Harnessing Synergy in Context and Emoji for Joint Detection of Harmful Online Content in Multi-turn Conversations

Feiyan Hu, Ciara Anne Byrne†, Mark Langan, Rena Maycock, Jiang Zhou
Chirp
DCU Alpha, Old Finglas Road, Glasnevin, Dublin 11, Ireland
†cdublin2023@gmail.com
{feiyan, mark, rena, jiang}@chirpfamily.com

Abstract

Detecting harmful content, such as cyberbullying, self-harm, and grooming, in self-generated content or conversations is an emerging research area with significant potential for positive social impact. However, challenges such as the scarcity of real-world conversational data, labor-intensive annotation processes, and inconsistent content policies hinder understanding and evaluating the performance of harmful content detection systems. In this study, we utilize openly available forum data to construct conversation proxies, facilitating the analysis and detection of harmful content. We undertook extensive efforts to label the conversational data using a consistent content policy developed by experts, with ten annotators contributing to the labeling process. Our experiments investigated the impact of context window size and found that performance in joint detection improved gradually up to a context window of 16 sentences, after which performance plateaued. Additionally, experiments with emojis demonstrated that using a tokenizer capable of decoding emojis yielded the best performance, while either removing emojis or converting them to text resulted in inferior outcomes.

Keywords: Cyberbullying detection, self-harm detection, online grooming detection, Deep Neural Network, conversational context, emojis

1. Introduction

The rise of digital platforms has transformed how children interact online, offering connectivity but also exposing them to risks like cyberbullying (CB), self-harm (SH), and online grooming (OG). Cyberbullying, the use of digital technologies to harass or harm others, can cause severe psychological impacts, including anxiety and depression (Kowalski and Limber, 2013). Online content promoting self-harm increases risks of self-injurious behavior (Lewis and Seko, 2016), while online grooming, where predators manipulate children, threatens their safety (Whittle et al., 2013). These interconnected harms often co-occur in digital spaces, amplifying psychological harm and exploitation risks (Milosevic, 2018). Early detection is critical for timely interventions and safer online environments.

Joint detection of CB, SH and OG is crucial due to their interconnected nature and compounding effects on vulnerable children. These threats often co-occur within the same digital spaces, amplifying psychological harm and increasing the risk of exploitation. However, joint detection systems and openly available datasets to support effective and holistic interventions have received limited research attention.

Our ultimate goal of developing a system for real-time detection of CB, SH and OG in instant messaging apps necessitates a novel dataset due to scarcity of real-world conversational data and inconsistent annotations. We address this by collecting social media forum data as a proxy for multi-turn conversations, annotated at the message level with consistent guidelines designed specifically for the three domains developed by experts. For example, CB is typically characterized as intentional and repeated aggression in an electronic context, targeting individuals who may struggle to defend themselves. In practice, however, it is difficult to capture repetition within lengthy intermittent conversations. To address this, the guidelines suggested that annotators focus on identifying intentional aggression in the conversations, helping to exclude accidental harm from being classified as CB. Unlike most public datasets with post-level annotations, comments and replies were annotated at the message level to simulate real-time chat filtering.

In multi-turn conversations prevalent on instant messaging apps, the context window, defined as the span of preceding messages fed into a detection model, plays a pivotal role in understanding the evolving intent behind CB, SH and OG, such as grooming's rapport-building or CB's sarcasm, which single-message classifiers often miss. We evaluate context window size (1 to 128 sentences) using a fine-tuned RoBERTa model, finding performance peaks at 16 sentences, particularly for OG, but plateaus beyond due to cognitive memory constraints (Kintsch, 1998).

Emojis serve as vital paralinguistic cues in digital conversations, conveying tone, intent, and emotional nuance that text alone may obscure. Their widespread use raises compelling questions about

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10082–10092
11–16 May 2026.

how people interpret and respond to these symbols, especially in contexts where tone and intent can be ambiguous. Understanding and integrating emoji semantics is thus essential for robust joint detection, our work addresses this by investigating how different emoji processing methods during training and inference affect detection performance.

Our key contributions are threefold:

- We collected and constructed a conversational dataset from openly available forum data, with annotations applied at the message level using a consistent policy developed by social scientists. This unique dataset enables the study of the targeted domains.[1]
- We are the first in the literature to develop and explore the joint detection of cyberbullying (CB), self-harm (SH), and online grooming (OG) using a message-level annotated conversational dataset from social media forums. The dataset was constructed with consistent guidelines from social scientists to simulate multi-turn conversations.
- Through extensive experiments, we demonstrate the impact of context window size (performance peaks at 16 sentences, with no further gains beyond this point) and the effect of emoji processing (inference with original emojis consistently outperforms masked or text-replaced variants, regardless of training strategy).

2. Related Work

The detection of CB, SH and OG has become a vital research area due to the growing prevalence of harmful online interactions. Early approaches treat these tasks separately, relying on traditional machine learning with hand-crafted features like Term Frequency-Inverse Document Frequency (TF-IDF) (Dinakar et al., 2011; Dadvar et al., 2012; Nahar et al., 2013; Chavan and Shylaja, 2015; Sugandhi et al., 2016) to detect CB.

Cyberbully Dinakar et al. (2011) used TF-IDF with Support Vector Machines (SVM) on YouTube comment datasets, achieving accuracies of 60–70%, constrained by the lack of contextual understanding. Similarly, Nahar et al. (2013) employed semi-supervised learning with TF-IDF on four different social networking sites, improving detection with unlabeled data and achieving F1-scores of about 0.35 using ensemble model while achieving 0.05 with baseline model that only using TF-IDF and SVM. The introduction of large-scale pretraining brought word embeddings like Word2Vec (Mikolov et al., 2013a,b), which captured semantic relationships and enhanced performance. Zhao and Mao (2017) used an autoencoder framework to improve feature representation over TF-IDF for CB detection. Rosa et al. (2018) combined embeddings with deep neural networks, achieving F1-scores of c. 0.84 on balanced Formspring datasets (Reynolds et al., 2011) using embedding pretrained on Twitter data. These datasets, often sourced from social media, suffer from class imbalance, with harmful instances comprising less than 10% of the data, challenging model generalization. Deep learning has revolutionized CB detection, as reviewed by Hasan et al. (2023). Transformer-based models like BERT achieved an state-of-the-art macro F1-score of 0.928 on Turkish Facebook activity content. It also performs better than other DNNs such as CNN, RNN + LSTM, and Bi-LSTM on multiple datasets. On the other hands RNN models like Bidirectional LSTMs (Bi-LSTMs), reported accuracies ranging from c. 82% to c. 96%.

Selfharm SH detection has evolved similarly, as reviewed by Abdulsalam and Alhothali (2024) and Arowosegbe and Oyelade (2023). Early lexical approaches on Reddit's r/SuicideWatch, with thousands of posts labeled for suicidal ideation, were limited by shallow features. Deep learning models, particularly transformers, have shown superior performance. Sawhney et al. (2018) used a C-LSTM network, achieving an F1-score of 0.827 on suicide forum data by modeling suicidal ideation, leveraging RNN's capability to classify, process and predict time series while capturing temporal dependencies in sentences. Abdulsalam and Alhothali (2024) reports that BERT and XLNet achieved accuracy scores above 73% on average across the six major categories in the test set. Both networks performed quite similarly and outperformed SVM. Using both structured and unstructured data can improve performance to facilitate passive observation of diagnosed individuals to effectively reduce suicide and self-harm incidence (Arowosegbe and Oyelade, 2023).

Online-grooming OG detection, though less studied, is critical for protecting vulnerable users. Bours and Kulsrud (2019) used PAN12 data with BoW and/or TF-IDF with SVM and/or Ridge classifier, achieving accuracies around 0.88 on PAN12 competetion. Muñoz et al. (2020) using CNN on PAN12 data, however the model resulted in a high number of false positives. Vogt et al. (2021) constructs PANC dataset from PAN12 and ChatCoder2 for early sexual predator detection, achieving an F1-

[1]Annotations are available upon request. Please email info@chirpfamily.com.

score of 0.89 with base BERT model. OG datasets are often scarce Mladenović et al. (2021) and sometimes synthetic due to privacy concerns, limiting model training. Street et al. (2024) used transformers (BERT and RoBERTa) and achieved an F1-score of up to 0.96 on PAN12 dataset, but the real-world applicability remains constrained. However, PAN12 and Perverted Justice (PJ) are the most widely used OG datasets, lacking nuanced behavioral cues present in authentic conversations.

To detect harmful content, a multitude of features have been explored, including lexical indicators (e.g., TF-IDF or n-grams), syntactic structures, and user metadata such as posting frequency or network centrality (Huang et al., 2014). While these elements provide useful signals, such as keyword-based aggression markers in CB or relational ties in graph-based models, they often fail to capture the dynamic, interactional nature of online harms, which manifest progressively across dialogues rather than in isolated sentences. In contrast, context windows or conversation segments emerge as a superior mechanism for modeling sequential dependencies, enabling models to integrate preceding messages or thread histories into harm prediction. This approach is particularly vital for OG, where predatory intent unfolds through rapport-building phases spanning multiple exchanges, and for SH, where subtle escalations from distress signals to explicit ideation require temporal awareness. As for CB, one important cue is repeated aggression (Van Hee et al., 2015) towards victims, which also needs a fairly large context window to detect this. Modeling interaction of conversations among different users is less studied in the literature. In the paper, we propose 3 different methods to aggregate sentence level embeddings as well as role/speaker of the sentence to better model the user-user interaction.

Non-verbal cues, such as gender information, can also improve detection performance Talat and Hovy (2016). Emojis, as non-verbal cues, are integral to online communication, conveying sentiment and intent that complement text, making them relevant for CB, SH and OG detection. Kralj Novak et al. (2015) curated a dataset of 1.6 million tweets in 13 European languages, annotating sentiment with emojis, highlighting their role in cross-cultural emotional expression. Felbo et al. (2017) used bidirectional LSTMs to predict emojis in tweets as a pretraining task, leveraging large-scale Twitter data to address the scarcity of labeled data, achieving strong performance in eight downstream tasks (F1 õ0.80), suggesting the potential for pretraining harm detection models. Eisner et al. (2016) projected emojis to the Word2Vec vector space, achieving robust performance with limited data. Chen et al. (2018) generated positive and negative emoji embeddings using FastText on Twitter data, enabling nuanced sentiment analysis to identify harmful intent. Yuan et al. (2022) employed graph-based approaches with EmojiNet (Wijeratne et al., 2017) to capture emoji semantics, relevant for OG detection by modeling predatory intent. Kimura and Katsurai (2017) created multidimensional sentiment vectors for emojis by analyzing co-occurrences with affective terms in a 415k tweet corpus, evaluated against the Emoji Sentiment Ranking, offering a framework for SH detection.

However, challenges include emoji ambiguity (e.g., 😊 can be sarcastic in CB), cultural variations in interpretation, and the difficulty of unifying emoji sentiment across languages (Kralj Novak et al., 2015). Multimodal models integrating emojis, text, and user metadata, along with cross-lingual emoji datasets, are needed to enhance joint detection robustness. Our research is the first in the literature to provide consistent message-level annotations for joint CB,SH and OG detection.

3. Data Collection

Posts related to CB, SH and OG were collected from multiple social media platforms to construct simulated conversations. Posts from Reddit and Imgur were first limited by search quires with keywords curated from academic papers. For example, "10basicfactsaboutme", "schoolmemories", "hypocrite" were used for searching cyberbullying related posts and self-harm related posts were confined by quires such as "secretsociety123", "selfharmmm", "fitspiration", etc. Since online grooming is very rare on public forums under heavy scrutiny, data from the PAN 2012 Sexual Predator Identification competition (Inches and Crestani, 2012) were used as the main data source.

The collected posts were human reviewed by two annotators and were filtered using less stringent criteria based on annotators' subjective judgment. Any posts that are not related to the three domains were excluded in the selection, which resulted in a manageable number of posts for message-level annotation in the next stage. Ten annotators with social science backgrounds and balanced perspectives were recruited for the annotation. Only posts containing messages related to the three domains were required to be annotated. Annotators were also instructed to consider the flow of a real conversation during the annotation of a post where every comment or reply was regarded as a conversation message. It was suggested that annotators estimated the intent of the current message using the context from previous messages, rather than looking ahead to understand the sentiment.

Five categories were designed in the annotation:

- Cyberbullying is assigned to messages where

potential signs of aggression that lead to or indicate the presence of cyberbullying have been identified. For example, a series of threatening messages such as "you'll regret crossing me" and "everyone hates you, and I'll make sure you know it". However, a message expressing disagreement or a harsh critique that does not involve threats or ongoing harassment will not be considered as cyberbullying.

- Selfharm labels are assigned to messages containing behaviors or expressions that could be reasonably interpreted as self-harmful or encouragining self-harm. For example, explicit expressions of self-harm or intent to harm oneself such as "I'm cutting myself again tonight" and "I like my cuts in my body".

- Online grooming is labeled to messages where an adult builds an emotional connection with a minor to exploit their vulnerabilities. For example, a message shows manipulation or emotional control to a minor such as "if you really liked me, you'd send me that pic. Don't make me feel bad, it's not a big deal".

- Other harm is assigned to messages that can be considered harmful but do not belong to the three harm categories above such as hate speech or abusive speech (Schöpke-Gonzalez et al., 2023).

- Neutral is given to messages that do not contain any harmful content.

A personal information scan was conducted on all messages in the annotated posts and sensitive information such as user names, locations, dates in the messages were removed. The final labels for each message include 1) a mask indicating if CB, SH and OG is labeled and 2) a probability indicating percentage of annotator labeled the message as positive for each class.

4. Methodology

We collected forum comments and corresponding replies in the trie data structure to facilitate traversal. Our baseline model is a RoBERTa uncased model that processes one sentence without conversation context and predicts CB, SH and OG 3 classes at the same time. Figure 1 shows the performance of the baseline model using AUC-ROC and Average Precision. Our target is to determine the 3 probability of text of Interest (ToI) and due to the conversation nature of the data, this is always the last sentence of a conversation segment. In the paper, we also refer to the conversation segment as a context window.

4.1. Pseudo labels

Pseudo labels are also contributed to training in the case annotators are making contradictory labels. It can also help to provide supervisory signal when messages are not labeled in a conversation. Pseudo labels are generated by fitting a classifier trained on intermediate labels acquired by using OpenAI moderation APIs, AWS semantic and moderation APIs. Intermediate labels generated by APIs are X_p instead of text message input X. In our proposed methods f_p is a random forest classifier that is trained only with the training data to predict CG, SH and CG. The pseudo label generator is defined in Equation 1:

$$f_p^* = \arg\min_{f_p} L_p(f_p(X_p), Y) \tag{1}$$

4.2. Loss function

We employ a composite of losses that combines ground-truth labels, pseudo labels and auxiliary supervision to address the challenges of annotation inconsistencies and data scacity. The total loss is a weighted combination of groundtruth loss (L_{gt}), pseudo-label loss (L_p) and auxiliary loss (L_{aux}) For a given message x and its corresponding labels $y \in \{0, 1\}$ for each harm class $c \in C = \{CB, SH, OG\}$ Multi-label loss is applied to groundtruth labels:

$$\begin{aligned} L_{gt} = \sum_{c \in C} \mathbb{1}_{x \in gt} \cdot [y \log f_\theta(x) + \\ (1-y) \log(1 - f_\theta(x))] \end{aligned} \tag{2}$$

To mitigate the limitation of sparse and inconsistent ground-truth labels, we incorporate pseudo-label loss, where X_p is generated probability of CB, SH and OG using 3 random forest classifiers:

$$\begin{aligned} L_p = \sum_{c \in C} f_p^*(x_p) \log f_\theta(x) + \\ (1 - f_p^*(x_p)) \log(1 - f_\theta(x)) \end{aligned} \tag{3}$$

The second pseudo label X_b is generated using the benchmark model f_b trained with message and only the ground truth label. L_b is defined as $L_b = \sum_{c \in C} f_b(x) \log f_\theta(x) + (1 - f_b(x)) \log(1 - f_\theta(x))$. Auxiliary loss, on the other hand, provides additional broader contextual cues that is not explicitly covered by groundth and pesudo labels. For a set of auxiliary classes aux_C (e.g., sentiment, hate speech indicator etc,.) is defined as:

$$\begin{aligned} L_{aux} = \sum_{c \in aux_C} x_p \log f_\phi(x) + \\ (1 - x_p) \log(1 - f_\phi(x))) \end{aligned} \tag{4}$$

The total loss combines the three losses with weights to balance their contributions:

$$L = w_1 L_{gt} + w_2 L_p + w_3 L_b + w_4 L_{aux} \tag{5}$$

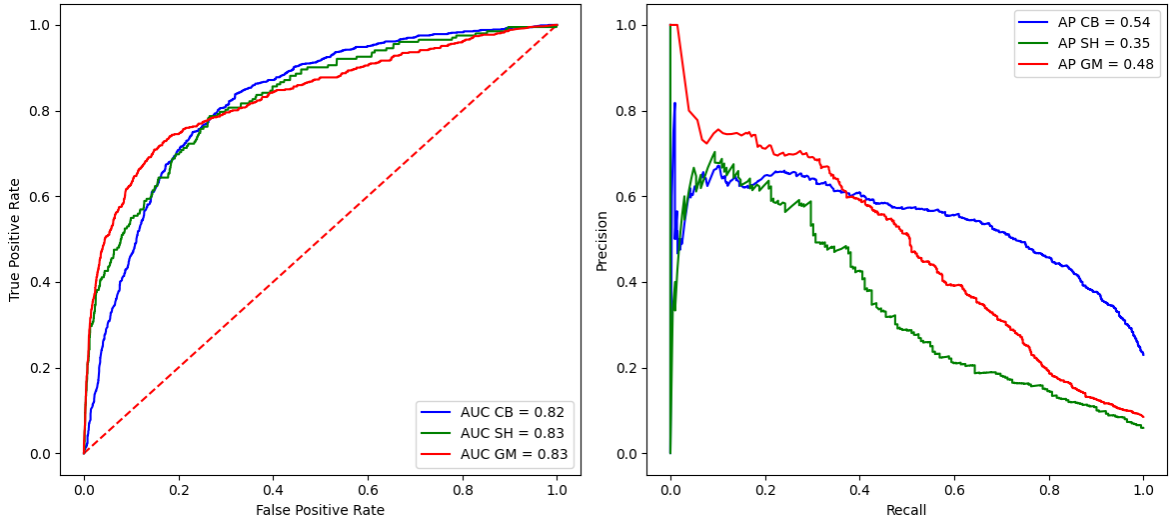

Figure 1: Baseline performance.

The weights are empirically tuned with optimal values set to $w1 = 0.2$, $w2 = 0.7$, $w3 = 0.05$ and $w4 = 0.05$.

4.3. Aggregating multi-turn sentences

Intuitively, context plays a role in how an immediate message is perceived in a conversation. To examine this, 3 different models were proposed to process reconstructed conversations of which, if a message has multiple replies, multiple conversations are constructed. The general structure of each model and approach is demonstrated in Figure 2. All the cross-attention modules used to merge context and ToI are the standard multi-head attention used in (Vaswani et al., 2017). All messages that contain only image URLs are masked out to prevent them from being attended to.

Model A The last sentence/message of a conversation segment will be used as query and treated as text of interest (ToI), while the rest of the messages will be used as context. For this model/approach, we use 2 separate RoBERTa models for the purpose of context and query encoding. After both context and ToI are encoded by language model (LM), we compute the cross-attention between all tokens from ToI branch and all tokens from context branch such that correlation between ToI and context is modeled, and the corresponding token of the query is used to predict 3 harmful classes.

Model B All messages within a conversation segment/context window are encoded by a language model (RoBERTa). However, 2 separate modules of cross-attention are used to model the correlation between user-user and user-context interaction, which is different from the previous Model A. Suppose within a conversation segment, messages that are written by the user that writes the last massage are marked as 1 and messages that are not written by the user are marked as 0. This constitutes a user mask $M_u \in \mathbb{R}^{1\times n}$ for a conversation segment with n messages. A context mask is defined as $M_c = 1 - M_u$. A user-user mask M_{uu} for cross attention is computed as $M_{uu} = M_u^T M_u$ and A user-context mask M_{uc} is defined as $M_{uc} = M_u^T M_c$. The output from 2 independent cross-attention modules are concatenated and feed into linear layer to predict classes.

Model C Instead of modeling the context and ToI explicitly, only the RoBERTa encoder is used to capture the interaction between users implicitly. The default tokenizer and embedding layer is expanded to decode additional artificial user IDs in the dictionary. Before input text to tokenization, these artificial user IDs are manually inserted at the beginning of each message within the conversation segment to reflect the evolution of the conversation throughout the time. Apart from expanding the dictionary and inserting user IDs into the message, the rest of the process is the same as training a baseline model.

4.4. Emoji

Emojis are processed using three methods, as outlined in Table 1. For instance, given the input text "Is that squizzy 😫", Method 1 retains both text and emojis from the dataset. Method 2 removes emojis, keeping only the text. Method 3 replaces coded emojis with their text descriptions. To investigate how emoji processing during training and inference

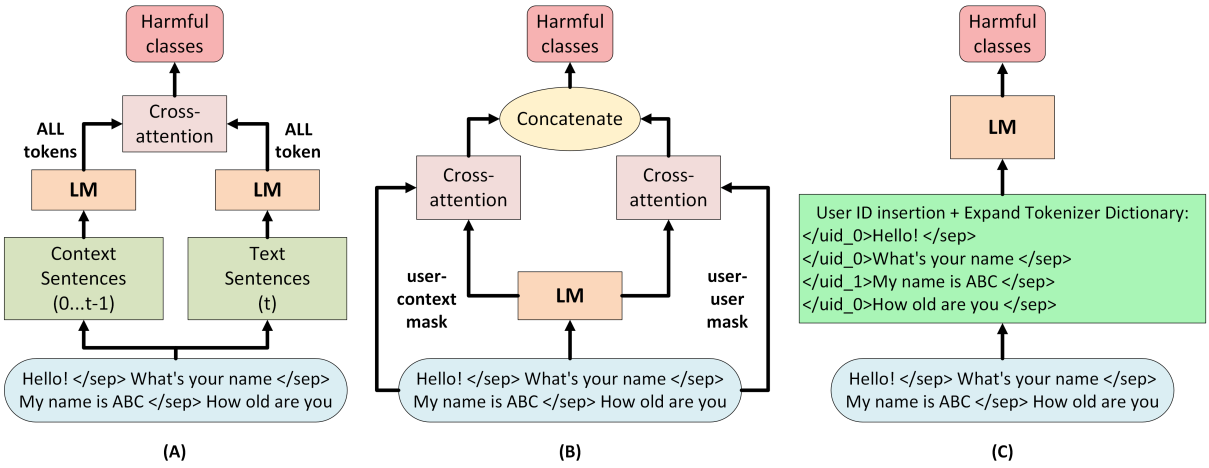

Figure 2: 3 proposed models that aggregate conversation context.

Table 1: 3 ways to process Emojis.

	Output Sentence
Method 1	Is that squizzy 🤮
Method 2	Is that squizzy
Method 3	Is that squizzy :face_vomiting:

affects model performance, we train and evaluate models using all three methods.

5. Experiments

For all experiments, we sampled posts, conversations and conversational segments to match the post distribution of each platform. We examined several sampling strategies and found that the oversampling strategy performed best. During training, posts are randomly sampled from different platforms, except for PAN12, where posts with ground truth labels were oversampled in order to compensate for the fact that the lack of labels in this dataset. 5 conversations are randomly sampled from each post, and a conversation snippet with certain context window length is also sampled randomly from each full conversation. AdamW optimizer is used with learning rate 2×10^{-6} and weight decaying 0.01. All models are trained with batch size 180 and 200 epochs. For a baseline model, the context window of one message is used as input to the text encoder. Models are trained on 2 NVIDIA L4 graphic cards, each with 24GB memory using models built by Pytorch. We balance the contribution from each platform so that in each batch the ratio of data from each platform is about the same.

5.1. Dataset summary

The dataset includes 1415 posts scraped from various social media platforms, 255 posts from Reddit, 74 posts from 4Chan, 884 posts from Pan12, 46 posts from Imgur and 156 posts from watchpeopledie.tv. 91102 messages were annotated with 4063 messages labeled as CB, 1278 messages labeled as SH and 7985 messages labeled as OG. The average message length is 17.02 tokens (± 29.62). There are 1288 posts in the training set, of which 86 posts are kept for validation and 173 posts are used for testing.

To compare messages with text & emojis and messages only with text, messages with emojis and labeled were filtered out for experiments. According to Interpreting and Translating Emojis from Safer Schools Together[2] where emojis were categorized into 8 categories, Figure 3 counts the number of emojis in each category for both neutral messages and messages rated as CB, SH and OG. It can be seen that harmful messages contain discernibly more toxic emojis than non-harmful messages.

5.2. Conversation aggregation

To find out which proposed model yields the best result, we compare 3 proposed models with the baseline model using a conversation segment/context window size of 8. The dataset is quite unbalanced with roughly 10% of the data considered harmful. Table 2 shows that model B performs the best among all proposed methods, suggesting that the independent fusion module seems to work better than the independent text encoder. For model A, we observe that using CLS token from ToI branch does not perform as well as using all tokens for cross attention. In the paper, we report AUC-ROC and Average Precision as performance metrics, since these metrics are indicators of the best overall classification performance without setting a specific

[2] https://resources.saferschoolstogether.com/view/294238754/

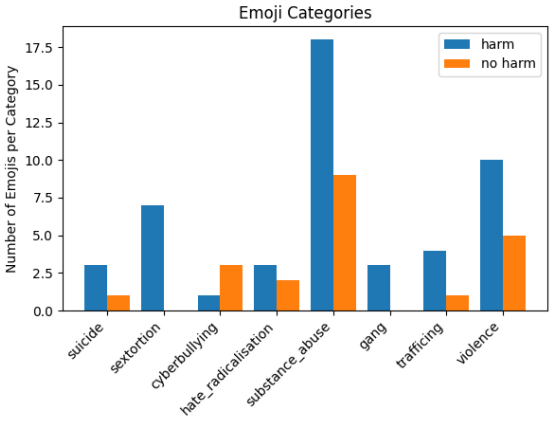

Figure 3: Number of emojis in both harmful messages and non-harmful messages.

Table 2: Comparison of Baseline and Proposed Approaches. Model A with ^{c} is trained and test with CLS token for ToI. AUC-ROC and Average Precision are reported on testing set.

	CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP
Baseline	0.82	0.54	0.83	0.35	0.83	0.48
Model A^c	0.82	0.53	0.82	0.34	0.78	0.42
Model A	0.82	0.54	0.84	0.39	0.83	0.47
Model B	0.82	0.53	0.83	0.40	0.87	0.53
Model C	0.83	0.54	0.83	0.38	0.86	0.54

threshold that can bias the result in a highly unbalanced dataset that normally less than 10% of data are positive for each class.

5.3. Ablation

The section conducts experiments to understand how each component can impact the performance of joint detection. The following experiments, unless specified, will use model B with context window size of 8 as the default conversation aggregation method.

Context window We selected context window sizes ranging from 1 to 128 sentences and used model B for training and testing. Table 3 shows that increasing window size up to 16 seems to gradually improve performance, and further increasing window size does not necessarily improve performance. This is probably due to more noise mixed into the data or fusion and/or encoding models are not able to handle context correlation that is too long.

Pseudo labels Performance is improved by introducing extra supervision to unlabeled data. Only using ground-truth gives obvious suboptimal results compared to using all pseudo- and axillary losses.

Table 3: Comparison of Context Window size of 4, 8, 16, 32 and 128. Extra weighted F1 metric is also reported.

Context	CB			SH			OG		
Window	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1
4	0.82	0.53	0.78	0.84	0.37	0.93	0.84	0.49	0.92
8	0.82	0.53	0.78	0.83	0.40	0.93	0.87	0.53	0.92
16	0.82	0.52	0.78	0.83	0.43	0.94	0.87	0.56	0.92
32	0.82	0.53	0.78	0.81	0.40	0.94	0.86	0.56	0.93
128	0.82	0.52	0.78	0.83	0.41	0.86	0.86	0.53	0.88

Table 4: Comparison of different losses using RoBERTa baseline model.

	CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP
L_{gt}	0.76	0.49	0.76	0.23	0.85	0.47
L_{all}	0.82	0.54	0.83	0.35	0.83	0.48

Table 4 shows that the improvement is particularly significant for cyberbully and self-harm conversations if all losses are included. We did not isolate the auxiliary loss in this case, because the auxiliary loss contributes weight $w_4 = 0.05$ in Equation 5, while pseudo losses dominate ($w_2 + w_3 = 0.75$). Auxiliary loss provides regularization instead of direct supervisory signal.

Emoji Table 5 shows that regardless of the emoji process used for training, if tested with unmodified Emoji, it is more likely to get good performance. The best performance occurs when both trained and tested with emojis decoded using tokenizer (M1 or Method 1). All the metrics AUC-ROC, AP, Point Biserial correlation and Brier loss score seem to support this, although improvement is not significant, due to limited number of sentences with Emoji. If we examine the overall performance in Table 6, how to process emojis does not really affect the performance. This might be due to the small number of sentences that actually contain emojis. Although having small advantages, M1 seems to have the best overall performance. The modest gains reflect the sparsity of emoji in forum data, which limits the impact on overall performance. With a higher percentage of emojis in chats, we could envisage greater benefits for decoding emojis with texts.

Cross-attention modules The depth of cross-attention layer does not really impact the performance that much. We gradually increase cross attention depth from 1 layer to 8 layers to demonstrate that increasing the fusion/cross attention layer depth does not increase the performance. The results are reported in Table 7 using Model B and a context window size of 8. It is reasonable to hypothesize that the quality of text embedding is more important than the increasing parameters of sentence aggregation module.

Sampling Different sampling strategies are also applied to sample messages from conversations

Table 5: Comparison of different methods to process emojis. Performance reported on validation and testing set to increase number of samples that contains emojis. Method 1/2/3 is implemented for training. M1/2/3 is used for testing. Since groundtruth does not contains positive samples, extra positive samples without emoji is added randomly for OG. Apart from AUC-ROC AND AP, point biserial correlation coefficient (pb) is computed. For brier loss (br) $n = 10,000$ is used for permutation to compute p values. For both correlation and brier loss, value with $p < 0.05$ is reported. To compare the performance: AUC↑ AP↑ and pb↑ br↓

	Method 1						Method 2						Method 3					
	CB		SH		OG		CB		SH		OG		CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
M1	**0.77**	**0.71**	**0.75**	**0.46**	**0.80**	**0.67**	**0.78**	**0.69**	**0.77**	**0.44**	**0.74**	0.60	**0.79**	**0.69**	**0.80**	**0.56**	0.73	0.56
M2	0.75	0.69	0.75	0.44	**0.80**	**0.67**	0.76	0.67	0.76	0.42	0.73	0.57	0.77	0.67	0.78	0.53	0.72	0.55
M3	0.73	0.66	0.70	0.43	0.76	0.66	0.76	0.68	0.70	0.42	0.72	**0.64**	0.78	0.69	0.79	0.47	**0.76**	**0.60**
	pb	br	pb	br	pb	br	pb	br	pb	br	pb	br	pb	br	pb	br	pb	br
M1	**0.47**	**0.22**	**0.33**	0.27	**0.41**	**0.22**	**0.50**	**0.22**	**0.36**	0.27	**0.36**	**0.21**	**0.51**	**0.22**	**0.40**	0.25	0.35	0.20
M2	0.44	**0.22**	0.32	**0.26**	0.37	0.24	0.47	**0.22**	0.34	**0.26**	0.32	0.23	0.48	**0.22**	0.39	**0.24**	0.31	0.22
M3	0.40	0.24	0.29	0.28	0.35	0.26	0.45	0.24	0.31	0.29	0.32	0.24	0.49	0.23	0.39	0.26	**0.39**	**0.18**

Table 6: Impact of Emoji processing to general performance.

	CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP
$M1$	0.82	0.53	0.83	0.40	0.87	0.53
$M2$	0.82	0.53	0.83	0.40	0.84	0.49
$M3$	0.82	0.53	0.81	0.38	0.85	0.52

Table 7: Impact of cross-attention layer depth to detection performance.

	CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP
$\times 1$	0.82	0.53	0.83	0.40	0.87	0.53
$\times 2$	0.82	0.52	0.82	0.39	0.85	0.51
$\times 4$	0.82	0.54	0.82	0.37	0.85	0.53
$\times 8$	0.82	0.54	0.85	0.39	0.82	0.47

Table 8: Impact of sampling strategy to performance.

	CB		SH		OG	
	AUC	AP	AUC	AP	AUC	AP
$S1$	0.82	0.53	0.81	0.39	0.85	0.49
$S2$	0.82	0.53	0.83	0.40	0.87	0.53
$S3$	0.82	0.52	0.83	0.38	0.81	0.48

and posts. S1 (random sampling) is when we randomly sample all posts, conversations and conversation segment. S2 (over-sampling) is the sampling strategy where we random sample all posts except pan12. For pan12, posts with labels are over-sampled 20 times in order to increase the positive OG samples, since it is more scarce than other harms. The S3 (depth sampling) sampling strategy is similar to S1, except that conversations are sampled by the maximum depth of it, which means that the longer a conversation, the higher chance it will be sampled. Table 8 shows that oversampling S2 is probably the best strategy among all.

6. Discussion

The observation that a context window of 16 sentences yields optimal performance in detecting CB, SH and OG in multi-turn conversations, with diminishing returns beyond this length, may stem from a combination of computational, linguistic, and cognitive factors. While the exact reasons depend on the dataset, model architecture, and task specifics, this length likely aligns with computational and human cognitive constraints to track conversational context. In multi-turn dialogues, the most recent messages typically carry the strongest signal for intent or emotional state, as earlier messages become less relevant to the current context. Human short-term and working memory have limited capacity, often cited as handling 7 ± 2 items (e.g., words or concepts) at a time (Miller, 1956), although sentence-level discourse processing reduces this slightly. In conversational contexts, humans typically track recent exchanges before older details fade unless reinforced by emotional salience or repetition (Kintsch, 1998). For OG, manipulative patterns (e.g., rapport-building transition to isolation) often unfold within a short sequence of exchanges, while CB and SH rely on immediate emotional or aggressive cues. We also observe similar phenomenon, which leads to better performance gain in OG than CB and SH. This could also impact the capacity of annotator when labeling messages, where annotators labeling CB, SH or OG rely on recent conversational context to infer intent, and a 16-sentence window may align with their cognitive capacity to recall and synthesize relevant cues. Beyond this, memory overload or fading recall may lead to inconsistent annotations, indirectly affecting model performance by introducing noisy labels.

7. Conclusion

While forum data serve as proxy for multi-turn conversations, they may not fully capture real-time instant messaging dynamics. This study advances the joint detection of cyberbullying, self-harm, and online grooming using a novel message-level annotated conversational dataset from social media forums. Experiments show that a context window of 16 sentences optimizes performance, with diminishing returns beyond due to noise or model limita-

tions, aligning with human cognitive constraints. Retaining original emojis during training and inference slightly improves detection, although their overall impact is limited. Model B, incorporating user-user and user-context cross-attention, outperforms the baseline, enhanced by pseudo-labels and auxiliary losses. These findings inform efficient context-aware detection systems, with future work needed to address class imbalance and integrate multimodal cues using efficient models.

8. Ethical Statement

The research presented in this paper was conducted in accordance with the ethical standards of the research community and was originally underpinned the auspices of the Dublin City University (DCU) Research Ethics Committee. To enable broader dataset utilization beyond initial approval, the research proceed while upholding core principles of privacy, consent and responsible AI development.

Data Privacy and Consent All data collection and processing involving human participants were performed with informed consent. Any personally identifiable information (PII) was anonymized or pseudonymized to protect participant privacy.

Implication of Automated Detection Harmful content detection is inherently high-stakes with risks of over moderation and/or under moderation. For example, emotional support for SH or sarcasm used in CB could be treated as harmful content, while subtly orchestrated OG rapport-building extend for a long period of time could evade detection. Culture-specific factors could exacerbate this, especially for paralinguistic cues such as Emojis. For example, some emojis could indicate sarcasm in a western CB context while meaning irony elsewhere. Also, linguistic norms differ by platforms for instance (e.g. slangs used in forums vs. private chat). At the moment training data are dominated by English content, which could introduce further demographic biases and thus may disadvantage non-Western content.

Mitigation of Annotation Bias Consistent instruction and policy which emphasizing intent (e.g. manipulation for OG) are developed by an expert group and given to annotators with a social science background. The current pipeline is designed to reduce individual biases; however inter-annotator variability still highlights subjectivity in harm perception.

Usage Restrictions and Safeguards Access to the dataset should be limited to researchers and practitioners working on harmful content detection, moderation, or related NLP tasks. The dataset should not be used to generate harmful content or deployed in any context that could cause harm to individuals or communities. A protocol is followed to appropriately handle the sensitive content in the data and to prevent distress to those exposed to it. Any real-word deployment of moderation models should always be guided by human oversight.

Data Access Due to the sensitive nature of the dataset containing cyberbullying (CB), self-harm (SH), and online grooming (OG) content, to prevent misuse, we kindly ask researchers to email access request with brief introduction of their project and statement that project complies with institutional ethical guidelines using their affiliated email address to info@chirpfamily.com.

9. Limitations

The research is designed to detect online harms with SoA approaches, however, there are some limitations.

Dataset Proxies and Generalizability Public forum data effectively simulates multi-turn conversations but diverges from real-world instant messaging. Lacking ephemerality, multimedia beyond emojis, or dynamic learning mechanism, it may underperform on apps where harms evolve rapidly. However acquiring such data from the public domain deemed to be a very difficult task.

Annotation Challenges Despite consistent policies, class imbalance and subjectivity could lead to inconsistencies in labels. We addressed this via pseudo and auxiliary losses but not fully resolved. Our labeling process captures the nuance of joint harms, but complicates threshold-setting in imbalanced real-word scenarios.

Model and Context Constraints Performance peaks at 16-sentence windows, then plateaus due to noise accumulation, cognitive memory limits , and transformer capacity for long dependencies. Emoji retention aids slightly, but ambiguities (e.g., toxic emojis more prevalent in harms) remain underexplored multimodally.

Broader Deployment Risks When the model is deployed on scale, low-resource languages are not tested, and ethical drift from proxy data could amplify societal biases without diverse and longitudinal validation. Future work must prioritize cross-lingual models, multimodal fusion, long-term validation and stakeholder co-design for impact.

10. Acknowledgements

This research is supported by the Disruptive Technologies Innovation Fund (DTIF) under grant No. 178866/RR from the Department of Enterprise, Trade and Employment in Ireland and administered by Enterprise Ireland (EI). We would like to thank Dublin City University for their support to this study.

11. Bibliographical References

Asma Abdulsalam and Areej Alhothali. 2024. Suicidal ideation detection on social media: A review of machine learning methods. *Social Network Analysis and Mining*, 14(1):188.

Abayomi Arowosegbe and Tope Oyelade. 2023. Application of natural language processing (nlp) in detecting and preventing suicide ideation: a systematic review. *International Journal of Environmental Research and Public Health*, 20(2):1514.

Patrick Bours and Halvor Kulsrud. 2019. Detection of cyber grooming in online conversation. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE.

Vikas S Chavan and SS Shylaja. 2015. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2354–2358. IEEE.

Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 117–125.

Maral Dadvar, Franciska MG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, pages 23–25. Universiteit Gent.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Md Tarek Hasan, Md Al Emran Hossain, Md Saddam Hossain Mukta, Arifa Akter, Mohiuddin Ahmed, and Salekul Islam. 2023. A review on deep-learning-based cyberbullying detection. *Future Internet*, 15(5):179.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-aware Multimedia*, pages 3–6.

Giacomo Inches and Fabio Crestani. 2012. Overview of the international sexual predator identification competition at PAN-2012. In *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mayu Kimura and Marie Katsurai. 2017. Automatic construction of an emoji sentiment lexicon. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 1033–1036, New York, NY, USA. Association for Computing Machinery.

Walter Kintsch. 1998. *Comprehension: A paradigm for cognition*. Cambridge university press.

Robin M Kowalski and Susan P Limber. 2013. Psychological, physical, and academic correlates of cyberbullying and traditional bullying. *Journal of adolescent health*, 53(1):S13–S20.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Stephen P Lewis and Yukari Seko. 2016. A double-edged sword: A review of benefits and risks of online nonsuicidal self-injury activities. *Journal of clinical psychology*, 72(3):249–262.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Tijana Milosevic. 2018. *Protecting children online?: Cyberbullying policies of social media companies*. The MIT Press.

Miljana Mladenović, Vera Ošmjanski, and Staša Vujičić Stanković. 2021. Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. *ACM Computing Surveys (CSUR)*, 54(1):1–42.

Fabián Muñoz, Gustavo Isaza, and Luis Castillo. 2020. Smartsec4cop: smart cyber-grooming detection using natural language processing and convolutional neural networks. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 11–20. Springer.

Vinita Nahar, Xue Li, Chaoyi Pang, and Yang Zhang. 2013. Cyberbullying detection based on text-stream classification. In *The 11th Australasian Data Mining Conference (AusDM 2013)*.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Hugo Rosa, David Matos, Ricardo Ribeiro, Luisa Coheur, and João P Carvalho. 2018. A "deeper" look at detecting cyberbullying in social networks. In *2018 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

Angela Schöpke-Gonzalez, Siqi Wu, Sagar Kumar, Paul J. Resnick, and Libby Hemphill. 2023. How we define harm impacts data annotations: Explaining how annotators distinguish hateful, offensive, and toxic comments.

Jake Street, Isibor Ihianle, Funminiyi Olajide, and Ahmad Lotfi. 2024. Enhanced online grooming detection employing context determination and message-level analysis. *arXiv preprint arXiv:2409.07958*.

Rekha Sugandhi, Anurag Pande, Abhishek Agrawal, and Husen Bhagat. 2016. Automatic monitoring and prevention of cyberbullying. *International Journal of Computer Applications*, 8(8):17–19.

Zeerak Talat 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.

Cynthia Van Hee, Ben Verhoeven, Els Lefever, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Guidelines for the fine-grained analysis of cyberbullying, version 1.0. *LT3 Technical Report Series*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matthias Vogt, Ulf Leser, and Alan Akbik. 2021. Early detection of sexual predators in chats. 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)*, pages 4985–4999.

Helen Whittle, Catherine Hamilton-Giachritsis, Anthony Beech, and Guy Collings. 2013. A review of online grooming: Characteristics and concerns. *Aggression and violent behavior*, 18(1):62–70.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. Emojinet: An open service and api for emoji sense discovery. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 437–446.

Xiaowei Yuan, Jingyuan Hu, Xiaodan Zhang, and Honglei Lv. 2022. Pay attention to emoji: feature fusion network with emograph2vec model for sentiment analysis. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1529–1535. IEEE.

Rui Zhao and Kezhi Mao. 2017. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, 8(3):328–339.