# SpectrumNet: Detecting LGBTQ+ Cyberbullying with Dynamic Context-Aware Attention

**Arslan Bisharat[1], Manuel Sandoval[1], Mujtaba Nazari[1], Deborah L. Hall[2], Mohammed Abuhamad[1], Yasin N. Silva[1]**
[1]Loyola University Chicago
[2]Arizona State University
{marslan, msandovalmadrigal, mnazari, mabuhamad, ysilva1}@luc.edu
d.hall@asu.edu

**Abstract**

Cyberbullying remains a critical societal issue, with LGBTQ+ individuals disproportionately affected. Although previous work proposed general cyberbullying detection models, LGBTQ+-targeted cyberbullying detection remains relatively unexplored. SpectrumNet, a novel transformer-based model introduced in this paper, goes beyond conventional cyberbullying detection by adding conversational context and identity-aware modeling. SpectrumNet freezes the RoBERTa backbone and adds three key components: a hierarchical attention network to capture linguistic nuance, a GRU-based encoder to better capture comment history, and a dynamic fusion module to effectively weigh contextual signals. To address dataset imbalance, we apply focal loss and weighted sampling. Trained on a large, annotated Instagram dataset, SpectrumNet effectively differentiates between non-bullying, general bullying, and LGBTQ+-targeted bullying. In particular, it achieves strong recall on targeted content and excels at detecting subtle forms of discrimination often missed in isolation but evident within threaded interactions.

## 1 Introduction

Cyberbullying—the use of digital media to transmit intentionally "negative, harmful, false, or mean content about someone else" (StopBullying.gov 2024)—is a growing problem. Recent findings indicate just under half of U.S. teens have experienced cyberbullying or harrassment online (Pew Research Center 2022) and cyberbullying has been directly linked to negative mental health outcomes, including an increased risk of suicide (Cook 2024; Gini and Espelage 2014; Holt et al. 2015). For LGBTQ+ individuals, rates of cyberbullying victimization are comparably higher and the risk of suicide is particularly severe (TheTrevorProject 2024b). Indeed, the relation between cyberbullying and negative mental health outcomes is especially concerning for this vulnerable population, given the cumulative impact that repeated exposure to acts of peer aggression can have (Sophie. et al. 2022).

Traditional machine learning approaches for cyberbullying detection (Cheng et al. 2019a, 2021a; Singh, Ghosh, and Jose 2017; Nandhini and Sheeba 2015; Perera and Fernando 2021) have largely followed a one-size-fits-all paradigm, failing to account for the unique characteristics of LGBTQ+-targeted cyberbullying. This oversight is particularly concerning given that LGBTQ+ individuals experience fundamentally different forms of online harassment compared to their non-LGBTQ+ peers (Plöderl and Tremblay 2015; Abreu and Kenny 2017). That is, LGBTQ+-targeted cyberbullying is distinct not only due to its higher prevalence, with rates nearly 50% greater than those among non-LGBTQ+ youth (TheTrevorProject 2024a) (31.7% vs. 21.8%) (Hinduja and Patchin 2021), but also in light of the identity-focused nature of the attacks that specifically target an individual's sexual orientation, gender identity, or gender expression (Abreu and Kenny 2018; Powell, Stratton, and Cameron 2022; Sophie. et al. 2022). For instance, LGBTQ+ identity-based attacks can include using slurs such as "*What a f*gg!*", gendered harassment such as "*Must be on ur period haven a bi**h fit man up h*e cake*", intentional misgendering or "deadnaming" (using a transgender person's former pronouns or name), "outing" (disclosing someone's LGBTQ+ identity without consent), or invalidation of one's LGBTQ+ identity *such as "You're a d*ke cause yo uglya*s can't get any man"* (Social Media Victims Law Center 2024; NordVPN 2024). Moreover, these incidents increasingly involve technology-facilitated abuse, including doxxing (publicly exposing private information), sexually explicit harassment, and physical threats that exploit LGBTQ+ individuals' vulnerabilities (Powell, Stratton, and Cameron 2022).

Notably, LGBTQ+-targeted cyberbullying creates additional psychological harm beyond typical cyberbullying by contributing to minority stress, the chronic stress experienced by marginalized groups due to prejudice, discrimination, and social rejection (Meyer 2003). Put differently, when LGBTQ+ individuals face cyberbullying that explicitly attacks their core identity, it reinforces societal stigma and can trigger internalized shame, isolation, and mental health struggles beyond what non-LGBTQ+ victims may experience from general forms of cyberbullying (Fisher, Tao, and Ford 2024). Adding to these challenges, many LGBTQ+ victims experience intersectional targeting, where multiple forms of discrimination, such as transphobia combined with racism or homophobia combined with sexism, are directed simultaneously, resulting in particularly harmful effects (Bauer et al. 2021; Angoff and Barnhart 2021; NAACP 2024). The compounded effects of such targeted aggression,

combined with pre-existing minority stress and disparities in mental health support (Duarte et al. 2018; Tanni et al. 2024), underscore the urgent need for specialized detection models. Recent research has made significant strides in addressing bias in cyberbullying detection (Cheng et al. 2021b) and using deep learning techniques to improve detection performance (Hosseinmardi et al. 2015; Arslan et al. 2024). However, there remains a critical gap in developing solutions that specifically address the unique linguistic patterns and contextual cues related to LGBTQ+-targeted cyberbullying.

In this paper, we introduce SpectrumNet, a novel transformer-based framework specifically designed to identify cyberbullying directed at LGBTQ+ individuals. Our approach moves beyond generic cyberbullying detection to create more nuanced and effective tools for protecting LGBTQ+ users from targeted online cyberbullying. Trained on a large annotated Instagram dataset comprising over $106,618$ comments with $11,310$ cyberbullying instances (including $1,053$ LGBTQ+-targeted comments), SpectrumNet adds a hierarchical attention mechanism with historical context processing while freezing the RoBERTa backbone for computational efficiency. By focusing on the specific patterns and language used in LGBTQ+ cyberbullying, our work addresses a critical gap in online safety measures for this vulnerable community.SpectrumNet's architecture aims at addressing the challenges of LGBTQ+ cyberbullying by combining three synergistic components. The frozen RoBERTa backbone provides robust language understanding to detect coded language and microaggressions. Hierarchical attention captures both word-level slurs and sentence-level sarcasm, critical for identifying subtle harassment (Cheng et al. 2019a). GRU-based encoding models sequential patterns in comment threads that are evident only in context (Chung et al. 2014). Dynamic contextual fusion adaptively weighs post content, comments, and historical context to handle complex, identity-based harassment. This combination, validated by an ablation study in Section 4, ensures high accuracy for LGBTQ+-targeted bullying while maintaining computational efficiency. Experimental results, detailed in Section 4, show SpectrumNet outperforms state-of-the-art models in detecting context-dependent LGBTQ+ bullying.

The contributions of our work are as follows.

① We introduce SpectrumNet, a novel transformer-based architecture specifically designed to detect LGBTQ+-targeted cyberbullying through the synergistic integration of hierarchical attention mechanisms with historical context processing. This combination of components directly addresses the multi-layered nature of LGBTQ+-targeted harassment, where detection requires simultaneous analysis of linguistic subtlety, conversational context, and sequential patterns unique to identity-based attacks. We show that freezing parts of pre-trained large models while adding specialized processing modules provides an effective balance between preserving pre-trained linguistic knowledge and capturing domain-specific patterns. The model's source code will be made publicly available with this publication.

② We implement a dynamic contextual fusion approach that adaptively weighs the importance of posts' content, comments, and history to improve detection accuracy.

③ We provide a comprehensive performance evaluation of our approach and compare with existing cyberbullying detection models, showing SpectrumNet's advantages in detecting LGBTQ+-targeted content. The extended dataset used for training and evaluation will be made publicly available with this publication.

**Organization.** The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 formally defines the problem, describes our dataset, and presents the methodology behind SpectrumNet. Section 4 evaluates the performance of our model against state-of-the-art cyberbullying detection models. Section 5 discusses the results, insights, and limitations. Section 6 concludes the paper.

## 2 Related Work

**Psychological Impact of LGBTQ+-targeted Cyberbullying.** Previous research has demonstrated unequivocally that LGBTQ+ individuals experience cyberbullying at significantly higher rates than their non-LGBTQ+ peers (Centre 2020; Duarte et al. 2018). Cyberbullying targeting LGBTQ+ victims can be particularly detrimental, given the well-documented mental health risks faced by these marginalized populations. Indeed, studies have shown that cyberbullying can compound existing minority stress and negatively impact identity development during formative years (Dürrbaum and Sattler 2020; Plöderl and Tremblay 2015). The increased rates of depression, anxiety, and suicidal ideation among LGBTQ+ youth who are cyberbullied (Tanni et al. 2024; Holt et al. 2015; Gini and Espelage 2014) highlight the critical need for specialized detection and intervention approaches.

**General Cyberbullying Detection Models.** The field of automated cyberbullying detection has evolved substantially over the past decade. Early approaches relied primarily on lexicon-based methods and traditional machine learning classifiers like Support Vector Machines and Naive Bayes (Nandhini and Sheeba 2015; Sathya and Harin Fernandez 2024; Özel et al. 2017). These models typically analyzed text for the presence of explicit slurs and offensive language but struggled with more subtle forms of cyberbullying. More recent advances have used deep learning architectures for improved detection, including convolutional neural networks and recurrent neural networks for toxic content detection (Badjatiya et al. 2017; Cao, Lee, and Hoang 2020; Dadvar and Eckert 2018). Hierarchical attention networks have been introduced specifically for cyberbullying detection, capturing both linguistic content and temporal interaction patterns (Cheng et al. 2019b, 2021a). Despite these advances, most general cyberbullying detection models are developed using datasets that lack sufficient representation of LGBTQ+-targeted content, which limits their effectiveness for this specific type of cyberbullying. Recent studies have explored the use of a variety of large-scale pretrained models for cyberbullying and hate speech detection. These in-

clude Twitter BERT (Devlin et al. 2019), CT-BERT (Müller, Salathé, and Kummervold 2023), HateBERT (Piot, Martín-Rodilla, and Parapar 2024), DeBERTa (He et al. 2021), XL-Net (Yang et al. 2020), and MetaHateBERT (Piot, Martín-Rodilla, and Parapar 2024), which is specifically tailored for abusive language and hate speech detection. These models are used as baseline approaches in cyberbullying research.

**LGBTQ+-Targeted Cyberbullying Detection.** While general cyberbullying detection has received substantial attention, specialized approaches for LGBTQ+-targeted cyberbullying remain relatively unexplored in the machine learning community. Standard language models often fail to capture the nuanced ways LGBTQ+ individuals are targeted online, as these models inherit and amplify societal biases present in training data. Conventional detection methods struggle with implicit forms of offensive content that rely on context, metaphors, and subtle microaggressions rather than explicit slurs. Previous work (Arslan et al. 2024) provided the first comparative analysis of various machine transformer models for detecting LGBTQ+ instances of cyberbullying, testing the performance of different language models on this specific task. However, this work primarily focused on evaluating existing models rather than developing specialized architectures for LGBTQ+ cyberbullying detection. Additional resources, such as (Chakravarthi et al. 2021; Hamlett et al. 2022), have contributed datasets aimed at LGBTQ+ hate speech detection.

**Automated Moderation for LGBTQ+ Content.** Effective moderation systems are crucial for creating safe online spaces for LGBTQ+ individuals. However, automated systems must carefully balance detection accuracy with the risk of over-flagging legitimate LGBTQ+ content. Context-aware moderation systems have shown significant improvements over keyword-based approaches (Wu, Zhao, and Yang 2022), but still struggle with the nuanced nature of LGBTQ+-targeted cyberbullying. The dynamic and evolving nature of cyberbullying tactics presents ongoing challenges for moderation systems (Bin Zia et al. 2022; Almerekhi, Jansen, and Kwak 2020). SpectrumNet addresses these challenges by providing more accurate detection of LGBTQ+-targeted cyberbullying through its context-aware architecture, potentially enabling more effective automated moderation while reducing false positives that might otherwise silence legitimate LGBTQ+ discourse.

# 3 Methods

## 3.1 Problem Definition

The task of LGBTQ+-aware cyberbullying detection targets to identify whether a given social media interaction contains bullying content, and if so, whether the content targets LGBTQ+ individuals. This problem is framed as a multi-class classification task over conversational threads.

We define the dataset as: $\mathcal{D} = \{(p_a, c_{a,b}, h_{a,b}, y_{a,b}) \mid a \in \{1, \ldots, A\}, b \in \{1, \ldots, B_a\}\}$. Each tuple in the dataset represents a comment and its associated context, where: $A$ denotes the total number of unique posts; $B_a$ represents the number of comments associated with post $a$; $p_a = (w_1, \ldots, w_l)$ is the tokenized text of post $a$, with $l$ tokens; $c_{a,b} = (v_1, \ldots, v_m)$ is the tokenized text of the $b$-th comment on post $a$, with $m$ tokens; $h_{a,b} = (h_1, \ldots, h_k)$ contains the flattened sequence of $k$ tokens from all comments preceding comment $b$ in the thread of post $a$; and $y_{a,b} \in \{0, 1, 2\}$ is the class for comment $b$ on post $a$.

The output label corresponds to: 0: Non-bullying, 1: LGBTQ+-targeted bullying, 2: Non-LGBTQ+ bullying. The model's objective is to learn a function $f_\Theta$ with parameters $\Theta$, such that: $f_\Theta(p_a, c_{a,b}, h_{a,b}) \mapsto y_{a,b}, \quad \forall a \in \{1, \ldots, A\}, b \in \{1, \ldots, B_a\}$. This task presents several modeling challenges: (1) Capturing subtle linguistic markers of identity-based bullying, (2) Adding contextual information from both the post content and comment text, as well as historical comment threads, (3) Handling significant class imbalance due to the rarity of LGBTQ+-targeted samples, and (4) Discriminating between LGBTQ+ and Non-LGBTQ+ bullying, which often differ in tone and intent.

## 3.2 Dataset

We used an improved version of the Instagram dataset originally collected by Hosseinmardi et al. (2015), which was composed of 2,219 sessions and 158,201 comments. While the dataset was originally labeled at the session level for general cyberbullying detection, we extended it with comment-level annotations specifically targeting LGBTQ+ cyberbullying instances. Using our own programmatic framework that converts each session into an HTML survey interface, we collected new fine-grained annotations for the presence of LGBTQ+ related cyberbullying, cyberbullying roles, severity levels, and thematic categorization of bullying content.

The Instagram platform provides several unique characteristics that make it particularly suitable for studying LGBTQ+ cyberbullying compared to other social media datasets. Instagram's visual-centric and highly interactive nature facilitates nuanced discussions around personal identity, gender expression, and sexual orientation, creating an environment where both implicit and explicit forms of cyberbullying targeting LGBTQ+ individuals can emerge. Unlike Twitter, where character limits constrain extensive interactions, Instagram supports extended conversational threads within comment sections, enabling the capture of contextual dynamics crucial for detecting subtle forms of harassment. In contrast to Reddit, where higher anonymity may allow users to evade personal accountability, Instagram's profile-linked system means interactions are more often tied to known identities. Research indicates that bullying by known perpetrators can have stronger psychological effects for some victims (Martínez Soler 2022), though reduced anonymity might also limit bullying frequency due to fear of consequences (Kim, Ellithorpe, and Burt 2023). Additionally, studies have found complex relationships between anonymity and online confrontations, where social vigilantism behaviors can emerge when users feel compelled to police others' actions (Lawless 2023). Given that Instagram is estimated to be the third largest social media platform in terms of users (Statista 2025), and considering how social media platforms foster relationships with one's self-identity

(Jeyanthi 2022), particularly for LGBTQ+ users (Fisher, Tao, and Ford 2024), developing accurate detection tools for this platform addresses a critical need in protecting vulnerable communities. To ensure high-quality annotations, we implemented a rigorous data collection methodology using Amazon Mechanical Turk (MTurk), a crowdsourcing platform widely adopted by natural language processing and social science researchers for collecting reliable training data (Crowston 2012). Each session required annotation by five distinct Master workers, who represent Amazon's highest-performing annotators with demonstrated superior accuracy compared to regular workers. We followed established best practices for MTurk data collection (Saravanos et al. 2021), requiring workers to have a prior approval rate of 95% or higher, be located in the United States, and have completed at least 1,000 previous tasks on the platform.

To maintain annotation quality, we incorporated three attention-check mechanisms throughout each survey. The first was an integrity question asking annotators to commit to providing reliable responses. The remaining two were disguised as regular comments but required careful reading to provide obvious answers, such as "Do you agree with the following statement? Most people are born with three legs." Workers who failed to answer the integrity question affirmatively or missed both disguised questions had their responses excluded, with the survey redistributed to new annotators.

To optimize annotation efficiency while maintaining data quality, we focused our efforts on sessions where three or more of the five original Hosseinmardi et al. (2015) annotators had identified cyberbullying content. This strategic filtering reduced our annotation scope to 439 sessions containing 106,618 comments. Following the annotation process, 21,255 comments were initially flagged as containing cyberbullying by at least one annotator. Using a majority voting criterion requiring agreement from three of the five MTurk workers, we identified 11,310 comments as containing cyberbullying behavior, with 1,053 specifically categorized as targeting gender identity and sexual orientation. Our annotation instructions explicitly distinguished between gender identity/sexual orientation harassment and general sexual harassment to ensure precise categorization.

The main statistics of the dataset are presented in Table 1.

| Statistic | Value |
|---|---|
| Total Comments | 106618 |
| Cyberbullying Comments | 11310 |
| LGBTQ+-targeted Comments | 1053 |
| Max Length of Comments | 2110 |
| Max Length of Cyberbullying Comments | 1941 |
| Max Length of LGBTQ+-targeted Comments | 857 |

Table 1. Statistics of the Instagram dataset, including total comments and class-specific comment lengths.

The dataset is available at Dataverse (https://doi.org/10.7910/DVN/MUVBRH).

### 3.3 SpectrumNet

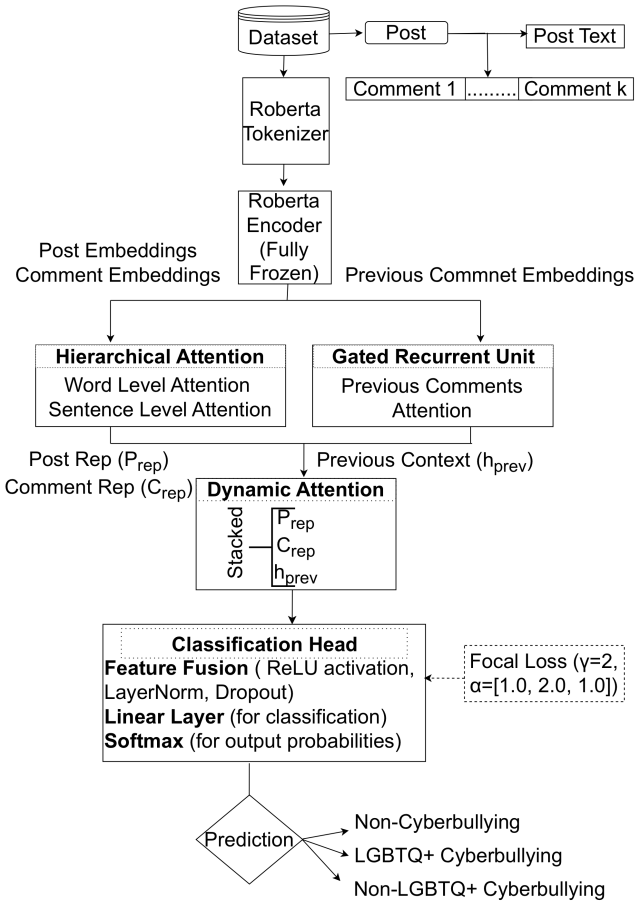


Figure 1: SpectrumNet architecture: frozen RoBERTa with hierarchical attention, GRU-based context encoding, and dynamic attention fusion for cyberbullying detection.

**Overview.** SpectrumNet has a transformer-based architecture designed to detect cyberbullying targeting LGBTQ+ individuals in social media posts. Our approach has the following objectives: recognizing subtle forms of identity-based harassment, analyzing contextual relationships between posts and comments, and adding historical conversation context to identify patterns of targeted behavior. The architecture of SpectrumNet is shown in Figure 1. The architecture of SpectrumNet incorporates these key components:

- A frozen RoBERTa encoder
- Multi-head attention with 12 attention heads
- GRU-based sequential encoding for historical comments
- Dropout rates of 0.2 (feature fusion) and 0.3 (classifier)
- Dynamic context weighting to adapt influence of each information source

**SpectrumNet: RoBERTa Encoder Configuration.** The selection of RoBERTa as the backbone encoder is motivated


Figure 2: Hierarchical attention visualization showing how SpectrumNet identifies and weights key words and phrases in cyberbullying detection, with brighter areas indicating stronger influence on classification decisions.

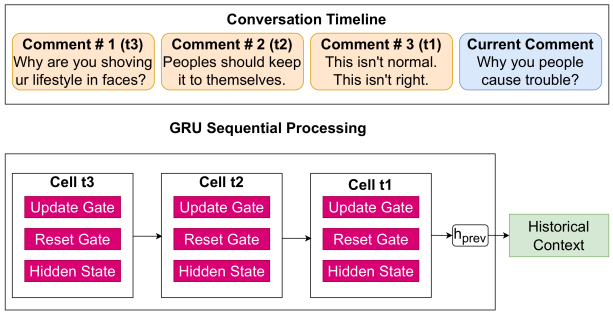


Figure 3: Gated Recurrent Unit (GRU) encoder architecture showing how SpectrumNet processes the sequential flow of historical comments to capture conversational context for cyberbullying detection.

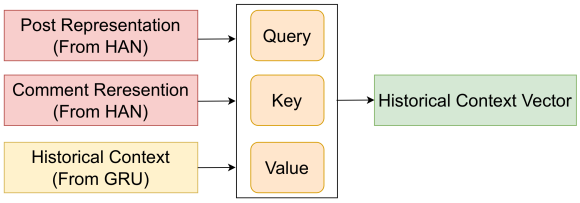


Figure 4: Attention fusion mechanism illustrating how SpectrumNet integrates multiple information sources (post content, current comment, and historical context) with weighted importance to improve LGBTQ+ cyberbullying detection.

by both theoretical and empirical considerations specific to LGBTQ+ cyberbullying detection. RoBERTa (Liu et al. 2019) removes the Next Sentence Prediction objective and employs dynamic masking during pretraining, leading to improved contextual understanding compared to BERT. Prior empirical work (Arslan et al. 2024) directly supports this choice for LGBTQ+ cyberbullying detection, showing that RoBERTa achieved the highest F1-score (0.733) compared to BERT and GPT-2 on this specific task. The RoBERTa encoder is fully frozen, and we add specialized modules such as hierarchical attention and dynamic fusion mechanisms so that SpectrumNet achieves improved contextual understanding without the computational complexity typically associated with fine-tuning. The input is processed by encoding the post, comment, and historical context separately:

$$\mathbf{x}_a^p = \text{RoBERTa}(p_a) \in \mathbb{R}^{l \times d},$$
$$\mathbf{x}_{a,b}^c = \text{RoBERTa}(c_{a,b}) \in \mathbb{R}^{m \times d},$$
$$\mathbf{Y}_{a,b} = \text{RoBERTa}(h_{a,b}) \in \mathbb{R}^{k \times d},$$

where $l$, $m$, and $k$ are the lengths of the post, comment, and historical context token sequences as defined in our problem formulation, and $d$ is the hidden state dimension (768 for RoBERTa-base).

The historical context $h_{a,b}$ is constructed by concatenating previous comments with the RoBERTa separator token:

$$h_{a,b} = c_{a,1} \| [\text{SEP}] \| c_{a,2} \| [\text{SEP}] \| \ldots \| c_{a,b-1}$$

For long sequences exceeding the model's token limit (*i.e.,* 512 tokens), we adopt a sliding window strategy to retain contextual cues at segment boundaries, ensuring that important conversational signals are preserved across multiple comments in a thread.

**SpectrumNet: Hierarchical Attention Network.** As shown in Figure 2, to capture both word-level and sentence-level semantics, building on (Cheng et al. 2019a), we implement a hierarchical attention mechanism. For post embeddings $\mathbf{x}_a^p \in \mathbb{R}^{l \times d}$ and comment embeddings $\mathbf{x}_{a,b}^c \in \mathbb{R}^{m \times d}$ from our RoBERTa encoder, we apply multi-head attention:

$$\mathbf{H}_a^p = \text{MultiHeadAttn}\left(\mathbf{x}_a^p, \mathbf{x}_a^p, \mathbf{x}_a^p\right) \in \mathbb{R}^{l \times d}$$
$$\mathbf{H}_{a,b}^c = \text{MultiHeadAttn}\left(\mathbf{x}_{a,b}^c, \mathbf{x}_{a,b}^c, \mathbf{x}_{a,b}^c\right) \in \mathbb{R}^{m \times d}$$

The token representations are then aggregated to produce sentence-level summaries:

$$\mathbf{r}_a^p = \frac{1}{l} \sum_{i=1}^{l} \mathbf{H}_{a,i}^p \in \mathbb{R}^d,$$
$$\mathbf{r}_{a,b}^c = \frac{1}{m} \sum_{i=1}^{m} \mathbf{H}_{a,b,i}^c \in \mathbb{R}^d,$$

where $\mathbf{H}_{a,i}^p$ refers to the $i$-th token representation in the post attention output and $\mathbf{H}_{a,b,i}^c$ refers to the $i$-th token representation in the comment attention output.

Following (Cheng et al. 2019a), the hierarchical attention mechanism enables the model to capture information at both word and comment levels, with attention mechanisms that can differentiate the importance of words and comments based on their context, as different elements are differentially informative depending on the specific social media session context.

**SpectrumNet: Historical Context Encoding.** As shown in Figure 3, we encode the historical context $h_{a,b}$ using token embeddings from previous comments:

$$\mathbf{Y}_{a,b} = \text{RoBERTa}(h_{a,b}) \in \mathbb{R}^{k \times d}$$

To capture sequential patterns in this historical context, we use a Gated Recurrent Unit (GRU) (Chung et al. 2014):

$$\mathbf{h}_{a,b}^{\text{enc}} = \text{GRU}\left(\mathbf{Y}_{a,b}\right) \in \mathbb{R}^{d}$$

The GRU processes tokens sequentially, updating its hidden state to capture temporal dynamics across the conversation history. Formally, the GRU cell updates can be described as follows for each time step $t \in \{1, 2, \ldots, k\}$:

$$\begin{aligned}
\mathbf{z}_t &= \sigma\left(\mathbf{W}_z \mathbf{y}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z\right), \\
\mathbf{r}_t &= \sigma\left(\mathbf{W}_r \mathbf{y}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r\right), \\
\tilde{\mathbf{h}}_t &= \tanh\left(\mathbf{W}\mathbf{y}_t + \mathbf{U}\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right) + \mathbf{b}\right), \\
\mathbf{h}_t &= \left(1 - \mathbf{z}_t\right) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t,
\end{aligned}$$

where: (1) $\mathbf{y}_t \in \mathbb{R}^d$ is the $t$-th token embedding from $\mathbf{Y}_{a,b}$, (2) $\mathbf{z}_t \in \mathbb{R}^d$ is the update gate vector at time $t$, (3) $\mathbf{r}_t \in \mathbb{R}^d$ is the reset gate vector at time $t$, (4) $\tilde{\mathbf{h}}_t \in \mathbb{R}^d$ is the candidate hidden state, (5) $\mathbf{h}_t \in \mathbb{R}^d$ is the hidden state at time $t$, (6) $\mathbf{W}_z, \mathbf{W}_r, \mathbf{W} \in \mathbb{R}^{d \times d}$ and $\mathbf{U}_z, \mathbf{U}_r, \mathbf{U} \in \mathbb{R}^{d \times d}$ are weight matrices, (7) $\mathbf{b}_z, \mathbf{b}_r, \mathbf{b} \in \mathbb{R}^d$ are bias vectors, (8) $\sigma(\cdot)$ is the sigmoid activation function, (9) $\tanh(\cdot)$ is the hyperbolic tangent function, and (10) $\odot$ represents element-wise multiplication.

As shown in Figure 3, the final hidden state $\mathbf{h}_{a,b}^{\text{enc}} = \mathbf{h}_k \in \mathbb{R}^d$ serves as a compact representation of the sequential/contextual information present in the previous comments.

**SpectrumNet: Attention-based Contextual Fusion.** As shown in Figure 4 To integrate multiple sources of contextual information, we form a context matrix by stacking the representations:

$$\mathbf{C}_{a,b} = \begin{bmatrix} \mathbf{r}_a^p \\ \mathbf{r}_{a,b}^c \\ \mathbf{h}_{a,b}^{\text{enc}} \end{bmatrix} \in \mathbb{R}^{3 \times d}$$

For the dynamic fusion mechanism, and analogous to attention terms, we define:

$$\mathbf{K} = \mathbf{r}_a^p, \quad \mathbf{Q} = \mathbf{r}_{a,b}^c, \quad \mathbf{V} = \mathbf{h}_{a,b}^{\text{enc}},$$

where $\mathbf{K}$, $\mathbf{Q}$, and $\mathbf{V}$ represent the key, query, and value vectors, respectively. The fused representation $\mathbf{z}_{a,b} \in \mathbb{R}^d$ is computed as:

$$\mathbf{z}_{a,b} = \mathsf{softmax}\left(\frac{(\mathbf{Q} \cdot \mathbf{w}_1^T) \cdot (\mathbf{K} \cdot \mathbf{w}_2^T)}{\sqrt{d}}\right) \cdot (\mathbf{V} \cdot \mathbf{w}_3^T),$$

where $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3 \in \mathbb{R}^d$ are learnable weight vectors, $\cdot$ represents the dot product operation, and the softmax operation ensures that the attention scores are normalized and can be interpreted as the relative importance of each context source.

**SpectrumNet: Classifier Branch.** After dynamic fusion, the representation passes through a sequence of layers that applies a series of transformations:

$$\tilde{\mathbf{z}}_{a,b} = \text{Dropout}\Big(\text{LayerNorm}\big(\text{ReLU}(\mathbf{W}_f \mathbf{z}_{a,b} + \mathbf{b}_f)\big)\Big) \in \mathbb{R}^d,$$

where $\mathbf{W}_f \in \mathbb{R}^{d \times d}$ and $\mathbf{b}_f \in \mathbb{R}^d$ are the weight matrix and bias vector. This transformation consists of: (1) a linear projection ($\mathbf{W}_f \mathbf{z}_{a,b} + \mathbf{b}_f$) that maps the fused representation to a new feature space; (2) A ReLU activation function that introduces non-linearity; (3) a LayerNorm operation that normalizes features across the embedding dimension to stabilize training; and (4) a Dropout operation with rate 0.2 that randomly drops elements to prevent overfitting and improve generalization. This transformation goes to another layer that outputs logits for three cyberbullying categories (non-bullying, LGBTQ+ bullying, and non-LGBTQ+ bullying):

$$\mathbf{o}_{a,b} = \mathbf{W}_{\text{cyb}} \tilde{\mathbf{z}}_{a,b} + \mathbf{b}_{\text{cyb}} \in \mathbb{R}^3,$$

where $\mathbf{W}_{\text{cyb}} \in \mathbb{R}^{3 \times d}$ and $\mathbf{b}_{\text{cyb}} \in \mathbb{R}^3$ are the weight matrix and bias vector of the classifier, and the outputs correspond to Non-Bullying, LGBTQ+ Bullying, and Non-LGBTQ+ Bullying classes.

**SpectrumNet: Loss Function and Optimization.** To address class imbalance, a key characteristic of cyberbullying datasets, we use focal loss (Lin et al. 2018). Focal loss has proven effective in improving model performance on imbalanced classification tasks by down-weighting easy examples and focusing training on hard, minority-class samples. Previous studies in the detection of online abuse and hate speech have successfully applied focal loss for similar reasons (Lu et al. 2023). The loss is defined as:

$$\mathcal{L}_{\text{cyb}} = -\alpha_i (1 - p_i)^{\gamma} \cdot \log(p_i),$$

where $i = \arg\max_j y_{a,b,j}$ represents the target class index, $\gamma = 2$ is the focusing parameter that reduces the relative loss for well-classified examples, and $\alpha_i$ is the class weight (we use $\alpha = [1.0, 2.0, 1.0]$ to emphasize LGBTQ+ bullying).

We define the softmax output as:

$$\mathbf{p}_{a,b} = \mathsf{softmax}(\mathbf{o}_{a,b}) \in \mathbb{R}^3,$$

where $\mathbf{o}_{a,b} \in \mathbb{R}^3$ is the logits vector for comment $b$ on post $a$, and $\mathbf{p}_{a,b}$ is the resulting softmax probability vector. The predicted probability for class $i$ is:

$$p_i = \mathbf{p}_{a,b,i} = \frac{\exp(\mathbf{o}_{a,b,i})}{\sum_{j=0}^{2} \exp(\mathbf{o}_{a,b,j})}$$

Here: (1) $\mathbf{o}_{a,b} \in \mathbb{R}^3$: logits for each of the three classes; (2) $\mathbf{p}_{a,b} \in \mathbb{R}^3$: softmax-normalized probabilities. and (3) $p_i$: the predicted probability for class $i$, used in the focal loss.

**SpectrumNet: Training Strategy.** We trained and tested our model using GroupKFold cross-validation with 10 folds. Each fold was split into 80% for training and 20% for testing. The model is optimized using AdamW (Loshchilov and Hutter 2019), which separates weight decay from gradient updates and is known to improve generalization in

transformer-based models. AdamW has consistently shown strong results when fine-tuning large pre-trained language models (Zhuang et al. 2022). We set the learning rate to $1 \times 10^{-5}$ with a batch size of 8. Since our data was imbalanced, we applied a weighted random sampler that gave higher weights to underrepresented classes. The input sequence length is limited to a maximum of 512 tokens. To support stable training, we also used a linear learning rate warmup during the first 10 percent of the training steps (Pareja et al. 2024).

The coding files have been made available at https://github.com/ysilva/SpectrumNet.

# 4 Experiments and Results

## 4.1 Baseline Models

To evaluate the effectiveness of SpectrumNet, we compare its performance with several state-of-the-art models used in cyberbullying and hate speech detection. These models span a diverse set of architectures and pre-training strategies:

**LLaMA.** A family of large language models developed by Meta AI based on the Transformer architecture, designed to efficiently generate and understand human-like text across multiple languages. LLaMA models range from 7 billion to 65 billion parameters and are scalable for research and applications in natural language processing. For our analysis we are using Llama-3.1-8B-Instruct. LLaMA 3.1 8B tested on 2 folds instead of 10 due to computational limitations.(Touvron et al. 2023).

**BERT.** A bidirectional transformer-based language model that set the foundation for many later variants, including RoBERTa, CT-BERT, and HateBERT. Pre-trained on large-scale English corpora with masked language modeling, BERT has been widely applied to text classification and hate speech detection tasks (Devlin et al. 2019).

**MetaHateBERT.** A BERT-based model retrained on large-scale hate speech datasets including those from banned communities, crafted specifically for abusive language and hate speech detection in English. MetaHateBERT builds upon BERT by further pretraining on millions of hateful text samples to enhance detection capabilities (Piot, Martín-Rodilla, and Parapar 2024).

**Twitter BERT.** Pre-trained on Twitter data, this variant of BERT (Devlin et al. 2019) captures social media-specific language, slang, and abbreviations, making it suitable for toxic content detection (Qudar and Mago 2020).

**CT-BERT.** A domain-adapted BERT model trained on COVID-19-related tweets, which demonstrates strong generalization across noisy, real-world social media text (Müller, Salathé, and Kummervold 2023).

**HateBERT.** Fine-tuned on the r/roastme subreddit, HateBERT is designed to detect offensive language and sarcasm, particularly useful for detecting aggressive tones and implicit bullying (Caselli et al. 2021).

**DeBERTa.** An enhanced BERT variant with disentangled attention and improved masking strategies, used in many NLP classification tasks (He et al. 2021).

**RoBERTa.** A transformer-based model that is widely used in hate speech and toxicity detection tasks (Liu et al. 2019).

**XLNet.** A permutation-based language model known for outperforming BERT in many tasks, though less robust on short social media texts (Yang et al. 2020).

**XGBoost.** A gradient-boosted decision tree classifier trained on TF-IDF and sentiment-based features, included to assess the gap between traditional and transformer-based methods (Chen and Guestrin 2016).

## 4.2 Experimental Setup

We tested all the models using the same Instagram dataset. All models were tested using 10-fold cross-validation to ensure fair comparison. Hyperparameters were optimized based on validation set performance, with models trained for a maximum of 50 epochs per fold. In performance comparison, we report the precision, recall, and F1-score across all folds.

## 4.3 Results

Table 2 show the performance for each model. The results show that SpectrumNet outperforms other models in detecting LGBTQ+-targeted cyberbullying, where it achieves the highest F1-score. This superior performance on LGBTQ+ class detection is further presented in Table 6, which highlights SpectrumNet's stronger accuracy, precision, recall, and F1-score. The results reveal that existing models struggle significantly with LGBTQ+-targeted cyberbullying detection. While baseline models achieve reasonable performance on general cyberbullying detection, their performance drops dramatically for LGBTQ+ cases. This indicates that conventional approaches fail to capture the linguistic and contextual patterns specific to identity-based harassment.

## 4.4 Ablation Study

To better understand the impact of our architectural design, we conduct an ablation study where components of SpectrumNet are incrementally added to a RoBERTa baseline. Table 3 reports performance across five configurations, culminating in the full model. The results highlight that each component makes a meaningful contribution to LGBTQ+ cyberbullying detection. *Historical context modeling* with the GRU yields the largest performance gain, underscoring the importance of capturing conversational dynamics for detecting identity-based harassment. *Hierarchical attention* provides consistent improvements by emphasizing subtle linguistic cues that are easily overlooked in flat representations. Finally, *dynamic fusion* further improves the performance over simple aggregation, showing the advantage of adaptively weighting contextual signals rather than treating them uniformly.

## 4.5 Qualitative Analysis

To understand SpectrumNet's detection capabilities, we selected two representative examples from each class and

| Model | Non-Bullying | | | LGBTQ+ Bullying | | | Non-LGBTQ+ Bullying | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| LLaMA 3.1 8B | 0.94±0.00 | 0.97±0.00 | 0.96±0.00 | 0.65±0.00 | 0.61±0.00 | 0.63±0.00 | 0.70±0.00 | 0.54±0.00 | 0.61±0.00 |
| BERT | 0.94±0.01 | **0.98±0.01** | **0.96±0.01** | 0.61±0.14 | 0.14±0.04 | 0.23±0.05 | 0.66±0.06 | 0.46±0.04 | 0.54±0.04 |
| MetaHateBERT | 0.94±0.01 | 0.97±0.01 | 0.95±0.01 | 0.64±0.18 | 0.08±0.05 | 0.14±0.08 | 0.58±0.08 | 0.49±0.04 | 0.53±0.05 |
| Twitter BERT | 0.94±0.01 | **0.98±0.00** | **0.96±0.01** | 0.70±0.31 | 0.05±0.03 | 0.09±0.05 | 0.69±0.05 | 0.49±0.04 | 0.57±0.04 |
| CT-BERT | 0.95±0.01 | 0.97±0.01 | **0.96±0.01** | 0.72±0.09 | 0.46±0.05 | 0.56±0.06 | 0.68±0.06 | 0.57±0.05 | 0.61±0.04 |
| HateBERT | 0.95±0.01 | 0.97±0.00 | **0.96±0.01** | 0.72±0.10 | 0.43±0.12 | 0.53±0.10 | 0.67±0.06 | 0.52±0.04 | 0.59±0.04 |
| DeBERTa | 0.93±0.01 | **0.98±0.00** | 0.95±0.01 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.61±0.06 | 0.36±0.03 | 0.45±0.03 |
| RoBERTa | 0.92±0.00 | **0.98±0.00** | 0.95±0.00 | 0.00±0.00 | 0.00±0.00 | 0.00±0.00 | 0.66±0.02 | 0.38±0.01 | 0.49±0.01 |
| XLNet | 0.93±0.01 | **0.98±0.00** | 0.95±0.01 | 0.56±0.13 | 0.09±0.06 | 0.16±0.08 | 0.62±0.05 | 0.39±0.04 | 0.48±0.04 |
| XGBoost | 0.94±0.01 | **0.98±0.00** | **0.96±0.01** | 0.71±0.06 | 0.60±0.07 | 0.65±0.06 | **0.72±0.06** | 0.44±0.05 | 0.54±0.04 |
| SpectrumNet | **0.96±0.00** | 0.92±0.00 | **0.96±0.00** | **0.77±0.13** | **0.75±0.02** | **0.80±0.01** | 0.70±0.01 | **0.88±0.01** | **0.78±0.01** |

Table 2. Detailed class-specific performance breakdown revealing SpectrumNet's balanced detection capabilities across all cyberbullying types, with particularly strong improvements in LGBTQ+ bullying detection. **Bold values** show the highest performance in each metric within each class; underlined values show the second-highest performance (mean ± standard deviation).

| Configuration | Bal. Acc. | Precision | Recall | F1 | AUROC |
|---|---|---|---|---|---|
| RoBERTa Baseline | 0.45±0.00 | 0.53±0.01 | 0.45±0.00 | 0.48±0.00 | 0.89±0.01 |
| + Hierarchical Attention | 0.66±0.04 | 0.60±0.04 | 0.66±0.04 | 0.66±0.03 | 0.90±0.02 |
| + Historical Context (GRU) | 0.55±0.02 | 0.60±0.03 | 0.55±0.02 | 0.57±0.02 | 0.91±0.01 |
| + Both (Simple Aggregation) | 0.70±0.03 | 0.65±0.03 | 0.70±0.03 | 0.67±0.02 | 0.92±0.02 |
| **Full SpectrumNet (Dynamic Fusion)** | **0.85±0.01** | **0.78±0.03** | **0.85±0.01** | **0.81±0.02** | **0.93±0.01** |

Table 3. Ablation study showing component contributions to LGBTQ+ cyberbullying detection across multiple evaluation metrics including Balanced Accuracy (Bal. Acc.), Precision, Recall, F1-score, and AUROC (mean ± standard deviation).

tested all baseline models on these same instances. The goal was to evaluate how different methods perform in the presence of nuanced real-world social media interactions.

Table 4 lists six representative examples from our test set, showing scenarios where SpectrumNet correctly identified different types of cyberbullying content, with particular focus on LGBTQ+-targeted bullying that baseline models frequently misclassified. Example $C1$ shows SpectrumNet's ability to detect subtle, gendered insults that contain LGBTQ+ targeted language (*on ur period*," *man up*"), which most baseline models failed to recognize as bullying. In Example $C2$, we observe that SpectrumNet was the only model to correctly identify the homophobic slur presented in a longer comment with other contextual elements, highlighting its superior semantic understanding. For general cyberbullying detection ($C3$ and $C4$), the baseline models showed mixed results, with several correctly classifying these examples but lacking consistency across different contexts. Examples $C5$ and $C6$ show SpectrumNet's ability to differentiate between actual bullying and comments that merely contain potentially offensive terms but lack bullying intent, a nuance that several baseline models struggled with, resulting in false positives.

Table 4 presents a performance comparison of different models across various types of comments, showing how well they classify LGBTQ+ bullying, non-LGBTQ+ bullying, and non-bullying content. As this table shows, SpectrumNet outperforms the baseline models, especially in LGBTQ+-targeted cases, particularly where most other models fail to pick up on subtle or context-dependent cues.

Based on the patterns observed in these examples, we identified several recurring characteristics of LGBTQ+ targeted bullying that SpectrumNet effectively captures:

- Identity erasure through deliberate misgendering, often detectable only through historical context.
- Coded language that appears neutral in isolation but becomes clearly targeted in context.
- Sequential escalation of cyberbullying across multiple comments, and implicit associations that avoid explicit slurs but convey prejudice.
- Context-dependent microaggressions that require understanding the full conversation.

These patterns are challenging for standard cyberbullying detection systems that lack contextual awareness or identity-specific training.

## 5 Discussion

**Limitations.** While SpectrumNet shows strong performance in detecting LGBTQ+-targeted cyberbullying, some limitations remain. For example, the model's reliance on frozen embeddings and linear GRU encoding may limit its ability to adapt to domain-specific or non-linear interaction patterns fully. Furthermore, the added architectural complexity introduces training overhead, which may impact scalability.

Another limitation is the model's inability to assess how the severity of bullying may evolve throughout a conver-

| ID | Comment | True Label |
|---|---|---|
| C1 | Must be on ur period haven a b**ch fit man up h*e cake | LGBTQ+ |
| C2 | @username why are you following him or stalking his pictures to talk sh*t? If you think he's a fag**t f**ck off he's livin his life while your wasting yours being a douchebag | LGBTQ+ |
| C3 | @username if you think he's such a f*g, why you on his d*ck? Hop off and go somewhere, sheesh. | Non-LGBTQ+ |
| C4 | All you t*ats keep kissing his a*s. He won't even respond and acknowledge half of you. Yet I'm the fag**t somehow. | Non-LGBTQ+ |
| C5 | And ur gonna call me a f*g and don't know anything about me, that makes complete sense @username | Non-Bullying |
| C6 | @username you know f*g means pile of sticks right | Non-Bullying |

Table 4. Representative test case examples showing challenging cyberbullying instances that show SpectrumNet's ability to correctly detect subtle LGBTQ+ targeted language (C1-C2), general cyberbullying (C3-C4), and differentiate non-bullying content containing potentially sensitive terms (C5-C6). LGBTQ+ = LGBTQ+ Bullying and Non-LGBTQ+ = Non-LGBTQ+ Bullying.

| Model | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| LLaMA 3.1 8B | TP | TP | FN | TP | TN | TN |
| BERT | FN | FN | TP | TP | TN | FP |
| MetaHateBERT | FN | TP | FN | TP | TN | TN |
| Twitter BERT | FN | FN | TP | TP | TN | FP |
| CT-BERT | TP | FN | TP | FN | TN | TN |
| HateBERT | FN | TP | TP | TP | FP | FP |
| DeBERTa | FN | FN | TP | FN | TN | FP |
| RoBERTa | FN | FN | FN | TP | TN | FP |
| XLNet | FN | FN | FN | TP | TN | FP |
| XGBoost | TP | FN | TP | FN | TN | TN |
| **SpectrumNet** | **TP** | **TP** | **TP** | **TP** | **TN** | **TN** |

Table 5. Performance comparison of cyberbullying detection models across different comment types. SpectrumNet shows superior performance in identifying LGBTQ+ targeted cyberbullying (C1, C2). TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

| Model | Bal. Acc. | Macro F1 | AUROC |
|---|---|---|---|
| LLaMA 3.1 8B | 0.70±0.00 | 0.73±0.00 | 0.95±0.00 |
| BERT | 0.51±0.02 | 0.54±0.03 | 0.92±0.01 |
| MetaHateBERT | 0.51±0.02 | 0.54±0.03 | 0.93±0.01 |
| Twitter BERT | 0.50±0.02 | 0.54±0.03 | 0.92±0.02 |
| CT-BERT | 0.67±0.02 | 0.71±0.02 | **0.95±0.01** |
| HateBERT | 0.64±0.04 | 0.69±0.04 | 0.94±0.01 |
| DeBERTa | 0.45±0.01 | 0.47±0.01 | 0.86±0.02 |
| RoBERTa | 0.45±0.00 | 0.48±0.00 | 0.89±0.01 |
| XLNet | 0.49±0.02 | 0.53±0.03 | 0.88±0.02 |
| XGBoost | 0.67±0.02 | **0.72±0.02** | 0.92±0.01 |
| SpectrumNet | **0.75±0.01** | 0.71±0.02 | 0.91±0.01 |

Table 6. Overall Model Performance including Balanced Accuracy (Bal. Acc.), Macro F1-score, and AUROC (mean ± standard deviation). LLaMA 3.1 8B tested on 2 folds instead of 10 due to computational limitations.

sation, despite its use of previous comments as context. That is, whereas SpectrumNet adds historical information through a GRU-based encoder, so that each comment is not classified in isolation, the model still treats severity as a static signal and does not explicitly track how severity escalates from one comment to the next within a thread. To illustrate, in Comment $C1$ from Table 7, we see a relatively low-severity insult. Later in the same thread, this is followed by Comment $C2$, which contains a much more severe and explicitly hateful LGBTQ+-targeted attack. Although SpectrumNet uses C1 when analyzing C2, it does not model how tone and severity escalate over time.

We also found examples where severity did not increase linearly but instead fluctuated across the thread. In the comment sequence in Table 8, the thread begins with a non-bullying comment $C1$, then sharply escalates to an LGBTQ+-targeted slur $C2$. After that, it transitions to a more general insult $C3$, softens to a vague expression of frustration $C4$, and finally returns to a non-bullying tone $C5$.

While SpectrumNet uses previous comments to predict the classification of a given comment, tracking the way severity increases or decreases across a conversation could enable the detection of key signals for proactive moderation.

**Future Work.** Future work could explore more flexible context encoders, adaptive sequence truncation strategies, and fine-tuning under low-resource settings to better generalize across diverse online cyberbullying scenarios. Additionally, integrating psychological frameworks for measuring bullying severity would improve the model's practical utility. This could include developing severity scoring mechanisms based on established psychological bullying and harm metrics, particularly those validated for LGBTQ+ adolescents who face disproportionate mental health impacts from targeted harassment. Adding sequence prediction to forecast how initial bullying comments might trigger escalation patterns represents another promising direction. Such models could help platform moderators intervene proactively rather than reactively. Future research should also consider the multimodal nature of LGBTQ+ cyberbullying, where text, images, and platform-specific features combine to create psychologically harmful content. Finally, cultural and lin-

| ID | Comment | True Label |
|---|---|---|
| C1 | Who cares the story line is a rip off and if u didn't see this coming Ur a ta*d | Non-LGBTQ+ |
| C2 | bully a s**t head f*get who screwed everybody and that's why his fi*y hates him. I wish the g**yses and f**ets would go die and burn in hell | LGBTQ+ |

Table 7. Cyberbullying severity escalation pattern showing how initial mild bullying (C1) can rapidly intensify into severe LGBTQ+ targeted hate speech (C2) within the same conversation thread, highlighting the need for contextual awareness in detection systems. Non-LGBTQ+ means Non-LGBTQ+ Bullying, and LGBTQ+ means LGBTQ+ Bullying.

| ID | Comment | True Label |
|---|---|---|
| C1 | Roors don't s**k u do lol | Non-Bully |
| C2 | What a f*gg! | LGBTQ+ |
| C3 | What an a**hat. | Non-LGBTQ+ |
| C4 | F**kin people | Non-LGBTQ+ |
| C5 | That guy on top of that guy that I'm writing under, he's depressed that you didn't read his comment, poor guy | Non-Bully |

Table 8. Conversation thread showing shifting cyberbullying severity: non-bullying (C1), LGBTQ+ slurs (C2), general insults (C3–C4), then non-bullying (C5), illustrating contextual modeling challenges. (Non-Bully = Non-Bullying; LGBTQ+ = LGBTQ+ Bullying; Non-LGBTQ+ = Non-LGBTQ+ Bullying.)

guistic variations in cyberbullying expression merit attention, as psychological impacts and bullying manifestations differ significantly across global contexts and marginalized communities.

## 6 Conclusion

In this work, we introduced SpectrumNet, a novel transformer-based model specifically designed to detect LGBTQ+-targeted cyberbullying by adding hierarchical attention, GRU-based historical encoding, and dynamic contextual fusion over a frozen RoBERTa backbone. Our results show that SpectrumNet significantly outperforms strong baselines, particularly in identifying subtle, context-dependent, and identity-specific forms of cyberbullying that are often missed by generic detection models. Using conversational history and contextual relationships, SpectrumNet enables a deeper understanding of cyberbullying dynamics that affect LGBTQ+ individuals. We show that structured, identity-aware context modeling is critical to improving online safety for marginalized communities. Our findings highlight the importance of building inclusive and fair moderation tools that go beyond surface-level text analysis. SpectrumNet provides a flexible foundation that can be extended to other forms of targeted abuse, integrated into real-time moderation pipelines, or adapted to multilingual and multimodal contexts.

## Ethical Statement

We believe our work has significant potential to protect vulnerable LGBTQ+ individuals from targeted online harassment and create safer digital spaces. However, we acknowledge potential risks of misuse. Malicious actors could potentially adapt our methods for surveillance or censorship of marginalized communities. Content moderation systems, if implemented without proper oversight, may disproportionately restrict legitimate speech from the communities they aim to protect. We encourage researchers and platforms to use such technologies with human oversight and community input into moderation policies.

## Acknowledgements

This work was supported by National Science Foundation Awards No. 2435164 and No. 2435165, a Google Award for Inclusion Research, and a Lambda Research Grant. We would also like to thank Dr. Hong-Long Ji for his valuable guidance and support.

## References

Abreu, R. L.; and Kenny, M. C. 2017. Cyberbullying and LGBTQ Youth: A Systematic Literature Review and Recommendations for Prevention and Intervention. *Journal of Child & Adolescent Trauma*, 11(1): 81–97.

Abreu, R. L.; and Kenny, M. C. 2018. Cyberbullying and LGBTQ Youth: A Systematic Literature Review and Recommendations for Prevention and Intervention. *Journal of Child & Adolescent Trauma*, 11(1): 81–97.

Almerekhi, H.; Jansen, B. J.; and Kwak, H. 2020. Investigating toxicity across multiple reddit communities. *Information Processing & Management*, 57(6).

Angoff, N. R.; and Barnhart, W. R. 2021. Bullying, Cyberbullying, and LGBTQ Students from an Intersectional Perspective. *Journal of LGBT Issues in Counseling*, 15(2): 127–142.

Arslan, M.; Madrigal, M. S.; Abuhamad, M.; Hall, D. L.; and Silva, Y. N. 2024. Detecting LGBTQ+ Instances of Cyberbullying. arXiv:2409.12263.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 759–760.

Bauer, G. R.; et al. 2021. Application of an intersectional lens to bias-based bullying among LGBTQ+ youth of color in the United States. *SSM-Population Health*, 16: 100991.

Bin Zia, H.; Raman, A.; Castro, I.; and Tyson, G. 2022. Toxicity in the decentralized web and the potential for model sharing. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 1445–1449.

Cao, J.; Lee, R. K.-W.; and Hoang, T.-A. 2020. Detection of cyberbullying on social media using deep learning approaches. *Information Processing & Management*, 57(6).

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. arXiv:2010.12472.

Centre, C. R. 2020. Bullying, Cyberbullying, and LGBTQ Students. https://www.humanitarianlibrary.org/resource/bullying-cyberbullying-and-lgbtq-students. Accessed: May 28, 2024.

Chakravarthi, B. R.; Priyadharshini, R.; Ponnusamy, R.; Kumaresan, P. K.; Sampath, K.; Thenmozhi, D.; Thangasamy, S.; Nallathambi, R.; and McCrae, J. P. 2021. Dataset for Identification of Homophobia and Transphobia in Multilingual YouTube Comments. Details not provided.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. ACM.

Cheng, L.; Guo, R.; Silva, Y.; Hall, D.; and Liu, H. 2019a. Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, 235–243. Society for Industrial and Applied Mathematics.

Cheng, L.; Guo, R.; Silva, Y.; Hall, D.; and Liu, H. 2019b. Hierarchical attention networks for cyberbullying detection on the instagram social network. *Proceedings of the 2019 SIAM International Conference on Data Mining*, 235–243.

Cheng, L.; Guo, R.; Silva, Y. N.; Hall, D.; and Liu, H. 2021a. Modeling Temporal Patterns of Cyberbullying Detection with Hierarchical Attention Networks. *ACM/IMS Transactions on Data Science*, 2(2).

Cheng, L.; Mosallanezhad, A.; Silva, Y.; Hall, D.; and Liu, H. 2021b. Mitigating Bias in Session-Based Cyberbullying Detection: A Non-Compromising Approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2158–2168. Association for Computational Linguistics.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555.

Cook, S. 2024. Cyberbullying data, facts and statistics for 2018 – 2024.

Crowston, K. 2012. Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars. In Bhattacherjee, A.; and Fitzgerald, B., eds., *Shaping the Future of ICT Research. Methods and Approaches*, 210–221. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-35142-6.

Dadvar, M.; and Eckert, K. 2018. Cyberbullying Detection in Social Networks Using Deep Learning Based Models: A Reproducibility Study. arXiv preprint arXiv:1812.08046.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Duarte, C.; Pittman, S. K.; Thorsen, M. M.; Cunningham, R. M.; and Ranney, M. L. 2018. Correlation of Minority Status, Cyberbullying, and Mental Health: A Cross-Sectional Study of 1031 Adolescents. *Journal of Child & Adolescent Trauma*, 11(1): 39–48.

Dürrbaum, T.; and Sattler, F. A. 2020. Minority Stress and Mental Health in Lesbian, Gay Male, and Bisexual Youths: A Meta-Analysis. *Journal of LGBT Youth*, 17(3): 298–314.

Fisher, C. B.; Tao, X.; and Ford, M. 2024. Social media: A double-edged sword for LGBTQ+ youth. *Computers in Human Behavior*, 156: 108194.

Gini, G.; and Espelage, D. L. 2014. Peer Victimization, Cyberbullying, and Suicide Risk in Children and Adolescents. *JAMA*, 312(5): 545–546.

Hamlett, M.; Powell, G.; Silva, Y. N.; and Hall, D. 2022. A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1251–1258.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv:2006.03654.

Hinduja, S.; and Patchin, J. W. 2021. Cyberbullying Statistics 2021: Age, Gender, Sexual Orientation, and Race.

Holt, M. K.; Vivolo-Kantor, A. M.; Polanin, J. R.; Holland, K. M.; DeGue, S.; Matjasko, J. L.; Wolfe, M.; and Reid, G. 2015. Bullying and Suicidal Ideation and Behaviors: A Meta-Analysis. *Pediatrics*, 135(2): e496–e509.

Hosseinmardi, H.; Mattson, S. A.; Rafiq, R. I.; Han, R.; Lv, Q.; and Mishra, S. 2015. Detection of Cyberbullying Incidents on the Instagram Social Network. Details not provided.

Jeyanthi, M. 2022. Social Media and Identity Formation – The Influence of Self-Presentation and Social Comparison. *Mind and Society*, 11: 138–144.

Kim, M.; Ellithorpe, M.; and Burt, S. 2023. Anonymity and its role in digital aggression: A systematic review. *Aggression and Violent Behavior*, 72: 101856.

Lawless, T. J. 2023. *Social (media) vigilantes: Effects of social vigilantism and anonymity on online confrontations of prejudice*. Phd dissertation, Kansas State University.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Lu, J.; Lin, H.; Zhang, X.; Li, Z.; Zhang, T.; Zong, L.; Ma, F.; and Xu, B. 2023. Hate Speech Detection via Dual Contrastive Learning. arXiv:2307.05578.

Martínez Soler, C. 2022. *Anonymity and Cyberbullying on Social Media: Research into the influence of anonymity and the types of negative messages on the self-esteem and body appreciation of cyberbullying victims*. Master's thesis, Tilburg University. Available at http://arno.uvt.nl/show.cgi?fid=159041.

Meyer, I. H. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological bulletin*, 129(5): 674–697.

Müller, M.; Salathé, M.; and Kummervold, P. E. 2023. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *Frontiers in Artificial Intelligence*, 6: 1023281.

NAACP. 2024. Curing the Epidemic of Black Youth Suicides.

Nandhini, B. S.; and Sheeba, J. I. 2015. Cyberbullying Detection and Classification Using Information Retrieval Algorithm. In *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, ICARCSET '15. New York, NY, USA: ACM.

NordVPN. 2024. How to Stay Safe Online: A Guide Against LGBTQ+ Cyberbullying.

Pareja, A.; Nayak, N. S.; Wang, H.; Killamsetty, K.; Sudalairaj, S.; Zhao, W.; Han, S.; Bhandwaldar, A.; Xu, G.; Xu, K.; Han, L.; Inglis, L.; and Srivastava, A. 2024. Unveiling the Secret Recipe: A Guide For Supervised Fine-Tuning Small LLMs. License: CC BY 4.0, arXiv:2412.13337.

Perera, A.; and Fernando, P. 2021. Accurate Cyberbullying Detection and Prevention on Social Media. *Procedia Computer Science*, 181: 605–611.

Pew Research Center. 2022. Teens and Cyberbullying 2022. Technical report, Pew Research Center.

Piot, P.; Martín-Rodilla, P.; and Parapar, J. 2024. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. volume 18, 2025–2039.

Plöderl, M.; and Tremblay, P. 2015. Mental health of sexual minorities: a systematic review. *International Review of Psychiatry*, 27(5): 367–385. PMID: 26552495.

Powell, A.; Stratton, G.; and Cameron, R. 2022. A Phenomenological Investigation into Cyberbullying as Experienced by People Identifying as Transgender or Gender Diverse. *International Journal of Environmental Research and Public Health*, 19(11): 6560.

Qudar, M. M. A.; and Mago, V. 2020. TweetBERT: A Pretrained Language Representation Model for Twitter Text Analysis. arXiv:2010.11091.

Saravanos, A.; Zervoudakis, S.; Zheng, D.; Stott, N.; Hawryluk, B.; and Delfino, D. 2021. The Hidden Cost of Using Amazon Mechanical Turk for Research. In Stephanidis, C.; Soares, M. M.; Rosenzweig, E.; Marcus, A.; Yamamoto, S.; Mori, H.; Rau, P.-L. P.; Meiselwitz, G.; Fang, X.; and Moallem, A., eds., *HCI International 2021 - Late Breaking Papers: Design and User Experience*, 147–164. Cham: Springer International Publishing. ISBN 978-3-030-90238-4.

Sathya, J.; and Harin Fernandez, F. M. 2024. Effective Automatic Cyberbullying Detection Using a Hybrid Approach SVM And NLP. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems*.

Singh, V. K.; Ghosh, S.; and Jose, C. 2017. Toward Multimodal Cyberbullying Detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2090–2099. ACM.

Social Media Victims Law Center. 2024. Cyberbullying and the LGBTQ+ Youth Community.

Sophie., E.; Clancy., E. M.; B., K.; and R., T. 2022. A Phenomenological Investigation into Cyberbullying as Experienced by People Identifying as Transgender or Gender Diverse. *International Journal of Environmental Research and Public Health*, 19(11): 6560.

Statista. 2025. Biggest social media platforms by users 2025| Statista — statista.com. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

StopBullying.gov. 2024. What Is Cyberbullying. https://www.stopbullying.gov/cyberbullying/what-is-it. Last reviewed October 7, 2024; accessed [put the date you accessed it].

Tanni, T. I.; Akter, M.; Anderson, J.; Amon, M. J.; and Wisniewski, P. J. 2024. Examining the Unique Online Risk Experiences and Mental Health Outcomes of LGBTQ+ Versus Heterosexual Youth. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM.

TheTrevorProject. 2024a. Bullying and Suicide Risk among LGBTQ Youth.

TheTrevorProject. 2024b. National Survey on LGBTQ+ Youth Mental Health.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Wu, T.; Zhao, X.; and Yang, Q. 2022. Automated content moderation in the fediverse. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 1201–1211.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.

Zhuang, Z.; Liu, M.; Cutkosky, A.; and Orabona, F. 2022. Understanding AdamW through Proximal Methods and Scale-Freeness. arXiv:2202.00089.

Özel, S. A.; Saraç, E.; Akdemir, S.; and Aksu, H. 2017. Detection of cyberbullying on social media messages in Turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*, 366–370.

## Paper Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this research helps protect LGBTQ+ individuals from cyberbullying without violating privacy norms. Our model uses anonymized data and aims to reduce discrimination rather than perpetuate it.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, our abstract and introduction clearly state our three main contributions: (1) introducing SpectrumNet with hierarchical attention mechanisms, (2) implementing a dynamic contextual fusion approach, and (3) providing comprehensive performance evaluation.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in Section 3, we provide a detailed description of our methodology, including problem definition, dataset characteristics, model architecture, and training strategy, all of which support our claims about improved LGBTQ+ cyberbullying detection.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in our dataset section we acknowledge the class imbalance issue and discuss how our focal loss and weighted sampling approach addresses this challenge. We also note the representativeness of Instagram data for studying LGBTQ+ cyberbullying.

(e) Did you describe the limitations of your work? Yes, in Section 5 (Discussion), we explicitly outline several limitations, including architectural constraints, inability to fully assess psychological severity evolution, and challenges with non-linear interaction patterns.

(f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss the risk of overflagging legitimate LGBTQ+ content in the Related Work section, noting that automated systems must carefully balance detection accuracy with avoiding false positives that might silence legitimate discourse.

(g) Did you discuss any potential misuse of your work? Yes, we implicitly address this by discussing the importance of balance in moderation systems, acknowledging that overly aggressive flagging could lead to censorship of legitimate LGBTQ+ content. Additionally, we note the need for human review in conjunction with automated systems.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we describe our annotation process with quality control measures, use of anonymized data, and implementation of focal loss to reduce bias. Our future work section also suggests improvements for better contextual understanding to reduce false positives.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, our research focuses on protecting a vulnerable population (LGBTQ+ individuals), uses anonymized data, and aims to reduce harm rather than create it, conforming to ethical guidelines for AI research.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes, we state our assumptions that: (1) conversational context is critical for detecting subtle forms of cyberbullying, (2) LGBTQ+-targeted bullying has distinctive linguistic patterns, and (3) hierarchical attention can capture these patterns effectively.

(b) Have you provided justifications for all theoretical results? Yes, we justify our architectural choices in Section 3.3, explaining why each component (hierarchical attention, GRU-based encoding, dynamic fusion) contributes to improved detection performance for LGBTQ+ targeted content.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, we compare our approach with eight baseline models representing different theoretical approaches to cyberbullying detection, discussing their relative strengths and limitations.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, in comparing baseline models, we examine alternative mechanisms for cyberbullying detection and evaluate whether simpler approaches could achieve similar results.

(e) Did you address potential biases or limitations in your theoretical framework? Yes, we discuss the limitations of our approach in Section 5, including challenges with severity assessment and the potential need for more flexible context encoders.

(f) Have you related your theoretical results to the existing literature in social science? Yes, we connect our work to social science research on LGBTQ+ cyberbullying in the Introduction and Related Work sections, citing studies on psychological impact and minority stress.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, we discuss implications for creating safer online spaces for LGBTQ+ individuals and suggest future directions including psychological frameworks for measuring bullying severity.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? Our paper does not include formal theoretical proofs.

(b) Did you include complete proofs of all theoretical results? Not applicable to our work.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results

(either in the supplemental material or as a URL)? Yes, our dataset and full code have been provided with the submission for reviewer evaluation, including the model architecture, training scripts, and evaluation code. Upon paper acceptance, we will publicly release the complete codebase, along with data processing scripts, to ensure full reproducibility.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in Section 3.3 we detail our training strategy, including optimizer choice (AdamW), learning rate (1e-5), batch size (8), weighted sampling approach, and sequence length limits.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we used 10-fold cross-validation for all experiments and report the mean $\pm$ standard deviation performance metrics, which accounts for statistical variation.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, experiments were conducted using NVIDIA RTX 3090 GPUs with 24GB memory, RTX 5090, and Quadro RTX 8000 hardware across distributed computing environments including Lambda Labs cloud infrastructure.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, our evaluation compares SpectrumNet against eight state-of-the-art baselines on the same dataset, focusing specifically on LGBTQ+ cyberbullying detection metrics that directly address our claims.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance)? Yes, we discuss the differential costs of misclassification in Section 5, noting that false negatives can lead to continued harassment while false positives might silence legitimate discourse.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, we cited Hosseinmardi et al. (2015) as the creators of the original Instagram dataset which we further annotated for LGBTQ+ cyberbullying detection, not created.

(b) Did you mention the license of the assets? No, but the original Instagram dataset we used is publicly available for research purposes through the cited publication.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. We have submitted our extended Instagram dataset, which includes LGBTQ+ cyberbullying annotations, along with the paper. Upon acceptance, these specialized annotations will be publicly released through a research repository to ensure reproducibility and to advance research on LGBTQ+ cyberbullying.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, we discuss our annotation process using Amazon Mechanical Turk with 5 master-level annotators per session and the consent procedures for annotators under formal IRB approval. The original dataset by Hosseinmardi et al. (2015) used public Instagram posts.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we acknowledge that the dataset contains offensive content (masked in our examples) and explain that usernames and personally identifiable information were anonymized in the original dataset.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes, we intend to release our annotations (not a new dataset) with clear documentation, metadata, and accessibility through a research repository with appropriate citation methods to ensure FAIR principles. The original dataset was created by Hosseinmardi et al. (2015), and our contribution is the additional LGBTQ+ cyberbullying annotations which we will share upon paper acceptance.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes, we have prepared a detailed documentation for our annotation extension including annotation procedures, quality control measures, class distributions, content warnings, annotator qualifications, and intended research uses. This datasheet follows established guidelines and will accompany the public release of our LGBTQ+ cyberbullying annotations.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, we included a detailed description (without screenshots). The annotation task was conducted under formal IRB approval with proper informed consent procedures.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes. Our annotation task was conducted under formal IRB approval.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. We stated the rate per comment and calculated the total session compensation as the product of this rate and the number of comments.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, we mention that the data was anonymized prior to annotation and discuss our multi-stage validation process to ensure quality while protecting privacy. Our IRB-approved protocol specified that all collected data was kept in private (password-protected) folders, and that an anonymized version of the labels would only be shared with researchers who commit to not

sharing the data directly and using the labels only for research purposes.