Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Timeliness Matters: Leveraging Reinforcement Learning on Social Media Data to Prioritize High-Risk Conversations for Promoting Youth Online Safety

Ashwaq Alsoubai,[1] Jinkyung Katie Park, [2] Gianluca Stringhini, [3] Meiyi Ma, [4] Munmun De Choudhury, [5] Pamela J. Wisniewski [6]

[1] King AbdulAziz University
[2] Clemson University
[3] Boston University
[4] Vanderbilt University
[5] Georgia Institute of Technology
[6] The Socio-Technical Interaction Research Lab
atalsoubai@kau.edu.sa, jinkyup@clemson.edu, gian@bu.edu, meiyi.ma@vanderbilt.edu, munmund@gatech.edu, pamwis@stirlab.org

Abstract

Ensuring the online safety of youth has motivated research towards the development of machine learning (ML) methods capable of accurately detecting social media risks after-the-fact. However, for these detection models to be effective, they must proactively identify high-risk scenarios (e.g., sexual solicitations, cyberbullying) to mitigate harm. This 'real-time' responsiveness is a recognized challenge within the risk detection literature. Therefore, this paper presents a novel two-level framework that first uses reinforcement learning to identify conversation stop points to prioritize messages for evaluation. Then, we optimize state-of-the-art deep learning models to accurately categorize risk priority (low, high). We apply this framework to a time-based simulation using a rich dataset of 23K private conversations with over 7 million messages donated by 194 youth (ages 13-21). We conducted an experiment comparing our new approach to a traditional conversation-level baseline. We found that the timeliness of conversations significantly improved from over 2 hours to approximately 16 minutes with only a slight reduction in accuracy (0.88 to 0.84). This study advances real-time detection approaches for social media data and provides a benchmark for future training reinforcement learning that prioritizes the timeliness of classifying high-risk conversations.

Introduction

Approximately 90% of U.S. youth between the ages of 13 and 17 have a social media account, with platforms like Instagram being used nearly daily (Vogels 2023; Laborde 2023). Although this widespread usage offers youth the chance to acquire knowledge, explore identity, and interact with others, it also opens them up to online dangers, like sexual solicitations and cyberbullying, as well as mental health risks that can manifest both on and offline (Nesi 2020; Reid and Weigle 2014; O'reilly 2020). The adverse impacts of these online-offline risk interactions can be long-term, significantly affecting teens' mental health, self-esteem, and overall well-being (Nesi 2020). Therefore, Social Computing and Human-Centered Machine Learning (HCML) researchers have highlighted the urgency for effective ways to identify these risks, as well as interventions tailored to youth to proactively protect them online (Chancellor 2023; Kaluarachchi 2021). Significant strides have been made that reflect a shift in focus from technology alone to a more holistic view that includes human elements in the design and implementation of risk detection models (Razi et al. 2021; Kim et al. 2021). Nevertheless, there is still a critical area for improvement, particularly in existing training methods that detect the risks without accounting for timing, highlighting the need for enhanced real-time risk detection to ensure timely and immediate response to potential threats.

In particular, advancements in risk detection need to be *responsive* to the rapid evolution of online interactions and how online risk manifests, necessitating immediate or real-time interventions that are context-specific to reduce harm, particularly among youth (Ansary 2020). First, the current reliance on entire conversations might be inefficient for timely risk detection, while the reliance on single messages could be insufficient for a comprehensive understanding of context (Sawhney et al. 2022). This limitation highlights the need for more accurate context-sensitive analyses to identify risks that may be overlooked when only considering individual elements or the entirety of conversations (Razi et al. 2023). Secondly, traditional methods that evaluate online conversations for risk detection without differentiating between their varying levels of risk severity often struggle to achieve responsiveness. Thus, it is crucial to adopt a triage approach that focuses on identifying interactions that pose higher risks by directing the right type of attention to those who need it the most (Alsoubai et al. 2024).

Lastly and most importantly, for risk detection models to be effective in the dynamic realm of online interactions, their evaluation must extend beyond accuracy to include *timeliness* (Sevtiyun, Oktadini, and Bardadi 2020). Timeliness as a measure is often evaluated based on the time elapsed between the initial occurrence of a potentially harmful interaction and the moment a response or intervention is initiated (Vosoughi, Mohsenvand, and Roy 2017). The shorter this duration, the more timely the response is considered, indicating a high efficiency in recognizing and addressing risks promptly. Thus, the ability of a model to quickly prioritize high-risk conversations or detect potential risks is as important as its accuracy in identifying them because harm can significantly increase over time if the sources of risk are not addressed quickly. To this end, in this paper, we present a novel real-time two-level algorithmic framework for prioritizing high-risk conversations within youth online interactions. To fulfill this objective, we address the following research questions:

- **RQ1:** *How can we identify optimal stopping points for evaluating conversation level of risk priority?*
- **RQ2:** *How can we optimize deep learning models to accurately prioritize high-risk conversations?*
- **RQ3:** *How is timeliness impacted by the prioritization approaches?*

To answer these research questions, we obtained access to the Instagram Data Donation (IGDD) dataset collected by Razi et. al 2022, which contains 23K risk-flagged private conversations with over 7 million messages of youth (aged 13-21) on Instagram. We leveraged youths' ground truth risk annotations of their risk experiences to train classifiers used for the real-time conversation prioritization algorithm.

In this study, we developed a two-level algorithmic framework. The first level involves training a reinforcement learning agent to determine the appropriate point to stop reading a conversation's messages and forward them for evaluation to answer RQ1. At the second level, a deep learning model assesses whether the conversation is of high or low priority based on the risk severity in the messages to address RQ2. The primary contribution of this paper is on the novel approach for conversation prioritization; therefore, we integrated pre-existing and pre-trained models for risk detection from published works by (Ali et al. 2023; Razi et al. 2023; Alsoubai et al. 2022a) to round out our framework's capabilities. This includes models for identifying sexual content and cyberbullying. For RQ3, we evaluated our proposed real-time conversation prioritization framework with two key goals: firstly, to demonstrate the efficacy of our RL agent decisions of the evaluation points in contrast to a baseline approach that processes conversations in their entirety; and secondly, to compare the overall processing time required for assessing all conversations against an approach that concentrates exclusively on conversations deemed high-risk.

In this paper, we presented a benchmark for training reinforcement learning algorithms to identify the optimal moment for halting and assessing messages in the context of conversation priority detection. Additionally, we found that identifying conversation priority at the conversation level achieved superior accuracy (Acc. 0.88) compared to an individual message level (Acc. 0.82). However, this method could potentially delay the allocation of resources. To address potential delay, we showcased the integration of a benchmarked reinforcement learning model with evaluation points, effectively balancing the determination of when to send a conversation for a priority check. Our model, which included evaluation points for prioritization, achieved a higher accuracy rate (Acc. 0.84) than the message-level model (Acc. 0.82), but lower than the conversation-level model (Acc. 0.88). it offered faster detection responses than waiting for the complete conversation. Therefore, through this work, we shed light on the empirical and practical implications of efficient conversation prioritization and the balance between accuracy and timeliness in determining conversation priority. In summary, this work makes the following novel research contributions:

- We created a novel two-level framework with the ability to identify high-risk conversations, which represents a significant advance in the field of real-time risk detection for prompt risk mitigation of high-risk online conversations of youth.
- We establish a benchmark for training a reinforcement learning agent in the context of dynamic online environments, setting a foundation for future advancements that would broaden the scope and effectiveness of real-time risk detection mechanisms.
- We demonstrate the need for balanced solutions that effectively manage the dual requirements of precision and quick response in risk detection by empirically illustrating the trade-offs between accuracy and timely identification of conversation priority.

Next, we will synthesize the related work that motivated the need for the creation of our conversation priority framework.

Related Works

Automated Risk Detection for Youth

Research on online risk detection for youth is well-established and marked by significant advancements. Numerous experts in machine learning and computational social science have developed and evaluated automated detection algorithms to detect various risks in social media such as sexual risks, cyberbullying, and mental illness (Alsoubai et al. 2022b; Obaidat et al. 2023; McHugh et al. 2023; Ali et al. 2023). In prior work, scholars have mainly aimed to create and present accurate automated detection models for these risks by heavily relying on traditional machine learning approaches such as supervised and semi-supervised models and ensemble methods. For instance, Ali et al. (Ali et al. 2023) leveraged an ensemble approach to detect unsafe conversations of youth and showed that classifiers that were trained based on metadata and relationship types performed better in terms of classifying conversations as safe or unsafe, with 87% accuracy. As the field matured, limitations have emerged, notably how these detection algorithms could overlook the nuances of human interactions and fail to adapt to the evolving nature of online risks. This led to a paradigm shift towards HCML, an approach that places humans, youth in our case, at the core of algorithm design and development (Razi et al. 2021; Kim et al. 2021; Chancellor, Baumer, and De Choudhury 2019).

A crucial element in creating human or youth-centered risk detection models is highly dependent on the quality of the dataset utilized in their training (Kim et al. 2021). Consequently, HCML scholars have placed a significant emphasis on establishing ecologically valid, ethical, and trauma-informed data collection practices designed precisely for

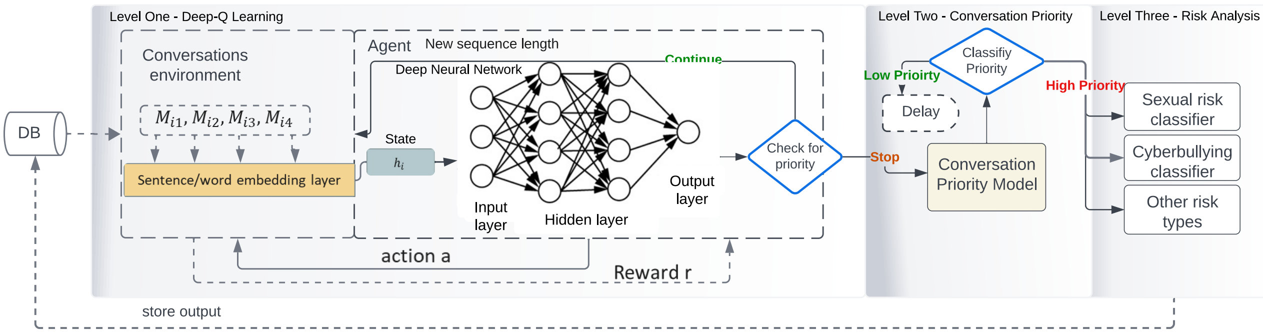

Figure 1: Structure of the conversation priority framework. The trained agent determines the optimal moment to route a specific conversation for priority evaluation. Conversations labeled as low-priority were deferred for risk analysis, whereas those identified as high-risk were directed to a dedicated high-priority pipeline for more in-depth analysis by the risk detection algorithms within the third level that is out of the scoop for this paper.

vulnerable populations like youth. The Instagram Data Donation project (IGDD) (Razi et al. 2022) represents a significant recent effort in this area, focusing on a human-centered methodology to gather and analyze private social media conversations from young individuals aged 13 to 21 years to advance risk classification research. This initiative highlights both ethical challenges (e.g., reporting child abuse for an immediate risk to minors) and technical hurdles (e.g., enhancing system efficiency for the upload of Instagram data files), all pivotal in the collection of such challenging-to-access datasets. Given the sensitive nature of the data involved, the researchers prioritized protecting the privacy of participants by restricting access to this sensitive data and allowing data sharing only through collaborative partnerships under a data usage agreement that legally protects the data from unauthorized access and/or misuse. Through these collaborative efforts, this rich dataset has been distributed among researchers, allowing them to build automated models to detect risks targeted to youth based on a human-centered understanding of their behaviors and risk perspectives (Ali et al. 2023; Razi et al. 2023; Park et al. 2023; Ali et al. 2022). For example, Park et al. (Park et al. 2023) used this dataset to conduct qualitative content analyses on media content flagged as risky by youth to inform the development of semi- and self-supervised vision transformers for media risk detection. The study found that vision transformers were able to effectively learn complex image features for automated detection of contextualized media labels (e.g., harassment, screenshots, and personally targeted). Although such initiatives have been beneficial in developing human-centered algorithmic methods, current models were geared towards recognizing risks post-occurrence, instead of using advanced techniques to proactively identify potential risks early enough to prevent victimization or minimize the damage. Therefore, we address this gap by focusing on providing a framework that would enhance the timeliness of risk detection models by utilizing the IGDD data.

Towards Real-Time Risk Detection

In recent research, scholars have highlighted the need for more timely approaches in detecting and addressing online risksby examining time-sensitive risk detection across various issues such as depression, self-harm, pathological gambling, and eating disorders using social media datasets (Parapar et al. 2023, 2022; Crestani, Losada, and Parapar 2022). The majority of the research in real-time risk detection relied on deep learning models such as Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM), particularly suited for the sequential understanding and representation of data. For instance, Liu and Wu (Liu and Wu 2020) used CNN for real-time fake news detection by adding an attention mechanism, which was used to learn how much attention should be given to the data points in the sequence. The essence of such solutions was captured in performance-based evaluations, notably, accuracy performance over time (Dodhiawala et al. 1989), underlining the importance of timely responses in risk mitigation strategies.

Yet, a key computational challenge has been acknowledged, where real-time models struggle with the vast scale of social media data, impacting their responsiveness and effectiveness (Rafiq et al. 2018). To address this challenge, a few studies have introduced a method for prioritizing conversations that would invoke risky interactions (Rafiq et al. 2018; Zhang, Yamanaka, and Li 2020). Reducing the number of conversations or messages that need to be classified or queued for risk assessment decreases the computational burden associated with feature calculation, which in turn, leads to quicker decision-making in risk detection (Zhang, Yamanaka, and Li 2020). For instance, Rafiq et al. (Rafiq et al. 2018) introduced a Dynamic Priority Scheduler (DPS) that actively allocated high priority to Instagram sessions (i.e., posts and their comments) requiring immediate review by the detection model, while assigning lower priority to others that can be deferred until new comments emerged.

While this scheduling approach demonstrated superior responsiveness compared to conventional methods, there is still a need for adaptive and more generalized allocation approaches that allocate resources based on the volume of conversations and the urgency of risk detection. One way to achieve this is by Reinforcement Learning(RL) (Bassen et al. 2020) that could learn when potentially risky conversations would need the detection algorithms. RL is a prominent technique in the field of machine learning, and it has been notably leveraged in other fields such as healthcare (Yu et al. 2021), finance trading (Liu et al. 2020), and autonomous systems (Wang et al. 2023). These prior studies have demonstrated how RL has emerged as a powerful tool for addressing diverse problems and optimizing complex systems that could contribute to technology and society advancements (Shuford 2024). Thus, while prior works have introduced prioritization methods that rely on fixed thresholds, we expanded the RL application to the online risk detection systems by applying one of the state-of-the-art RL methods, a deep Q-learning model (Mnih et al. 2013), that dynamically learns optimal decision-making policies to prioritize youths' high-risk conversations for real-time online risk detection. Our approach focuses on incorporating rewards based on youths' real experiences, making it a more youth-centered and context-sensitive approach in risk detection methodologies and the HCML field.

Additionally, this paper will be the first or one of the first papers that leverages reinforcement learning to create a novel algorithmic framework that focuses on high-risk conversations, as illustrated in Figure 1. This framework is structured around two key levels: 1) An agent responsible for determining whether to route a specific conversation for priority evaluation or to maintain it within the standard conversation stream and 2) In scenarios where the conversation is flagged for priority check, the second level, a pre-trained deep learning model, assesses the conversation's priority level. High-priority conversations are shifted to a specialized high-priority pipeline, while low-priority ones are tagged accordingly and retained within the normal conversation flow. To complete the algorithmic cycle, advanced risk detection algorithms thoroughly analyze the conversation's content within the high-priority conversations pipeline, ensuring continuous monitoring and evaluation of these high-priority interactions. Introducing this framework for prioritization, decision-making, and content analysis could enhance risk detection timeliness, effectiveness, and responsiveness in dynamic online environments.

Methods

Instagram Data Donation Dataset

In this work, we utilized data from a youth Instagram Data Donation (IGDD) project by Razi et al. (Razi et al. 2022), which was approved by their Institutional Review Board (IRB). Instagram is one of the most popular platforms among young individuals (Laborde 2023), which was why we selected the dataset.

The dataset included Instagram Direct Messages (DMs) from 194 U.S.-based English-speaking participants aged 13-21. These individuals had active Instagram accounts for over three months during their teens (13-17 years), interacted through DMs with at least 15 people, and had a minimum of two uncomfortable or unsafe DM conversations. Participants downloaded and uploaded their Instagram data and flagged their conversations as safe or unsafe. They also categorized unsafe messages according to risk types (i.e. sexual solicitations, harassment, etc.) derived from Instagram's reporting feature risk categories [1] and risk levels (i.e., high, medium, and low) grounded from existing research (Wisniewski et al. 2016). **Low Risk** messages were those that caused discomfort to the participant but were not likely to lead to emotional or physical harm. **Medium Risk** encompassed messages that had the potential to cause emotional or physical harm if they were to continue or escalate. **High Risk** messages were identified as dangerous, having already caused emotional or physical harm to the participant.

Characteristics of Dataset and Participants. The dataset was gathered from 194 young individuals aged 13 to 21 years, with the mean age being 17 years and a standard deviation of 2.21. Among them, 68% identified as female, 22% as male, and 10% as either non-binary or did not specify their gender identity. The dataset contained a total of 23,089 Instagram private conversations, with a range from a minimum of 17 to a maximum of 1477 conversations. These conversations encompassed over 7 million messages. Participants labeled 2,760 of these conversations as unsafe, causing feelings of discomfort or unsafety. These unsafe conversations contained 205,187 messages. Out of these messages, 3642 messages were labeled by participants for risk levels, with 2,025 being labeled as low risk, 1,127 as medium risk, and 490 as high risk. In this paper, our emphasis was on conversation priority for the triage approach, which was dependent on the levels of risk severity. Consequently, we limited our analysis to these risk levels, reserving the examination of specific risk types for a forthcoming paper.

Data Processing

In the data preprocessing stage, we eliminated punctuation, hyperlinks, stop words, non-Latin words, single or numeric characters, and conversations with less than three words. To retain the semantic value conveyed by emojis, we converted them into their word equivalents using the 'demoji' Python library [2]. This process was critical to maintaining the meaningful content of the conversations while filtering out irrelevant elements.

Our training approach for both the conversation prioritization and the RL model utilized the unsafe conversations only with its 205,187 messages. Messages lacking risk labels were classified as safe. For training the models in levels one and two of the framework, risk level labels were numerically encoded (safe:0, low:1, medium:2, high:3). We merged the 'low' and 'safe' labels into a '0' label to indicate low-level risk, and combined 'medium' and 'high-risk' levels into a '1' label for high-level risk messages. This decision was made because the actions for both safe and low-risk level messages would be the same, involving a delay in the evaluation of messages. Conversely, messages labeled as medium to high risk required immediate attention by the risk classifiers. Therefore, combining these categories streamlined the

[1] https://www.facebook.com/help/instagram/192435014247952

[2] demoji - https://pypi.org/project/demoji/

response process, aligning actions with the appropriate level of urgency needed for each risk category. In our approach, we focused on the textual features of the messages to establish a benchmark for training the reinforcement learning algorithm by demonstrating the algorithm's effectiveness using basic input, without the need to create complex features. The dataset was split 80% for training, 10% out of the 80% training data for validation, and 20% for testing.

We conducted experiments with transfer learning using Google Research's BERT-based ELECTRA model and word2vec. It was observed that using a pre-trained ELECTRA model, which had been trained on general English language data for generating message-level sequence embeddings, improved the classification accuracy more than the word2vec embedding. Due to memory constraints, we utilized the HuggingFace Transformers library for an ELECTRA-based implementation but could not deploy the larger model variant. The ELECTRA model generates a sequence embedding for each input message that represents the words in the message within the context of other words in the same message. Through transfer learning, we leveraged the outputs of these models as inputs for our RL and conversation prioritization model.

Models

Two-Level Conversation Prioritization Algorithmic Framework The objective of early conversation prioritization was to identify the conversation priority as high or low based on the likelihood of having unsafe messages as early as possible while keeping an acceptable accuracy performance. We employed three models: 1) an RL model that identified the stopping point for the model to review messages and trigger the prioritization model, 2) a deep learning model that classified the conversation priority, and 3) pre-trained risk detection algorithms. Unlike Rafiq et al.'s work (Rafiq et al. 2018), we did not rely on a threshold to set the priority category for the conversations as this threshold was identified based on their dataset which lacked generalizability to other datasets. Instead, we relied on first-persons' (i.e., youth) perspective of the risk severity as ground truth, which could be considered as a more ecologically valid and human-centered practice to rank the conversation priority (Razi et al. 2021).

Reinforcement Learning Model-Level One (RQ1). We utilized an RL model to determine the point at which messages needed to be evaluated for priority. The RL model was designed to reward based on the actual priority labels of the messages provided by the youth and apply a minor penalty when it failed to trigger the classifier. This approach allowed our prioritization algorithm framework to learn the balance between accuracy and timeliness in decision-making.

For the RL model, we chose the Deep Q-Network (DQN) (Mnih et al. 2013). A significant hurdle in Q-learning involves finding a reliable action-value approximation, enabling precise mapping from (state-action) pairs to expected returns in real-time. This becomes particularly challenging when dealing with nonlinear system dynamics or reward patterns, similar to our case as the conversations can have varying lengths and content, making the state space complex, and the relationship between the number of messages in a conversation and its prioritization to be nonlinear and intricate. Therefore, the DQN, a deep reinforcement learning method, was the best option to be used because of its ability to handle high-dimensional state representations and its ability to model complex and nonlinear relationships, which may not be achievable with simpler and off-policy algorithms.

The training of DQN was done using kerasRL2[3] and TensorFlow 2 (Abadi et al. 2016) with integration of Gym (Brockman et al. 2016) to build our custom environment. The action space consists of two options: 0, signifying the action of continuing to read messages without assessment, and 1, signifying the action of halting to forward the messages that have been previously read and trigger the conversation prioritization model. For a given step t, the agent reads the messages that were sent from the environment and takes action. Once the agent made an action for the given step, the environment calculated the reward. The reward was determined based on the risk severity levels labels when the agent took an action whether it was continuing to read messages or stopping for evaluation, where r_i can be expressed as follows:

$$r_i = \begin{cases} \epsilon & \text{if correct continue,} \\ -\epsilon & \text{if incorrect continue,} \\ \log(M) & \text{if correct stop,} \\ -P & \text{if incorrect stop.} \end{cases} \tag{1}$$

If it was a *continue* action and the risk levels in the messages were either safe or low risk, the ϵ was a minor reward applied since the data was imbalanced and most of the messages were either safe or low. $-\epsilon$ was a delayed penalty for having high-risk messages while choosing to delay the evaluation. If the decision was *stop* for evaluation and there were high-risk messages, then $\log(M)$ was a reward where M was the number of correct stops that have been made so far. This logarithmic scaling motivated the agent to accumulate more experience, foster continuous learning, and enhance its decision-making process by diminishing returns in terms of reward (Hester et al. 2018; Sun et al. 2018). When the decision was *stop* and the messages were low-risk or safe messages, a penalty of $-P$ was applied since the agent asked for the evaluation model to review the messages when it was unneeded.

DQN leverages the adherence of the optimal action-value function $Q^*(s,a)$ to the Bellman equation (Sutton and Barto 2018) to update the neural network weights θ_i for minimizing a specific loss function during each iteration i, which occurs after each time step. Once the DQN was trained, and $Q_\theta(s,a)$ was obtained, the action taken by the agent was determined as follows:

$$a(s) = \arg\max_a Q_\theta(s,a)$$

This equation calculates the action $a(s)$ that maximizes the estimated action-value function $Q_\theta(s,a)$. We conducted hyperparameter tuning using the validation data to optimize

[3]https://github.com/inarikami/keras-rl2

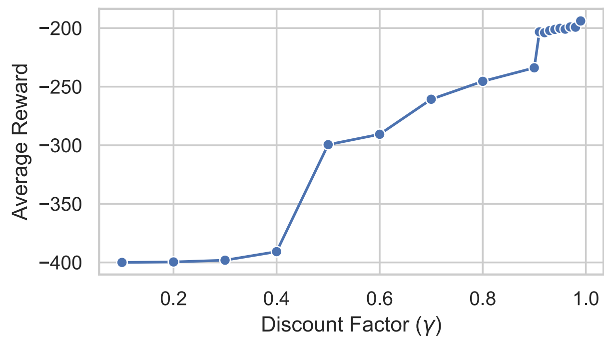

Figure 2: Impact of the discount factor (γ) on average reward: the average reward values obtained over 1000 validation episodes as the discount factor (γ) increased.

the agent's performance. As illustrated in Figure 2, an increase in the discount factor (γ) generally enhanced the reward values, reaching its peak at an optimal discount factor of $\gamma = 0.99$, the value used for training the DQN model. The network learning rate (θ) was set to 0.01, while the exploration rate (ϵ) initiated at 1 at the onset of training and then gradually decreased. The Adam optimizer with a learning rate of 0.001 was employed for stable training, which was commonly used in prior research (Kingma and Ba 2014).

Conversation Prioritization Model-Level Two (RQ2). Our approach involved training various deep learning models to determine the most effective one for conversation prioritization. This included LSTM, Bi-LSTM, CNN, and Deep Neural Network (DNN) (see Appendix for the detailed architecture of each model).

Both conversation prioritization and DQN models were trained jointly, where the DQN model generated a sequence of instances and the conversation prioritization model acted as the discriminator. The training process was done alternatingly, where training one model and keeping the parameter for the other model fixed and vice versa. The model converges when the reward value stabilizes between consecutive episodes. Additionally, to demonstrate the effectiveness of our framework, we created a set of sub-conversations using random stopping points that had similar mean and variance to the subsets extracted by our trained DQN agent. We employed uniform random sampling to choose random stopping points from each conversation while ensuring that each message had an equal chance of being selected (Munoz et al. 2019). For each randomly chosen stopping point, we extracted the conversation subset and adjusted its mean length and the risk-level distribution to match the subsets extracted from the learned stopping points.

Timeliness Assessments (RQ3)

To determine the performance of our conversation prioritization model, we compared its response time to a model that treated all conversations equally through a simulated scenario employing both approaches the baseline and the prioritization model. Additionally, we employed the time elapsed between the timestamp of the message that activates the prioritization model and the timestamp of the high/medium risk level message. The time gaps within these timestamps were exponentially distributed, with a median of 10 seconds. This distribution follows real-world social media messaging time gaps distributions (Balakrishnan 2019; Zhou et al. 2014), where the majority of the time gaps were relatively small, demonstrating back-and-forth communications. Using these metrics, we gauged the lag between the occurrence of the high/medium risk message and the point at which the trigger was activated to evaluate the conversation.

To understand the impact of the conversation prioritization approach versus treating all conversations equally, we ran a simple simulation using Python code where we sent API requests to AWS servers. The hardware configurations of these servers are listed in Table 3 in the Appendix. We used various virtual CPUs, memory, and storage configurations to illustrate the comparisons between the two approaches. We hosted pre-trained models in these servers, including the DQN agent model, Bi-LSTM for priority identification with the same architecture that was described previously in the models' section, and CNN for sexual and cyberbullying risk detection. The messages-level model was trained using the same dataset to detect risks, including sexual and cyberbullying, with detection performances of prior works (Ali et al. 2023; Razi et al. 2023; Alsoubai et al. 2022a) (Table 2). Then, our system recorded the total processing time the server took to process these conversations. We randomly sampled 20K conversations from the dataset that were not part of the training process. Then, we ran the trained conversation prioritization model (with the stopping point) against them, resulting in 1,875 predicted high-priority conversations. Through this test, we were able to compare the total processing time when the system had a prioritization approach versus when treating all conversations equally. In the next section, we present the results of this paper.

Results

Reinforcement Learning Model (RQ1)

The agent's training was done over a set of 16,000 episodes. Figure 3, provides insights into the average reward for episodes. Notably, the reward values started at around -900, which was a significantly high negative value. This suggests that the initial strategy or actions chosen by the DQN were far from optimal, which means it initially stopped reading the messages within a given conversation too early or too late, resulting in negative reinforcement. Then, the rewards increased (become less negative) to around -100 to -250. This improvement indicates that the DQN began to learn a more effective strategy for deciding when to stop and evaluate messages within a conversation. The DQN's actions started aligning better with the optimal points for stopping, as evidenced by the reduced penalty. We also found that the fluctuation in the reward values around -100 to -250 signifies that the DQN was fine-tuning its strategy.

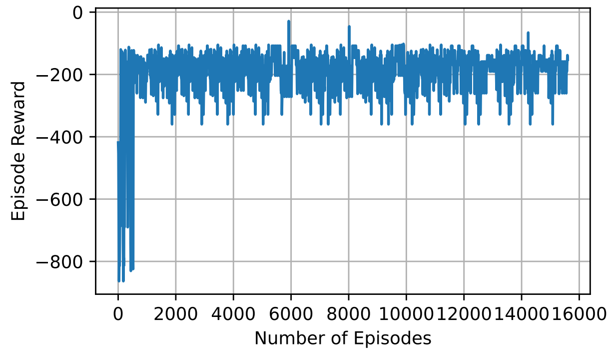

Figure 3: DQN performance: Average reward over 16,000 episodes. The rewards started at -900 and then improved to be between -100 and -200.

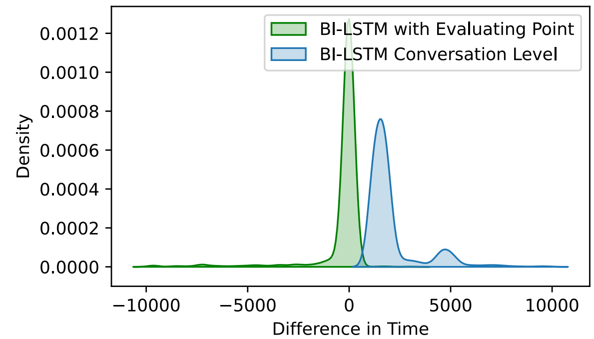

Figure 4: Density distribution to compare the timeliness performance between the stopping point model and the best-performing baseline at the conversation level.

Conversation Prioritization (RQ2)

In this section, we reported the evaluation of the accuracy of models in classifying conversation priorities. While all baseline models and benchmark models classify the conversation priorities using all messages from the conversation, Razi-CNN (Razi et al. 2023) performed the best on classifying the risk severity levels at the individual message level, using the same IGDD dataset. Additionally, for our presented model, we illustrated three models' performances: (1) the baseline model, which utilized the entire set of messages within a conversation for the classification of conversation priority,(2) a model that classified the subsets of conversations that were extracted using random stopping points, and (3) our model, which classified the conversations using a selected subset of messages determined by the DQN. Table 1 summarizes the performance metrics of the models on the dataset.

Overall, having the entire conversation as input improved the classification of the risk levels as we found that the Bi-LSTM model at the conversation level outperformed all other models, including state-of-the-art models at the conversation level, the messages level, and the conversations with stopping points. Bi-LSTM at the conversation level showed a high level of performance with the best F1 score and AUC at 0.88, and the highest accuracy at 0.89. Therefore, the Bi-LSTM model was then deployed at the message level and the conversations with random stopping points to demonstrate its effectiveness across these cases. Bi-LSTM with the conversations that had stopping points identified by the trained DQN agent performed the best compared with Bi-LSTM at the messages level and the conversations with random stopping points, yielding an averaged F1 score of 0.84. There was a marginal increase in the precision of the Bi-LSTM model when applied at the stopping point with a precision of 0.89 for the low risk class and 0.82 for the high risk class, compared with the recall and other metrics experienced a decrease in performance, with the AUC and accuracy both reporting at 0.83 and 0.84, respectively.

Evaluation of Timeliness Performances (RQ3)

While our prioritization model determines the stopping point dynamically, we compared the timeliness measures of this model against the baseline model that required the availability of all messages, which resulted in the best accuracy performance.

We found notable differences in the BI-LSTM model with the stopping point and with the conversation level performance in terms of timeliness as shown in Figure 4. The baseline model that processed entire conversations demonstrated a markedly different distribution. Its values were less concentrated around zero difference and tended more towards positive values. The positive skew in this model's distribution illustrated a delay gap between the occurrence of high-risk-level messages and the end of the conversations of between 0 to approximately 5K minutes (i.e., 3 and half days), showing its tendency to identify the high-risk conversation later. BI-LSTM with the stopping point showed a distribution predominantly centered around zero difference in the time elapsed between the message that triggered the stopping action and the timestamp of the risk messages with a slight shift towards negative values. This shift indicates the presence of risk messages occurring after the model decided to stop, in this case around 16 hours before the risks occurred.

Table 4 in Appendix shows the time processing differences between our conversation prioritization approach and the approach that treats all conversations equally across three types of systems configurations. Overall, the enhancements in computational power, memory, and storage have contributed to decreased response times. However, it is consistently observed that prioritizing high-risk data led to more efficient processing, which was particularly evident in the m5.xlarge instance, where it consistently outperformed the equal risk prioritization approach. In the high-risk prioritization system using m5.xlarge as shown in Figure 5, a significant concentration of conversations was found within a time frame of 0 to 9 minutes, roughly less than 500,000 milliseconds. This means most high-priority conversations were processed faster, with the entire batch of high-risk pri-

Level	Model	Class	Precision	Recall	F1-Score	F1-(Avg.)	AUC	Accuracy
Messages	**Razi-CNN***	Safe	0.80	0.83	0.82	0.82	**0.88**	0.82
		Low Risk	0.79	0.77	0.78			
		High Risk	0.88	0.86	0.87			
	Bi-LSTM	Low Risk	0.80	0.85	0.82	0.78	0.79	0.80
		High Risk	0.75	0.74	0.74			
Conversations	**Bi-LSTM**	Low Risk	0.90	0.89	0.89	**0.88**	**0.88**	**0.89**
		High Risk	0.86	0.86	0.86			
	LSTM	Low Risk	0.90	0.88	0.88	0.85	0.85	0.85
		High Risk	0.80	0.83	0.81			
	CNN	Low Risk	0.87	0.85	0.86	0.84	0.84	0.84
		High Risk	0.81	0.81	0.81			
	DNN	Low Risk	0.88	0.85	0.86	0.81	0.81	0.82
		High Risk	0.75	0.76	0.75			
Stopping Points	**Bi-LSTM (DQN)**	Low Risk	0.89	0.85	0.87	0.84	0.83	0.84
		High Risk	0.82	0.79	0.80			
	Bi-LSTM (Random)	Low Risk	0.80	0.79	0.78	0.73	0.76	0.77
		High Risk	0.65	0.71	0.67			

Table 1: Accuracy performances comparison of our state–of–the–art conversation prioritization model. '*' denotes values taken from the original publications.

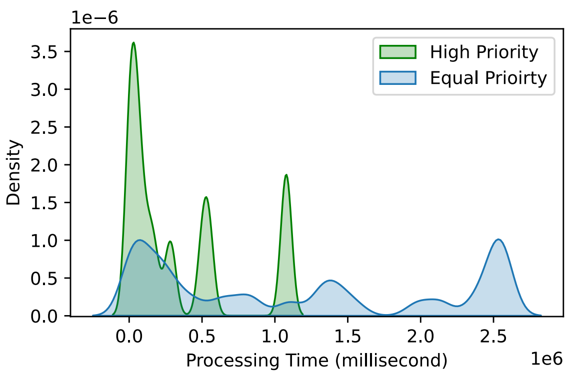

Figure 5: Density distribution of processing times for models with conversation priority versus with equal priority using m5.xlarge EC2 instance.

ority conversations being processed in the server within a 25-minute window.

Conversely, when the system treated all conversations equally, a different pattern was shown on the density graph. We found a lower intensity observed across most of the graph, suggesting a more even distribution of conversation processing times without any significant peaks. However, a notable spike in density was found over the one-hour mark (over 2.5 million milliseconds), indicating a concentration of conversations being processed around this duration. This peak reveals a slower overall processing time than prioritizing high-risk conversations.

Discussion

Considerations for Risk Detection Research

Our work highlights that by adopting conversation prioritization methods, timely and efficient decision-making for high-risk conversations is achievable, utilizing less time and fewer resources compared to the analysis of entire conversations. While not all social media interactions need to be monitored equally (Rafiq et al. 2018), current practices in the real-time risk detection literature focus on classifying all social media conversations without having a process to prioritize high-risk conversations. Most data processing approaches in prior research include segmenting the dataset into chunks of data of equal size, predefined by a fixed time window (Li et al. 2022; Lu et al. 2022) or by a fixed number of posts (Leiva and Freire 2017; Zhao et al. 2021), to be fed to the training models sequentially, which are not representative of real-world peoples' interactions. Our work was the first to apply RL and deep learning approaches to optimize conversation prioritization for real-time risk detection with acceptable accuracy performance (0.84 accuracy). Our results confirmed that with RL, the model finetunes itself to find optimal points to identify priority evaluation points within conversations. With optimal evaluation points, we showed that our deep-learning model can significantly increase the timeliness of the detection, which is critical for providing timely and customized interventions to mitigate high-risk interactions on social media. Therefore, instead of relying on pre-defined chunks of data without consideration of risk priorities, we call for future research to apply our RL approaches and the algorithmic framework as benchmarks so that the real-time risk detection models can respond to high-risk social media interaction promptly.

The observed learning curve of our RL agent within the conversation environment, characterized by initial high negative rewards followed by a gradual increase and subsequent fluctuation, is indicative of a typical reinforcement learning process (Mnih et al. 2016). This pattern highlights the significant impact of the structure of our reward function on the agent's learning trajectory, which was designed to balance various factors: minor rewards for 'continue' actions in low-risk scenarios due to data imbalance, penalties for delaying evaluation in the presence of high-risk messages, and substantial rewards for stopping for evaluation at appropriate junctures, particularly in high-risk situations. This nuanced approach to reward structuring could guide future research in real-time risk detection as it highlights the importance of carefully considering the reward mechanism to align with the environments characterized by data imbalances and levels of risk. Therefore, future research is encouraged to explore deeply how different reward function structures could optimize learning for real-time conversation prioritization, especially in environments with similar complexities. Our results were based on computational matrices (i.e., average reward); thus, we recommend future research to further examine tuning our algorithm's parameters using human evaluation by looking deeper into the context of the conversations. Understanding these dynamics could lead to the development of more sophisticated RL models that are better suited for the nuanced and dynamic nature of real-time risk detection in social media platforms.

Another key contribution of this paper is that we empirically illustrated an important dichotomy between achieving high accuracy and attaining fast response times in prioritizing conversations. We observed that when the entire conversation was processed by the model, it achieved higher accuracy in identifying the conversation priority. However, this comprehensive analysis came at the cost of speed, leading to slower prioritization of high-risk conversations. Conversely, our prioritization model with the stopping points, designed to identify the conversation priority earlier in the conversation, demonstrated a marked increase in response speed, with a slight reduction in accuracy. While many scholars have called for timely and accurate risk detection algorithms (Caddle et al. 2023), we highlight that this balance could be difficult to achieve because accuracy forces the full context (messages) to be available while timeliness forces the work with an incomplete context. Therefore, future research, especially within the youth online safety area, is urged to identify the cases where accuracy could be prioritized over timeliness and vice versa.

Specifically, online youth safety literature needs to identify the cases where a rapid response may be essential to prevent immediate harm, while in others, a more accurate understanding of the context might be needed to avoid false alarms or inappropriate actions. Whether to prioritize the slower conversation-focused model or the faster dynamically chosen subconversation-focused model could depend on what type of intervention we seek to provide. For instance, if the goal of the intervention is content moderation or shadow banning, slower models with more caution are warranted as wrongfully banning accounts, removing content, or hiding content for certain audiences could adversely impact user experience and the notion of freedom of speech (Kozyreva et al. 2023; Vaccaro, Sandvig, and Karahalios 2020). On the other hand, if the goal is to provide immediate support to potential victims, then faster models would be necessary. For instance, faster sub-conversation-focused risk detection algorithms can be used to trigger real-time nudges or warnings for potential victims of harassment messages or for potential perpetrators, before they send out offensive content (Agha, Badillo-Urquiola, and Wisniewski 2023). Even if the model makes inaccurate decisions (e.g., false positives), such proactive intervention approaches would still be more beneficial than slower models given the persistent trauma that the victims can experience (Koenen et al. 2017; Chen et al. 2022).

Considerations for Real-World Implementation

Our simulation results showcased the enhanced responsiveness of leveraging a conversation prioritization technique that focused on conversations with potentially high-risk messages over treating all conversations equally. Content moderation in online platforms is a resource-intensive task, often limited by constraints such as time and workforce availability (Jhaver et al. 2019). Thus, implementing an even imperfect system of conversation prioritization could alleviate the burden on human moderators by streamlining the process and focusing their attention on the most critical cases first. Yet, implementing this approach in the real world requires careful consideration. The incorrect prioritization of conversations by an algorithm could have several significant consequences, The consequences of such prioritization errors may vary, including delayed responses to critical high-risk scenarios, or conversely, the misdirection of resources to low-risk conversations erroneously identified as high-risk. Additionally, users whose conversations are mistakenly flagged as high-risk could experience undue stress or anxiety, which could also lead to undeserved reputational harm, especially if actions are unjustly taken based on these incorrect classifications resulting from false positives (Haimson et al. 2021; Završnik 2021).

A key aspect for a safer real-world implementation of these algorithms could be through incorporating a human-agent reinforcement learning approach (Monarch 2021; Li et al. 2019; Alsoubai et al. 2022a), where human insights and decision-making are integrated with the RL algorithm prioritization decisions, especially in cases where they believe a conversation has been misclassified. Through this approach, developers could use this feedback loop as an ongoing learning process for reinforcement learning and the conversation priority classification models from real-world interactions, which could significantly improve its accuracy over time. In addition, leveraging their feedback and communicating with them about how their feedback is being used to improve the algorithms could build trust and encourage more user engagement, as highlighted in prior research (Hamon et al. 2022). More importantly, their involvement can facilitate the building of a more dynamic prioritizing system that could adapt and evolve based on real-world feedback and changing scenarios, rather than being static and potentially becoming outdated.

At the same time, this human-in-the-loop approach can encounter the inherent challenge of balancing timely intervention with thorough review. In situations requiring immediate interventions, the integration of human feedback with the RL algorithm could introduce delays. Therefore, we suggest the need for differentiating the timing and level of human involvement based on varying levels of priority. For instance, the high-priority cases that demand immediate action, automated systems can first provide rapid initial responses, potentially followed by more nuanced human evaluation and intervention, while for the lower-priority conversations, a more in-depth human review could be more appropriate to ensure it does not impede the real-time nature of the platform. As such, the real-world implementation of the human-in-the-loop approach needs careful structuring to maintain the essential real-time nature of communication platforms. A possible way to achieve this structuring could be through developing tiered response mechanisms where the level of human involvement is scaled according to the urgency and risk level of the conversation, similar to Meta's (Facebook) three-tiered system for managing its dangerous organizations and individuals policy (Díaz and Hecht-Felella 2021).

Another critical component for ensuring a safer real-world implementation of the conversation prioritization approach could be the involvement of experts in auditing algorithms and overseeing the prioritization decisions (Sandvig et al. 2014). Auditing this prioritization approach would provide valuable insights about its impacts on users and society by identifying and rectifying biases, inaccuracies, or unintended consequences that may not be apparent to the algorithm or its developers. In addition, regular auditing by experts would ensure that the prioritization decisions are aligned with ethical standards and societal norms, thereby maintaining trust among youth and stakeholders. Thus, this audition would add a layer of accountability and transparency – essential in sensitive areas such as youth online safety and content moderation, where decisions can have significant ramifications for individuals and communities.

A key strength of our prioritization framework is that it has the potential to address a wide range of challenges in diverse risk scenarios across industries such as cybersecurity. For instance, large companies invest in powerful tools to detect malware that produce an overwhelming number of alarms, making it difficult for analysts to examine and respond to these alarms. Therefore, companies such as Target suffered the most prominent retail hack in U.S. history as the malware was detected and then escalated, but was ignored by analysts without any action (Riley et al. 2014). Customizing and training our prioritization framework could prioritize malware detection alerts based on their severity, potential impact, and relevance to the organization's security posture. This framework could assign priority levels to alerts to allow the security analysts to focus their attention on the most critical threats first to provide a more effective response to these threats.

Limitations and Future Work

While our study offers significant insights into identifying conversations with high-risk priorities, its focus exclusively on Instagram poses a challenge to its wider applicability. To better understand the broader relevance and effectiveness of our approach, future research is encouraged to investigate other popular social media platforms among youth, like TikTok. This study aimed to introduce our novel framework, utilizing the DQN agent, to assess its effectiveness in identifying stopping points for prioritizing high-risk conversations. Thus, we suggest that future researchers employ our framework to compare the performance of other reinforcement learning (RL) algorithms, such as Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC), against our DQN algorithm. In this study, we focused primarily on textual features to demonstrate the efficacy of our models, avoiding more complex features. Our future work will include contextual factors such as age, gender, and the nature of relationships, which have been shown to influence the performance of sexual content classifiers (Razi et al. 2023). Although we employed pre-trained risk detection models in this study, we plan to explore integrated multi-risk detection within this framework to assess how context identified by the RL agent affects risk detection accuracy and responsiveness in future work. Additionally, we plan to engage youth in evaluating the quality of our algorithmic classifiers by implementing a web-based risk detection system. This step will close the Human-Centered Machine Learning (HCML) loop, allowing direct user feedback on our algorithm's performance, essential for further refinement and real-world application and impact.

Conclusion

In today's digitally-driven world, youth are confronted with various online risks that affect both themselves and society. The necessity for timely and accurate risk identification is key for effective intervention by stakeholders like governments and online platforms. Our research marks a significant step in improving real-time responses, introducing a framework that emphasizes prioritizing high-risk situations. This innovative framework balances the urgency of addressing high-risk scenarios with resource constraints, offering a more focused approach to online risk management. This is particularly crucial in platforms with large volumes of interactions, where traditional automatic risk detection and moderation methods may struggle to keep pace. By prioritizing and continuously monitoring high-risk conversations, our framework could not only enhance the effectiveness of risk detection and intervention strategies but also significantly improve the overall safety of the younger generation.

Acknowledgements

This research was supported by the U.S. National Science Foundation under grants IIP-2329976, IIS-2333207, and the William T. Grant Foundation grant 187941. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Agha, Z.; Badillo-Urquiola, K.; and Wisniewski, P. J. 2023. " Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–32.

Ali, S.; Razi, A.; Kim, S.; Alsoubai, A.; Gracie, J.; De Choudhury, M.; Wisniewski, P. J.; and Stringhini, G. 2022. Understanding the digital lives of youth: Analyzing media shared within safe versus unsafe private conversations on Instagram. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.

Ali, S.; Razi, A.; Kim, S.; Alsoubai, A.; Ling, C.; De Choudhury, M.; Wisniewski, P. J.; and Stringhini, G. 2023. Getting Meta: A Multimodal Approach for Detecting Unsafe Conversations within Instagram Direct Messages of Youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–30.

Alsoubai, A.; Caddle, X. V.; Doherty, R.; Koehler, A. T.; Sanchez, E.; De Choudhury, M.; and Wisniewski, P. J. 2022a. MOSafely, Is that Sus? A Youth-Centric Online Risk Assessment Dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, 197–200.

Alsoubai, A.; Song, J.; Razi, A.; Naher, N.; De Choudhury, M.; and Wisniewski, P. J. 2022b. From'Friends with Benefits' to'Sextortion:'A Nuanced Investigation of Adolescents' Online Sexual Risk Experiences. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–32.

Alsoubai, A., Ashwaq; RAZI; AGHA, Z.; ALI, S.; STRINGHINI, G.; DE CHODHURY, M.; and WISNIEWSKI, P. J. 2024. Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety.

Ansary, N. S. 2020. Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention. *Aggression and violent behavior*, 50: 101343.

Balakrishnan, K. 2019. *Exponential distribution: theory, methods and applications*. Routledge.

Bassen, J.; Balaji, B.; Schaarschmidt, M.; Thille, C.; Painter, J.; Zimmaro, D.; Games, A.; Fast, E.; and Mitchell, J. C. 2020. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Caddle, X. V.; Naher, N.; Miller, Z. P.; Badillo-Urquiola, K.; and Wisniewski, P. J. 2023. Duty to Respond: The Challenges Social Service Providers Face When Charged with Keeping Youth Safe Online. *PACM-HCI*, 7(GROUP): 1–35.

Chancellor, S. 2023. Toward Practices for Human-Centered Machine Learning. *Comm. ACM*, 66(3): 78–85.

Chancellor, S.; Baumer, E. P.; and De Choudhury, M. 2019. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *PACM-HCI*, 3(CSCW): 1–32.

Chen, J. X.; McDonald, A.; Zou, Y.; Tseng, E.; Roundy, K. A.; Tamersoy, A.; Schaub, F.; Ristenpart, T.; and Dell, N. 2022. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, 1–20.

Crestani, F.; Losada, D. E.; and Parapar, J. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Díaz, Á.; and Hecht-Felella, L. 2021. Double standards in social media content moderation. *Brennan Center for Justice at New York University School of Law. https://www. brennancenter. org/our-work/research-reports/double-standards-socialmedia-content-moderation*.

Dodhiawala, R. T.; Sridharan, N.; Raulefs, P.; and Pickering, C. 1989. Real-Time AI Systems: A Definition and An Architecture. In *IJCAI*, 256–264. Citeseer.

Haimson, O. L.; Delmonaco, D.; Nie, P.; and Wegner, A. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Hamon, R.; Junklewitz, H.; Sanchez, I.; Malgieri, G.; and De Hert, P. 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine*, 17(1): 72–85.

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jhaver, S.; Birman, I.; Gilbert, E.; and Bruckman, A. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5): 1–35.

Kaluarachchi, T. 2021. Human-centered machine learning.

Kim, S.; Razi, A.; Stringhini, G.; Wisniewski, P. J.; and De Choudhury, M. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–34.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koenen, K.; Ratanatharathorn, A.; Ng, L.; McLaughlin, K.; Bromet, E.; Stein, D.; Karam, E. G.; Ruscio, A. M.; Benjet, C.; Scott, K.; et al. 2017. Posttraumatic stress disorder in the world mental health surveys. *Psychological medicine*, 47(13): 2260–2274.

Kozyreva, A.; Herzog, S. M.; Lewandowsky, S.; Hertwig, R.; Lorenz-Spreen, P.; Leiser, M.; and Reifler, J. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7): e2210666120.

Laborde, S. 2023. Teenage social media usage statistics in 2023.

Leiva, V.; and Freire, A. 2017. Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, 428–436. Springer.

Li, G.; Gomez, R.; Nakamura, K.; and He, B. 2019. Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems*, 49(4): 337–349.

Li, K.; Guo, B.; Liu, J.; Wang, J.; Ren, H.; Yi, F.; and Yu, Z. 2022. Dynamic probabilistic graphical model for progressive fake news detection on social media platform. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(5): 1–24.

Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; and Talwalkar, A. 2018. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185): 1–52.

Liu, Y.; Liu, Q.; Zhao, H.; Pan, Z.; and Liu, C. 2020. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2128–2135.

Liu, Y.; and Wu, Y.-F. B. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)*, 38(3): 1–33.

Lu, M.; Huang, Z.; Li, B.; Zhao, Y.; Qin, Z.; and Li, D. 2022. Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 429–442.

McHugh, C. M.; Ho, N.; Iorfino, F.; Crouse, J. J.; Nichles, A.; Zmicerevska, N.; Scott, E.; Glozier, N.; and Hickie, I. B. 2023. Predictive modelling of deliberate self-harm and suicide attempts in young people accessing primary care: a machine learning analysis of a longitudinal study. *Social psychiatry and psychiatric epidemiology*, 1–13.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Monarch, R. M. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Munoz, D.-J.; Oh, J.; Pinto, M.; Fuentes, L.; and Batory, D. 2019. Uniform random sampling product configurations of feature models that have numerical features. In *Proceedings of the 23rd International Systems and Software Product Line Conference-Volume A*, 289–301.

Nesi, J. 2020. The impact of social media on youth mental health: challenges and opportunities. *North Carolina medical journal*, 81(2): 116–121.

Obaidat, I.; Al-zou'bi, A.; Mughaid, A.; and Abualigah, L. 2023. Investigating the cyberbullying risk in digital media: protecting victims in school teenagers. *Social Network Analysis and Mining*, 13(1): 139.

O'reilly, M. 2020. Social media and adolescent mental health: the good, the bad and the ugly. *Journal of Mental Health*, 29(2): 200–206.

Parapar, J.; Martín-Rodilla, P.; Losada, D. E.; and Crestani, F. 2022. Overview of erisk 2022: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, 233–256. Springer.

Parapar, J.; Martín-Rodilla, P.; Losada, D. E.; and Crestani, F. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 294–315. Springer.

Park, J.; Gracie, J.; Alsoubai, A.; Stringhini, G.; Singh, V.; and Wisniewski, P. 2023. Towards Automated Detection of Risky Images Shared by Youth on Social Media. In *Companion Proceedings of the ACM Web Conference 2023*, 1348–1357.

Rafiq, R. I.; Hosseinmardi, H.; Han, R.; Lv, Q.; and Mishra, S. 2018. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1738–1747.

Razi, A.; AlSoubai, A.; Kim, S.; Ali, S.; Stringhini, G.; De Choudhury, M.; and Wisniewski, P. J. 2023. Sliding into My DMs: Detecting Uncomfortable or Unsafe Sexual Risk Experiences within Instagram Direct Messages Grounded in the Perspective of Youth. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–29.

Razi, A.; AlSoubai, A.; Kim, S.; Naher, N.; Ali, S.; Stringhini, G.; De Choudhury, M.; and Wisniewski, P. J. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social Media Data for the Purpose of Adolescent Online Risk Detection. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–9.

Razi, A.; Kim, S.; Alsoubai, A.; Stringhini, G.; Solorio, T.; De Choudhury, M.; and Wisniewski, P. J. 2021. A human-centered systematic literature review of the computational approaches for online sexual risk detection. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–38.

Reid, D.; and Weigle, P. 2014. Social media use among adolescents: Benefits and risks. *Adolescent Psychiatry*, 4(2): 73–80.

Riley, M.; Elgin, B.; Lawrence, D.; and Matlack, C. 2014. Target missed warnings in epic hack of Credit Card Data.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014): 4349–4357.

Sawhney, R.; Agarwal, S.; Neerkaje, A. T.; Aletras, N.; Nakov, P.; and Flek, L. 2022. Towards suicide ideation detection through online conversational context. In *Proc. SIGIR*, 1716–1727.

Sevtiyun, P. E.; Oktadini, N. R.; and Bardadi, A. 2020. Information risk assessment model of accuracy and timeliness dimensions. In *Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)*, 634–639. Atlantis Press.

Shuford, J. 2024. Deep Reinforcement Learning Unleashing the Power of AI in Decision-Making. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, 1(1).

Sun, F.-Y.; Chang, Y.-Y.; Wu, Y.-H.; and Lin, S.-D. 2018. Designing non-greedy reinforcement learning agents with diminishing reward shaping. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 297–302.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Vaccaro, K.; Sandvig, C.; and Karahalios, K. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *PACM-HCI*, 4(CSCW2): 1–22.

Vogels, E. A. 2023. Teens and social media: Key findings from Pew Research Center Surveys.

Vosoughi, S.; Mohsenvand, M. and Roy, D. 2017. Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4): 1–36.

Wang, J.; Yang, S.; An, Z.; Han, S.; Zhang, Z.; Mangharam, R.; Ma, M.; and Miao, F. 2023. Multi-agent reinforcement learning guided by signal temporal logic specifications. *arXiv preprint arXiv:2306.06808*.

Wisniewski, P.; Xu, H.; Rosson, M. B.; Perkins, D. F.; and Carroll, J. M. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3919–3930.

Yu, C.; Liu, J.; Nemati, S.; and Yin, G. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1): 1–36.

Završnik, A. 2021. Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology*, 18(5): 623–642.

Zhang, J.; Yamanaka, J.; and Li, L. 2020. Early Automatic Detection of False Information in Twitter Event Considering Occurrence Scale and Time Series. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services*, 282–289.

Zhao, T.; Ni, B.; Yu, W.; Guo, Z.; Shah, N.; and Jiang, M. 2021. Action sequence augmentation for early graph-based anomaly detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2668–2678.

Zhou, X.; Zhao, Z.; Li, R.; Zhou, Y.; Palicot, J.; and Zhang, H. 2014. Understanding the nature of social mobile instant messaging in cellular networks. *IEEE Communications Letters*, 18(3): 389–392.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, as discussed in Sections "Introduction" and "Discussion," our work advances the field of real-time risk detection by presenting a novel two-level framework with the ability to identify high-risk conversations, which represents the potential for prompt risk mitigation of high-risk online conversations.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, We have carefully reviewed the claims made in abstract and Introduction accurately reflect our contributions.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, please refer to Sections "Background" and "Methods." We justified the appropriateness of our methods by referring to prior works that used them for similar tasks/objectives. We describe the appropriateness of the usage of the Deep-Q learning model instead of other reinforcement learning models.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, Please refer to Section "Instagram Data Donation Dataset" and "Characteristics of Dataset and Participants" where we describe the characteristics of our data such as, number of conversations, messages, and risk-levels distribution. We also elaborate on our data cleaning process in Section "Data Processing."
   (e) Did you describe the limitations of your work? we discuss limitations of our work in Section "Limitations and Future Work."
   (f) Did you discuss any potential negative societal impacts of your work? Yes, please refer to the "Considerations for Real-World Implementation" subsection
   (g) Did you discuss any potential misuse of your work? Yes, please refer to the "Considerations for Real-World Implementation" subsection
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we used a dataset that followed best practices in human-centered research as stated in the original paper that described the dataset (Razi et al. 2022) and we worked with de-identified data.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we care-

fully read the ethics guidelines and ensured that the paper conforms to them.

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, we will open-source the code using our lab GitHub account upon acceptance to avoid de-anonymization. The dataset is highly sensitive (unsafe youth private conversations), so the authors of the dataset decided to give access only through collaborations.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, please refer to the two sections ”Characteristics of Dataset and Participants” and ”Data Processing.”
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because our work was not compute intensive and we did not use any external storage/compute resources. We just used an AWS server for API calls.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. Please refer to Section “Methods” where we relate our proposed evaluation procedures to the research questions. Also, refer to Section “Discussion” in which we justify that the evaluation/findings are appropriate with respect to our claims, research goals, and prior work.
   (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? Yes, please refer to the section ”Considerations for Real-World Implementation.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, we cited the paper that we used their dataset throughout the paper.
   (b) Did you mention the license of the assets? N/A
   (c) Did you include any new assets in the supplemental material or as a URL? N/A
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? No, because we utilized a pre-collected dataset and the cited paper has all the details about the consent process.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No, because we utilized the deanonymized version of the dataset, and the cited paper has all the details about this process.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? N/A
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? N/A
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? N/A
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, because we used a secondary dataset, but the original paper described that the authors of the dataset got IRB approval for the collection of their dataset.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A
   (d) Did you discuss how data is stored, shared, and deidentified? N/A

Appendices

Models’ Architecture for Conversation Prioritization Model-Level Two (RQ2)

The architecture of our models was broadly similar, starting with the input layer that takes the text sequences obtained from the ELECTRA embeddings, which then this output was pooled using GlobalPooling.

- **LSTM:** The model contained an LSTM layer that returned the full sequence of hidden states for each input. Then, the output went through multiple dense layers with the Rectified Linear Unit (ReLU) activation function, batch normalization to stabilize the training process, and dropout regularization to avoid overfitting.

- **Bi-LSTM:** To process the input text sequences from the input layer bidirectionally, we used the Bi-LSTM layer. Then, the model included a series of dense layers with the ReLU activation function followed by batch normalization and dropout regularization.
- **CNN:** The input layer was followed by convolutional blocks designed to extract features from the input sequences. These convolution blocks consisted of a convolutional layer with filters of size 3 and 100 filters, followed by batch normalization, ReLU activation function, and MaxPooling that was used to downsample the feature maps. Then, we added a dropout regularization to the output of the convolutional blocks, which were then fed to a series of dense layers that incorporated the ReLU activation function.
- **DNN:** The model was built with multiple dense layers designed to process and abstract features from the input data with the ReLU activation function. These layers were followed by batch normalization and dropout regularization. Then, we added additional dense layers with decreasing size patterns, with the size of each layer being half of the previous layer's size.

The final dense layer consisted of a single neuron with a sigmoid activation function, producing a continuous output value between 0 and 1. The models were compiled using the Nadam optimizer and binary cross-entropy loss function. Hyperparameter tuning was achieved using Hyperband, which is a search algorithm that demonstrated better and faster results than other methods (Li et al. 2018). We built the Hyperband with Keras Tuner to search for the best hyperparameter configuration for the models above, optimizing their performance on the validation dataset.

Timeliness Assessments (RQ3)

Risk	**Prec.**	**Rec.**	**F1**	**AUC**	**Acc.**
Cyberbullying	0.79±0.02	0.79±0.04	0.79±0.03	0.80±0.13	0.82±0.05
Sexual Risk	0.85±0.02	0.88±0.02	0.86±0.02		

Table 2: The accuracy performances of the risk detection classifier at the message level from prior work (baseline).

Instance Type	**vCPU**	**Memory (GiB)**	**Instance Storage (GB)**	**Network Bandwidth (Gbps)**
t3.medium	2	4	16	up to 5
m5.large	2	8	32	up to 10
m5.xlarge	4	16	64	up to 10

Table 3: The hardware configurations of the EC2 instances used for the simulation to evaluate the response times of two approaches: (1) a system that prioritized high-risk conversations and (2) a system that processed all conversations with equal risk priority.

Approach	**t3.medium**	**m5.large**	**m5.xlarge**
High-risk prioritization	7 million	1.5 millions	500 K
Equal risk prioritization	-	5 millions	2.5 millions

Table 4: Simulation results, measured in milliseconds, comparing the performance of the models across three types of EC2 instances. The findings indicate consistent improvements in response time, with the high-risk prioritization approach consistently outperforming the equal-risk prioritizing approach.