A Multi-Task Benchmark for Abusive Language Detection in Low-Resource Settings

Fitsum Gaim **Hoyun Song** **Huije Lee**
Changgeon Ko **Eui Jun Hwang** **Jong C. Park**

Korea Advanced Institute of Science and Technology (KAIST)
{fitsum.gaim,hysong,huijelee,pencaty,ehwa20,jongpark}@kaist.ac.kr

Abstract

Content moderation research has recently made significant advances, but remains limited in serving the majority of the world's languages due to the lack of resources, leaving millions of vulnerable users to online hostility. This work presents a large-scale human-annotated multi-task benchmark dataset for abusive language detection in Tigrinya social media with joint annotations for three tasks: abusiveness, sentiment, and topic classification. The dataset comprises 13,717 YouTube comments annotated by nine native speakers, collected from 7,373 videos with a total of over 1.2 billion views across 51 channels. We developed an iterative term clustering approach for effective data selection. Recognizing that around 64% of Tigrinya social media content uses Romanized transliterations rather than native Ge'ez script, our dataset accommodates both writing systems to reflect actual language use. We establish strong baselines across the tasks in the benchmark, while leaving significant challenges for future contributions. Our experiments demonstrate that small fine-tuned models outperform prompted frontier large language models (LLMs) in the low-resource setting, achieving 86.67% F1 in abusiveness detection (7+ points over best LLM), and maintain stronger performance in all other tasks. The benchmark is made public to promote research on online safety.[1]

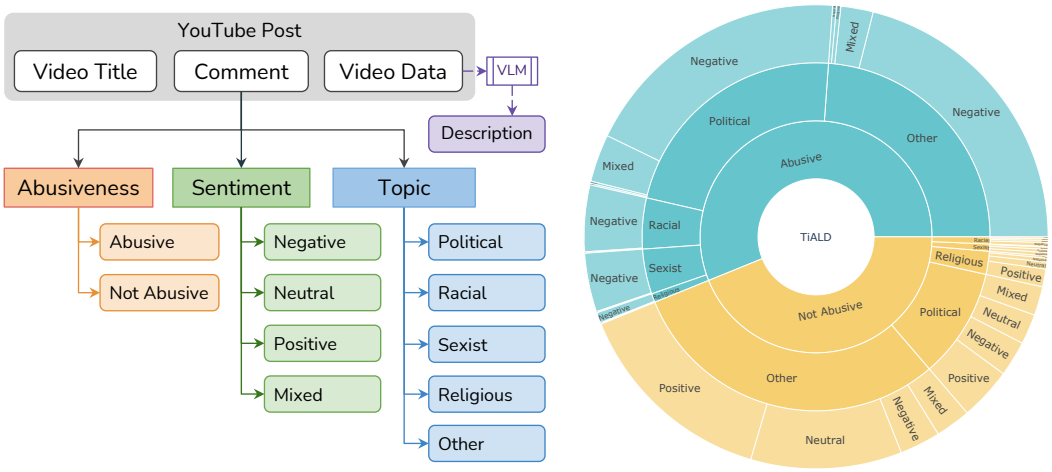
1 Introduction

The proliferation of social media has revolutionized global communication, enabling unprecedented connectivity while simultaneously creating new vectors for harm through abusive content [1]. Online hostility and harassment affect millions of users, particularly vulnerable groups including minors and minority communities, often causing physical and psychological harm while reinforcing social marginalization [2, 3]. Although significant progress has been made in automated detection of abusive content for high-resource languages such as English [4–6], the majority of the world's low-resourced languages, such as those spoken in Africa, remain understudied [7], creating an alarming disparity in online safety and protection.

Tigrinya, a language with approximately 10 million speakers mainly in Eritrea and Ethiopia, exemplifies this technological divide [8]. Despite its significant speaker population, Tigrinya remains computationally under-resourced with minimal datasets, tools, and models [9]. In particular, there is a lack of well-established benchmarks to gauge progress in content moderation research. This gap exposes the Tigrinya-speaking communities to unchecked online abuse. More broadly, it highlights

[1] TiALD Resources (Dataset, Code, Models): https://github.com/fgaim/TiALD

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.



(a) Dataset Schema and Features

(b) Task-wise Class Distributions

Figure 1: An overview of the TiALD Dataset Design and Annotated Class Distributions.

the urgent need for dedicated computational resources and methods of content moderation research in underrepresented languages.[2]

In this paper, we address the critical gap by introducing the **Ti**grinya **A**busive **L**anguage **D**etection (**TiALD**) dataset, a large-scale human-annotated multi-task benchmark for abusive language detection in the Tigrinya language. TiALD adopts a multifaceted approach, providing joint annotations for three tasks: abusiveness detection, sentiment analysis, and topic classification across 13,717 manually annotated YouTube comments. We further enrich the dataset by generating descriptions of the visual content of the corresponding videos using a Vision-Language Model (VLM), enabling the relational analysis against the user comments. Our annotation scheme enables richer contextual understanding of abusive content, supporting more nuanced analysis than that with the typical binary classification setups alone. Figure 1 shows an overview of the dataset design and class distributions. The contributions of this work can be summarized as follows:

- We present the first large-scale multi-task benchmark dataset for abusive language detection in Tigrinya based on user comments collected from YouTube channels in the community.
- We propose a data selection methodology for iterative semantic clustering of terms to address the inherent imbalance in abusive vs. non-abusive content on social media.
- We accommodate the sociolinguistic reality of Tigrinya social media by covering posts written in both the standard Ge'ez script and Latin transliterations, ensuring that the trained models handle actual language practices.
- We demonstrate that small, specialized models with joint multi-task learning of abusiveness, topic, and sentiment tasks outperform large frontier models, establishing a strong baseline.

2 Dataset Construction

In this section, we describe the construction of the TiALD dataset and an analysis of the annotations.

2.1 Data Collection and Preprocessing

As the source of data, we initially collected 4.1 million comments from 51 popular YouTube channels with more than 34.5K videos and a total of over 2.2 billion views at the time of collection. These channels cover various genres, including news, entertainment, music, educational, documentaries,

[2] Monitoring and Analyzing online content is a major commitment of the United Nations' *Strategy and Action Plan on Hate Speech* established in 2019: `https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf` (accessed on 2025-04-19).

vlogs, and more, ensuring a diverse and representative sample of the Tigrinya-speaking social media landscape. We then preprocessed the data by filtering out non-text comments, such as those that contain only emojis and also non-Tigrinya comments, such as those written fully in English, Arabic, Amharic, etc, using the GeezSwitch library [9]. Within the Tigrinya content, we observed that around 64% of the collected source comments contained Romanized text, where users employ improvised and non-standard transliteration schemes. To accommodate this, we cover comments written in both scripts in the annotation process, 70% Ge'ez and 30% Latin or mixed, which could help develop models that reflect a realistic usage of the language in social media.

2.2 Data Samples Selection for Annotation

Abusive language constitutes a minority of online content, making random sampling inefficient for dataset construction, while simple keyword searches yield lexically homogeneous datasets. Moreover, most low-resourced languages lack extensive curations of abusive terms for this purpose. Recognizing this gap, we propose a semi-automatic strategy that leverages a vector space of candidate samples and a small set of seed terms to effectively expand the selection criteria without heavily relying on the lexical search of manually curated terms. To this end, we trained word embeddings on the original 4.1 million comments from YouTube using word2vec [10] implemented in Gensim [11] with the CBOW architecture and a vector dimension of 300. We used cosine similarity to compute nearest neighbors and then applied an iterative process of term expansion and deduplication to construct a diverse and balanced annotation pool.

Seed Word Selection and Iterative Expansion. As a first step, we curated an initial set of seed words representing the target classes: abusive, non-abusive, political, religious, etc. These seed words included not only derogatory terms but also common words, political party names, ethnic groups, and religious terms, totaling 61 terms across all categories. We then expanded this seed set through a three-stage iterative search in the embedding space, designed to maximize lexical diversity while maintaining semantic relevance. In the first stage, for each seed term w_0, we retrieved its 50 nearest neighbors, retaining only those that are morphologically distinct from w_0 (i.e., not simple inflections). In the second stage, for each term obtained in Stage 1, we retrieved 25 additional nearest neighbors, filtering out simple derivations from the source terms. In the third stage, we retrieved 10 nearest neighbors for each term from Stage 2, applying the same distinctness criteria.

This iterative expansion process yielded 8,728 diverse and representative terms. We then selected 15K comments covering the expanded term list, with each term appearing in at least two comments, ensuring coverage across all categories. To this set, we added 5K randomly sampled comments from the remaining corpus as a control group. Our approach achieved a substantially higher type-to-token ratio of 0.28 compared to 0.13 for pure random sampling, resulting in a balanced pool of 20K comments ready for human annotation.

2.3 Data Annotation

We hired nine native speakers as annotators, four females and five males, between the ages of 22 to 48 years. The annotators were asked to label each comment for three tasks: abusiveness, sentiment, and topic. For **Abusiveness**, comments are categorized into two classes (abusive or not abusive), providing the primary classification target. The **Sentiment** dimension adds emotional context with four possible classifications (positive, neutral, negative, or mixed). Finally, the **Topic** classification assigns each comment to one of five categories (political, racial, sexist, religious, or miscellaneous topics), capturing the subject matter of the context. The annotation schema of the three tasks is depicted in Figure 1a. The annotation campaign was conducted in a controlled setting with informed consent from each participant, and a set of instructions and annotation guidelines was provided to ensure the quality and consistency of the dataset. By the end of the process, we collected annotations for 13,717 comments, and we measured the inter-annotator agreement as discussed in Section 2.4.

Generating Video Descriptions. To provide richer contextual information for analysis, we extended the dataset with descriptions of the visual content in the videos corresponding to comments in the evaluation splits. These descriptions enable researchers to investigate potential relationships between video content and abusive language, such as whether certain visual elements might trigger hostile comments. We generated these descriptions using the Qwen-2.5-VL 3B [12] and refined them with

Table 1: TiALD Dataset: Distribution of the Three Tasks and Dataset Splits.

Task	Label	Train	Test	Dev	Samples
Abusiveness	Abusive	6,980	450	250	7,680
	Not Abusive	5,337	450	250	6,037
Sentiment	Positive	2,433	226	108	2,767
	Neutral	1,671	129	71	1,871
	Negative	6,907	474	252	7,633
	Mixed	1,306	71	69	1,446
Topic	Political	4,037	279	159	4,475
	Racial	633	113	23	769
	Sexist	564	78	21	663
	Religious	244	157	11	412
	Others	6,839	273	286	7,398
	Total	**12,317**	**900**	**500**	**13,717**

GPT-4o [13] to ensure quality and consistency, creating a unique multimodal dimension for studying contextual factors in online abuse detection. See Appendix D for the model instructions and other details used in this step.

2.4 Inter-Annotator Agreement (IAA)

To compute IAA scores, we randomly sampled 100 comments from each annotator's contributions (a total of 900 samples) and then asked each of the nine annotators to provide secondary labels to comments that were not initially annotated by them. Then each comment in the sample was rated by two different annotators, resulting in a total of 1,800 annotations for each of the three tasks. This approach enabled us to compute the agreement between the pairs of contributors using Cohen's Kappa [14]. The aggregate scores for each task are: $\kappa = 0.758$ for *Abusiveness*, $\kappa = 0.649$ for *Sentiment*, and $\kappa = 0.603$ for *Topic* annotations. According to Cohen's interpretation, these scores indicate a *substantial agreement* for abusiveness detection and sentiment analysis annotations (i.e., $0.61 \leq \kappa \leq 0.80$) and a *moderate agreement* for topic classification (i.e., $0.41 \leq \kappa \leq 0.60$). We assessed 25 random comments that were assigned different Topic labels by the annotators and found that most disagreements were due to the potential applicability of the comments to multiple topics.

2.5 Gold-label Adjudication for Evaluation

To construct a high-quality test set, we extended the double-annotation process to three annotators and determined a gold label for each of the 900 samples. Our analysis showed that the initial two annotators achieved a full agreement on 546 comments, while they disagreed on at least one of the three tasks for the remaining 354 samples. To adjudicate the differences, we hired two additional experts who reviewed the inconsistent annotations of the initial annotators and decided on the final labels, which are considered as the *gold* labels in our test set.

Finally, we partitioned the remaining single-annotated samples into two splits for training and validation (500 samples), maintaining stratified proportions of the classes for abusiveness, sentiment, and topic in each split. Table 1 presents the detailed sample distribution across the splits, and Figure 1b depicts the overall class distribution of the TiALD dataset for the three tasks.

3 Experimental Setup of Baselines

To establish strong baselines for abusive language detection in Tigrinya, we evaluate three complementary approaches. First, we conduct single-task fine-tuning by training and evaluating several pre-trained language models (PLMs) on each classification task individually. Second, we implement a multi-task joint learning framework, where a shared encoder model simultaneously learns all three tasks through task-specific output heads. Finally, we assess the zero- and few-shot capabilities of state-of-the-art generative LLMs on the abusiveness detection task using prompts.

3.1 Single-task Fine-tuning

For each task in TiALD, we fine-tune several monolingual and multilingual PLMs that offer varying levels of adaptation to Tigrinya and other African languages. These include: monolingual models **TiRoBERTa** (125M) and **TiELECTRA** (14M) [15] trained exclusively on Tigrinya texts; multilingual models **AfriBERTa-base** (112M) [16] and **AfroXLMR-Large-76L** (560M) [17], pre-trained on 11 and 76 African languages, respectively; and **XLM-RoBERTa-base** (279M) [18], a general-purpose multilingual model pre-trained on 100 languages, as a control.

3.2 Joint Multi-task Training

In our joint learning setup, we employ a single transformer encoder shared across the three tasks that simultaneously learns to categorize the input content according to all relevant labels. Formally, let $\mathbf{h} = \mathrm{Encoder}(x) \in \mathbb{R}^d$ denote the contextualized representation of an input comment x, given by the final hidden state corresponding to the model's classification token (e.g., '[CLS]' for BERT or '<s>' for RoBERTa). A single linear classification head then maps $\mathbf{h}$ to a vector of logits:

$$\mathbf{z} = W\,\mathbf{h} + \mathbf{b}, \quad W \in \mathbb{R}^{L\times d},\ \mathbf{b} \in \mathbb{R}^{L},$$

where $L = 2 + 4 + 5 = 11$ is the total number of labels covering (i) abusiveness (binary), (ii) sentiment (4-way), and (iii) topic (5-way) tasks in the TiALD dataset. Each logit z_j is passed through a sigmoid to produce the probability of label j; a threshold of 0.5 is applied during inference to obtain binary predictions:

$$\hat{y}_j = \sigma(z_j), \quad j = 1, \ldots, L.$$

Training minimizes the average binary cross-entropy (BCE) loss over all label-example pairs:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\Big[y_{ij}\,\log\hat{y}_{ij}\ +\ (1-y_{ij})\,\log\big(1-\hat{y}_{ij}\big)\Big],$$

where N is the total number of examples in the training set and $y_{ij} \in \{0, 1\}$ indicates the presence of label j in the ith example.

This hard-parameter-sharing approach treats each label as an independent binary predictor while leveraging a shared representation across all three tasks, encouraging the model to capture features that benefit abusive language detection, topic classification, and sentiment analysis simultaneously. In our baseline setup, all labels contribute equally to the loss; future work may explore per-task or per-class weighting to address label imbalance.

Model training settings. For single-task and multi-task fine-tuning experiments, we set the maximum input length to 256 tokens when using comment text only and 384 tokens when incorporating video titles. We use a learning rate of $2e^{-5}$, a batch size of 16, and train for a maximum of six epochs with early stopping based on validation macro F1 score (patience of 3). We employ the AdamW optimizer [19] and implement our training system with PyTorch [20] and the Hugging Face Transformers library [21].

3.3 In-context Learning of LLMs

To assess the capabilities of state-of-the-art generative LLMs on Tigrinya abusive language detection, we employ prompt-based zero- and few-shot in-context learning [22, 23]. We design prompt templates that include either no examples (zero-shot) or a small set of annotated examples (few-shot) randomly sampled from the training set. We evaluate two commercial frontier models, **GPT-4o** [13] and **Claude Sonnet 3.7** [24],[3] and two smaller open-weight models, **LLaMA-3.2 3B** [25] and **Gemma-3 4B** [26], for comparison. To account for variability, we run the predictions twice and compute the average scores. All input comments are preserved in their original script (Ge'ez, Latin, or mixed), and the prompt explicitly mentions that the comment is in Tigrinya. See Appendix C for the full template of the instructions used in our experiments.

[3] OpenAI's GPT-4o (gpt-4o-2024-08-06) and Anthropic's Sonnet 3.7 (claude-3-7-sonnet-20250219).

Table 2: Performance of fine-tuned encoder models (single and multi-task) and prompted generative LLMs (zero-shot and few-shot) evaluated on user comments across all three tasks. The *TiALD Score* is the average macro F1 across the three tasks. Overall task-level best scores are in **bold**; category-best scores are underlined.

Model	Abusiveness	Sentiment	Topic	TiALD Score
Fine-tuned Single-task Models				
TiELECTRA-small	82.33	42.39	26.90	50.54
TiRoBERTa-base	**86.67**	52.82	54.23	64.57
AfriBERTa-base	83.42	50.81	53.20	62.48
Afro-XLMR-Large-76L	85.20	**54.94**	51.42	63.86
XLM-RoBERTa-base	81.08	30.17	43.97	51.74
Fine-tuned Multi-task Models				
TiELECTRA-small	84.21	43.44	29.27	52.30
TiRoBERTa-base	86.11	53.41	**54.91**	**64.81**
AfriBERTa-base	83.66	50.19	53.49	62.45
Afro-XLMR-Large-76L	85.44	54.50	52.46	64.13
XLM-RoBERTa-base	79.87	45.40	35.50	53.59
Zero-shot Prompted LLMs				
GPT-4o	71.05	20.55	26.25	39.28
Claude Sonnet 3.7	59.20	22.64	25.25	35.70
Gemma-3 4B	59.35	29.47	35.24	41.35
LLaMA-3.2 3B	49.98	25.30	16.55	30.61
Few-shot Prompted LLMs				
GPT-4o	72.06	21.88	27.56	40.50
Claude Sonnet 3.7	79.31	23.39	27.92	43.54
Gemma-3 4B	58.37	30.46	39.49	42.78
LLaMA-3.2 3B	45.65	19.94	21.68	29.09

Evaluation Metrics. We report Macro F1 as the primary task-level metric. We prioritize F1 over accuracy due to the inherent class imbalance and the multi-class nature of the sentiment (4-way) and topic (5-way) classification tasks. To facilitate holistic comparison at the benchmark level, we introduce the *TiALD Score*, defined as the average of the task-level macro F1 scores. We supplement these metrics with per-class F1 scores to enable granular analysis of model performance, particularly on minority classes.

Our experimental setup provides strong baselines for comparing single and multi-task fine-tuning against in-context learning of LLMs for abusive language detection under low-resource settings.

4 Results and Analysis

4.1 Performance of Fine-tuned Encoder Models

Our experimental results demonstrate that jointly training models on all three tasks consistently enhances performance over the single-task approaches. This improvement suggests that abusiveness, sentiment, and topic share complementary linguistic features that benefit from unified representation learning. As shown in Table 2, Tigrinya-specific models outperform general multilingual alternatives. TiRoBERTa-base achieves the highest macro F1 scores across most settings (86.67% for abusiveness detection, 54.23% for topic classification), demonstrating the value of language-specific pre-training. The Africa-centric AfroXLMR-76L model performs competitively, particularly for sentiment analysis, where it reaches the highest F1 score of 54.94%, suggesting that well-adapted multilingual models can reach monolingual performance.

Multi-task joint learning improves performance across almost all models and tasks, with the most substantial gains observed for TiELECTRA-small (+1.76 percentage points in overall TiALD score) and XLM-RoBERTa-base (+1.85 points). The consistent improvement across diverse model architectures confirms that the complementary signals in the TiALD tasks can be leveraged through parameter sharing. Notably, XLM-RoBERTa-base consistently underperformed compared to both the Tigrinya-specific and Africa-adapted models, with a substantial 12 percentage point average gap

Table 3: Performance of models with video title as context. Fine-tuned models were trained on concatenation of user comment and video title. LLMs were prompted with both comment and video title. Overall task-level best scores are in **bold**; category-best scores are underlined.

Model	Abusiveness	Sentiment	Topic	TiALD Score
Fine-tuned Single-task Models				
TiELECTRA-small	81.67	39.40	27.81	49.62
TiRoBERTa-base	**86.17**	**54.97**	**54.55**	**65.23**
AfriBERTa-base	82.44	51.33	52.10	61.96
Afro-XLMR-Large-76L	84.20	52.64	54.11	63.65
XLM-RoBERTa-base	75.09	43.47	41.60	53.39
Zero-shot Prompted LLMs				
GPT-4o	75.59	41.03	55.52	57.38
Claude Sonnet 3.7	67.64	44.39	50.10	54.05
Gemma-3 4B	58.41	29.27	34.44	40.71
LLaMA-3.2 3B	44.13	21.85	15.91	27.30
Few-shot Prompted LLMs				
GPT-4o	75.89	45.50	58.59	59.99
Claude Sonnet 3.7	80.29	48.01	59.45	62.58
Gemma-3 4B	59.39	30.43	39.60	43.14
LLaMA-3.2 3B	48.29	20.19	20.20	29.56

on the aggregate macro F1 scores. This performance disparity highlights the limitations of general multilingual models when applied to low-resource languages with unique linguistic characteristics.

4.2 Performance of Large Language Models

The results in Table 2 reveal that even state-of-the-art LLMs struggle to match the performance of small fine-tuned models on Tigrinya abusive language detection. GPT-4o achieves 71.05% F1 with zero-shot prompting, which, while impressive, still falls 15 percentage points behind the tuned TiRoBERTa-base. The performance gap between zero-shot and few-shot settings varies dramatically across models. Claude Sonnet 3.7 shows substantial improvement (59.20% $\rightarrow$ 79.31%), demonstrating high sensitivity to in-context examples. By contrast, the smaller open-weight models exhibit severe limitations in understanding Tigrinya text. LLaMA-3.2 3B shows classification bias that reverses across prompting conditions: in zero-shot settings, it classified 68% of comments as *abusive*, while in few-shot settings it conversely assigned 77% of them to *not abusive*, resulting in F1 scores of 49.98% and 45.65%, respectively. This inconsistency highlights the fundamental challenges current LLMs face when processing low-resource languages outside their primary training distribution.

Task-Specific Performance Gaps. Further analysis of the scores in Table 2 reveals a critical finding: while fine-tuned models maintain strong performance across all tasks (52-87% F1), LLMs exhibit severe degradation on multi-class sentiment and topic classification. The best LLM achieves only 30.46% F1 on sentiment and 39.49% on topic, showing significant deficits of 24 and 15 percentage points respectively compared to fine-tuned models. This disparity persists despite competitive LLM performance on binary abusiveness detection (71-79% F1), suggesting that current LLMs fundamentally struggle with fine-grained multi-class classification in low-resource settings. Interestingly, the smaller Gemma-3 4B outperforms frontier models on sentiment and topic tasks, indicating that model scale alone cannot overcome these limitations.

4.3 Impact of Contextual Information on Performance

Social media comments often respond to the original post, making contextual information valuable for understanding them. When video titles are added as context, the performance improves for most models, as shown in Table 3. TiRoBERTa-base trained on comments and video titles achieves the highest overall performance (65.23% TiALD score), with a significant gain (+1.5 points) observed in sentiment classification. This suggests that the video context provides topical cues that help disambiguate the intent and subject of the comments. While the generated video descriptions provided

Table 4: Performance of LLMs on Abusiveness Detection with Cross-Modality Contextual Information: user `comment` augmented with `video_title` and auto-generated `video_description`. Best scores for each prompting approach are in **bold**; highest scores within model category are underlined.

	Comment Only		Video Title + Comment		Video Title + Desc. + Comment	
	Zero-shot	**Few-shot**	**Zero-shot**	**Few-shot**	**Zero-shot**	**Few-shot**
Closed Frontier Models						
GPT-4o	**71.05**	72.06	**75.59**	75.89	**74.70**	74.53
Claude Sonnet 3.7	59.20	**79.31**	67.64	**80.29**	72.02	**78.21**
Open-weight Models						
Gemma-3 4B	59.35	58.37	58.41	59.39	54.84	50.95
LLaMA-3.2 3B	49.98	45.65	44.13	48.29	48.64	29.44

valuable contextual signals for LLMs with their large context windows (Table 4), incorporating this long-form text into fine-tuned encoder models was not feasible due to their limited input length constraints of 256-512 tokens.

More detailed class-level breakdown of model performances can be found in Appendix. Tables 5 and 6 present per-class F1 scores across all experimental settings.

4.4 Analysis and Insights

Cross-Task Performance Analysis. Performance analysis reveals that models achieve higher F1 scores for detecting abusive content (79-86%) compared to sentiment analysis (30-54%) and topic classification (26-54%). This pattern likely reflects the multi-class nature of sentiment and topic annotations, increasing the difficulty of the tasks, as evidenced by the lower inter-annotator agreement scores. The best-performing model, TiRoBERTa-base with the joint learning setup, demonstrates a relatively balanced increase across the sentiment (+0.6) and topic (+0.7) tasks, but both remain challenging with only 53.41% and 54.91% macro F1 scores, respectively. This performance gap presents an opportunity for future research to develop more specialized approaches to sentiment and topic understanding in morphologically complex languages like Tigrinya.

Effectiveness of Iterative Seed-Expansion Sampling. Compared to conventional fixed-vocabulary methods, our iterative seed-expansion sampling approach generates a more diverse and representative annotation pool while preserving lexical diversity, which is a crucial factor for languages with highly inflectional morphology where words can take numerous surface forms. Quantitative analysis reveals that our approach yielded a higher type-to-token ratio (27.6%) compared to keyword-based sampling using existing toxic word lists (18.2%) and the source corpus (7.2%) of all the 4.1M comments. Furthermore, we analyzed the ratio of *abusive* class annotations for the comments from our iterative sampling against those in the control group via random sampling, and we observed a significant difference, 65.2% vs. 14.3%, respectively. The random sampling is biased towards the majority type and hence leads to more benign non-abusive comments, while the keyword sampling is biased towards the seed words and fails to produce a diverse pool of samples.

Cross-Modality Context for Abusiveness Detection in LLMs. As shown in Table 4, the performance of the frontier LLMs improves when the user comments are enriched with contextual information (i.e., video titles and the auto-generated video descriptions). GPT-4o performs the best in the zero-shot settings, gaining 3.65 percentage points. Similarly, Claude Sonnet 3.7 shows substantial improvement in zero-shot performance (+12.82 points) when provided with video context. These gains underscore the importance of contextual understanding in accurately identifying abusive content, particularly when the language itself presents challenges for the models. The consistent improvement across settings confirms our hypothesis that supplementary video context provides valuable signals for content moderation in low-resource languages.

5 Related Work

Datasets for Abusive Language Detection. Numerous datasets have been created for the purpose of training and evaluating models for abusive language detection in English, typically sourced from

online platforms such as Twitter, Reddit, YouTube, and Wikipedia [27–31]. Researchers have also developed datasets for languages other than English, such as the Dutch-Bully-Corpus [32] consisting of over 85,000 abusive posts, and the Arabic dataset by Mubarak et al. [33] containing 1,100 tweets and 32,000 YouTube comments. Moreover, shared tasks such as OffensEval [5, 34], GermEval 2018 [35], and HASOC 2019 [36] have provided multilingual datasets for such tasks. Pavlopoulos et al. [37] created the Greek-Gazzetta-Corpus, consisting of 1.6 million comments, and Song et al. [38] proposed a comprehensive abusiveness detection dataset with multifaceted labels from Reddit. Furthermore, Tonneau et al. [39] introduced a dataset for hate speech detection containing 35,976 tweets, comprising instances of mixed languages such as English, Pidgin, Hausa, and Yoruba.

Approaches to Abusive Language Detection. Approaches to abusive language detection have evolved from statistical models to deep neural architectures, with transformer-based models like BERT, RoBERTa, and the GPT-family, establishing the state-of-the-art performance across multiple benchmarks [40–45]. For low-resource languages, cross-lingual transfer from multilingual models offers significant benefits [46, 34]. Recent advances applied generative Large Language Models (LLMs) with in-context learning and data augmentation of abusive language [22, 47]. For instance, Shin et al. [48] demonstrated that GPT-generated synthetic data enhanced offensive language detection in Korean. Zhang et al. [49] proposed an approach of Bootstrapping and Distilling LLMs for toxic content detection using a novel Decision-Tree-of-Thought prompting. Jaremko et al. [50] evaluated the performance of LLMs for implicitly abusive language detection using zero-shot and few-shot learning approaches, and analyzed the models' ability to extract relevant linguistic features.

Multi-Task Learning for Abusiveness Detection. Recent studies have demonstrated the efficacy of multi-task learning (MTL) in enhancing abusive language detection by jointly modeling related tasks. For instance, Dai et al. [51] employed a BERT-based MTL framework to simultaneously address offensive language detection, categorization, and target identification, achieving promising results. Similarly, Zhu et al. [52] integrated Prompt tuning [22, 53] with MTL to improve detection performance across multiple datasets. Mnassri et al. [54] leveraged MTL to jointly model hate speech and emotions, showing that sentiment information improves hate speech detection performance. In a related approach, Rajamanickam et al. [55] demonstrated that joint learning of toxicity and sentiment leads to more robust models than single-task approaches.

Tigrinya Language Processing. Tigrinya (ISO 639-3: `tir`) is a Semitic language of the Afro-Asiatic family that shares linguistic features with Amharic and Tigre, and uses the Ge'ez script as a writing system [9]. There is a growing interest in computational approaches to Tigrinya, such as machine translation, part-of-speech tagging, sentiment analysis, text classification through transfer learning [56–61, 9]. More recent progress on question answering and named entity recognition [62, 63] has been enabled by datasets and pre-trained language models [64, 15, 16].

Resources for African Languages. In a related effort, Ayele et al. [65] developed a dataset for Amharic, which contains 8,258 tweets annotated for hate/offensive category, target type, and intensity on a continuous scale. Contemporary to our work, Muhammad et al. [7] introduced AfriHate, a large-scale Twitter-based benchmark dataset for hate and abusive language in 15 African languages, including a Tigrinya (tir) subset with 5,072 tweets annotated for abusiveness and target types. The Tigrinya slice shows a class imbalance typical of keyword-retrieval pipelines and a moderate inter-annotator agreement. Baseline experiments on AfriHate demonstrate that multilingual fine-tuned models reach a macro F1 of 74.5%, outperforming few-shot prompting of GPT-4o [13] by 18 points. In TiALD, we employ a semi-automated data-driven sampling strategy designed to diversify the annotation pool, resulting in a more challenging and representative benchmark. We sourced data from YouTube due its substantially higher popularity among Tigrinya speakers in Eritrea and Ethiopia compared to Twitter (X) and other social media platforms.[4] Furthermore, YouTube comments are accompanied by rich context, such as video title and description, beneficial for modeling abusive content detection. Thus, TiALD and AfriHate are complementary resources for the research community with different characteristics. To the best of our knowledge, there is no prior resource that simultaneously addresses abusiveness, sentiment, and topic under low-resource settings.

[4] YouTube vs. Twitter (X) usage in Eritrea (36.4% vs. 9.1%) and Ethiopia (13.1% vs. 6.8%) since 2018: `https://gs.statcounter.com/social-media-stats/all/eritrea/#monthly-201801-202503` and `https://gs.statcounter.com/social-media-stats/all/ethiopia/#monthly-201801-202503`.

6 Conclusion

In this work, we introduced the first large-scale multi-task benchmark dataset for abusive language detection in Tigrinya. **Ti**grinya **A**busive **L**anguage **D**etection (TiALD) dataset comprises 13,717 user comments from YouTube videos, annotated by native speakers for three tasks: abusiveness, sentiment, and topic classification, providing a rich resource for understanding and detecting harmful content in this understudied language's social media. Our comprehensive experiments establish strong baselines while revealing substantial room for improvement across all tasks, with performance gaps of 15-45 percentage points from perfect classification, indicating that TiALD will serve as a challenging benchmark for future research. Our analysis reveals three key insights: first, we demonstrate that small, specialized Tigrinya models substantially outperform the current frontier models in the low-resource setting; second, we show that joint multi-task learning in aggregate outperforms single-task approaches, indicating that the abusiveness, sentiment, and topic tasks share complementary signals; third, incorporating auxiliary visual content descriptions further enhances abusiveness detection performance. We make the TiALD dataset and trained models publicly available to advance content moderation research for the Tigrinya-speaking community and to serve as a blueprint for similar efforts in other low-resource languages, promoting more inclusive and effective online safety systems.

Limitations

Explicit *vs.* Implicit Abusiveness: As the first study of its kind for Tigrinya, our work focuses on overt forms of offensive language. We acknowledge that implicit forms of toxicity, such as microaggressions and subtle forms of prejudice, also contribute to online harassment and should be addressed in future work. **Granular Annotation of Abusiveness:** Our dataset includes a single label for abusiveness, which may not capture the full range of abusive language. Future work could look into a more granular annotation scheme that captures nuanced subtypes of abusive language.

Ethics Statement

This research adheres to the academic and professional ethics guidelines of our institution, obtaining its Institutional Review Board (IRB) approval.[5] All data collection and annotations were conducted with informed consent of the participants. While the development of abusive content detection systems has the potential to improve the online experience of millions of social media users worldwide, it is crucial to consider the possible societal and ethical implications of such research. **Fairness and bias mitigation:** We carefully designed data collection and annotation procedures to minimize biases and avoid reinforcing stereotypes in the dataset and baseline models. **Respecting privacy:** We adhere to strict privacy guidelines while collecting and using user-generated data, ensuring that any personal information remains anonymized and protected. **Balancing moderation and expression:** We recognize the tension between detecting harmful content and protecting free expression, emphasizing that systems built on our dataset should incorporate transparent, accountable processes to minimize over-censorship. By addressing these considerations and making our dataset publicly available, we aim to contribute to safer online environments for Tigrinya speakers while providing a foundation for ethical content moderation research.

Acknowledgments and Disclosure of Funding

We would like to appreciate the language communities that contributed to this work through annotation, validation, and feedback. We also thank the anonymous reviewers for their valuable input. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208054). This work was also supported by the GeezLab Research Program funded by GeezLab.com (No. 2023-001).

[5] Approval number: KH2022-133

References

[1] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL https://aclanthology.org/W17-1101/.

[2] Maeve Duggan. Online Harassment. Technical report, Pew Research Center, Washington, DC, July 2017. URL https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2017/07/PI_2017.07.11_Online-Harassment_FINAL.pdf. Accessed on 2025-04-14.

[3] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), July 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

[4] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May 2017. doi: 10.1609/icwsm.v11i1.14955. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

[5] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL https://aclanthology.org/S19-2010/.

[6] Holly Lopez and Sandra Kübler. Context in abusive language detection: On the interdependence of context and annotation of user comments. *Discourse, Context and Media*, 63:100848, 2025. ISSN 2211-6958. doi: https://doi.org/10.1016/j.dcm.2024.100848. URL https://www.sciencedirect.com/science/article/pii/S2211695824000941.

[7] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, Samuel Rutunda, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Seid Muhie Yimam, and Nedjma Ousidhoum. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.92/.

[8] David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 28 edition, 2025. Online version: www.ethnologue.com.

[9] Fitsum Gaim, Wonsuk Yang, and Jong C. Park. GeezSwitch: Language identification in typologically related low-resourced East African languages. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6578–6584, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.707/.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems*, 2013. URL https://api.semanticscholar.org/CorpusID:16447573.

[11] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[12] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. Technical report, Alibaba Cloud, 2025. URL https://arxiv.org/abs/2412.15115.

[13] OpenAI. GPT-4o System Card. Technical report, OpenAI, 2024. URL https://arxiv.org/abs/2410.21276. arXiv:2410.21276.

[14] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960. URL https://api.semanticscholar.org/CorpusID:15926286.

[15] Fitsum Gaim, Wonsuk Yang, and Jong C. Park. Monolingual pre-trained language models for tigrinya. *WiNLP 2021*, 2021.

[16] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11/.

[17] David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.14/.

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

[19] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. In *ICLR*, 2018.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Mark Hesse, Eric Chen, Mateusz Sigler, Scott Litwin, Benjamin Gray, Jack Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[23] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[24] Anthropic. Claude 3.7 Sonnet model. Technical report, Anthropic, 2025. Retrieved April 20, 2025.

[25] Meta AI. The LLaMA 3 Herd of Models. Technical report, Meta AI, 2024. URL arXivpreprintarXiv:2407.21783.

[26] Gemma Team. Gemma 3 Technical Report. Technical report, Google, 2025. URL https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[27] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Proceedings of the Ninth Conference on Innovative Applications of Artificial Intelligence (IAAI-97)*, pages 1058–1065, Providence, Rhode Island, USA, 1997. AAAI Press.

[28] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://aclanthology.org/N16-2013/.

[29] Noé Cécillon, Vincent Labatut, Richard Dufour, and Georges Linarès. WAC: A corpus of Wikipedia conversations for online abuse detection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1382–1390, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.173/.

[30] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 29–30, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2742760. URL https://doi.org/10.1145/2740908.2742760.

[31] Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors, *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.nlperspectives-1.11/.

[32] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Automatic detection of cyberbullying in social media text. *PLoS ONE*, 13, 2018. URL https://api.semanticscholar.org/CorpusID:2317624.

[33] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on Arabic social media. In Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors,

Proceedings of the First Workshop on Abusive Language Online, pages 52–56, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3008. URL https://aclanthology.org/W17-3008/.

[34] Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.188. URL https://aclanthology.org/2020.semeval-1.188/.

[35] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of the GermEval 2018 Workshop, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10, Vienna, Austria, 2018.

[36] Thomas Mandl, Sandip J Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandalia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, 2019. URL https://api.semanticscholar.org/CorpusID:263876502.

[37] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. Deeper attention to abusive user content moderation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1117. URL https://aclanthology.org/D17-1117/.

[38] Hoyun Song, Soo Hyun Ryu, Huije Lee, and Jong Park. A large-scale comprehensive abusiveness detection dataset with multifaceted labels from Reddit. In Arianna Bisazza and Omri Abend, editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 552–561, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.conll-1.43. URL https://aclanthology.org/2021.conll-1.43/.

[39] Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. NaijaHate: Evaluating hate speech detection on Nigerian Twitter using representative data. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.488. URL https://aclanthology.org/2024.acl-long.488/.

[40] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. A Unified Deep Learning Architecture for Abuse Detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 105–114, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362023. doi: 10.1145/3292522.3326028. URL https://doi.org/10.1145/3292522.3326028.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[42] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[43] Kunze Wang, Dong Lu, Caren Han, Siqu Long, and Josiah Poon. Detect all abuse! toward universal abusive language detection models. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6366–6376, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.560. URL https://aclanthology.org/2020.coling-main.560/.

[44] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PLoS ONE*, 15, 2020. URL https://api.semanticscholar.org/CorpusID:221136077.

[45] Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13:20871–20892, 2025. doi: 10.1109/ACCESS.2025.3532397.

[46] Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. Aggression-annotated corpus of Hindi-English code-mixed data. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1226/.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Edouard Grave, Armand Joulin, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. URL https://arxiv.org/abs/2302.13971.

[48] Jisu Shin, Hoyun Song, Huije Lee, Fitsum Gaim, and Jong Park. Generation of Korean offensive language by leveraging large language models via prompt design. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 960–979, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.62. URL https://aclanthology.org/2023.ijcnlp-main.62/.

[49] Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. Efficient toxic content detection by bootstrapping and distilling large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21779–21787, 2024.

[50] Julia Jaremko, Dagmar Gromann, and Michael Wiegand. Revisiting implicitly abusive language detection: Evaluating LLMs in zero-shot and few-shot settings. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3879–3898, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.262/.

[51] Wenliang Dai, Tiezheng Yu, Zihan Liu, and Pascale Fung. Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2060–2066, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.272. URL https://aclanthology.org/2020.semeval-1.272/.

[52] Jian Zhu, Yuping Ruan, Jingfei Chang, and Cheng Luo. Deep prompt multi-task network for abuse language detection. *ArXiv*, abs/2403.05268, 2024. URL https://api.semanticscholar.org/CorpusID:268296964.

[53] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

[54] Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. Hate speech and offensive language detection using an emotion-aware shared encoder. In *ICC 2023 - IEEE International Conference on Communications*, pages 2852–2857, 2023. doi: 10.1109/ICC45041.2023.10279690.

[55] Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Joint modelling of emotion and abusive language detection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.394. URL https://aclanthology.org/2020.acl-main.394/.

[56] Yemane Tedla and Kazuhide Yamamoto. The effect of shallow segmentation on english-tigrinya statistical machine translation. In *2016 International Conference on Asian Language Processing (IALP)*, pages 79–82. IEEE, 2016.

[57] Yemane Keleta Tedla. Tigrinya morphological segmentation with bidirectional long short-term memory neural networks and its effect on english-tigrinya machine translation. *Nagaoka University of Technology: Niigata, Japan*, 2018.

[58] Alp Öktem, Mirko Plitt, and Grace Tang. Tigrinya neural machine translation with transfer learning for humanitarian response. *arXiv preprint arXiv:2003.11523*, 2020.

[59] Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. An exploration of data augmentation techniques for improving english to tigrinya translation. *ArXiv*, abs/2103.16789, 2021.

[60] Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. Transferring monolingual model to low-resource language: The case of tigrinya, 2020.

[61] Awet Fesseha, Shengwu Xiong, Eshete Derb Emiru, Moussa Diallo, and Abdelghani Dahou. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Inf.*, 12:52, 2021.

[62] Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.661. URL https://aclanthology.org/2023.acl-long.661/.

[63] Sham K. Berhane, Simon M. Beyene, Yoel G. Teklit, Ibrahim A. Ibrahim, Natnael A. Teklu, Sirak A. Bereketeab, and Fitsum Gaim. Towards neural named entity recognition system in tigrinya with large-scale dataset. *Springer Journal of Language Resources and Evaluation*, 2025. URL https://doi.org/10.21203/rs.3.rs-4485676/v1.

[64] Fitsum Gaim, Wonsuk Yang, and Jong C. Park. TLMD: Tigrinya Language Modeling Dataset. *Zenodo*, July 2021. doi: 10.5281/zenodo.5139094. URL https://doi.org/10.5281/zenodo.5139094.

[65] Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. Exploring Amharic hate speech data collection and classification approaches. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.6/.

Appendix

A Class-level Performance of Baseline Models on the TiALD Benchmark

We analyze model classification biases through the lens of performance variance across the three tasks in TiALD and data subsets, revealing critical insights for content moderation in low-resource settings. The class-level results expose disparities that have important implications for real-world deployment.

A.1 Analysis of Fine-tuned Encoder Models

Table 5 presents the per-class F1 scores for all three tasks. For abusiveness detection, performance is relatively balanced between *abusive* and *not abusive* classes, with TiRoBERTa-base achieving the highest scores (86.52% and 86.81% in the single-task setting). However, sentiment and topic classification reveal severe class imbalances that reflect both natural data distribution and the inherent difficulty of nuanced content classification.

For sentiment analysis, all models exhibit strong bias toward the *negative* class (achieving up to 81.32% F1) while dramatically underperforming on *neutral* and *mixed* classes (as low as 1.53% and 0% F1). This disparity suggests that models default to negative sentiment when uncertain, potentially over-flagging neutral content in production systems.

Topic classification shows similar patterns, with models achieving strong performance on *political* content (up to 71.21% F1) but substantially weaker results on minority classes. Most concerning are the near-zero F1 scores for *racial* (6.50%), *sexist* (0%), and *religious* (0%) categories in some single-task configurations, indicating complete failure to identify these sensitive content types.

Critically, multi-task joint learning demonstrates significant bias mitigation. TiRoBERTa-base's F1 score on *sexist* content improves dramatically from 31.78% to 46.30% (+14.52 points) with joint learning. Similarly, performance on *neutral* sentiment increases from 40.0% to 42.75%, and *religious* content shows recovery from complete failure in TiELECTRA (0% to 7.36%). These improvements demonstrate that the complementary signals across tasks help models better recognize minority classes, a crucial benefit for equitable content moderation.

Table 5: Class-level Performance of Single and Multi-task settings in F1 score. Models are trained and evaluated on the comment text only. The highest class-level scores for each approach are underlined, and the overall best scores are in **bold**. Multi-task learning yields significant performance improvements for the minority classes (e.g., *sexist*, *religious*).

	Abusiveness		**Sentiment**				**Topic**				
Model	Abusive	Not Abusive	Positive	Neutral	Negative	Mixed	Political	Racial	Sexist	Religious	Other
Single-task Models											
TiELECTRA-small	82.35	82.31	62.84	4.48	80.88	21.36	67.48	06.50	00.00	00.00	60.51
TiRoBERTa-base	**86.52**	**86.81**	68.68	40.00	81.18	21.43	70.86	**46.49**	31.78	56.90	65.15
AfriBERTa-base	84.00	82.85	63.25	33.78	81.03	25.17	69.64	40.46	34.29	56.89	64.71
Afro-XLMR-Large-76L	85.74	84.66	69.44	41.06	**81.32**	27.94	70.71	39.75	28.00	54.38	64.29
XLM-RoBERTa-base	80.32	81.84	46.03	01.53	73.11	00.00	67.32	33.33	02.53	52.53	64.12
Multi-task Joint Learning											
TiELECTRA-small	84.67	83.75	62.68	14.39	81.10	15.58	65.17	06.84	07.41	07.36	59.56
TiRoBERTa-base	86.13	86.10	62.98	**42.75**	79.83	28.07	**71.21**	45.96	**46.30**	47.44	63.65
AfriBERTa-base	83.93	83.39	65.39	27.43	81.13	26.79	70.59	44.16	44.00	43.35	**65.36**
Afro-XLMR-Large-76L	85.16	85.71	**71.79**	33.88	80.96	**31.34**	69.02	36.60	35.64	**57.78**	63.27
XLM-RoBERTa-base	80.43	79.31	67.06	15.47	79.92	19.15	67.51	16.26	16.47	15.29	61.95

A.2 Performance Disparity in Large Language Models

Class-level evaluation on the TiALD tasks reveals striking performance disparities in the generative LLMs, as shown in Table 6. Despite the balanced test set (50% samples per class of abusiveness), models show severe classification imbalances. Claude Sonnet 3.7 exhibits a highly asymmetric zero-shot performance (44.18% F1 on *abusive* vs. 72.58% on *not abusive*), suggesting a bias toward classifying content as non-abusive. However, it shows dramatic improvements in the few-shot setting, achieving 79.26% and 80.69% F1 for *abusive* and *not abusive* classes, respectively, but still falls short compared to fine-tuned small models. Adding contextual information partially mitigates the

disparities but also yields mixed results. GPT-4o's F1 score for detecting *abusive* content jumps from 69.04% to 76.66% when provided with video context.

LLaMA-3.2 3B exhibits dramatic classification instability across prompting conditions, predicting 68% of comments as *abusive* in zero-shot and inversely 77% as *not abusive* in few-shot settings. This erratic behavior stems from the model's fundamental limitation to comprehend Tigrinya text. Tokenization analysis reveals LLaMA-3.2 requires 2.31 tokens per character for Tigrinya versus only 0.20 for English (an 11.5× increase), significantly impacting both accuracy and inference cost for low-resource language deployment.

Table 6: Class-level Performance of LLMs across all tasks in TiALD evaluated on the user comment. The highest class-level scores for each approach are underlined, and the overall best scores are in **bold**. Most models show severe classification biases on the balanced test set. Reported in F1 score.

	Abusiveness		**Sentiment**				**Topic**				
Model	Abusive	Not Abusive	Positive	Neutral	Negative	Mixed	Political	Racial	Sexist	Religious	Other
Zero-shot Prompted LLMs											
GPT-4o	69.04	75.89	51.03	14.55	76.16	**22.68**	62.96	33.70	27.78	75.07	63.04
Claude Sonnet 3.7	44.18	72.58	**65.85**	29.80	77.94	07.55	**66.90**	21.33	19.78	**79.00**	**65.48**
Gemma-3 4B	59.25	60.94	35.95	00.00	69.49	12.00	55.03	03.45	06.19	64.31	49.62
LLaMA-3.2 3B	59.81	42.43	09.60	13.20	64.01	13.00	00.70	10.20	00.00	22.22	48.47
Few-shot Prompted LLMs											
GPT-4o	74.82	71.41	54.40	23.92	74.32	22.38	60.85	37.12	37.87	78.72	61.02
Claude Sonnet 3.7	**79.26**	**80.69**	65.75	**33.18**	**79.65**	8.55	63.59	**43.26**	**39.05**	78.23	55.07
Gemma-3 4B	52.52	64.26	25.81	22.22	57.56	17.72	55.97	09.16	15.84	60.90	56.70
LLaMA-3.2 3B	28.30	60.16	26.67	20.29	15.52	16.13	21.14	06.76	11.43	23.40	48.76

A.3 Script-Based Robustness and Joint Annotation Benefits

Our dataset's accommodation of both Ge'ez script and Romanized text (reflecting the 64% Romanized usage in real Tigrinya social media) enables models to develop script-agnostic representations. Initial qualitative analysis indicates that models trained on this mixed-script data show more robust performance across both writing systems, though comprehensive quantitative evaluation is reserved for future work.

Furthermore, the joint annotations in TiALD enable nuanced analysis beyond binary classification. Comments labeled as both *Abusive* and *Political* can be interpreted as political hate speech, while *Abusive+Sexist* combinations identify misogynistic content. This multi-dimensional labeling provides pathways for fine-grained content moderation without requiring extensive re-annotation, addressing reviewer concerns about granularity while maintaining high annotation quality.

A.4 Implications for Low-Resource Content Moderation

The demonstrated performance disparities have critical implications for deploying content moderation systems in low-resource settings. The severe underperformance on minority classes means that certain types of harmful content (particularly sexist and religious abuse) may go undetected. Our results demonstrate that:

1. Multi-task learning provides a practical approach to mitigate these biases without additional data collection
2. Current LLMs, despite their impressive capabilities in high-resource languages, require fundamental architectural changes (particularly in tokenization) to serve low-resource languages effectively
3. Fine-tuned specialized models significantly outperform general-purpose LLMs, achieving 86.67% F1 compared to 79.26% for the best LLM configuration

B TiALD Dataset Features

Table 7 presents the descriptions of the fields in the TiALD dataset. Figure 2 depicts the task-wise class distribution across the tasks of Abusiveness, Sentiment, and Topic.

Table 7: An overview of the features included in the TiALD Dataset.

Feature	Data Type	Description
`sample_id`	String	Unique identifier for the sample in the dataset.
`comment_id`	String	Unique identifier for the comment.
`comment_original`	String	Original comment text as posted by user.
`comment_cleaned`	String	Pre-processed version of the comment text.
`abusiveness`	Categorical	Abuse label (*Abusive* or *Not Abusive*).
`sentiment`	Categorical	Sentiment label (*Positive*, *Neutral*, *Negative*, or *Mixed*).
`topic`	Categorical	Topic label (*Political*, *Racial*, *Sexist*, or *Religious*, *Other*).
`annotator_id`	String	Identifier of the annotator.
`comment_script`	Categorical	Script used in the comment (*Ge'ez*, *Latin*, or *Mixed*).
`comment_publish_date`	String	Year and month the comment was published.
`video_id`	String	Identifier of the video the comment was posted under.
`video_title`	String	Title of the associated video.
`video_num_views`	Numeric	Number of views the video received.
`video_publish_year`	Numeric	Year the video was published.
`video_description`	String	Auto-generated description of the video content.
`channel_id`	String	Identifier of the YouTube channel the video belongs to.
`channel_name`	String	Name of the YouTube channel the video belongs to.

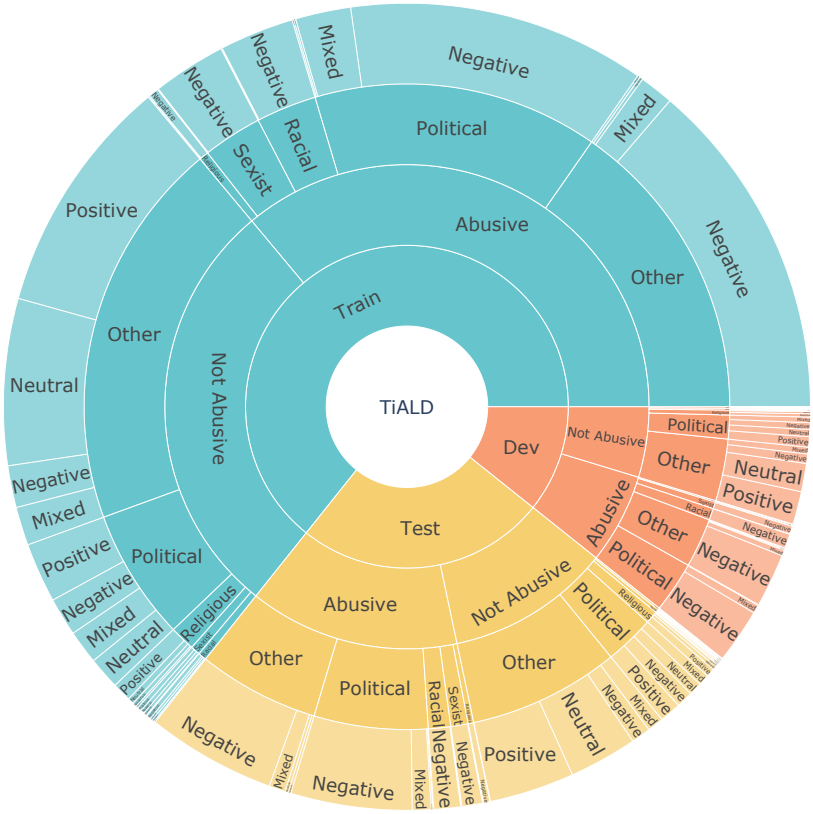


Figure 2: TiALD Class Distribution across the Splits and Tasks of Abusiveness, Sentiment, and Topic.

C Evaluation Instructions for LLMs

We conducted experiments on two commercial frontier models and two open-weight smaller models under both zero-shot and few-shot settings. All input comments were preserved in their original script (Ge'ez, Latin, or mixed), with an explicit instruction that the comment was written in Tigrinya. For the few-shot evaluation, we included four balanced examples (two abusive, two non-abusive) randomly sampled from the training set and arranged in alternating order as follows:

> [Instruction]
> `Classify the following user comment written in Tigrinya as "Abusive" or "Not Abusive". Do not provide any explanation or additional text.`
>
> [Optional In-context Examples]
> `Here are some examples:`
> `<comment text>: Abusive`
> `<comment text>: Not Abusive`
> `<comment text>: Abusive`
> `<comment text>: Not Abusive`
>
> [Comment]
> Comment: `<comment text>`

D Generating Video Content Descriptions

We used a vision-language model, Qwen2.5-VL-3B [12], to generate detailed descriptions of video content corresponding to comments in the evaluation splits of the TiALD dataset. Qwen2.5-VL handles long-form videos up to several hours by applying dynamic resolution processing and absolute time encoding. The TiALD dataset's videos average 28.1 minutes (1,686 seconds) in length and can run as long as 334.75 minutes (20,085 seconds) at 30 FPS. To ensure high-quality descriptions, we trimmed videos exceeding 20 minutes to their first 20-minute segment. Each resulting clip was then passed to the model with the following instruction:

> [Instruction]
> `Describe the content of this video frame in detail. Focus on the people, objects, actions, and settings visible in the image. Provide a comprehensive description that could help understand what the video is about.`
>
> [Video]
> `<video frames>`

To enhance the quality and consistency of the generated video descriptions, we revise them using a larger, more capable model, GPT-4o [13], with the following instruction:

> [Instruction]
> `Revise the following automatically generated video description to make it clearer and consistent, while preserving the key information. Keep your response to a moderate length, up to 150 words, and focus only on the video content.`
>
> [Video Title]
> `Video Title: <video title>`
>
> [Video Description]
> `Video Description: <video description>`

Finally, the resulting video descriptions were included in the dataset as auxiliary features to enable deeper analysis of potential relationships between video content and the abusiveness of comments.

We also empirically show the benefit of using contextual information in our experiments, as shown by the results in Table 4.

E Annotation Guidelines for TiALD Dataset

This section outlines the detailed annotation guidelines provided to the native Tigrinya speakers who participated in the TiALD dataset creation. Annotators were instructed to classify each YouTube comment across three dimensions: Abusiveness, Sentiment, and Topic, while following specific protocols for comment eligibility.

Task 1: Abusive Language Detection

Annotators were asked to determine whether a comment contained abusive language according to the following criteria:

- **Abusive**: The comment contains language that attacks, insults, demeans, or threatens an individual or group. This includes hate speech, profanity directed at others, derogatory terms, threats of violence, severe personal attacks, or language intended to humiliate or degrade.
- **Not Abusive**: The comment does not contain language that attacks an individual or group. It may express disagreement, criticism, or negative opinions without using abusive language toward others.

Task 2: Sentiment Analysis

Annotators classified the emotional tone of each comment into one of four sentiment categories:

- **Positive**: The comment expresses primarily positive emotions, approval, praise, gratitude, happiness, or optimism. This includes congratulatory messages, expressions of joy, or positive feedback.
- **Neutral**: The comment does not express a clear positive or negative sentiment. This includes factual statements, questions without emotional content, or balanced objective comments.
- **Negative**: The comment expresses primarily negative emotions, disapproval, criticism, anger, sadness, or pessimism. This includes expressions of disappointment, frustration, or negative judgments.
- **Mixed**: The comment contains a relatively balanced mix of both positive and negative sentiments, with neither clearly dominating. This includes comments expressing contrasting emotions or evaluating different aspects both positively and negatively.

Task 3: Topic Classification

Annotators classified each comment into one of five topical categories:

- **Political**: Comments discussing political figures, parties, governments, policies, elections, or expressing political opinions. This includes references to specific political events, governance issues, or politically divisive topics.
- **Racial**: Comments referring to racial or ethnic identity, characteristics, or relationships between racial/ethnic groups. This includes discussions about cultural identity tied to ethnicity.
- **Sexist**: Comments discussing gender roles, gender identity, or containing gendered language. This includes content related to expectations based on gender or discussions about gender relations.
- **Religious**: Comments discussing religious beliefs, practices, institutions, or figures. This includes references to religious texts, doctrines, religious communities, or spirituality.
- **Other**: Comments that don't primarily fall into any of the above categories. This includes everyday conversations, entertainment, personal updates, or general information sharing.

Comment Eligibility Criteria

Annotators were instructed to exclude comments from annotation if they met any of the following conditions:

1. Comments written entirely in a language other than Tigrinya (e.g., English, Amharic, Arabic)
2. Comments containing no actual Tigrinya words (e.g., consisting only of repeated characters, symbols, or emojis)
3. Comments that were unintelligible or lacked meaningful content

However, annotators were instructed to retain comments if they:

- Contained at least some Tigrinya words, even if mixed with words from other languages
- Used Romanized Tigrinya (Latin script) rather than the native Ge'ez script
- Contained code-switching between Tigrinya and other languages

These guidelines were designed to create a dataset that accurately reflects the linguistic and cultural nuances of abusive language in Tigrinya social media content. Figure 3 shows the annotation system we developed to annotate the TiALD dataset.

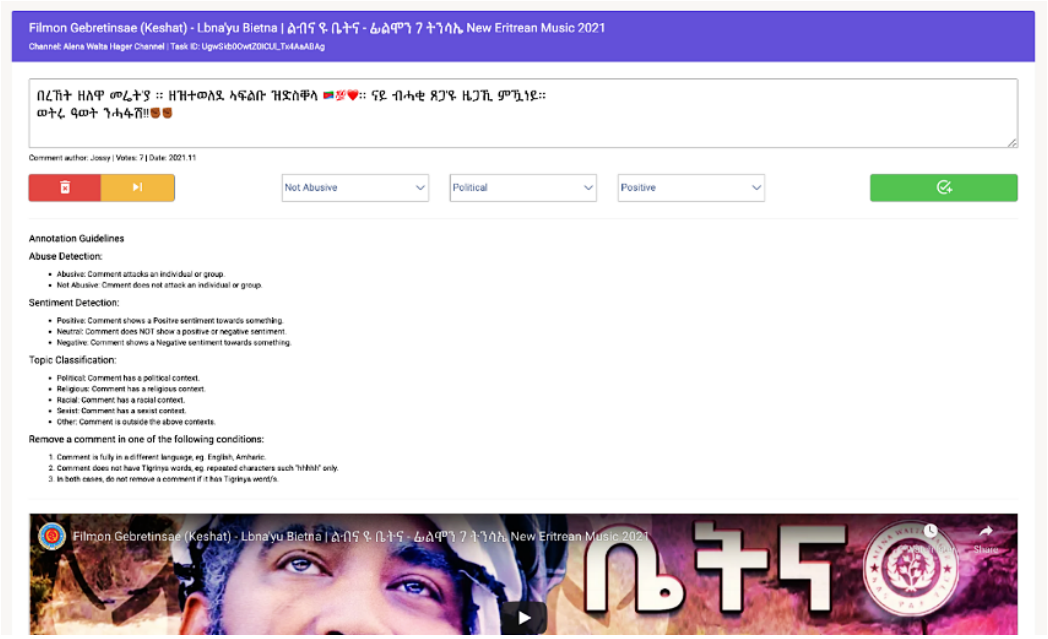


Figure 3: The annotation system we developed for the TiALD dataset. A summary of the annotation guidelines is provided on screen to encourage consistent annotations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in our abstract and introduction accurately reflect the paper's contributions and scope. The paper clearly states in the introduction that we present a multi-task benchmark dataset for abusive language detection in Tigrinya with three annotation tasks, demonstrates the effectiveness of joint learning approaches, and provides strong baselines for future research.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a dedicated "Limitations" section that acknowledges the focus on explicit rather than implicit forms of abusive language and discusses the current limitations of granularity in our abuse annotations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper focuses on dataset creation and empirical evaluations rather than presenting theoretical results that require mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide complete details on our dataset construction methodology, annotation process, and model implementation. The dataset and baseline models are available at the GitHub repository mentioned in the footnote (anonymized for review).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will make the TiALD dataset and all code for baseline models publicly available in a GitHub repository (currently anonymized in the paper for blind review). The repository includes data preprocessing scripts, model training code, and evaluation scripts with clear instructions.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 (Experimental Setup) provides comprehensive details on model architectures, hyperparameters, training procedures, and evaluation metrics. We specify all training settings including learning rates, batch sizes, number of epochs, and optimization algorithms.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experiments report inter-annotator agreement scores using Cohen's Kappa with clear explanation of their interpretation. Performance metrics (F1 scores and accuracy) are reported across all experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments (task-specific and multitask model fine-tuning) were conducted using single NVIDIA RTX A6000 GPUs with 48GB memory. Each model's training time ranged from 2-6 hours depending on model size and complexity. The total compute for all experiments reported in the paper was approximately 750 GPU hours.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and ensured our research conforms to it. Our paper includes an Ethics Statement section detailing our adherence to ethical guidelines, including obtaining IRB approval and respecting privacy considerations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: "Ethics Statement" discusses potential societal impacts of our work, including the tension between detecting harmful content and protecting free expression. We acknowledge the importance of developing transparent, accountable content moderation systems.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our released dataset has been carefully anonymized to protect user privacy by removing personally identifying information from comments. The dataset is intended for research purposes only, and we include usage guidelines that prohibit using the models for automated content removal without human oversight.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all existing assets used in our research, including pre-trained language models and evaluation metrics. All external tools and libraries are properly credited with appropriate citations in the references section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our released dataset is accompanied by comprehensive documentation including data format, class distributions, annotation guidelines, and intended uses and limitations. We provide clear license information (CC BY-NC-SA 4.0) for our dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include details about our annotation process in Section 3.2, including annotator demographics, training procedures, and compensation details. The full annotation guidelines are available in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: Our research received formal Institutional Review Board approval, as mentioned in the Ethics Statement section (approval number: KH2022-133). All data collection and annotations were conducted with the informed consent of participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM Usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLMs as part of our core methodology or experimental development. LLMs were only used for generating video descriptions as mentioned in Section 3.3, but this was an extended feature and not part of the core methodology or experiments.

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.