

Triaging Casual from Critical: Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media

Sarvech Qadir, Mrs Ashwaq Alsoubai, Jinkyung Park, Naima Samreen Ali,
Munmun De Choudhury, Pamela Wisniewski

Submitted to: JMIR Mental Health
on: April 15, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Triaging Casual from Critical: Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media

Sarvech Qadir¹; Mrs Ashwaq Alsoubai²; Jinkyung Park³; Naima Samreen Ali¹; Munmun De Choudhury⁴; Pamela Wisniewski⁵

¹ Vanderbilt University Nashville US

² King Abdulaziz University Jeddah SA

³ School of Computing Clemson University Clemson US

⁴ Georgia Institute of Technology Atlanta US

⁵ Socio-Technical Interaction Research Lab Orlando US

Corresponding Author:

Sarvech Qadir

Vanderbilt University
2301 Vanderbilt Place
Nashville
US

Abstract

Background: This study aims to detect self-harm and/or suicidal ideation (SH/S) language used by youth (ages 13–21) in their private Instagram conversations. While automated mental health tools have shown promise, there remains a gap in understanding how nuanced youth language around SH/S can be effectively identified. Our work focuses on developing interpretable models that go beyond binary classification to recognize the spectrum of SH/S expressions.

Objective: Our work focuses on developing interpretable models that go beyond binary classification to recognize the spectrum of SH/S expressions.

Methods: We analyzed a dataset of Instagram private conversations donated by youth. A range of traditional machine learning models (SVM, Random Forest, Naïve Bayes, XGBoost) and transformer-based architectures (BERT, DistilBERT) were trained and evaluated. In addition to raw text, we incorporated contextual, psycholinguistic (LIWC), sentiment (VADER), and lexical (TF-IDF) features to improve detection accuracy. We further explored how increasing conversational context—from message-level to sub-conversation level—affected model performance.

Results: DistilBERT demonstrated a good performance in identifying the presence of SH/S behaviors within individual messages, achieving an accuracy of 99%. However, when tasked with a more fine-grained classification—differentiating among “Self” (personal accounts of self-harm or suicide), “Other” (references to SH/S experiences involving others), and “Hyperbole” (sarcastic, humorous, or exaggerated mentions not indicative of genuine risk)—the model's accuracy declined to 89%. Notably, by expanding the input window to include a broader conversational context, the model's performance on these granular categories improved to 91%, highlighting the importance of contextual understanding when distinguishing between subtle variations in SH/S discourse.

Conclusions: Our findings underscore the importance of designing SH/S automatic detection systems sensitive to the dynamic language of youth and social media. Contextual and sentiment-aware models improve detection and provide a nuanced understanding of SH/S risk expression. This research lays the foundation for developing inclusive and ethically grounded interventions, while also calling for future work to validate these models across platforms and populations.

(JMIR Preprints 15/04/2025:76051)

DOI: <https://doi.org/10.2196/preprints.76051>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://www.jmir.org/preprint/76051>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/76051>



Original Manuscript

Triaging Casual from Critical: Leveraging Machine Learning to Detect Self-Harm and Suicide Risks for Youth on Social Media

ABSTRACT

Background:

This study aims to detect self-harm and/or suicidal ideation (SH/S) language used by youth (ages 13–21) in their private Instagram conversations. While automated mental health tools have shown promise, there remains a gap in understanding how nuanced youth language around SH/S can be effectively identified.

Objective:

Our work focuses on developing interpretable models that go beyond binary classification to recognize the spectrum of SH/S expressions.

Methods:

We analyzed a dataset of Instagram private conversations donated by youth. A range of traditional machine learning models (SVM, Random Forest, Naive Bayes, XGBoost) and transformer-based architectures (BERT, DistilBERT) were trained and evaluated. In addition to raw text, we incorporated contextual, psycholinguistic (LIWC), sentiment (VADER), and lexical (TF-IDF) features to improve detection accuracy. We further explored how increasing conversational context—from message-level to sub-conversation level—affected model performance.

Results:

DistilBERT demonstrated a good performance in identifying the presence of SH/S behaviors within individual messages, achieving an accuracy of 99%. However, when tasked with a more fine-grained classification—differentiating among “Self” (personal accounts of self-harm or suicide), “Other” (references to SH/S experiences involving others), and “Hyperbole” (sarcastic, humorous, or exaggerated mentions not indicative of genuine risk)—the model's accuracy declined to 89%. Notably, by expanding the input window to include a broader conversational context, the model's performance on these granular categories improved to 91%, highlighting the importance of contextual understanding when distinguishing between subtle variations in SH/S discourse.

Conclusions:

Our findings underscore the importance of designing SH/S automatic detection systems sensitive to the dynamic language of youth and social media. Contextual and sentiment-aware models improve detection and provide a nuanced understanding of SH/S risk expression. This research lays the foundation for developing inclusive and ethically grounded interventions, while also calling for future work to validate these models across platforms and populations.

Keywords: suicide/self-harm; machine learning; mental health; youth; natural language processing

Introduction

Youth increasingly turn to social media platforms to express their feelings—even on highly sensitive topics such as self-harm and suicide (SH/S) [1], [2], [3]. These platforms provide an

outlet where young individuals can seek support and solidarity by connecting with others who share similar struggles [4], [5]. The anonymity and expansive reach of these networks allow users to express themselves more freely than in offline contexts; however, such openness can also expose them to potentially harmful content and triggering material that may exacerbate mental health issues [6]. In response, there have been growing efforts to predict SH/S behavior on social media using automated techniques [7], [8]. The integration of machine learning (ML) and natural language processing (NLP) into mental health applications shows promising potential for early detection and intervention [9]. For instance, artificial intelligence (AI) chatbots implemented in schools to support student mental health have demonstrated how ML-powered technologies can address gaps where human counselors are unavailable [10]. However, when applied to the informal and unstructured language of social media, these systems face substantial challenges as youth often employ dynamic slang, hyperbolic expressions, and indirect language cues [11], [12], [13], which can lead automated models to misinterpret benign or humorous exaggerations as genuine distress signals [14]. The resulting risk of overreaction or underreaction underscores the critical need for context-sensitive approaches that balance timely intervention with the preservation of safe expressive spaces [15], [16], [17], [18].

Prior Work

Social media platforms are increasingly recognized as pivotal in identifying individuals exhibiting signs of self-harm and suicide (SH/S) [19]. Significant strides have been made by ML and social computing scholars in addressing mental health issues like SH/S through automated approaches [20]. Common practices in SH/S detection on social media include advanced ML and NLP techniques—such as BERT-based transformer architectures—which capture semantic meanings even when explicit keywords are absent, enabling the identification of at-risk individuals by detecting subtle linguistic markers associated with mental distress [8], [20], [21]. Despite these advances, such approaches have not been fully adapted to the specific language patterns used by youth on social media [22]. Youth share their experiences with self-harm/suicide (SH/S) in diverse ways, such as through hyperbolic expressions (without intent), discussions of others' SH/S experiences, and explicit self-disclosures. A qualitative study by [23] revealed that youth disclosures range from hyperbolic expressions (without intent) to discussions of others' SH/S experiences and explicit self-disclosures. Moreover, prior research has identified distinctions within SH/S, such as differentiating between suicidal ideation, the likelihood of a suicide attempt, and shifts toward suicidal ideation [7]. Most existing studies have employed binary classification methods for SH/S detection, failing to capture the multidimensionality of youth disclosures [24]. Given that online risks exist on a spectrum and can intensify over time, models need to differentiate among varying levels of risk rather than simply categorizing interactions as risky or not risky [25]. While substantial research has provided insights into SH/S prediction on public forums [26], [27], there is limited understanding of how such risky discussions unfold in youth private discourses [22]. Our work addresses this multi-dimensional nature of youth SH/S disclosures in private conversation by performing multi-class classification. Moreover, previous work identified a key limitation of existing SH/S detection systems is that, while they effectively capture textual content using features such as n-grams, bag-of-words, and deep learning based word embeddings [28], they often overlook critical contextual factors and user characteristics, such as age and mental health history. This gap can lead to misclassifications

that either trigger inappropriate interventions or miss genuine signals of distress [29], [30]. To address these challenges, recent work has emphasized a Human-Centered Machine Learning (HCML) approach, which integrates insights from individuals with lived experiences and those providing support to ensure that models are both accurate and interpretable [31]. Incorporating first-person annotations has proven particularly beneficial, as models trained on insider labels tend to outperform those based on third-party annotations [32], [33],[27]. Moreover, transparent and explainable AI methods—such as attention-based models and SHAP (Shapley Additive explanations) [34], [35]—are being leveraged to highlight key features driving model decisions, thereby enhancing trust and accountability. Finally, integrating both linguistic and psychological signals, including sentiment shifts, interaction patterns, and user history [36], is crucial for accurately distinguishing between different forms of SH/S language, such as humorous references, self-disclosures, and discussions of others. By leveraging HCML principles, we aim to develop a system that not only detects risk with high accuracy but also respects the agency and lived experiences of individuals expressing distress online.

Goal of This Study and Research Questions

This study aims to develop a context-sensitive framework for detecting and classifying diverse types of SH/S language in youth private communications. Utilizing data from the Instagram Data Donation (IGDD) project [37], where youth (aged 13–21) self-labeled their private direct messages as safe or unsafe, we build upon prior annotations of 2,019 sub-conversations from 151 participants [22]. The annotated data are divided into three categories: 1) **Self-Disclosures of SH/S**: Explicit personal disclosures of self-harm or suicidal ideation, 2) **SH/S Experiences of Others**: Discussions involving self-harm or suicide incidents of others, and 3) **Hyperbolic Representations of SH/S**: Metaphorical or humorous expressions of SH/S-related language.

The study is driven by the following research questions:

- **RQ1:** *What is the best approach to distinguish between different types of self-harm and suicide language (humorous, disclosures about others, self-disclosures) at the message and sub-conversation levels?*
- **RQ2:** *What are the contextual and linguistic characteristics that help distinguish between the three types of SH/S discourse?*

To address these questions, we evaluated several ML models—including transformer-based architectures (BERT, DistilBERT) and classical classifiers (SVM, Random Forest, Naive Bayes, and XGBoost)—and enriched them with features such as contextual, psycholinguistic (LIWC), sentiment (VADER), and lexical (TF-IDF) indicators. DistilBERT achieved the highest accuracy (99%) in distinguishing SH/S messages. Expanding the context from single messages to sub-conversations improved model accuracy to 91%. For RQ2, results highlighted that males tended to use hyperbolic SH/S language, females discussed others' experiences, and non-binary individuals more often shared personal SH/S experiences.

Methods

This section outlines the dataset, preprocessing steps, feature engineering, classification models,

and evaluation strategies employed to address our research questions.

Dataset

We utilized a subset of the Instagram Data Donation (IGDD) project dataset by Razi et al. [36], specifically scoped and annotated by Ali et al. [22]. In Razi et al. [36], researchers collected ecologically valid social media data through the IGDD project to study adolescent online safety. They recruited 195 English-speaking adolescents (ages 13–21) in the U.S. who had active Instagram accounts during their teenage years and had experienced at least two direct message conversations that made them or someone else feel unsafe or uncomfortable. The dataset includes 32,055 private conversations, providing valuable insights into youth social media interactions and supporting machine learning models for online risk detection. Ali et al. [22] manually annotated the original IGDD dataset for self-harm and suicide-related (SH/S) conversations, which resulted in 1,224 self-harm and suicide-related (SH/S) conversations from 151 youth. Since individual conversations could cover multiple topics, they further segmented them into distinct, topic-specific segments, referred to as sub-conversations. This process resulted in **2,019** sub-conversations and **35,963** messages from the original **1,224** self-harm and suicide-related conversations.

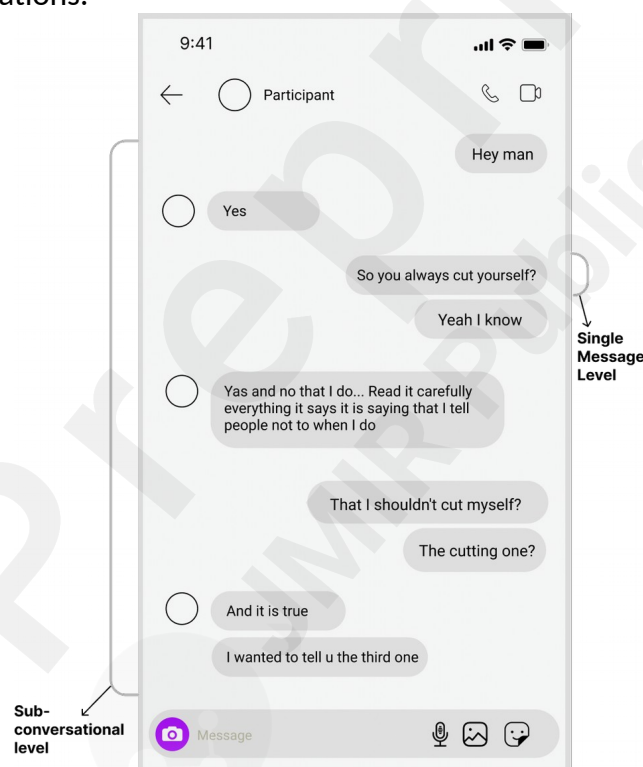


Figure 1: Message level and sub-conversational level

Annotation Categories

Each sub-conversation in the SH/S dataset ($n = 2,019$) was annotated to capture context-specific interactions and categorized into three primary classes:

- **Self-Disclosures of SH/S:** Instances where users explicitly disclose personal self-harm or suicidal ideation (e.g., “I’ve been cutting again. I don’t know how to stop.”). This category comprises **1,262** sub-conversations (62.5%) and a total of **10,407** messages (28.9%).

- **SH/S Experiences of Others:** Conversations discussing self-harm or suicide incidents involving others (e.g., “Did you know she cut themselves?”). This category includes **259** sub-conversations (12.8%) and a total of **7,507** messages (20.9%).
- **Hyperbolic Representations of SH/S:** Instances where SH/S-related language is used humorously without serious intent (e.g., “This homework is killing me!”). This category accounts for **498** sub-conversations (50.2%) and a total of **18,049** messages (35.7%).

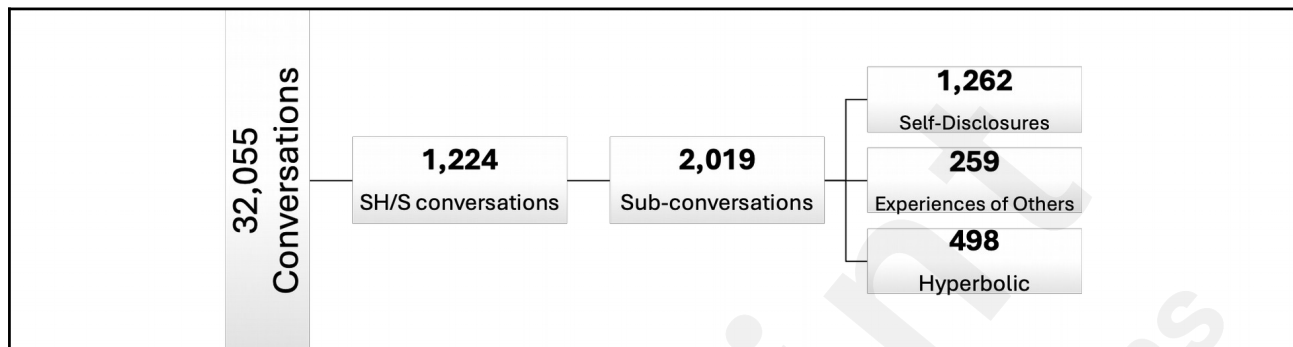


Figure 2: Overview of dataset distribution

Data Pre-processing

We applied a multi-step pre-processing pipeline to clean the dataset while preserving essential linguistic cues:

- **Noise Removal:** Punctuation, hyperlinks, stop words, non-Latin words, and isolated numeric/single characters were removed. Emojis were converted to text using the Python demoji library to retain emotional cues.
- **Sensitive Lexicon Development:** A lexicon of SH/S and violence-related keywords was constructed based on existing literature [38], guiding the filtering of messages containing SH/S disclosures.
- **Text Embedding and Filtering:** Sub-conversations were embedded using Sentence-BERT (SBERT), and cosine similarity was computed against the sensitive lexicon. SBERT's proven effectiveness in sentence-level retrieval and paraphrase detection [39] allowed us to identify indirect SH/S references.
- **Data Augmentation:** BertAug [40] was applied to increase the size and diversity of the training data by generating semantically coherent variations, thereby enhancing model robustness [41].

The table below represents the final number of sub-conversations for each category of SH/S:

Category	Original Instances	Post-Augmentation
Self-Disclosures	498	996
SH/S Experience of Others	259	518
Hyperbolic Representations	1262	2,592
Total	2,019	4,106

Table 1: The number of sub-conversation instances before and after augmentation.

The final dataset contained **4,106** sub-conversations, offering a more balanced distribution across the three categories. To assess the impact of conversational context, we evaluated our best-performing model on the SH/S vs. Non-SH/S dataset at both the **message-level** (isolated messages) and the **sub-conversational** level (15-20 messages of holistic units). Prior research indicates that SH/S language is context dependent [42].

Automatic Self-harm and Suicide Detection Approaches

To address RQ1, we implemented a diverse set of models:

- **Transformer-Based Models:** Fine-tuned pre-trained BERT [43] and DistilBERT [44] models, leveraging their deep contextual understanding [45] and ability to capture subtle SH/S cues [46].
- **Sequential Models:** LSTM [47], CNN-BiLSTM, and GRU [48] architectures were employed to model long-term dependencies and temporal patterns in sub-conversations.
- **Classical ML Models:** SVM, Random Forest, Naive Bayes, and XGBoost [49] were trained on engineered features to provide interpretable baselines [41].

Analysis

Feature Engineering

For RQ2, we employed a traditional non-end-to-end model (XGBoost, selected for its binary classification performance [49]) using features extracted from the annotated sub-conversations:

- **Contextual Features:** Demographic data i.e. gender from IGDD survey responses) to explore influences on SH/S discourse.
- **Psycholinguistic Attributes:** LIWC scores [50] covering affective processes (e.g., sadness, anger), cognitive mechanisms (e.g., causation, certainty), and social/personal concerns, normalized by subconversation length.
- **Sentiment Analysis:** Polarity scores computed using VADER [51] to capture the emotional tone.
- **Lexical Features:** TF-IDF [52] scores for the top 1,000 terms to measure term importance.
- **Additional Linguistic Scores:** Toxicity, politeness, humor, empathy, and hate speech scores extracted using pre-trained models from Hugging Face [53].

Model Evaluation

Models were evaluated using standard metrics: **accuracy**, **precision**, **recall**, **F1-score**, and **AUC-ROC**. To mitigate overfitting, we employed stratified k-fold cross-validation with k=5, which maintains the original class distribution across each fold. This method provided reliable performance estimates, especially in imbalanced datasets [54]. For deep learning models, validation sets were further used to apply early stopping based on validation loss to prevent

overfitting [55]. Each model underwent hyperparameter optimization to maximize performance. For classical ML models (SVM, Random Forest, XGBoost), hyperparameters (e.g., SVM kernel type, XGBoost learning rate, max depth, number of estimators) were optimized using grid search with cross-validation which ensures that the hyperparameter evaluation is reliable and generalizable, while grid search exhaustively identifies the best combination of hyperparameters [56]. For deep Learning models (LSTM, CNN-BiLSTM, GRU), the Hyperparameters such as hidden layer sizes, dropout rates, learning rates, and batch sizes were optimized using a random search combined with early stopping. For transformer models (BERT, DistilBERT), fine-tuning was performed with learning rate schedulers (e.g., linear decay), with hyperparameters like learning rate, batch size, and warm-up steps optimized based on validation loss.

Unpacking Results Qualitatively

To gain deeper insights beyond quantitative metrics, we conducted a qualitative reading [27] focused on unpacking model misclassifications, feature importance, and demographic influences on SH/S discourse.

1. We examined instances where models failed at the **message level but succeeded at the **sub-conversational level** in RQ1.**

2. Unpacking Psycholinguistic Feature Contributions and additional features: We conducted a detailed analysis of the **Linguistic Inquiry and Word Count (LIWC)** features [50], evaluating the influence of its **64 categories** on model predictions. This involved both quantitative and qualitative steps, such as utilizing SHAP values, we identified which LIWC categories most significantly influenced the model's classification decisions. To validate these findings, we manually reviewed sub-conversations with high activations in key LIWC categories. Moreover, to understand the significance of each of the additional linguistic features i.e, toxicity [57], politeness, humor, empathy, and hate speech [58], we utilized ANOVA tests [59] to help understand if there was any significance in prediction.

3. Demographic Analysis: Gender and Discourse Types: To provide additional context on the dataset, we include demographic statistics from the IGDD survey responses, which describe the characteristics of the **151 youth participants** whose conversations were analyzed in this study. Most participants identified as **female (71%)**, followed by **male (20%)**, and **non-binary or preferred not to self-identify (9%)**. For RQ2, we also explored potential demographic influences on SH/S discourse, we conducted a χ^2 [60] to examine the relationship between **gender** and different **types of discourse** (self-disclosure, SH/S experiences of others, hyperbolic representations). The test revealed statistically significant differences ($p < 0.05$) in discourse types across gender groups.

Results

In this section, we first present the results of the classifiers that predict SH/S risks at the message level (i.e. binary classification). The classifiers were further evaluated with the best performing end-to-end model at the message and sub-conversational level (RQ1). Next, we added different features to the best non-end-to-end model (RQ2) and the results were presented highlighting the unpacking of contextual analysis of our model.

Automated Classification of Suicide/Self-harm in Private Conversation of Youth (RQ1)

Binary classification (message level)

For the classification task to identify SH/S vs non-SH/S messages (binary classification), the DistilBERT (end-to-end learner) (accuracy = 0.99, precision = 0.99, F1 score = 0.99, precision = 0.94, recall = 0.98, and an AUC of 0.99) model outperformed the other models used across all assessed metrics (Table 2).

Models	Accuracy	Precision	Recall	F1 Score	AUC	Type
Random Forest	0.85	0.88	0.73	0.80	0.83	Non-E2E
SVM	0.89	0.88	0.84	0.86	0.88	Non-E2E
Naïve Bayes	0.80	0.87	0.57	0.69	0.91	Non-E2E
XGBoost	0.94	0.91	0.94	0.93	0.98	Non-E2E
CNN	0.88	0.88	0.83	0.85	0.95	E2E
CNN-BiLSTM	0.89	0.85	0.88	0.87	0.90	E2E
LSTM	0.85	0.83	0.79	0.81	0.90	E2E
GRU	0.84	0.82	0.78	0.80	0.92	E2E
BERT	0.97	0.92	0.98	0.95	0.99	E2E
DistilBERT	0.99	0.99	0.98	0.99	0.99	E2E

Table 2: Performance metrics of various models in binary classification with non-SH/S data, categorized as End-to-End Learners (E2E) or Non-End-to-End Learners (Non-E2E).

This high level of accuracy was maintained even when keywords associated with S/SH were removed from the dataset. The results from the keyword removal process using S-BERT, which was a methodological approach also indicated a notable performance with the DistilBERT model achieving high metrics: accuracy at 0.88, F1 score at 0.87, precision at 0.80, recall at 0.95, and an AUC of 0.94.

Three-class classification (message and sub-conversational level)

Next, we compared the DistilBERT model's performance on a single message and a sub-conversational level. As seen from Table 3, the inclusion of context at the sub-conversational level resulted in a marked improvement in the model's performance, with all the accuracy metrics showing an increase.

Level	Precision	Recall	F1 Score	Accuracy
Message				
Overall	0.89	0.89	0.89	0.89
Hyperbole	0.90	0.97	0.94	
Other	0.86	0.78	0.82	
Self	0.86	0.73	0.79	
Sub-conv				
	0.91	0.91	0.91	0.91

Overall			
Hyperbole	0.94	0.94	0.94
Other	0.90	0.95	0.92
Self	0.83	0.79	0.81

Table 3: Performance metrics at the Message and Sub-conversation levels

Our qualitative analysis of classification results confirmed that end-to-end models like DistilBERT, which learn to predict directly from raw text to classification labels, benefited from larger context windows. Specifically, we observed an overall improvement across all performance metrics when moving from **message-level** to **sub-conversation-level** classification. The overall **accuracy** increased from **0.89** to **0.91**, reflecting more accurate predictions with broader conversational context. This trend was even more pronounced within individual classes. For the **"Other"** class, F1-score improved from **0.82** to **0.92**, and for the **"Self"** class, from **0.79** to **0.81**. The model was unable to determine the true intent behind potentially alarming situations when analyzing single messages in isolation. Specifically, the model initially categorized self-disclosure messages indicating imminent risk as hyperbole, likely due to the common use of exaggerated language in casual conversations, particularly among youth. However, responses from conversation partners often provided critical interpretive signals that helped the model correct its classification when additional context was introduced.

For example, in the following sub-conversation, the model initially classified the message "And want to kill myself" as hyperbole risk. Yet, when the model processed the entire sub-conversation, it identified contextual cues such as "No don't kill yourself" indicating that the situation was serious rather than exaggerated and highlighting an ongoing struggle and the urgency of the distress being expressed.

P: And want to kill myself

O: I can't believe you have to deal with this everyday

O: No don't kill yourself, It's not a good

idea P: I'm killing myself

For two-way conversations, 'P' refers to the primary participant, whose conversation is being analyzed. 'O' denotes the other individual involved in the exchange. For group conversations, we appended numbers 'O1', 'O2', 'O3', etc., to denote the other individuals participating in the group conversation. Conversely, the opposite scenario occurred when the model misclassified messages as immediate "Self" risk. Upon further examination of the context, it became clear that youth was either exaggerating their emotions or discussing the struggles of someone else. This was likely because the messages appeared to directly address an individual who may be struggling with suicidal thoughts or because they contained words like "cut" and "I", which are often associated with personal distress and self-harm. The pronoun "I" typically signals self-referential language, which, when combined with distress-related terms, may lead the model to interpret the message as expressing suicidal intent. However, additional context—such as the presence of laughter such as "HAHAHAHAHA" provided crucial disambiguation, shifting the interpretation toward humor or hyperbole. In the following sub-conversation, the initial classification of Self risk was corrected when the broader exchange revealed that the discussion was playful rather than

an actual self-harm disclosure:

O: Can you tell me The frick, Did you just cut, HAHAAHAHAHA

P: It's my middle

O: I DON'T CARE, WHYD YOU CUT, NOO, Are you kidding me. COME ON, WIPE IT UP, And I'll tell you, And take a picture of it cleaned

O:Put a bandaid on your cut

P: I didn't cut myself, It was food I KNEW IT. I knew I used a different finger to draw red

In addition, when additional context from the broader sub-conversation was added, it became evident that the participants were not discussing their own distress but rather engaging in third-person discussions about distressing events. For example, in the following sub-conversation, the message was initially classified as "Self," but with added context, it became clear that the participants were reacting to the news of someone else's passing.

P: How are you feeling

O: Still in shock but, life goes on. Gotta keep moving forward

P: And you should know, suicide is not the answer, think about the good things in life

O: Yeah, you have to try and find peace

P: Yeah I couldn't believe it. This is a sad news

O: May he rest in peace. I knew this guy

The improved performance at the sub-conversational level demonstrates the benefit of incorporating context providing a richer semantic information than isolated messages.

Contextual and Psycholinguistic Characteristics of Youth Suicide/Self-harm Conversation (RQ2)

Psycholinguistic Characteristics of Youth Suicide/Self-harm Conversation

We further unpacked the linguistic features and presented the results from a qualitative analysis evaluating how the psycholinguistic characteristics were associated with classification results for each category (self, other, hyperbole). We completed this analysis at the sub-conversation level rather than the message due to more context as confirmed by RQ1 results. Our results show that psycholinguistic words in each subconversation helped predict each category. For example, the words associated with tentativeness ('tentat') showed high importance for prediction in the "self" category. This suggested that individuals speaking from personal experience expressed uncertainty or hesitation, which could reflect the complexity of their feelings or situations. This was seen in the use of words like 'maybe,' 'or' 'anyone,' 'if,' and 'anything.' For example, in the following conversation, tentative words helped with the prediction of a participant reminiscing about something from their past:

P: i was all like "oh man what if this all gets so bad that i like self harm or something i'm going to cry" because i wanted attention and then fifth grade came and i was just fine for no reason

O: haha so i remember i was like "if i don't get this many likes in this many days i'm leaving"

P: i remember posting this thing about how i was all “depressed”

Moreover, words associated with space such as 'spatial' and 'time' language were prominent for the prediction of the “hyperbole” category, which implied a discussion about suicide/self-harm that involves words related to physical spaces or distances. This may be relevant in conversations where individuals were talking about other people or situations removed from themselves. For example, in the following subconversation, spatial terminology played a crucial role in amplifying the description of a strenuous physical routine. The words underlining physical locations and movements, such as “in the heat,” “up and down hills,” “on our track,” “on the hot hard ground,” and “on the track,” constructed a scenario rife with exertion. The hyperbolic essence is further enhanced by the repetitive mention of these terms, intensifying the narrative. The speaker employed this spatial language to dramatically overstate and exaggerate, typical of hyperbolic discourses. For instance, phrases like “weird ass suicides on the track” and the exaggerated claim of being ready to “pass out” further illustrated the intensity of the workout for dramatic effect.

*O1: nahhh we be running with full pads in the heat up and **down** hills and the sun shine directly **on** our track*

*O2: **in** the afternoon we run a **mile** do drills **where** we gotta lay **down on** the hot hard ground*

*O3: we do bleachers we do laps if our group came in last we do mad push ups crutches etc.. we do weird ass suicides **on** the track.*

In addition, language focused on family-related words ('home') and social processes ('social') aided in the prediction of “other” category which implied that discussions about others' situations in a home or family context or interactions when discussing others' experiences with S/SH. For instance, the following conversation talks about how social and family keywords i.e “roommate”, “person”, “person1”, “names” helped in prediction to models when talked about the experiences of others.

*P: The one that never showers has a new kid **roommate** and everyone in the suite beside her is new. Besides some that had accidents the rest r ppl that no one wanted to room with.*

*O: Most returners with new kid **roommates**? Watch out.*

P: Yeahhhhhh. And it makes [person] hecka sad and he's been staying in my room with me cause it's too much for him.

O: Cause [person] has him and he's dying cause [person1] won't talk about anything except suicide and gays.

P: At least his room is clean and he showers better than some annoying ppl.

As such, the psycholinguistics features aided the model in prediction based on the higher frequency of specific features in individual categories.

Identifying Key Differentiators in SH/S Conversations

To evaluate the role of additional linguistic features in distinguishing between categories (Hyperbole/Humor, Other, and Self), a one-way ANOVA tests was conducted. Table 4 presents the results of the ANOVA analysis.

Feature	F-value	p-value
EmpathyScore	15.32	< 0.001
Toxicity	6.87	< 0.01
PolitenessScore	3.92	< 0.05
HumorScore	8.45	< 0.01
HateSpeechScor	1.65	0.12 (n.s.)
e		

Table 4: ANOVA results for distinguishing conversational categories. Statistically significant features ($p < 0.05$) indicate meaningful variations across categories.

Among the features analyzed, **EmpathyScore** ($p < 0.001$) was the strongest differentiator, particularly in *Self* messages. This suggests that self-harm disclosures often contain highly empathetic language, either from the individual expressing distress or from responders offering support.

P: It's not very good please don't get mad. It's how I calm myself. It helpss.

O: Oh, I am not mad. What matters to me right now is that you're feeling better...Though, is it okay if I request that we attempt to find a safer method in the future? There are endless ways, we can find one!

Here, the *Self* message contains **hesitation** and a **plea for non-judgment**, while the response demonstrates high empathy by validating the user's feelings before gently encouraging an alternative. These findings reinforce the importance of empathy detection in SH/S identification—messages with high “EmpathyScore” often signal distress and should be carefully analyzed rather than dismissed as neutral or unimportant. Additionally, **Toxicity** ($p < 0.01$) was more prevalent in *Hyperbole/Humor* messages, indicating that sarcasm and exaggeration often introduce aggressive or negative wording. However, this poses a **challenge for SH/S detection**, as some genuine disclosures may also be flagged as toxic despite containing cries for help.

P: I sent this to her and said you included [person] before meeee Fucking kill me, LOL legit and we will always be novices in their eyes even when we are not.

Here, the phrase “Fucking kill me” demonstrates self-harm references with violent references increasing the overall toxicity of the sub-conversation. Interestingly, **Toxicity and EmpathyScore were positively correlated** ($r = 0.41$), highlighting a key challenge identified in content moderation systems; messages flagged as toxic **also contain empathy**. This indicates that toxicity does not always equate to harmful intent—rather, some individuals express distress in a way that appears toxic but is a plea for help.

P: he would call me a fat whore and a bitch, and would tell me to kill myself, and i stayed with him for six months and cried everyday.

In this message, the high toxicity score is due to the offensive language describing verbal abuse, yet the high empathy score reflects the emotional weight of the experience being shared. If an AI system were to automatically filter high-toxicity messages as humor, it might wrongly remove important self-harm disclosures, preventing individuals from receiving support. Therefore, there is a need for context-aware explainable moderation rather than strict toxicity-based filtering.

Contextual Characteristics of Youth Suicide/Self-harm Conversation

Next, we unpacked the contextual factors and presented the results from the relationship between 'gender' feature and discourse categories ('hyperbole', 'other', and 'self', as captured in Figure 5, highlighting noteworthy patterns in the communication styles among different genders using a χ^2 test.

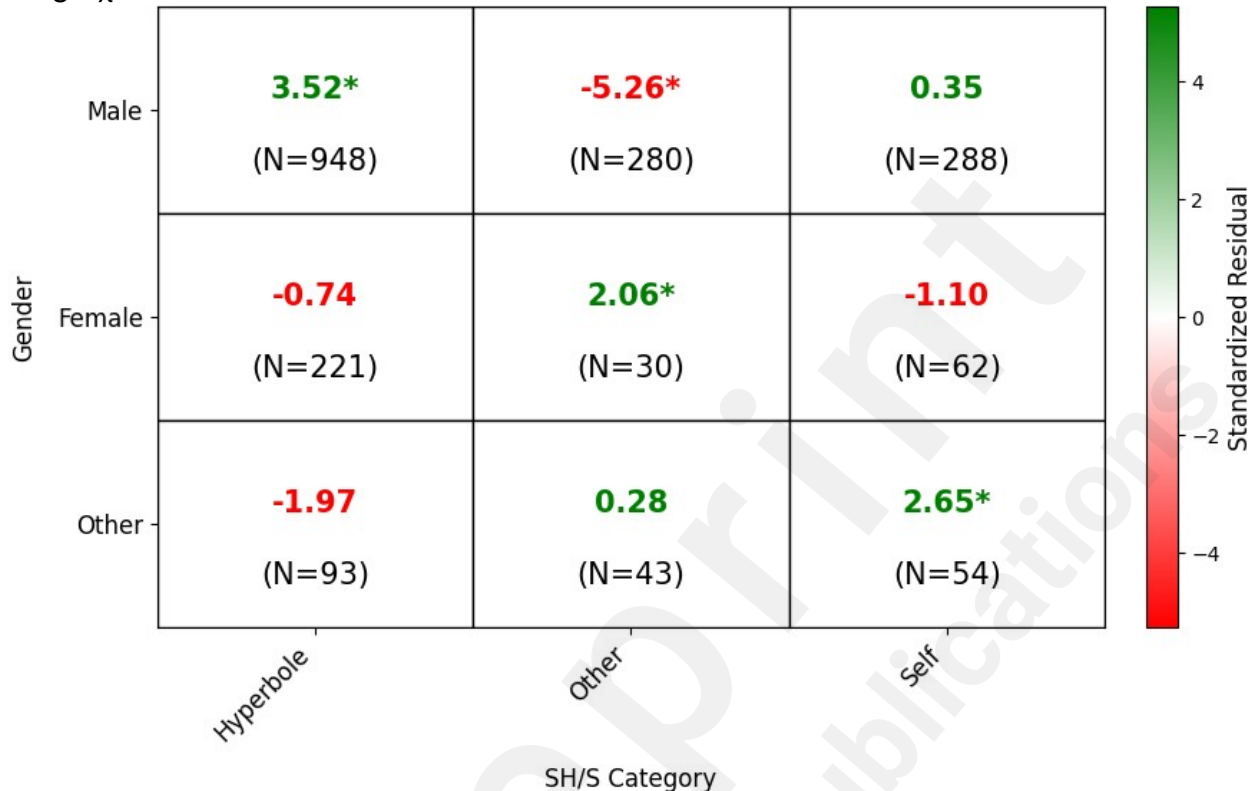


Figure 5. Residuals: Gender vs Category. The magnitude of the standardized residual (e.g., above 2 or below -2) indicates the strength of this deviation denoted with (*) with the value.

To examine gender-specific patterns in SH/S discourse, we conducted a chi-squared test of independence between gender identity (Female, Male, Other) and SH/S discourse categories (Hyperbole, Other, Self). The analysis yielded a statistically significant association, $\chi^2 (df = 4, N = 2,019) = 57.16, p < 0.001$, indicating that the distribution of SH/S expression types significantly varied by gender. Standardized residuals revealed notable patterns in language use across gender groups (Figure X). **Male participants** were substantially more likely than expected to use **hyperbolic SH/S language** (residual = 3.52*), reflecting a frequent use of exaggerated or non-literal references to suicide or self-harm. At the same time, they were significantly **less likely** to discuss the **SH/S experiences of others** (residual = -5.26*), suggesting that third-person narratives were underrepresented in male-authored conversations. **Participants identifying as non-binary** exhibited a distinct linguistic profile. They were **less likely** to use **hyperbole** (residual = -1.97) and **more likely** to disclose **personal SH/S experiences** (residual = 2.65*), reflecting a stronger inclination toward emotionally vulnerable and direct expression of mental health challenges. **Female participants**, on the other hand, were more likely to reference **others' SH/S experiences** (residual = 2.06*), indicating a tendency to discuss SH/S in the context of third parties, such as in gossiping situations.

Discussion

Principal Findings

Our study highlights the effectiveness of using transformer-based models, particularly DistilBERT, for detecting self-harm and suicidal ideation (SH/S) in youth's private Instagram conversations. DistilBERT achieved 99% accuracy in binary SH/S classification and 91% accuracy in multi-class categorization when extended to sub-conversational context. Incorporating features such as sentiment, psycholinguistic cues, and contextual windows significantly improved model performance. We also uncovered meaningful gender-based patterns in SH/S language: males were more likely to use hyperbolic expressions, females tended to discuss others' experiences, and non-binary individuals predominantly shared personal disclosures.

The Importance of Context in Detecting SH/S

The classification findings (RQ1) highlight the importance of employing our context-sensitive approach, which addressed the limitations of traditional binary classification and the reliance on single messages that constrained the model's ability to capture the nuanced meanings and intentions of SH/S discourses. This approach enabled us to develop an automatic detection model that distinguished between casual expressions and critical self-harm/suicide (SH/S) situations, achieving an accuracy of 91%. Although this level of accuracy is noteworthy, we advocate for a paradigm shift in future research. Rather than focusing solely on improving model accuracy or optimizing algorithms [61], greater emphasis should be placed on identifying and understanding the nuanced variations within self-harm and suicide-related content, as well as determining the optimal conversational context necessary to inform and enhance the detection capabilities. By incorporating these dimensions into detection models, researchers could differentiate low-risk expressions from those that signal imminent danger and reduce both false negatives—where serious risk may be overlooked—and false positives—where benign expressions are mistakenly flagged as we showed in our error analysis. Recognizing these distinctions supports more targeted and ethically responsible interventions, ensuring that individuals in genuine need receive appropriate care, while minimizing unnecessary responses. Future models could advance this framework by adopting more refined annotation schemes, as prior work has identified various proxies indicative of mental health challenges online. For example, online harassment has been shown to precede declines in mental health [62], and groups of teens who engaged in self-harm offline were found to participate in high-risk sexual conversations with strangers online [63]. These findings, along with the granular approach used in this study, underscore the complexity of identifying SH/S (self-harm/suicidal) cases in online conversations—far beyond what a simple binary classification can capture. In addition, we recommend future research to leverage advanced techniques, such as reinforcement learning (RL) [59], to dynamically identify and integrate the most informative conversational context. Such efforts will advance the field toward a more nuanced, context-aware, and human-centered approach to SH/S detection across both clinical and online environments.

Contextual and Linguistic Implications for Automated Self-harm and Suicide Detection

The prominence of tentative words (e.g., “maybe,” “if,” “anyone,” “anything”) in self-harm and suicidal ideation (SH/S) disclosures suggests that individuals may exhibit uncertainty or hesitation when expressing their personal experiences [64]. This observation aligns with prior psychological research, which indicates that ambivalence is common among individuals contemplating self-harm or suicide—reflecting an internal struggle or emotional distress [65]. Consequently, AI

models for SH/S detection must be designed to contextualize such uncertainty rather than dismiss it outright. Flagging potential risk cases based solely on tentative language may lead to overlooking individuals in distress; therefore, future work should consider incorporating longitudinal analyses to determine whether repeated use of tentative expressions correlates with escalating distress over time [66]. Moreover, the model's reliance on spatial references (e.g., "on the track," "up and down hills") and temporal references (e.g., "in the heat") for classifying hyperbolic language highlights the role of vivid, metaphorical descriptions in exaggerated narratives. In social media conversations, hyperbole is frequently used as a form of dark humor, exaggeration, or emphasis rather than as an indication of genuine self-harm intent [22]. Prior studies have demonstrated that language not only conveys information but also shapes cognitive processes; for example, spatial and temporal metaphors influence how individuals conceptualize and articulate their experiences [67]. Given the prevalence of such hyperbolic expressions among youth, detection systems must adapt to evolving linguistic trends. The integration of sentiment analysis and contextual embeddings is recommended to enhance classification accuracy by distinguishing between distress driven hyperbole and casual or humorous exaggeration. The frequent mention of social and familial words (e.g., "roommate," "person," "home") suggests that discussions regarding others' SH/S experiences are often framed within interpersonal and domestic contexts. This linguistic pattern reflects how individuals process and externalize their concerns by embedding them in familiar social environments [68]. Research indicates that discussing traumatic events, such as another's self-harm or suicide, can serve as a coping mechanism, helping individuals process their emotions through interpersonal narratives [69]. For automated detection systems, it is critical to differentiate between self-reports and third-party observations. While discussions about others may not signal personal risk, they can indicate a user's concern for someone at risk [70]. Future research should explore networked conversational analysis to identify patterns in how information about SH/S is shared, potentially guiding the design of interventions that support indirect reporting in real-time crisis situations.

Gender-Based Patterns in SH/S Detection

Our analysis reveals distinct gender-based patterns in SH/S communications. Specifically, males in our dataset more frequently employed hyperbolic or exaggerated language, females tended to discuss self-harm and suicide in the context of others' experiences, and non-binary youth were more likely to share personal self-disclosures. One potential explanation is that these differences mirror varying psychological needs and coping mechanisms where males are using violent language [71]. For example, males may use humor or hyperbole as a form of emotional regulation or deflection [72], while females may rely more on social or communal support, thus centering their discussions on peers, friends, or family members [73]. Meanwhile, non-binary youth could find direct self-disclosure to be a more authentic way to articulate distress, potentially reflecting lived experiences tied to identity-related stressors [74]. Individual communication styles are shaped by intersecting factors such as cultural background, social context, and personal history, which means no single gender group is homogeneous in its expression of SH/S [75]. Additionally, differences in disclosure may stem from how comfortable individuals feel discussing mental health in private online spaces, rather than from any inherent gender-based communication pattern [76]. Future research should therefore investigate not only what different gender groups share online but also the broader context—how, when, and why they choose to share. An intersectional approach that considers overlapping identities (e.g., race, ethnicity, sexual orientation) may offer deeper insights and help prevent models from

misclassifying or overlooking at-risk behaviors. Moreover, detection systems need to account for these diverse linguistic and psychological signals without reinforcing biases. By doing so, we can develop more inclusive, accurate, and ethically grounded tools for identifying and responding to SH/S discourse across varying demographic groups.

Limitations and Future Work

A primary limitation of our study is the reliance on human annotations from [22], where the instances analyzed were not flagged by the target participants themselves but were instead annotated by researchers. Additionally, the challenges of obtaining datasets in our domain may limit the generalizability of our findings to other platforms. Another limitation is that, due to the sensitive nature of the data, the original researchers have only shared it on a limited basis with trusted collaborators. This restricted access means that the dataset cannot be made publicly available, thus limiting opportunities for independent validation and replication of our findings. Another limitation identified in our error analysis was that existing models like DistilBERT do not explicitly account for slang and may struggle with the informal language used by youth. Additionally, the presence of automated Instagram messages in the dataset may have impacted the model's accuracy. Lastly, while our model performed well in simulated environments, we did not explore the responsiveness and effectiveness of different RL algorithms or system configurations. This might affect the robustness of our conclusions, especially in real-world applications where timeliness is critical.

Future studies should apply our approach to other youth datasets where the data is flagged either by the participants themselves or by clinicians. This would help ensure that the models are responsive to the subjective experiences of individuals at risk. Secondly, incorporating insights from clinical experts in reviewing flagged data could provide a more comprehensive understanding of risk levels and improve the classifiers' accuracy. Further work should focus on enhancing ML models by integrating slang and other colloquial expressions to improve their accuracy in detecting SH/S language. Also, it is important to consider cleaning the data to exclude automated content, thereby reducing noise and improving the precision of SH/S detection. We also recommend investigating the performance of alternative RL algorithms to benchmark against our current PPO-based approach. Finally, researchers should work on assessing the timeliness of detection, ensuring that the models classify risks accurately in a time-sensitive manner.

Conclusion

This study demonstrates the potential of machine learning models—particularly transformer-based architectures like DistilBERT—to accurately detect self-harm and suicidal ideation (SH/S) within youth's private social media conversations. By moving beyond binary classification and incorporating contextual, psycholinguistic, sentiment, and lexical features, our approach captures the nuanced spectrum of SH/S expressions, from hyperbole to personal disclosures. Importantly, we find that expanding the context window to sub-conversations significantly improves classification accuracy, underscoring the critical role of conversational context in understanding youth mental health language. Moreover, gender-specific patterns in SH/S expression highlight the need for inclusive models that account for diverse linguistic behaviors. As digital platforms become central to youth communication, our findings emphasize the importance of context-aware, ethically designed interventions that can support timely and sensitive mental health responses.

Acknowledgments

Will be provided upon manuscript acceptance.

Conflicts of Interest

None declared.

Abbreviations

AI: artificial intelligence

ML: Machine Learning

SH/S: suicide and/or self-harm

LLM: large language model

NLP: natural language processing

References

- [1] A. Shensa, J. E. Sidani, C. G. Escobar-Viera, G. E. Switzer, B. A. Primack, and S. Choukas-Bradley, "Emotional support from social media and face-to-face relationships: Associations with depression risk among young adults," *J. Affect. Disord.*, vol. 260, pp. 38–44, 2020.
- [2] P. Best, R. Manktelow, and B. J. Taylor, "Social work and social media: Online help-seeking and the mental well-being of adolescent males," *Br. J. Soc. Work*, vol. 46, no. 1, pp. 257–276, 2014.
- [3] C. Pretorius, D. Chambers, B. Cowan, D. Coyle, and others, "Young people seeking help online for mental health: cross-sectional survey study," *JMIR Ment. Health*, vol. 6, no. 8, p. e13524, 2019.
- [4] T. Kim, H. Hong, and others, "Understanding university students' experiences, perceptions, and attitudes toward peers displaying mental health-related problems on social networking sites: online survey and interview study," *JMIR Ment. Health*, vol. 8, no. 10, p. e23465, 2021.
- [5] A. Marchant *et al.*, "A systematic review of the relationship between internet use, self-harm and suicidal behavior in young people: The good, the bad and the unknown," *PloS One*, vol. 12, no. 8, p. e0181722, 2017.
- [6] A. Naz *et al.*, "Exploring lived experiences of adolescents presenting with self-harm and their views about suicide prevention strategies: a qualitative approach," *Int. J. Environ. Res. Public Health*, vol. 18, no. 9, p. 4694, 2021.
- [7] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *NPJ Digit. Med.*, vol. 3, no. 1, p. 43, 2020.
- [8] H. Ghanadian, I. Nejadgholi, and H. Al Osman, "Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models," *IEEE Access*, 2024.
- [9] K. D. Kannan, S. K. Jagatheesaperumal, R. N. Kandala, M. Lotfaliany, R. Alizadehsanid, and M. Mohebbi, "Advancements in machine learning and deep learning for early detection and management of mental health disorder," *ArXiv Prepr. ArXiv241206147*, 2024.
- [10] W. S. Journal, "When There's No School Counselor, There's a Bot," *Wall Str. J.*, 2025, [Online]. Available: https://www.wsj.com/tech/ai/student-mental-health-ai-chat-bots-school-4eb1ba55?utm_source=chatgpt.com
- [11] X. Li, F. Chen, and L. Ma, "Exploring the potential of artificial intelligence in adolescent suicide prevention: current applications, challenges, and future directions," *Psychiatry*, pp. 1–14, 2024.
- [12] C. Nuñez-Rola and N. J. Ruta-Canayong, "Social Media Influences to Teenagers," *Int. J. Res. Sci. Manag.*, vol. 6, no. 6, pp. 38–48, 2019.
- [13] G. McCulloch, *Because internet: Understanding the new rules of language*. Penguin, 2020.

- [14] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015.
- [15] A. Lavis and R. Winter, "# Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support around self-harm on social media," *J. Child Psychol. Psychiatry*, vol. 61, no. 8, pp. 842–854, 2020.
- [16] S. Choukas-Bradley, S. R. Roberts, A. J. Maheux, and J. Nesi, "The perfect storm: A developmental–sociocultural framework for the role of social media in adolescent girls' body image concerns and mental health," *Clin. Child Fam. Psychol. Rev.*, vol. 25, no. 4, pp. 681–701, 2022.
- [17] M. C. Keller and R. M. Nesse, "The evolutionary significance of depressive symptoms: different adverse situations lead to different depressive symptom patterns.," *J. Pers. Soc. Psychol.*, vol. 91, no. 2, p. 316, 2006.
- [18] B. Sapiro and A. Ward, "Marginalized youth, mental health, and connection with others: A review of the literature," *Child Adolesc. Soc. Work J.*, vol. 37, no. 4, pp. 343–357, 2020.
- [19] A. M. Memon, S. G. Sharma, S. S. Mohite, and S. Jain, "The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature," *Indian J. Psychiatry*, vol. 60, no. 4, p. 384, 2018.
- [20] K. Zeberga, M. Attique, B. Shah, F. Ali, Y. Z. Jembre, and T.-S. Chung, "A novel text mining approach for mental health prediction using Bi-LSTM and BERT model," *Comput. Intell. Neurosci.*, vol. 2022, 2022.
- [21] C. Crema, G. Attardi, D. Sartiano, and A. Redolfi, "Natural language processing in clinical neuroscience and psychiatry: A review," *Front. Psychiatry*, vol. 13, p. 946387, 2022.
- [22] A. Forte, G. Sarli, L. Polidori, D. Lester, and M. Pompili, "The role of new technologies to prevent suicide in adolescence: a systematic review of the literature," *Medicina (Mex.)*, vol. 57, no. 2, p. 109, 2021.
- [23] N. S. Ali, S. Qadir, A. Alsoubai, M. De Choudhury, A. Razi, and P. J. Wisniewski, "'I'm gonna KMS:' From Imminent Risk to Youth Joking about Suicide and Self-Harm via Social Media," 2024.
- [24] R. Haque, N. Islam, M. Islam, and M. M. Ahsan, "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning," *Technologies*, vol. 10, no. 3, p. 57, 2022.
- [25] E. Savoia, N. W. Harriman, M. Su, T. Cote, and N. Shortland, "Adolescents' exposure to online risks: Gender disparities and vulnerabilities related to online behaviors," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 11, p. 5786, 2021.
- [26] C. Brennan *et al.*, "Self-harm and suicidal content online, harmful or helpful? A systematic review of the recent evidence," *J. Public Ment. Health*, vol. 21, no. 1, pp. 57–69, 2022.
- [27] L. Soldaini, T. Walsh, A. Cohan, J. Han, and N. Goharian, "Helping or Hurting? Predicting Changes in Users' Risk of Self-Harm Through Online Community Interactions," *ArXiv Prepr. ArXiv180407253*, 2018.
- [28] A. Razi *et al.*, "Sliding into my DMs: Detecting uncomfortable or unsafe sexual risk experiences within Instagram direct messages grounded in the perspective of youth," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, pp. 1–29, 2023.
- [29] A. Atmakuru *et al.*, "Artificial Intelligence-based Suicide Prevention and Prediction: A Systematic Review (2019-2023)," *Inf. Fusion*, p. 102673, 2024.
- [30] S. Ali, "Youth, social media, and online safety: a holistic approach towards detecting and mitigating risks in online conversations," PhD Thesis, 2024.
- [31] S. Chancellor, E. P. Baumer, and M. De Choudhury, "Who is the "human" in human-centered machine learning: The case of predicting mental health from social media," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–32, 2019.

- [32] G. Waller *et al.*, “The barriers and facilitators to the reporting and recording of self-harm in young people aged 18 and under: a systematic review,” *BMC Public Health*, vol. 23, no. 1, p. 158, 2023.
- [33] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, “You Don’t Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2021, pp. 290–302.
- [34] N. Nordin, Z. Zainol, M. H. M. Noor, and L. F. Chan, “An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach,” *Asian J. Psychiatry*, vol. 79, p. 103316, 2023.
- [35] U. Naseem, M. Khushi, J. Kim, and A. G. Dunn, “Hybrid text representation for explainable suicide risk identification on social media,” *IEEE Trans. Comput. Soc. Syst.*, 2022.
- [36] N. P. Toliya and N. Nagarathna, “Leveraging Online Social Content for Early Detection of Suicidal Ideation: A Multi-Modal Deep Learning Approach,” in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, IEEE, 2024, pp. 1–7.
- [37] A. Razi *et al.*, “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*, 2022, pp. 1–9.
- [38] J. Robinson *et al.*, “Development of a self-harm monitoring system for Victoria,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, p. 9385, 2020.
- [39] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *ArXiv Prepr. ArXiv190810084*, 2019.
- [40] S. Kobayashi, “Contextual augmentation: Data augmentation by words with paradigmatic relations,” *ArXiv Prepr. ArXiv180506201*, 2018.
- [41] G. Ansari, M. Garg, and C. Saxena, “Data augmentation for mental health classification on social media,” *ArXiv Prepr. ArXiv211210064*, 2021.
- [42] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, “Natural language processing of social media as screening for suicide risk,” *Biomed. Inform. Insights*, vol. 10, p. 1178222618792860, 2018.
- [43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [44] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *ArXiv Prepr. ArXiv191001108*, 2019.
- [45] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “Mentalbert: Publicly available pretrained language models for mental healthcare,” *ArXiv Prepr. ArXiv211015621*, 2021.
- [46] D. Kodati and R. Tene, “Identifying suicidal emotions on social media through transformer-based deep learning,” *Appl. Intell.*, vol. 53, no. 10, pp. 11885–11917, 2023.
- [47] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [48] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Gated feedback recurrent neural networks,” in *International conference on machine learning*, PMLR, 2015, pp. 2067–2075.
- [49] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [50] P. Boot, H. Zijlstra, and R. Geenen, “The Dutch translation of the linguistic inquiry and word count (LIWC) 2007 dictionary,” *Dutch J. Appl. Linguist.*, vol. 6, no. 1, pp. 65–76, 2017.
- [51] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the international AAAI conference on web and social*

- media*, 2014, pp. 216–225.
- [52] J. Ramos and others, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, Citeseer, 2003, pp. 29–48.
 - [53] S. M. Jain, “Hugging face,” in *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, Springer, 2022, pp. 51–67.
 - [54] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, S. Ridella, and others, “The ‘K’ in K-fold Cross Validation,” in *ESANN*, 2012, pp. 441–446.
 - [55] X. Ying, “An overview of overfitting and its solutions,” in *Journal of physics: Conference series*, IOP Publishing, 2019, p. 022022.
 - [56] A. Mehday, A. Chehri, A. Jakimi, and R. Saadane, “Hyperparameter optimization with genetic algorithms and xgboost: A step forward in smart grid fraud detection,” *sensors*, vol. 24, no. 4, p. 1230, 2024.
 - [57] Z. Hu, J. Piet, G. Zhao, J. Jiao, and D. Wagner, “Toxicity detection for free,” *ArXiv Prepr. ArXiv240518822*, 2024.
 - [58] S. Frenda and others, “Sarcasm and implicitness in abusive language detection: a multilingual perspective,” 2022.
 - [59] L. St, S. Wold, and others, “Analysis of variance (ANOVA),” *Chemom. Intell. Lab. Syst.*, vol. 6, no. 4, pp. 259–272, 1989.
 - [60] T. M. Franke, T. Ho, and C. A. Christie, “The chi-square test: Often used and more often misinterpreted,” *Am. J. Eval.*, vol. 33, no. 3, pp. 448–458, 2012.
 - [61] C. Sweeney, E. Ennis, M. D. Mulvenna, R. Bond, and S. O’Neill, “Insights derived from text-based digital media, in relation to mental health and suicide prevention, using data analysis and machine learning: systematic review,” *JMIR Ment. Health*, vol. 11, p. e55747, 2024.
 - [62] S. Kim, A. Razi, A. Alsoubai, P. J. Wisniewski, and M. De Choudhury, “Assessing the Impact of Online Harassment on Youth Mental Health in Private Networked Spaces,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2024, pp. 826–838.
 - [63] A. Alsoubai *et al.*, “Profiling the offline and online risk experiences of youth to develop targeted interventions for online safety,” *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW1, pp. 1–37, 2024.
 - [64] P. Rowley, *Responding to Young People who Disclose Self-Harm: a Discourse Analysis of an on-line Counselling Service*. Lancaster University (United Kingdom), 2019.
 - [65] D. Galasiński and J. Ziółkowska, “The end of ambivalence. A narrative perspective on ambivalence in the suicidal process,” *Suicide Life. Threat. Behav.*, vol. 54, no. 5, pp. 888–899, 2024.
 - [66] F. D. Fincham, J. H. Grych, and L. N. Osborne, “Does marital conflict cause child maladjustment? Directions and challenges for longitudinal research,” *J. Fam. Psychol.*, vol. 8, no. 2, p. 128, 1994.
 - [67] D. Gentner, “Spatial metaphors in temporal reasoning,” in *Spatial schemas and abstract thought*, 2001, pp. 203–222.
 - [68] A. Wray and G. W. Grace, “The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form,” *Lingua*, vol. 117, no. 3, pp. 543–578, 2007.
 - [69] R. Alharbi, S. Langer, C. Hunter, N. Husain, F. Varese, and P. J. Taylor, “‘My entire life has moulded the person that I am’: narrations of non-suicidal self-injury and complex trauma in individuals with complex posttraumatic stress experiences,” *Psychol. Psychother. Theory Res. Pract.*, 2025.
 - [70] M. De Choudhury and E. Kiciman, “The language of social support in social media and its effect on suicidal ideation risk,” in *Proceedings of the international AAAI conference on web and social media*, 2017, pp. 32–41.
 - [71] N. Yıldırım and others, “The bullying game: sexism based toxic language analysis on online

- games chat logs by text mining,” *J. Int. Womens Stud.*, vol. 24, no. 3, p. 7, 2022.
- [72] E. L. Cameron, J. D. Fox, M. S. Anderson, and C. A. Cameron, “Resilient youths use humor to enhance socioemotional functioning during a day in the life,” *J. Adolesc. Res.*, vol. 25, no. 5, pp. 716–742, 2010.
- [73] D. C. Kerr, L. J. Preuss, and C. A. King, “Suicidal adolescents’ social support from family and peers: gender-specific associations with psychopathology,” *J. Abnorm. Child Psychol.*, vol. 34, pp. 99–110, 2006.
- [74] T. I. Tanni, M. Akter, J. Anderson, M. J. Amon, and P. J. Wisniewski, “Examining the unique online risk experiences and mental health outcomes of lgbtq+ versus heterosexual youth,” in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–21.
- [75] E. Arik *et al.*, “Communication dynamics and media interactions of young adults who have attempted suicide: a qualitative thematic analysis,” *Front. Psychol.*, vol. 15, p. 1460348, 2024.
- [76] E. Klineberg, M. J. Kelly, S. A. Stansfeld, and K. S. Bhui, “How do adolescents talk about self-harm: A qualitative study of disclosure in an ethnically diverse urban population in England,” *BMC Public Health*, vol. 13, pp. 1–10, 2013.