topic coherence is not determined by the number of comments. P2, the Crisis & Blackout phase, contains 33,684 comments (16.9% of the corpus) and produces 29 topics, yet it has the highest coherence by a wide margin ($C_v = 0.781$). Coherence in every other phase ranges only from 0.384 to 0.422. The largest phase, P4 (Post-Revolution), contains 69,536 comments (34.9%) but reaches only $C_v = 0.420$. The smallest phase, P1 (Pre-Escalation), contains 6,727 comments (3.4%) and reaches $C_v = 0.405$.

We call this result the *crisis coherence paradox*. During the blackout, limited information and immediate danger push people toward a small set of urgent concerns, making the discussion more semantically focused. After the crisis, attention spreads across more issues and coherence falls. Extreme conditions therefore crystallize the conversation instead of turning it into noise.

**Together, these patterns explain where models fail.** Errors cluster around the dominant topic and closely related political themes rather than appearing uniformly across the corpus. Politically charged topics, including quota reform and party politics, show three times the sentiment volatility of humanitarian topics, and 33% of all errors fall within T0's thematic orbit. Temporal robustness should therefore be measured separately for each topic. In the same way, confidence recalibration should use topic-level volatility instead of relying only on one aggregate confidence score.

### A.10 Prompting Techniques

After identifying where sentiment models struggle, we test whether LLMs can solve the task from instructions and examples alone. We compare two prompting settings: zero-shot and few-shot. The model parameters remain frozen in both settings, so the only difference is the information supplied in the prompt. This design separates in-context reasoning from any benefit that could come from task-

specific fine-tuning.

**Zero-shot prompting.** We begin with the harder setting: the model receives no labeled examples. The prompt contains only a short expert-role instruction (for example, "You are an expert in Bengali political sentiment analysis"), the parent `post`, the target `comment`, the allowed labels {positive, neutral, negative}, and an instruction to return one label. The resulting score shows how well the model can understand crisis-era Bangla using its pretrained knowledge alone. Success therefore depends on its Bangla vocabulary, cultural and political understanding, and ability to follow instructions.

**Few-shot prompting.** Next, we give the same model seven labeled examples from the training split of UNRESTSENT200K. Each example contains a `post`, a `comment`, and the correct `label`. Together, the examples cover all three sentiment classes, both supportive and critical views of crisis events, and difficult language such as sarcasm, rhetorical questions, emoji descriptions, Bangla–English code-mixing, and indirect political references. The model then labels a held-out post–comment pair under the same one-label rule. Comparing this result with zero-shot performance shows how much seven relevant examples can help without changing the model itself.

**Decoding and output format.** Finally, we keep the decision process identical across both settings. We use greedy decoding with temperature $= 0$ and require exactly one label from {positive, neutral, negative}. If a response does not match an allowed label, we count it as an error and do not ask the model again. The reported results therefore measure the reliability of the complete prompting process, including instruction following, rather than the quality of predictions after manual correction or extra parsing. The complete prompt templates appear in §B.1.

### A.11 Expert Prompting with LLMs

Table 3 reports the zero-shot and few-shot performance of open-source and proprietary LLMs using expert prompting with Bangla-native instructions on the UnRestSent200K benchmark.

## B Ethics Statement

All data used in this work were collected from publicly accessible Facebook and YouTube content in accordance with platform terms of service. Personally identifiable information, including usernames and profile links, was removed during preprocessing. The UNRESTSENT200K dataset will be released for research purposes under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

### B.1 Sample Prompts

### ফিউ-শট প্রম্পট: সেন্টিমেন্ট অ্যানোটেশন (Few-Shot: Sentiment Annotation — Bengali)

**System:** তুমি বাংলা ভাষার রাজনৈতিক sentiment analysis বিশেষজ্ঞ। তোমাকে একটি সংবাদ পোস্ট/শিরোনাম এবং সেই পোস্টের উপর করা একটি মন্তব্য দেওয়া হবে। তোমার কাজ হলো নির্ধারণ করা যে মন্তব্যটি উল্লিখিত পোস্ট, ঘটনা, বা প্রসঙ্গের প্রতি কী ধরনের sentiment প্রকাশ করছে।

**User:** নিচে কিছু উদাহরণ দেওয়া আছে। উদাহরণগুলো অনুসরণ করে শেষে দেওয়া প্রশ্নটির উত্তর দাও।

নির্দেশনা:

- post শুধুমাত্র contextual information হিসেবে ব্যবহার করবে।
- sentiment নির্ধারণের সময় sarcasm, rhetorical tone, slang, emoji description (যেমন: `কান্না জড়িত হাসিমুখ প্রতিক্রিয়া'), এবং mixed Bangla-English expression বিবেচনা করবে।
- রাজনৈতিক মতাদর্শ বা ব্যক্তিগত অবস্থান নয়, শুধুমাত্র comment-এর sentiment বিচার করবে।
- যদি comment সমর্থন, প্রশংসা, আশাবাদ, বা ইতিবাচক প্রতিক্রিয়া প্রকাশ করে → `positive`
- যদি comment সমালোচনা, হতাশা, রাগ, ব্যঙ্গ, আক্রমণ, বা নেতিবাচক প্রতিক্রিয়া প্রকাশ করে → `negative`
- যদি comment তথ্যভিত্তিক, প্রশ্নধর্মী, দ্ব্যর্থক, বা স্পষ্ট ইতিবাচক/নেতিবাচক sentiment না প্রকাশ করে → `neutral`

উদাহরণসমূহ:

উদাহরণ ১
পোস্ট: কোটাবিরোধী আন্দোলন : ঢাকা বিশ্ববিদ্যালয়ে মিছিল শুরু, মধুর ক্যানটিনে জড়ো হয়েছে ছাত্রলীগ
মন্তব্য: সাবধান, আবার রক্ত ঝরাবে।
উত্তর: `negative`

উদাহরণ ২
পোস্ট: অপরাধে সম্পৃক্ততায় যুবদল--ছাত্রদলের ১৫ জনকে বহিষ্কার
মন্তব্য: নাটক কম করো পিও, জনগণ সব বুঝে।
উত্তর: `negative`

উদাহরণ ৩
পোস্ট: আজ বৃহস্পতিবার দেশের প্রায় সব মোবাইল অপারেটরের ইন্টারনেট ধীরগতিতে চলছে।
মন্তব্য: শ্লা টাউট! কান্না জড়িত হাসিমুখ প্রতিক্রিয়া
উত্তর: `negative`

উদাহরণ ৪
পোস্ট: কোটাবিরোধী আন্দোলন : ঢাকা বিশ্ববিদ্যালয়ে মিছিল শুরু, মধুর ক্যানটিনে জড়ো হয়েছে ছাত্রলীগ
মন্তব্য: ছাত্রলীগও কোটা আন্দোলনে অংশ নিয়েছে? ভাবনা মুখ প্রতিক্রিয়া
উত্তর: `neutral`

উদাহরণ ৫
পোস্ট: অপরাধে সম্পৃক্ততায় যুবদল--ছাত্রদলের ১৫ জনকে বহিষ্কার
মন্তব্য: যাদের পদ নেই তাদের কি শাস্তি হবে?
উত্তর: `neutral`

উদাহরণ ৬
পোস্ট: কোটাবিরোধী আন্দোলন : ঢাকা বিশ্ববিদ্যালয়ে মিছিল শুরু, মধুর ক্যানটিনে জড়ো হয়েছে ছাত্রলীগ
মন্তব্য: কোটা সিস্টেম নিপাত যাক। মেধাবীরা চাকরি পাক।
উত্তর: `positive`

উদাহরণ ৭
পোস্ট: অপরাধে সম্পৃক্ততায় যুবদল--ছাত্রদলের ১৫ জনকে বহিষ্কার
মন্তব্য: সঠিক সময়ের সঠিক সিদ্ধান্ত ধন্যবাদ
উত্তর: `positive`

এখন নিচের উদাহরণটির sentiment নির্ধারণ করো:
পোস্ট: `{post}`
মন্তব্য: `{comment}`
শুধুমাত্র নিচের একটি label প্রদান করো: `positive`, `neutral`, `negative`

## Few-Shot: Sentiment Annotation(English)

**System:** You are an expert in political sentiment analysis for the Bangla language. You will be given a news post or headline and a comment made on that post. Your task is to determine what kind of sentiment the comment expresses toward the mentioned post, event, or context.

**User:** Some examples are given below. Follow the examples and answer the final question.

**Instructions:**

- Use the post only as contextual information.
- While determining sentiment, consider sarcasm, rhetorical tone, slang, emoji descriptions such as "crying laughing face reaction", and mixed Bangla-English expressions.
- Judge only the sentiment of the comment, not political ideology or personal position.
- If the comment expresses support, praise, optimism, or a positive reaction, label it `positive`.
- If the comment expresses criticism, frustration, anger, sarcasm, attack, or a negative reaction, label it `negative`.
- If the comment is factual, question-like, ambiguous, or does not express a clear positive or negative sentiment, label it `neutral`.

**Examples:**
**Example 1**
Post: Anti-quota movement: A procession has started at Dhaka University, and Chhatra League has gathered at Madhur Canteen.
Comment: Be careful, they will shed blood again.
Answer: `negative`
**Example 2**
Post: Fifteen members of Jubo Dal and Chhatra Dal expelled for involvement in crimes.
Comment: Stop the drama, people understand everything.
Answer: `negative`
**Example 3**
Post: Today, Thursday, internet service is slow across almost all mobile operators in the country.
Comment: Damn fraud! Crying laughing face reaction.
Answer: `negative`
**Example 4**
Post: Anti-quota movement: A procession has started at Dhaka University, and Chhatra League has gathered at Madhur Canteen.
Comment: Has Chhatra League also joined the quota movement? Thinking face reaction.
Answer: `neutral`
**Example 5**
Post: Fifteen members of Jubo Dal and Chhatra Dal expelled for involvement in crimes.
Comment: Will those who do not hold any position also be punished?
Answer: `neutral`
**Example 6**
Post: Anti-quota movement: A procession has started at Dhaka University, and Chhatra League has gathered at Madhur Canteen.
Comment: Down with the quota system. Let meritorious students get jobs.
Answer: `positive`
**Example 7**
Post: Fifteen members of Jubo Dal and Chhatra Dal expelled for involvement in crimes.
Comment: The right decision at the right time. Thank you.
Answer: `positive`

**Now determine the sentiment of the following example:**
**Post:** `{post}`
**Comment:** `{comment}`
Provide only one of the following labels: `positive`, `neutral`, or `negative`.

**Zero-Shot Prompt: Political Expert (Bengali)**
তুমি একজন বাংলা রাজনৈতিক sentiment analysis বিশেষজ্ঞ। নিচের বাংলা পোস্ট এবং মন্তব্যটির sentiment নির্ধারণ করো।

পোস্ট: `{post}`
মন্তব্য: `{comment}`

শুধুমাত্র একটি লেবেল দাও: `positive`, `neutral`, `negative`

**Zero-Shot: Political Expert — English**

You are an expert in Bengali political sentiment analysis. Determine the sentiment of the following Bengali post and comment.

**Post:** `{post}`
**Comment:** `{comment}`

Respond with only one label: `positive`, `neutral`, or `negative`.