Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights

Célia Nouri[1,2] **Jean-Philippe Cointet**[2] **Chloé Clavel**[1,3]
[1]INRIA, ALMAnaCH
[2]Sciences Po, médialab
[3]Télécom Paris
{celia.nouri, chloe.clavel}@inria.fr, jeanphilippe.cointet@sciencespo.fr

Abstract

Detecting abusive language in social media conversations poses significant challenges, as identifying abusiveness often depends on the conversational context, characterized by the content and topology of preceding comments. Traditional Abusive Language Detection (ALD) models often overlook this context, which can lead to unreliable performance metrics. Recent Natural Language Processing (NLP) approaches that incorporate conversational context often rely on limited or overly simplified representations of this context, leading to inconsistent and sometimes inconclusive results. In this paper, we propose a novel approach that utilizes graph neural networks (GNNs) to model social media conversations as graphs, where nodes represent comments, and edges capture reply structures. We systematically investigate various graph representations and context windows to identify the optimal configurations for ALD. Our GNN model outperforms both context-agnostic baselines and linear context-aware methods, achieving significant improvements in F1 scores. These findings demonstrate the critical role of structured conversational context and establish GNNs as a robust framework for advancing context-aware ALD. Our code is available at this link.

Disclaimer: This paper contains discriminatory content that may be disturbing to some readers.

1 Introduction

The expansion of social media has facilitated global communication but has also amplified the spread of abusive language (AL), posing significant challenges (Duggan, 2017; Saveski et al., 2021). Abusive language refers to communication that demeans, offends, or marginalizes individuals or groups, encompassing hate speech, toxicity, offensive language, and cyberbullying (Vidgen et al., 2021; Bourgeade et al., 2024). However, most Abusive Language Detection (ALD) models classify comments in isolation, disregarding conversational context, which is crucial for accurate classification (Pavlopoulos et al., 2020; Menini et al., 2021; Vidgen et al., 2021).

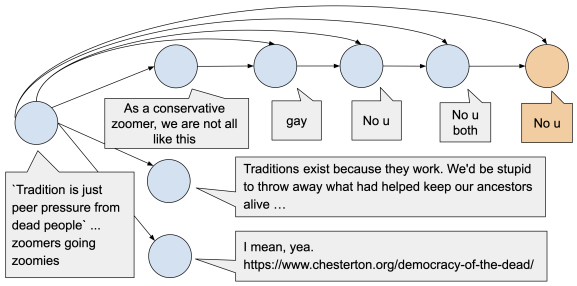

Figure 1: Example conversation from the Contextual Abuse Dataset (CAD), the graph was generated from our Affordance-based method. The target node is labeled abusive and colored in orange.

Figure 1 illustrates the challenge of context-aware ALD, where the goal is to classify a target comment (in orange) using preceding context without knowing the abusiveness of prior comments. The comment *"No u"* is labeled abusive, but its meaning is unclear in isolation. Examining the full conversation reveals that it occurs within a reply chain initiated by a homophobic insult, triggering a sequence of reactive comments. This pattern exemplifies the snowball effect of abusive speech, where insults propagate and reinforce toxicity. Understanding such interactions requires capturing conversation thread structures and contents beyond solely looking at the immediate preceding comment or initial post.

Despite efforts to incorporate context into ALD, existing methods remain limited. Many define context narrowly, considering only the previous comment or the original post (Yu et al., 2022; Anu-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18271–18286

chitanukul et al., 2022). Others treat context as a flat sequence, failing to model the conversation structure, leading to incomplete contextual understanding (Bourgeade et al., 2024).

To address these limitations, we introduce a graph-based framework for ALD that models conversations as graphs, where nodes represent comments, and edges capture reply relationships. This structure preserves the conversation topology and allows contextual information to propagate across multiple interaction levels. Through extensive experiments and analyses, we demonstrate that our graph-based models significantly outperform both context-agnostic approaches and prior context-aware models that rely on flat context representations. These results confirm the advantages of explicitly modeling the conversation topology and leveraging graph neural networks to capture contextual dependencies in ALD.

Contributions Our contributions are as follows:
(1) We propose a graph-based framework for ALD that effectively models Reddit conversations while preserving their structure and relevant content.
(2) We analyze the optimal amount of conversational context needed for graph-based models to maximize performance.
(3) We compare our graph-based approach with existing context-aware NLP models, highlighting the strengths and limitations of graph-based ALD.

Although our study focuses on Reddit, our approach is applicable to any platform featuring threaded discussions. Platforms such as X, Facebook comment threads, YouTube replies, or online forums like StackExchange and news comment sections all offer reply structures that can be translated into directed conversation graphs. For these platforms, our method requires only minimal adaptation, particularly in how the conversation graph is trimmed (see Section 3.2).

Paper Organization Section 2 reviews prior work on context-aware ALD. Section 3 presents our methodology, including problem formulation, graph construction, and model architecture. Section 5 describes our experimental setup and results, evaluating the impact of different context modeling strategies. Finally, Section 6 summarizes our findings and outlines directions for future research.

2 Related Work

This section examines the role of conversational context in human annotations, reviews context-aware ALD datasets, and discusses previous efforts to integrate conversational context into NLP models and graph-based approaches for ALD.

2.1 Impact of Context on Human Annotations

Previous studies have shown that the inclusion of context can significantly alter how comments are perceived and annotated for toxicity or abusiveness. For instance, Pavlopoulos et al. (2020) investigated the impact of context by annotating 250 Wikipedia Talk page comments under two conditions: in isolation and with context, where context included the post title and the previous comment. They found that 5% of the labels changed when context was provided, with most changes occurring from nontoxic to toxic. Similarly, Menini et al. (2021) re-annotated 8,000 tweets from the Founta dataset (Founta et al., 2018) with and without context, where context comprised all preceding messages in the thread. Conversely, their results showed a decrease in the percentage of abusive labels from 18% to 10% when context was provided, indicating that annotators perceived fewer tweets as abusive when they had additional contextual information. These contrasting findings highlight the complexity of incorporating context into ALD and underscore its influence on human perception. Further studies (Yu et al., 2022; Vidgen et al., 2021) similarly found that providing conversational context impacts the interpretation of abusiveness. These studies emphasized the need for context-aware datasets to improve ALD systems.

2.2 Datasets

Numerous datasets have been developed to support abusive language detection and related tasks such as identifying toxicity, hate speech, racism, and sexism (Waseem and Hovy, 2016; Davidson et al., 2017; Golbeck et al., 2017; Founta et al., 2018). However, most of these datasets focus on isolated comment instances, ignoring conversational context during both annotation and modeling.

Context-aware datasets have been developed across platforms such as Twitter (Menini et al., 2021; İhtiyar et al., 2023), Wikipedia (Pavlopoulos et al., 2020), and Reddit (Yu et al., 2022), but often rely on narrow definitions of conversational context. Pavlopoulos et al. (2020) included only

the post title and preceding comment, Yu et al. (2022) restricted context to a single prior comment, while Bourgeade et al. (2024) tried using the post title, preceding comment, or both, finding that even small changes in context definition affected model performance. Menini et al. (2021) explored variable context lengths on Twitter but observed diminishing returns, likely due to flat modeling limitations. These approaches are also limited by small sample sizes (Menini et al., 2021; Pavlopoulos et al., 2020), or suffer from inconsistent annotation quality (Hebert et al., 2024).

Our work focuses on Reddit, which offers rich and structured conversation threads well-suited for modeling abusive language in conversational context. We prioritize datasets with full conversation trees and high-quality annotations, such as the Contextual Abuse Dataset (CAD) (Vidgen et al., 2021), an English Reddit-based dataset, which meets these criteria. Further details on CAD are provided in Section 4.

2.3 Context-aware flat models for ALD

The value of conversational context in ALD has inspired various neural architectures designed to incorporate context into classification tasks. A common baseline approach, *Text-Concat*, concatenates the context (i.e., preceding comments and initial post) with the target text, processing the combined input through a transformer like BERT (Devlin et al., 2019). Studies such as Bourgeade et al. (2023); Menini et al. (2021); Anuchitanukul et al. (2022) have demonstrated the utility of this method for ALD for certain datasets. Another explored approach is *Embed-Concat*, which embeds the context and target text separately using distinct transformer encoders before combining the embeddings for classification (Bourgeade et al., 2024). These models serve as baselines in our work, enabling us to compare the performance of our graph-based method. Methods like history embedding (Anuchitanukul et al., 2022) attempt to preserve separate representations of context and target text but face limitations in modeling reply-relationships or multi-turn conversational structures. Moreover, these approaches show inconsistent performance across datasets, underlining the need for more robust techniques to integrate context effectively.

Large generative language models (GLMs) have shown strong performance in ALD and context-aware ALD through prompting strategies (Wei et al., 2022b; Chiu and Alexander, 2021), including Chain-of-Thought (CoT) prompting (Wei et al., 2022a) and Few-Shot prompting (Brown et al., 2020). However, these models face critical limitations in real-world scenarios. Proprietary models like GPT-4 (1.76T parameters) lack transparency, making them unsuitable for content moderation. Meanwhile, open-source alternatives such as LLaMA-2-13B (13B parameters) (Touvron et al., 2023), DeepSeek-V2 (236B total, 21B active) (DeepSeek-AI et al., 2024), and more recently DeepSeek-V3 (671B total, 37B active) (AI, 2024), remain impractical due to their large size and high inference latency. Compared to these models, our graph-based approach detailed in Section 3.3 offers a more parsimonious solution, combining BERT (110M parameters) with graph aggregation (less than 1M parameters per graph attention layer). Our approach offers a fast, computationally frugal, and interpretable alternative while preserving conversational structure (see Section 5.2 for a detailed Computational Cost and Runtime comparison).

2.4 Context-aware Graph Models for ALD

In this section, we review prior works leveraging graphs to represent online conversations, highlighting differences in graph construction and embedding generation, which are key aspects that set our method apart.

Graph Construction Graph-based approaches for ALD vary in how they represent relationships between social media comments. Some methods construct fully connected graphs, linking messages based on cosine similarity between text embeddings (Wang et al., 2020; Duong et al., 2022). While effective for propagating labels across datasets, these approaches overlook conversational structure and fail to prioritize messages within a thread. Other works reconstruct retweet paths using temporal and follower relationships (Beatty, 2020), but these methods are platform-specific and do not generalize well beyond Twitter. Temporal graphs have also been used in chat-based platforms, where context is defined by surrounding messages in the chat (Cecillon et al., 2021; Papegnies et al., 2019). However, this approach is unfit for structured threads, such as those on Reddit. More closely related to our work, several studies construct conversation graphs based on reply relationships, with nodes representing comments and edges denoting replies (Hebert et al., 2024; Agarwal et al., 2023; Meng et al., 2023; Zayats and Os-

tendorf, 2018). For example, Hebert et al. (2024) use such graphs but incorporate multimodal embeddings that combine post-image and text features as node attributes. While Zayats and Ostendorf (2018) also build reply-based graphs for Reddit threads, their goal is to predict the popularity rather than the abusiveness of a comment. Unlike previous methods, our approach trims conversation graphs to mimic what users see when writing a comment, leveraging the default Reddit rendering settings.

Context Embedding Generation Several methods generating embedding representations from conversation graphs focus on global conversation embeddings that summarize the structural properties of a conversation. For example, Meng et al. (2023) apply average pooling to node features across a conversation tree, encoding attributes such as the number of replies and overall tree shape. Similarly, Hebert et al. (2024, 2022) use Graphormer (Ying et al., 2021) to generate embeddings that capture global structural features such as node centrality and connectivity. While these approaches are effective at representing the overall conversation structure, they overlook localized interactions and the specific contextual nuances that can be critical for abusive language detection. Our work adopts a different perspective by focusing on the local conversation context that users directly interact with, rather than relying on global conversation summaries. Closer to our work, (Agarwal et al., 2023) propose GraphNLI, which generates context embeddings through random graph walks. Their method uses fixed probabilities to favor paths toward the root and applies discount factors to penalize nodes further away from the target comment. In contrast, our approach uses Graph Attention Layer (GATs) to dynamically learn the importance of contextual nodes, offering a more flexible and targeted mechanism to capture conversational nuances.

3 Methodology

This section details the problem formulation, and how it was instantiated for the task of detecting abusive language in Reddit conversations, leveraging graph-based modeling and GNNs.

3.1 Problem Formulation

The task of detecting abusive language in social media conversations can be formulated as a binary classification problem. Given a conversation thread T consisting of N comments, our objective is to classify whether a specific comment, c_i, within the thread is abusive ($y_i = 1$) or non-abusive ($y_i = 0$), incorporating its conversational context.

Let T represent a conversation thread of N comments:

$$T = \{c_1, c_2, \ldots, c_N\},$$

where c_i is the i-th comment to have been posted in the thread T ordered by posted time. Each comment c_i has an associated text u_i.

The thread T has a graph structure, with comments connected based on reply relationships. This structure is represented as a directed graph:

$$G(T) = (V, E),$$

where V is the set of nodes representing comments, and E is the set of edges representing reply relationships. An edge $(c_j, c_i) \in E$ exists if c_i is a reply to c_j. For each node $v_i \in V$, the feature vector $\mathbf{x}_i \in \mathbb{R}^d$ is derived from the [CLS] token embedding of its text u_i using a pre-trained language model, such as BERT.

For a target comment $c_i \in V$, the task is to predict its label $y_i \in \{0, 1\}$, where 1 denotes *abusive*. This prediction is made using the graph $G(T)$ and features derived from the comment texts u_i and the graph structure.

3.2 Affordance-based Graph Representation

Reddit conversations can span hundreds of comments, but users typically engage only with top-level replies or the direct sequence of comments leading to the one they are responding to. To improve computational efficiency while preserving user-relevant context, we implemented a graph trimming strategy based on Reddit's UI rendering logic. This *affordance-based* approach retains only the comments that would typically be visible to a user when composing a reply, thereby approximating the cognitive context available during interaction. We empirically evaluate this strategy against a simpler *most recent* baseline that retains the most recent preceding comments. While the affordance-based method achieves higher performance, the differences are not statistically significant (see Table 6 for full results). Although designed with Reddit in mind, this trimming strategy is adaptable to other platforms by leveraging platform-specific UI features or interaction cues—such as visibility, recency, or other relevance heuristics—to identify salient contextual information.

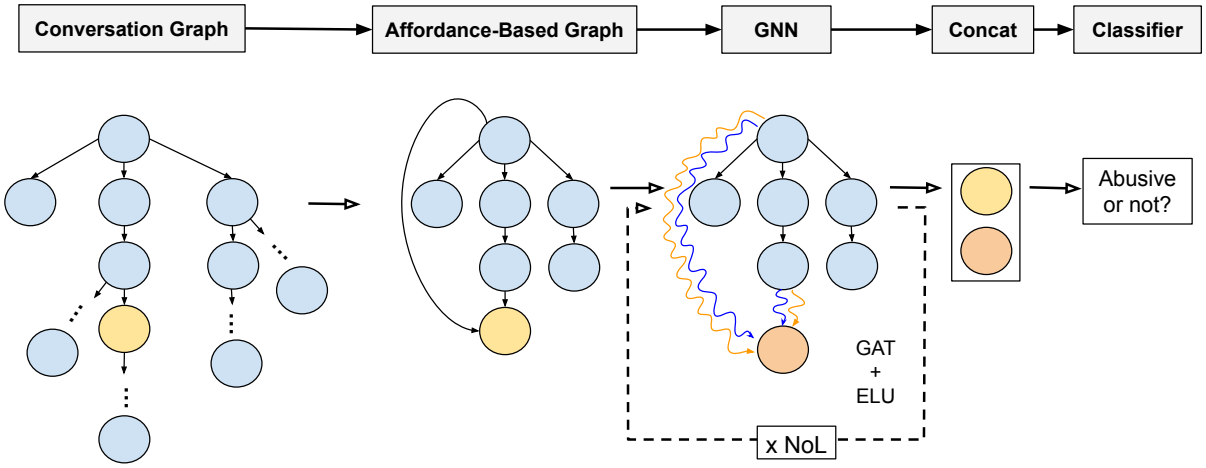

Figure 2: Overall Model Architecture. Nodes represent text embedding representations. The yellow node is for our target comment, while the orange node is for the conversation context. *x NoL* stands for *times Number of Layers*. For readability, we did not represent all the edges going from the post node to all other nodes.

The *affordance-based* trimming strategy can be formalized as follows. For each target comment c_i, we define the relevant conversational context as a subgraph G_i from $G(T)$. The subgraph $G_i = (V_i, E_i)$ includes nodes V_i and edges E_i corresponding to comments providing relevant context for c_i. This subgraph is trimmed to align with the default Reddit rendering algorithm, which determines the comments visible to a user when writing a reply. User scores correspond to the number of upvotes minus the number of downvotes for a given comment or post. The subgraph G_i includes: (1) The original post (c_1), (in blue in Figure 6); (2) The top 5 replies to the root post, ranked by user scores (in green in Figure 6); (3) The highest-scoring reply to each of these top-5 depth-1 comments (in yellow in Figure 6); (4) The full reply path leading to the target comment c_i (in red in Figure 6).

To model user interaction flow, each node connects to the original post. Formally, for a post $p \in V_i$, we add an edge (p, c_m) to E for every $c_m \in V_i \setminus \{p\}$. We also experimented with a trimmed variant where only the target node c_i connects to p via (p, c_i), but this approach did not improve performance (see Appendix A.3). Figure 6 illustrates a conversation trimmed using the affordance-based graph construction method.

3.3 Model Architecture

We employed a Graph Attention Network (GAT) (Velickovic et al., 2018) to model contextual relationships within the conversation graph. For each node v_m in the graph G, let $\mathbf{x}_m^{(l)} \in \mathbb{R}^{d_l}$ represent the node embeddings at layer l, where d_l is the embedding dimension. Each GAT layer updates the node embeddings as follows:

$$\mathbf{x}_m^{(l+1)} = \text{ELU}\left(\sum_{n \in \mathcal{N}(m)} \alpha_{mn} \mathbf{W}^{(l)} \mathbf{x}_n^{(l)}\right), \quad (1)$$

where $\mathcal{N}(m)$ denotes the neighbors of node m, $\mathbf{W}^{(l)}$ is a learnable weight matrix, and α_{mn} are normalized attention coefficients computed using $\mathbf{a}$—the learnable weight vector that calculates the attention scores between different nodes—as follows:

$$\alpha_{mn} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}^{(l)}\mathbf{x}_m^{(l)} \,|\, \mathbf{W}^{(l)}\mathbf{x}_n^{(l)}\right]\right)\right)}{\sum_{k \in \mathcal{N}(m)} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}^{(l)}\mathbf{x}_m^{(l)} \,|\, \mathbf{W}^{(l)}\mathbf{x}_k^{(l)}\right]\right)\right)} \quad (2)$$

After L GAT layers, the embedding of the target node $\mathbf{x}_i^{(L)}$ is concatenated with its text embedding $\mathbf{x}_i$, producing a final representation $\mathbf{z_i} \in \mathbb{R}^{2d}$. This representation is first passed through a fully connected layer, with parameters $\mathbf{W}_f$ and b_f, that reduces its dimensionality back to d, the original embedding size of the text model (768):

$$\mathbf{h} = \mathbf{W}_f \mathbf{z} + b_f. \quad (3)$$

The transformed representation $\mathbf{h}$ is then passed to the classifier layer, with parameters $\mathbf{W}_c$ and b_c, of the text model to predict y:

$$\hat{y} = \sigma(\mathbf{W}_c \mathbf{h} + b_c), \quad (4)$$

where σ denotes the sigmoid activation function.

The model parameters are optimized by minimizing the binary cross-entropy loss. The formulation for the loss is detailed in Appendix A.4. Details about hyper-parameters and training setup can be found in Appendix A.6.

4 Dataset

We use the Contextual Abuse Dataset (CAD) (Vidgen et al., 2021), the only high-quality dataset that provides full conversation threads to annotators for abusive speech classification.

General Description CAD is an English dataset, consisting of approximately 25,000 Reddit comments annotated using a target-based taxonomy: Identity-directed, Affiliation-directed, and Person-directed Abuse for the abusive class, and Neutral, Counter Speech, and Non-Hateful Slurs for the non-abusive class. The dataset spans 16 subreddits known for abusive content, with no single subreddit contributing more than 20% of the data. The CAD dataset is highly comprehensive, with approximately 90% of the original messages still available.

Reddit Conversation Description Reddit discussions are highly structured, with multiple parallel threads. Comment lengths range from couple of words for brief remarks to over 10,000 words, though 99.3% fit within the 512-token limit of BERT-based encoders. Conversations also vary in size, with some exceeding 400 comments, while the average training conversation contains 22 comments. Labeled comments appear at an average depth of 2.67, with a mean branching factor of 2.28, highlighting the nested and multi-threaded nature of discussions, albeit at a moderated level. Given computational constraints and the fact that users typically see only a subset of a conversation, we applied a trimming strategy based on affordances as described in Section 3.2. Trimmed graphs in our dataset contain a maximum of 25 nodes, and 9 at the median. The distribution of the number of nodes per graph is displayed in Figure 3.

Annotation Process CAD employs a rigorous annotation process, combining extensive conversational context, consensus-based adjudication, and expert supervision. Each entry was initially annotated by two trained annotators, with disagreements resolved through consensus adjudication under expert supervision, refining the annotation guidelines when necessary. A final expert review ensured consistency, yielding a Fleiss' Kappa of 0.583, indicating moderate agreement for a 6-way classification annotation. This score compares favorably with other abusive language datasets, considering the complexity of the six-class taxonomy and the inherent subjectivity of ALD (Davani et al., 2022; Vidgen et al., 2019; Sap et al., 2019; Röttger et al., 2022). Annotators had access to the full preceding thread, unlike our model, which only uses textual content and omits visual elements—a direction left for future work. Annotators recorded whether context was critical for classification, which was the case in one-third of abusive instances. We leverage this information to compare our model and baselines on context-sensitive *versus* context-free cases in Section 5.2.

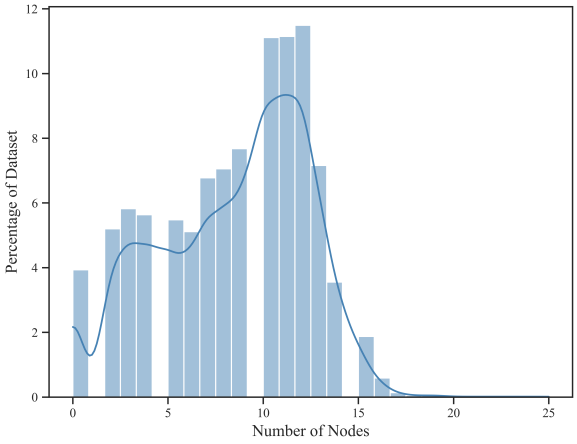

Figure 3: Node Distribution per Graph After Affordance-Based Trimming.

5 Experiments and Results

To evaluate our approach, we design experiments to address the following research questions:
- **RQ1**: What is the optimal amount of conversational context for ALD using GNN models?
- **RQ2**: How do graph-based approaches compare to other context-aware architectures for detecting abusive language?

This section details our experiments and findings, which allow us to answer the two above questions. The experimental setup for all experiments is described in Appendix A.6.

5.1 RQ1. Optimal Conversational Context

Experiments To determine the optimal conversational context for ALD, we evaluate graph models described in Section 3.3 with 1 to 5 GAT layers,

corresponding to 1-hop to 5-hop neighborhoods. Each layer aggregates node features from immediate neighbors, expanding the contextual radius. Formally, the node embedding update equation for each layer is described in Equation 1.

Results Table 1 presents the F1 scores for graph models with different numbers of GAT layers, along with the median and maximum number of nodes in their corresponding receptive fields. These values are derived from the conversation graphs in our experimental dataset. The best performance (F1 $= 76.24\%$) is obtained with three layers, where the receptive field contains a median of 5 nodes and a maximum of 12 nodes. This finding supports our hypothesis that the limited conversational context typically used in the literature is insufficient for ALD in online discussions. However, increasing the number of layers beyond three does not yield further gains and, in some cases, slightly reduces performance. This stagnation is likely due to the inclusion of less relevant distant comments in wider receptive fields. Moreover, deeper models introduce additional complexity without sufficient training data, potentially leading to overfitting and diminishing returns.

Notably, the three-hop neighborhood captures most nodes in the affordance-based graphs (see Figure 3), which, by design, are the most contextually relevant to the target comment. While extending context beyond immediate replies improves classification, the performance differences between models with two to five layers are not statistically significant. This suggests that additional conversational context beyond three hops does not provide significant gains in our dataset. Further evaluation on larger and more diverse datasets is necessary to determine whether deeper context windows can enhance performance or if they primarily introduce irrelevant information.

GAT Layers	(Max, Median) Nodes	Mean F1 $\pm$ CI
1	(3, 2)	75.37 ± 0.69
2	(7, 3)	76.13 ± 0.41
3	(12, 5)	$\mathbf{76.24 \pm 0.58}$
4	(13, 7)	75.92 ± 0.65
5	(14, 8)	76.09 ± 0.43

Table 1: Mean F1 scores (in %) with 95% confidence intervals for GAT models using different numbers of GAT layers, averaged over 10 runs. The (Max, Median) Nodes column shows the maximum and median number of comments in the corresponding k-hop neighborhoods.

5.2 RQ2. Graph-based vs. Flattened Models

Experiments We compare our graph-based models with three baselines. NO CONTEXT which classifies the target comment using BERT embeddings of the target text without considering the conversational context. TEXT-CONCAT which concatenates the target comment with preceding comments (trimmed to match the graph model's context) as a single input sequence, separated by [SEP] tokens. We use Longformer for its 4096-token limit which allows to consider and extended context. Finally, EMBED-CONCAT generates BERT embeddings for each comment, combines context embeddings pairwise through a fully connected layer to form a 768-dimensional vector, adds the target node embedding, and passes the result through a classification layer.

Results Table 2 reports the F1 scores for each model. GAT 3L achieves the highest performance (F1 $= 76.24\%$), surpassing all text-based baselines. In particular, the flattened-context models (TEXT-CONCAT and EMBED-CONCAT underperform compared to NO CONTEXT, reinforcing previous findings (Menini et al., 2021; Bourgeade et al., 2024) that naively concatenating context may introduce noise, limiting the performance for ALD.

Model	Mean F1 (%)
NO CONTEXT	74.53 ± 0.76
TEXT-CONCAT	74.17 ± 0.81
EMBED-CONCAT	74.88 ± 0.25
GAT 3L (*ours*)	$\mathbf{76.24 \pm 0.58}$

Table 2: Mean F1 score ($\pm$ 95% CI) in percentage for flattened text-based baselines and graph-based models, averaged over 10 runs.

To assess model performance in Context-Sensitive Samples (CSS), we examine instances where annotators explicitly indicated that prior conversational context was crucial for labeling. These context boolean labels are present almost uniquely for abusive samples, and the context-sensitive cases account for approximately one-third of the dataset's positive cases.

Table 3 shows that predicting Context-Sensitive Samples (CSS) is more challenging than Context-Free Samples (CFS) across all models. Our GAT model achieves the highest accuracy in both CSS and CFS settings, demonstrating the effectiveness of modeling conversational structure instead of treating context as a simple sequential input. Notably, the performance improvement of our GAT

Model	CSS PCP	CFS PCP
NO CONTEXT	70.71% ± 2.61	81.67% ± 2.89
TEXT-CONCAT	70.80% ± 3.61	82.33% ± 1.88
EMBED-CONCAT	70.97% ± 1.19	83.00% ± 1.98
GAT 3L (*ours*)	74.07% ± 1.12	84.21% ± 2.14

Table 3: Percentage of Correct Predictions (PCP) for different models on Context-Sensitive Samples (CSS) and Context-Free Samples (CFS). Results show the mean percentage of correct predictions (± 95% CI) across all predictions, averaged over 10 model runs.

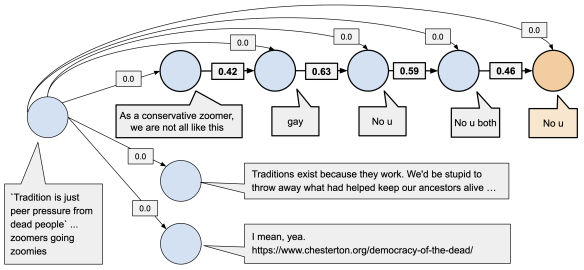

Figure 4: Example conversation graph with learned attention weights from the third layer of the best-performing GAT model. For readability, self-loop edges are omitted; their attention weights are one minus the sum of incoming edge weights.

model is more pronounced for CSS cases. Compared to the NO CONTEXT model, our GAT model achieves an improvement of 4.75% in CSS, while the improvement in CFS is 3.11%. Similarly, compared to TEXT-CONCAT, GAT 3L outperforms by 4.62% in CSS and 2.28% in CFS. This relative overperformance illustrates that our model particularly enhances predictions in cases where understanding conversational context is essential.

Computational Cost and Runtime We estimate and compare the runtime and resource usage of all evaluated models to assess their feasibility for deployment. Our GAT-based architecture demonstrates runtime, parameter count, and model size that are comparable to BERT-based baselines—such as NO CONTEXT, TEXT-CONCAT, and EMBED-CONCAT. Thanks to our *affordance-based* trimming strategy, the number of comments per conversation is typically modest, with 9 nodes at the median, and 25 at most. In the median case, GAT inference takes approximately 850 ms, closely matching the runtime of the baselines, while offering structural advantages through explicit graph modeling. By contrast, decoder-only language models such as MISTRAL-7B and LLAMA2-13B require substantially higher computational costs—ranging from 6 to 12 seconds per sample on CPU for median cases, due to full autoregressive decoding over long conversation prompts. These resource demands render them impractical for large-scale moderation pipelines. Full runtime estimates, parameter counts, and model sizes are detailed in Appendix A.5, Table 7.

5.3 Case Analysis

To better understand our model's behavior, we analyze a representative case from the test set—the conversation graph introduced in Section 1. This instance was incorrectly classified by the NO CONTEXT baseline but correctly identified as abusive by our graph-based model. The key contextual cue—a homophobic insult—appears three hops away from the target comment, a pattern common in large, multi-user conversations. As shown in Figure 1, abusive comments often trigger abusive responses, creating a snowball effect that requires tracing further up the conversation thread for proper interpretation.

To assess the impact of context depth, we conduct inference using the best-performing GAT model across 10 runs, varying the number of GAT layers from 1 to 5. As expected, models with only 1 or 2 layers fail to classify the target comment as abusive, whereas models with 3 or more layers succeed. This finding supports our hypothesis that short-context models lack the depth required to disambiguate meaning in threaded discussions, particularly in cases of abuse propagation where users reinforce or react to prior offensive content. Our results highlight the importance of sufficiently deep graph models-at least 3 GAT layers-for detecting abusive speech in complex, multi-user discussions such as those on Reddit.

To further interpret the model's predictions, we examine the learned attention weights of the GAT (3-LAYERS) model at inference time (Figure 4). The model effectively assigns attention to relevant contextual nodes within the conversation graph. Notably, edges connecting the original post to other comments receive minimal attention, indicating that the post content does not contribute to determining the abusiveness of the target comment. In contrast, edges along the reply chain leading to the target comment receive higher attention weights, highlighting their importance in contextualizing abusiveness. This analysis demonstrates that the GAT layer dynamically identifies and prioritizes key conversational cues, reinforcing the effective-

ness of structured context modeling in ALD.

6 Conclusion and Future Work

This study presents a novel methodology for incorporating conversational context using graph attention networks (GATs) for abusive language detection (ALD) in threaded discussions. Using the Contextual Abuse Dataset (CAD) from Reddit, we demonstrate that our graph-based model significantly outperforms existing context-aware baselines, particularly in instances where context is necessary to disambiguate meaning. While prior work often limits context to the root post and/or the immediately preceding comment, our findings suggest that broader and structurally informed context is essential for robust ALD in social media. By leveraging a trimming strategy grounded in Reddit's UI affordances, we incorporate extended conversational histories while maintaining compact and computationally efficient graph structures. Our approach remains scalable and lightweight, offering practical advantages for deployment in large-scale content moderation systems.

Our review of existing datasets reveals a significant lack of publicly available resources that fully captures threaded, multi-turn conversations for abusive language detection (ALD). Future research should focus on the development of such datasets to support cross-platform, multilingual, and large-scale evaluation. Extending our approach to other platforms and data sources will help assess its generalizability across diverse social and linguistic contexts. We also plan to incorporate richer contextual signals, particularly user-level information such as posting history or behavioral patterns, which have been shown to improve ALD and related tasks (Das et al., 2021; Ribeiro et al., 2018). This additional layer of context is especially important for disambiguating subtle forms of abuse, such as sarcasm, irony, or the reappropriation of degrading terms. Another promising direction is the integration of multimodal content—such as images, videos, or embedded links—into the analysis. Many abusive or harmful posts rely on non-textual cues that are essential for understanding user intent and conversational dynamics (Hebert et al., 2024). Incorporating these signals could enable more comprehensive and socially aware models of abusive language in online discourse.

Limitations

While this work highlights the effectiveness of graph-based methods for incorporating conversational context into Abusive Language Detection (ALD), it also exposes several limitations inherent to both our approach and the broader ALD research landscape.

Data Considerations Our findings are constrained by the characteristics of the Contextual Abuse Dataset (CAD), which limits the generalizability of our results in several ways. First, the dataset is restricted to English-language content from Reddit, preventing conclusions about multilingual or cross-platform applicability. Moreover, CAD was constructed via community-based sampling from 16 subreddits with a history of offensive content, which may introduce bias and reduce representativeness relative to broader social media discourse. The small dataset size also restricts the complexity and generalizability of our models. As prior work has shown, combining datasets across related abusive language tasks (e.g., hate speech, sexism, antisemitism, abuse, offensive) can improve generalization (Bourgeade et al., 2023; Swamy et al., 2019). We intend to explore this strategy alongside data augmentation strategies, such as synthetic data generation and perturbation-based augmentation, to increase training data diversity. To further validate our results, we plan to evaluate our approach on larger and more diverse corpora.

Defining Abuse, Subjectivity and Biases Abusive language detection lacks a universal definition, with studies adopting varying taxonomies for hate speech, toxicity, and offensiveness, leading to inconsistencies in annotation and evaluation (Vidgen et al., 2019; Fortuna et al., 2020). Existing datasets, such as CAD, reflect cultural and social biases, which can impact model predictions. Especially, CAD has been annotated by 12 annotators, mostly British English speakers, limiting generalizability. Additionally, Reddit-specific data with distinct language norms annotated by academic researchers introduces biases that can distort model predictions. For example, African American English (AAE) markers are often misclassified as abusive due to annotator bias (Sap et al., 2019). Vidgen et al. (2021) attempted to improve annotation quality and consistency through a consensus-based approach, but the lack of access to initial annotator disagreements prevents a deeper analysis of the subjective nature

of online abuse perceptions. Methods for subjectivity modeling include integrating multi-annotator models (Davani et al., 2022) and techniques addressing subjective annotation uncertainty (Rizos and Schuller, 2020; Helwe et al., 2024).

Scalability and Computational Efficiency While our approach demonstrates clear improvements over context-agnostic baselines, deploying such models on large-scale social media data requires optimizations to ensure efficiency without compromising performance. Our work introduced affordance-based pruning techniques to reduce the conversation graph size while focusing on relevant context. However, using graph networks still adds costs, energy consumption, and computational overhead which should be considered when scaling to real-time applications.

In summary, while graph-based methods advance ALD, challenges remain in mitigating biases, enriching contextual modeling, ensuring cross-platform generalizability, and improving scalability. Addressing these will be crucial for fair, efficient, and practical ALD systems.

Ethical Considerations

Potential Risks Our work contributes to the development of context-aware models for abusive language detection (ALD), which can aid in moderating harmful content on social media. However, automatic ALD systems present inherent risks, particularly when deployed without human oversight. False positives may result in the unjust removal or suppression of benign content, potentially restricting freedom of expression, while false negatives may fail to detect harmful speech, enabling the spread of abuse. Given these limitations, human oversight is essential, and users should retain the right to appeal algorithmic moderation decisions. Future work should focus on improving robustness, reducing errors, and mitigating biases to enhance the reliability of ALD systems. Additionally, the use of ALD models must align with ethical guidelines and platform policies to prevent misuse. Potential risks include weaponization for mass reporting, over-censorship, or the reinforcement of societal biases.

Data Privacy and Bias All experiments were conducted using the Contextual Abuse Dataset (CAD) (Vidgen et al., 2021), a publicly available dataset derived from Reddit. The dataset has been anonymized to remove personally identifiable information. However, ALD models trained on existing datasets may inherit biases from annotation processes, as discussed in Section 6. Biases related to cultural context, dialects (e.g., African American English) (Sap et al., 2019), or platform-specific discourse norms can lead to disproportionate misclassifications. Addressing these biases requires ongoing evaluation, diverse datasets, and improvements in annotation methodologies to mitigate unintended harms.

Transparency and Reproducibility We provide a detailed account of our methodology, dataset statistics, and hyperparameter settings to facilitate transparency and reproducibility. The code is released publicly to encourage further research and independent evaluations.

Acknowledgments

This work was conducted as part of the TIERED project, supported by the French government through funding managed by the National Research Agency (ANR) under the France 2030 program (ANR-22-EXES-0014). It was also partially funded by the SINNet project (ANR-23-CE23-0033-01). Additional support was provided by the ANR under the France 2030 program PRAIRIE (ANR-23-IACL-0008).

References

Vibhor Agarwal, Anthony P. Young, Sagar Joglekar, and Nishanth Sastry. 2023. A graph-based context-aware model to understand online conversations. *ACM Transactions on the Web*, pages 1–25.

DeepSeek AI. 2024. Deepseek-v3: Advancing scalable mixture-of-experts language models. *arXiv preprint*, arXiv:2412.19437.

Atijit Anuchitanukul, Julia Ive, and Lucia Specia. 2022. Revisiting contextual toxicity detection in conversations. *J. Data and Information Quality*, 15(1):6:1–6:22.

Arthur Heitmann, Stas Bekman. Arctic shift download tool. `https://arctic-shift.photon-reddit.com/download-tool`. Accessed: Aug. 30, 2024.

Matthew Beatty. 2020. Graph-based methods to detect hate speech diffusion on twitter. In *Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 502–506, The Hague, Netherlands. IEEE.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixin Sun. 2024. Humans need context, what about machines? investigating conversational context in abusive language detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8438–8452, Torino, Italia. ELRA and ICCL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, page 1877–1901. Curran Associates, Inc.

Noé Cecillon, Vincent Labatut, Richard Dufour, and Georges Linares. 2021. Graph embeddings for abusive language detection. *SN Computer Science*, 2(1):1–15. LIA (Laboratoire Informatique d'Avignon).

Ke-Li Chiu and Rohan Alexander. 2021. Detecting hate speech with GPT-3. *CoRR*, abs/2103.12407.

Mithun Das, Punyajoy Saha, Ritam Dutt, Pawan Goyal, Animesh Mukherjee, and Binny Mathew. 2021. You too brutus! trapping hateful users in social media: Challenges, solutions insights. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, HT '21, page 79–89, New York, NY, USA. Association for Computing Machinery.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 11, pages 512–515.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J.L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R.J. Chen, R.L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S.S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W.L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X.Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint*, arXiv:2405.04434.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maeve Duggan. 2017. Online harassment 2017.

Charles Duong, Lei Zhang, and Chang-Tien Lu. 2022. Hatenet: A graph convolutional network approach to hate speech detection. In *2022 IEEE International Conference on Big Data (Big Data)*, page 5698–5707.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association. Accessed: Jul. 30, 2024. [Online]. Available: https://aclanthology.org/2020.lrec-1.838.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos,

and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Jennifer Golbeck, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Zahra Ashktorab, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, Derek Michael Wu, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, and Quint Gregory. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*, pages 229–233. ACM Press.

Liam Hebert, Lukasz Golab, and Robin Cohen. 2022. Predicting hateful discussions on reddit using graph transformer networks and communal context. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 9–17.

Liam Hebert, Gaurav Sahu, Yuxuan Guo, Nanda Kishore Sreenivas, Lukasz Golab, and Robin Cohen. 2024. Multi-modal discussion transformer: Integrating text, images and graph transformers to detect hate speech on social media. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:20.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. Mafalda: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Musa İhtiyar, Ömer Özdemir, Mustafa Erengül, and Arzucan Özgür. 2023. A dataset for investigating the impact of context for offensive language detection in tweets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1543–1549, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of twitter conversation threads. *Know.-Based Syst.*, 275(C).

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *CoRR*, abs/2103.14916.

Etienne Papegnies, Vincent Labatut, Richard Dufour, and Georges Linares. 2019. Conversational networks for automatic online moderation. *IEEE Transactions on Computational Social Systems*. Also available as arXiv preprint: arXiv:1901.11281.

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. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4296–4305, Online. Association for Computational Linguistics.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(11).

Georgios Rizos and Björn W. Schuller. 2020. Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. In *Proceedings of the 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020)*, pages 42–55. Springer, Cham.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The structure of toxic conversations on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 1086–1097, New York, NY, USA. Association for Computing Machinery.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, page 940–950, Hong

Kong, China. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Benajiba, Rene Caudwell, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Jason Wang, Kaiqun Fu, and Chang-Tien Lu. 2020. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (BigData)*, pages 1699–1708. IEEE.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, page 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

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 M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34:28877–28888.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Victoria Zayats and Mari Ostendorf. 2018. Conversation modeling on reddit using a graph-structured lstm. *Transactions of the Association for Computational Linguistics*, 6:121–132.

A Appendix

A.1 Text Embeddings

We evaluated multiple text encoders for generating text embeddings in our models. Table 4 reports the F1 scores of NO CONTEXT classifiers fine-tuned on the CAD dataset using different base encoders. BERT achieved the highest performance and was selected for all experiments. Importantly, our results are independent of this choice, as the same encoder was used across all baselines and graph-based models to ensure a fair comparison.

Model	F1 Score (%)
BERT	74.5
ROBERTA	63.3
XLM-R	71.8
MODERN-BERT	66.7

Table 4: F1 scores (in %) for NO CONTEXT models fine-tuned on the adapted CAD dataset.

A.2 Graph Construction Methods

We tested various graph construction methods to determine the most effective approach for modeling conversational context. Directed graphs outperformed undirected ones, as they better capture reply relationships. Additionally, we experimented with temporal edges (Zayats and Ostendorf, 2018), linking sibling comments chronologically to model discussion flow. However, temporal edges did not improve performance and were excluded from the final model.

Figure 5 illustrates the different graph structures, and Table 5 reports their respective F1 scores. All methods were evaluated on the 3-layer GAT architecture (Section 3.3). The directed graph without

temporal edges achieved the highest performance and was used in all experiments.

Graph Type	F1 Score (%)
Directed	76.5
Undirected	75.7
Directed + Temporal Edges	76.1
Undirected + Temporal Edges	75.6

Table 5: F1 scores (in %) for different graph construction methods.

A.3 Trimming Strategies

To assess the effectiveness of different context selection methods, we compare multiple graph trimming strategies for reducing conversational graphs. Our primary method is an *affordance-based* strategy that replicates Reddit's UI rendering algorithm to determine which comments are visible to users when writing a comment (Section 3.2). As a baseline, we implemented a *most recent* trimming strategy, which retains the 25 most recent preceding comments while removing all subsequent ones. This threshold corresponds to the maximum subgraph size produced by the *affordance-based* method in our dataset, enabling a fair comparison in terms of graph size.

We further explored two edge configurations for affordance-based graphs: (1) $trim_{final}$, in which the root post is connected to all nodes, assuming that users always read the post before commenting, and (2) $trim_{alt}$, where the post is connected only to the target comment. These configurations are illustrated in Figure 6.

Table 6 reports the F1 scores for the different trimming strategies, evaluated using the 3-layer GAT architecture (Section 3.3) over 10 seeded runs. The *affordance-based* strategy with the $trim_{final}$ edge configuration slightly outperforms both the alternative setup and the *most recent* baseline, supporting the value of modeling context based on user-visible content. However, more data would be required to establish a statistically significant advantage for the *affordance-based* method.

Trimming Strategy	F1 Score (%)
Affordance-based with $trim_{final}$	**76.24 ± 0.58**
Affordance-based with $trim_{alt}$	76.11 ± 0.47
Most recent	75.94 ± 0.43

Table 6: Mean F1 scores (in %) with 95% confidence intervals for different trimming strategies, using the 3-layer GAT model averaged over 10 runs.

A.4 Objective Function

The classification task is formulated as estimating the conditional probability distribution:

$$P\left(y_i \mid G_i, \{u_j : c_j \in V_i\}\right),$$

where $\{u_i : c_i \in V_t\}$ denotes the textual content of comments within the subgraph G_t. The model parameters θ are learned by minimizing the binary cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{|D|} \sum_{(G_t, y_t) \in D} \Big[y_t \log f_\theta(G_t, \{u_i\}) + (1 - y_t) \log \left(1 - f_\theta(G_t, \{u_i\})\right) \Big], \tag{5}$$

where D represents the training dataset, and f_θ is the graph-based classification model that integrates both structural and textual features.

A.5 Computational Cost and Runtime

To provide insight into the computational efficiency of each model, Table 7 reports estimated CPU inference times, parameter counts, and model sizes. Estimates assume inference on a single sample with 9-comment Reddit threads (median case), each averaging 120 tokens. In BERT-based models, inference time is proportional to input length, with some parallelization used for multi-input architectures. GAT 3 L overhead remains moderate despite additional layers, making it suitable for real-world deployment. In contrast, decoder-only LLMs like MISTRAL-7B and LLAMA2-13B require full autoregressive decoding over long sequences, resulting in orders of magnitude higher latency and memory use.

These numbers are intended as indicative order-of-magnitude estimates. Actual runtimes depend heavily on system hardware, software stack, and deployment environment, and should be measured empirically in production settings.

Model	Params	Size	Runtime
NO CONTEXT	110M	420 MB	150 ms
TEXT-CONCAT	110M	420 MB	350 ms
EMBED-CONCAT	111M	424 MB	600 ms
GAT 3L (OURS)	112M	425 MB	850 ms
MISTRAL-7B	7B	28 GB	6 s
LLAMA2-13B	13B	52 GB	12 s

Table 7: Estimated CPU inference runtime (per sample, median 9-comment context), parameter count (params), and model size for each model.

A.6 Experiment Setup

Dataset Adaptation We evaluate models on the augmented and balanced Contextual Abuse Dataset

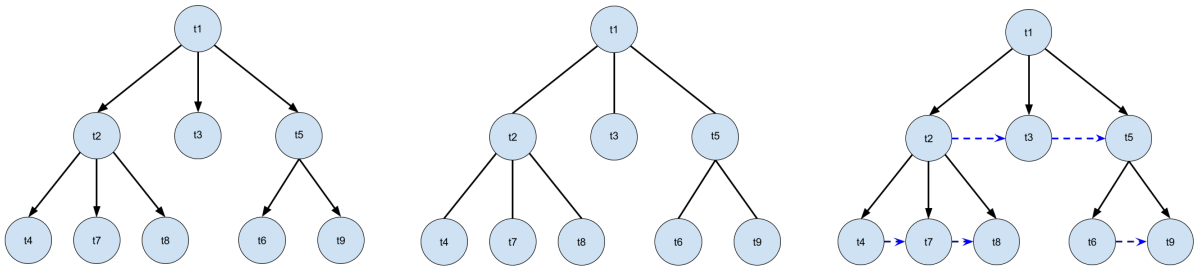

Figure 5: Diagram of Reddit conversation graphs constructed using different edge methods. Node labels t_i indicate comment publication times, with $t_i < t_j$ if $i < j$. From left to right: directed graph, undirected graph, and directed graph with temporal edges.

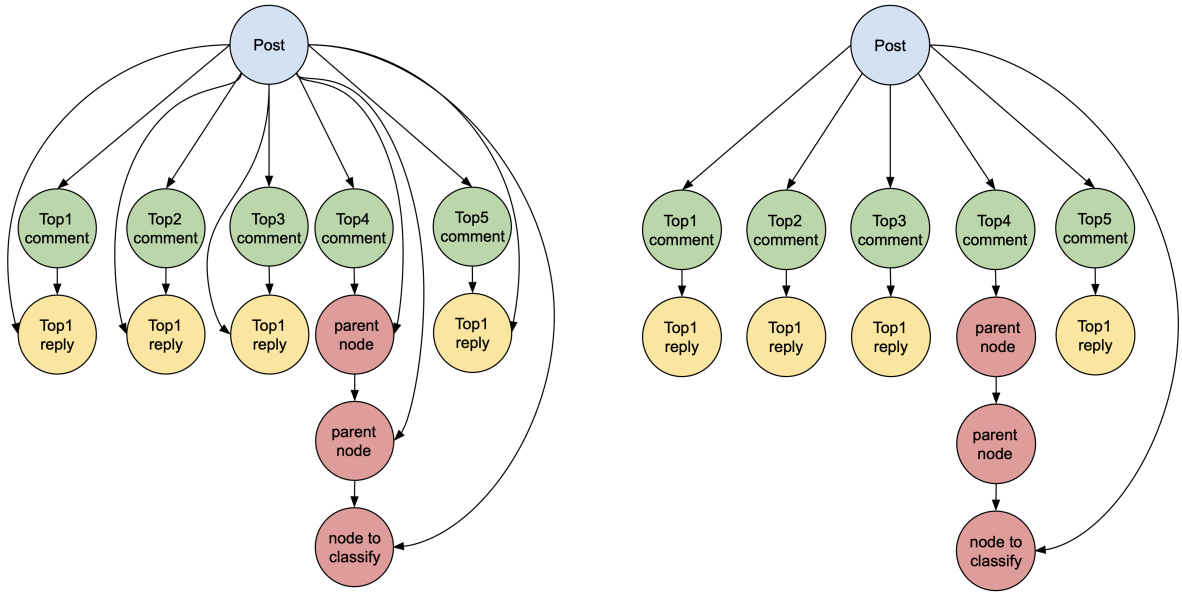

Figure 6: Diagram of a Reddit conversation processed using the affordance-based method. The left figure illustrates the edge option used in the paper ($trim_{final}$), while the right one represents the alternative option with fewer edges ($trim_{alt}$).

(CAD) (Section 4). To reconstruct conversation trees for labeled comments, we use Arctic-Shift (Arthur Heitmann, Stas Bekman), an API tool for retrieving past Reddit data. Due to class imbalance (81.2% non-abusive vs. 17.8% abusive), we apply under-sampling, resulting in a balanced dataset of 7,210 samples (3,605 per class). All abusive labels are merged into a single "abusive" category, and all non-abusive labels into "non-abusive." The dataset is split into 80% training (5,768 samples), 10% validation (721 samples), and 10% test (721 samples), ensuring class balance.

Hyperparameter Tuning and Model Training The model was trained with a learning rate of 3×10^{-6}, weight decay of 0.1, a dropout rate of 0.3 for BERT embeddings, and 0.4 for GAT layers. Due to memory constraints, training used a true batch size of 1 with gradient accumulation over 16 steps, resulting in an effective batch size of 16. Early stopping was applied after seven epochs without improvement. All models were trained on trimmed conversation graphs to ensure consistency, with ten seeded runs for reproducibility. We report mean F1 scores with 95% confidence intervals, assuming a two-tailed normal distribution ($n = 10$, $t = 2.228$).

Training was performed on Nvidia RTX 8000 GPUs and on Nvidia H100 GPUs (96 GB each). The training time for a full 20-epoch run ranged from approximately 10 to 16 hours depending on the model and hardware configuration.

Implementation Details We implemented our method using PyTorch (Paszke et al., 2019) and PyTorch Geometric (Fey and Lenssen, 2019). We

initialized with pre-trained BERT weights, and utilized the Hugging Face Transformers library (Wolf et al., 2020). For optimization, we employed the AdamW optimizer (Loshchilov and Hutter, 2019).