Systemization of Knowledge (SoK): Creating a Research Agenda for Human-Centered Real-Time Risk Detection on Social Media Platforms

Ashwaq Alsoubai
ashwaq.alsoubai@vanderbilt.edu
atalsoubai@kau.edu.sa
Vanderbilt University
Nashville, Tennessee, USA
King AbdulAziz University
Jeddah, Saudi Arabia

Jinkyung Park
jinkyung.park@vanderbilt.edu
Vanderbilt University
Nashville, USA

Sarvech Qadir
sarvech.qadir@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

Gianluca Stringhini
gian@bu.edu
Boston University
Boston, Massachusetts, U.S.A

Afsaneh Razi
afsaneh.razi@drexel.edu
Drexel University
Philadelphia, Pennsylvania, USA

Pamela J. Wisniewski
pamela.wisniewski@vanderbilt.edu
Vanderbilt University
Nashville, Tennessee, USA

ABSTRACT

Accurate real-time risk identification is vital to protecting social media users from online harm, which has driven research towards advancements in machine learning (ML). While strides have been made regarding the computational facets of algorithms for "real-time" risk detection, such research has not yet evaluated these advancements through a human-centered lens. To this end, we conducted a systematic literature review of 53 peer-reviewed articles on real-time risk detection on social media. Real-time detection was mainly operationalized as "early" detection after-the-fact based on pre-defined chunks of data and evaluated based on standard performance metrics, such as timeliness. We identified several human-centered opportunities for advancing current algorithms, such as integrating human insight in feature selection, algorithms' improvement considering human behavior, and utilizing human evaluations. This work serves as a critical call-to-action for the HCI and ML communities to work together to protect social media users before, during, and after exposure to risks.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

KEYWORDS

Online Risk, Human-Centered Machine Learning, Real-Time Risk Detection, Social Media, Literature Review

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642315

ACM Reference Format:
Ashwaq Alsoubai, Jinkyung Park, Sarvech Qadir, Gianluca Stringhini, Afsaneh Razi, and Pamela J. Wisniewski. 2024. Systemization of Knowledge (SoK): Creating a Research Agenda for Human-Centered Real-Time Risk Detection on Social Media Platforms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642315

1 INTRODUCTION

Within the evolving field of Human-Centered Machine Learning (HCML), scholars have highlighted the need to keep machine learning algorithms grounded in human social and psychological needs [21], to minimize bias by being more inclusive to adequately represent the myriad of individuals' experiences, and to incorporate transparency and interpretability to understand the potential harms that could be caused to people by these algorithms [14, 40, 54]. As such, there has been a shift in which scholars within the SIGCHI research communities have begun to apply a human-centered lens to synthesize and critique computational approaches for various forms of automated risk detection, including but not limited to online harassment, unwanted sexual solicitations, and mental health disclosures that occur via social media platforms (c.f., [3, 66, 120, 144, 146]). Indeed, social media has become a prominent part of people's lives that allows users to connect with others, share, create, and engage with various forms of digital content [2]. In 2022, there were 3.96 billion social media users who spent hours a day using various social media platforms (e.g., Facebook, Twitter, Instagram) [10], demonstrating how social media is now an irrevocable and important part of our daily lives.

While social media can undeniably be beneficial, it can also facilitate digitally-mediated risks (e.g., amplifying misinformation, mental health challenges, and interpersonal violence [50, 103, 111, 112, 131, 148]) that cannot be ignored. As a case in point, Meta (i.e., Facebook) recently faced significant criticism and legal scrutiny after the release of internal documents[1] suggested that the company

[1] https://www.wsj.com/articles/the-facebook-files-11631713039

took inadequate actions to mitigate risks related protecting youth from sexual exploitation, preventing human and drug trafficking, unfairly influencing political outcomes, and other harms to users and society-at-large [43]. Yet, Meta is not alone regarding such criticism [130], nor have they or other social media platforms been inactive in addressing these mounting concerns. Legislators are also grappling with how to make social media companies accountable for the harms that occur on their platforms and have introduced several bills (e.g., Kids Online Safety Act [32], Fight Online Sex Trafficking Act [33, 63]) to proactively protect social media users from informational, mental health, and physical threats that have been at the forefront of news media reports, and consequently, machine learning (ML) and Human-Computer Interaction (HCI) research. These issues are hard to tackle because of the complexities inherent in human behavior that may result in risk misidentification.

Timely and accurate risk identification is necessary to effectively prevent harm and to provide a safe environment for all social media users [48]. Therefore, recent advancements in ML and automated risk detection have grown beyond the traditional supervised learning paradigm to tackle more complex and dynamic problems, such as "real-time" risk detection [101]. The real-time aspect of detection is important as identification must occur as early as possible to prevent the spread of the risk (e.g., fake news) or to mitigate harm as a result of it (e.g., mental health problems). Therefore, several recent works [118, 122] on computational machine-learning algorithms for real-time risk detection on social media have called for more research to advance the *technical aspects* of real-time risk detection to optimize performance. These works provide valuable insights about advancing the computational approaches for real-time risk detection; yet, there is still a need to evaluate real-time algorithms for detecting risks from a human-centered perspective (e.g., whether and how such algorithms can impact users in the real world) to ensure that they can effectively be leveraged as a first defense to prevent online harm, as opposed to contributing to it.

To apply such a human-centered perspective, we leveraged and extended Razi et al.'s [120] generalizable framework established for conducting human-centered systematic reviews of computational risk detection research. We apply this framework to the novel context of real-time risk detection within social media and augment it by adding new dimensions (e.g., input prioritization, timeliness) related to 'real-time' risk detection, which has been conceptualized and operationalized in the literature in multiple, and at times, conflicting ways. In this paper, we considered both computational and human-centered aspects of this literature to create a forward-thinking research agenda that advances our capacity to proactively protect social media users from online harm as it unfolds. As such, we set forth to answer the following high-level research questions:

- **RQ1:** *How has 'real-time' social media risk detection been defined and operationalized in the literature?*
- **RQ2:** *What are the state-of-the-art computational trends for real-time risk detection on social media?*
- **RQ3:** *Using a human-centered lens, what are the potential gaps and areas for future research for real-time risk detection on social media?*

To answer these questions, we systematically reviewed 53 peer-reviewed papers published between 2015 and 2023 that tackled core aspects of 'real-time' risk detection using social media data. We broadly considered all types of social media risks that may result in individual-level harm (e.g., mental health, sexual solicitations) or community-level harm (e.g., fake news, misinformation). We qualitatively coded the articles to answer our research questions. Overall, we found that real-time risk detection has been predominantly operationalized as early risk detection after-the-fact, but as early as possible (RQ1). For RQ2, the computational trends in prior studies included utilizing publicly available large-scale datasets, using commonly known machine learning features, improving deep learning-based approach, presenting the performance evaluation metrics mainly using pre-defined chunks of data. For RQ3, we identified gaps and opportunities for future research to advance these computational approaches based on a human-centered perspective. Opportunities included placing humans at the center of data collection and model evaluation endeavors, with the aim of comprehending their behaviors to serve as the foundation for the selection of features and the optimization of real-time risk detection models. To synthesize our findings, we created a cohesive framework to direct future research on creating efficient and human-centered real-time risk detection algorithms for social media. Our systemization of knowledge makes the following novel contributions to the HCI, HCML, and ML research communities:

- Formally defined and expanded the term "real-time" within the context of social media risk detection by incorporating a spectrum of detection mechanisms to detect the risk before, early, and to mitigate harm after-the-fact.
- Highlighted trends and best technical practices of the existing state-of-the-art computational approaches for real-time risk detection.
- Extended Razi et al. [120] and discovered potential gaps within the literature and provided agendas toward human-centered real-time risk detection that goes beyond the current state-of-the-art in computational risk detection.
- Established a research agenda for advancing real-time computational risk detection to address both the computational challenges and human-centered gaps.

Next, we will synthesize the related work that motivated the need for this systematic literature review.

2 BACKGROUND

In this section, we synthesize the current literature on trends within computational approaches for social media risk detection, timeliness in social media risk detection, and human-centered lens to review risk detection algorithms.

2.1 Trends within Computational Approaches for Social Media Risk Detection

Prior literature reviews of computational risk detection on social media focused predominantly on evaluating the data collection and prepossessing techniques, feature engineering process, algorithms, and common machine learning metrics for benchmarking performance [23, 51, 114, 159]. A major number of these reviews have been centered around evaluating which machine learning algorithm

performed best for detecting risks on social media. Traditional algorithms such as Logistic Regression, Support Vector Machines, Random Forest, and Naive Bayes have been extensively used in detecting risks in various social media platforms [4, 9, 23, 44, 51, 110, 163]. However, given the massive scale of social media data, these algorithms were found to not adapt well to the evolving patterns of the risks and struggled to handle large volumes of data [13]. In recent years, there has been an increased interest in leveraging deep learning models, particularly suited for large-scale and complex datasets such as social media dataset [47, 90]. For instance, Dowlagar and Mamidi [41] found that transformers with selective translation demonstrated promising results compared to other common neural network-based models. As such, prior research has pointed to the importance of capturing the intricate patterns and evolving trends of risks on social media, which could mainly be accomplished through leveraging dynamic, novel, and specialized techniques for annotating datasets, crafting features, or enhancing models [4]. In fact, these techniques were shown to enhance risk detection capabilities and provide more reliable and effective risk management solutions in social media contexts. For instance, Yi and Zubiaga's [163] showed that novel models (i.e., MMCD [115] and XBully [29]) outperformed all pre-trained language models (e.g., BERT). Expanding beyond these works, in this review, we focused on identifying the trends within novel computational approaches, rather than off-the-shelf models, to detect the rapidly changing nature of risks within social media in real time.

2.2 Research on 'Real-Time' Risk Detection on Social Media

Research has called for current and future efforts on detecting and mitigating risks in social media to move towards building real-time risk detection systems [66, 120]. One of the main efforts toward presenting and evaluating timely risk detection was introduced by the Early Risk Prediction on the Internet (eRisk) group, to examine methodologies and metrics related to early risk detection. Based on their yearly events, several reviews [83–86, 106, 107] have been published to evaluate the timing risk detection for a myriad of issues (e.g., depression, self-harm, pathological gambling, and eating disorders) using social media data. Real-time solutions lie in performance-oriented evaluations; for example, detection time [39]; yet, a remaining question is what the real-time aspects of these models are that set them apart from more traditional and cross-sectional approaches for computational risk detection.

Ample literature within computer science has focused on specifying the definition of "real-time" problem-solving [39, 67, 104, 133]. Examples of common real-time definitions were "there is a strict time limit by which a system must have produced a response, regardless of the algorithm employed" [104] and "ability of the system to guarantee a response after a (domain defined) fixed time has elapsed [67]." These definitions carry flexibility, leaving room for varied interpretations based on application. Therefore, researchers have attempted to identify specific components of real-time systems [39, 133]. For example, Shin et al [133] presented three main components of real-time systems, which were *time*, arguably the most important aspect of real-time systems, *task* that must be accomplished before the deadline, and *message or response* that should be received in a timely manner. Given these efforts of identifying real-time components, a more unified and comprehensive definition of real-time was still needed to provide clarity and precision. Therefore, Bruda and Akl [18] presented a formal and unified theory of real-time definition that was generalizable across domains. This theory consisted of two concepts centered on real-time systems: "computing with deadlines, and input data that arrive in a sequential manner or real-time", which is the definition of our review.

2.3 Using a Human-Centered Lens to Review Real-Time Risk Detection Algorithms

The SIGCHI community has recently exhibited a growing interest in human-centered reviews, aiming to assess the effectiveness and impact of algorithms in real-world contexts (e.g., [21, 66, 120, 129]). Scholars have presented systematic reviews that underscore the importance of a human-centered approach to online risk detection, focusing on specific topics such as cyberbullying [66], sexual risks [120], child welfare system [129] or mental health [21]. In the context of misinformation detection, Das et al. [37], reviewed NLP approaches for fact-checking from different human-centered strategies. They suggested guiding technology development for human use and practical adoption, and human-centered design practices early in model development. In another work, using a three-prong human-centeredness algorithm design framework, Kim et al. [66] analyzed cyberbullying detection approaches and found a lack of human-centeredness in defining cyberbullying, establishing ground truth in data annotation, assessing detection models' performance, speculating the uses and users of the models, including potential negative consequences. These prior reviews highlighted that human-centered approaches to evaluating risk detection algorithms are pivotal to ensure that the algorithms are designed to benefit those who are negatively affected by the risks the most [120], and how their involvement in research can lead to more practical, widely accessible solutions catering to individuals' diverse needs [66].

In this paper, we adopted the human-centered framework proposed by Razi et al. [120] for systematically evaluating real-time risk detection algorithms in terms of: 1) characteristics of the **dataset**, 2) pre-processing and **model development**, 3) **evaluation**, and 4) **applications and interventions** (As shown in Table 1). While our work leverages Razi et al.'s framework for how to conduct a human-centered review of computational risk detection research, the scope of our research differs. Razi et al.'s work focused solely on online sexual risks, similar to the other human-centered reviews of computational risk detection that focused on singular risk types (e.g., online harassment [66], and mental health [21]). In contrast, our review synthesizes computational approaches *across multiple risk types*. This approach allows us to consider these risk detection algorithms' broader implications and potential consequences in real-world settings, ensuring the development of more effective and socially responsible solutions. Most importantly, while these systematic and human-centered reviews [21, 66, 120] covered many aspects of computational risk detection, in general, they did not specifically focus on the *real-time* aspect of risk detection, which is the novel focus and contribution of this paper.

Table 1: Codebook for RQ2 RQ3 (N = 53 articles) based on the Razi et al. framework [120] for performing a human-centered review of computational research. Note: * and bolded text in the table represents new dimensions and/or emerging codes we added to the Razi et al.'s framework to extend it to better account for research that focuses on 'real-time' computational risk detection.

Razi's et al. [120] Dimensions	Codes	Sub-Codes
Characteristics of the Datasets: *What were the sources for data collection? What was the privacy level of the dataset? Was the data collected from targeted users? How large were the datasets? What were the data types? How was the data annotated for training datasets? What was the distribution of classes?*	Data Source	Twitter (68%), Weibo (26%), Instagram (15%), Vine (9%), Reddit (9%), Meta (4%).
	Privacy Level	Public (100%), Private (0%)
	Selection Criteria*	Unidentifiable users (94%), Targeted users (6%)
	Dataset Size*	Large (47%), Medium (25%), Small (28%)
	Data Type	Text (100%), Meta (100%), Images (11%), Videos(4%)
	Ground Truth	Existing (74%), Third-party annotators (26%), Automatic (17%)
	Class Distribution*	Balanced (42%), Unbalanced (58%)
Pre-Processing and Model Development: *How was the data processed for simulating real-time? What were the features and how were they calculated for the model? How the data were prioritized to review and detect risk? What machine learning model (s) were used?*	**Data Processing***	Fixed chunks of data (83%), Dynamic input (17%)
	Feature Selection	Domain specific/ Theory Driven (32%), General ML features (100%)
	Feature Computation*	Straightforward (92%), Optimized (8%)
	Input Prioritizing*	Equal prioritization (96%), Prioritizing technique (4%)
	Algorithms	Deep learning (60%), Statistical (40%)
Evaluation: *What accuracy and timeliness metrics were used? What explainability analysis was incorporated to explain the models' performance?*	Accuracy	F1-score (70%), Accuracy (53%), Recall (58%), Precision (51%), AUC (11%), RMSE (4%)
	Timeliness*	Fixed chunks of input (53%), Fixed time window (21%), Time (21%)
	Explainability	Qualitative analysis (32%), Error analysis (13%), Case studies (13%), Models' fairness (2%), and Human evaluations (2%)
Application and Interventions: *What were the final artifacts? What interventions were provided for risk mitigation?*	Applications	Algorithm only (92%), Interfaces (4%), Deployment (6%)
	Interventions	Alerts (6%), Immunization (4%), Language alteration (2%)

3 METHODS

Below, we describe our systematic review of the literature and our qualitative synthesis of the articles in our dataset.

3.1 Systematic Literature Search Process

We selected five electronic databases (i.e., IEEE Xplore Digital Library, ACM Digital Library, ScienceDirect, Springer-link, and ACL Anthology) that ranged in computational and social science research approaches for the initial literature search to ensure broad coverage. We searched using combinations of keywords at the intersection of 1) social media (i.e., "social media", "Twitter', "Facebook", "Instagram", "YouTube"), 2) real-time detection (i.e., "real-time", "forecast", "early detection", "early prediction", "proactive prediction", or "proactive detection"), 3) online risks ("risk detection", "mental health", "cyberbullying", "sexual", "hate speech", "fake news", "incivility", "harassment", "abuse", or "spam"), and 4) computational approaches ("machine learning", "artificial intelligence", "deep learning", or "algorithm"). We did not restrict the filter to a given date range. The initial search resulted in 2,212 papers, where 48% of the papers were from the ACM Digital Library. To confirm relevancy, we read through the papers' titles, abstracts, methods, results, and conclusions based on the following inclusion criteria:

(1) The paper was a peer-reviewed published work. Journal articles and conference proceedings were both included.
(2) The paper should not be a purely theoretical analysis or summarize or evaluate existing studies (e.g., reviews).
(3) The paper focused on social media risk detection. We used a wide angle of prevalent social media risks since this literature review focuses on real-time detection approaches, not online risks. Social media was selected due to the affordance of open-source data, which made these platforms a popular choice for researchers to apply risk detection approaches [136].

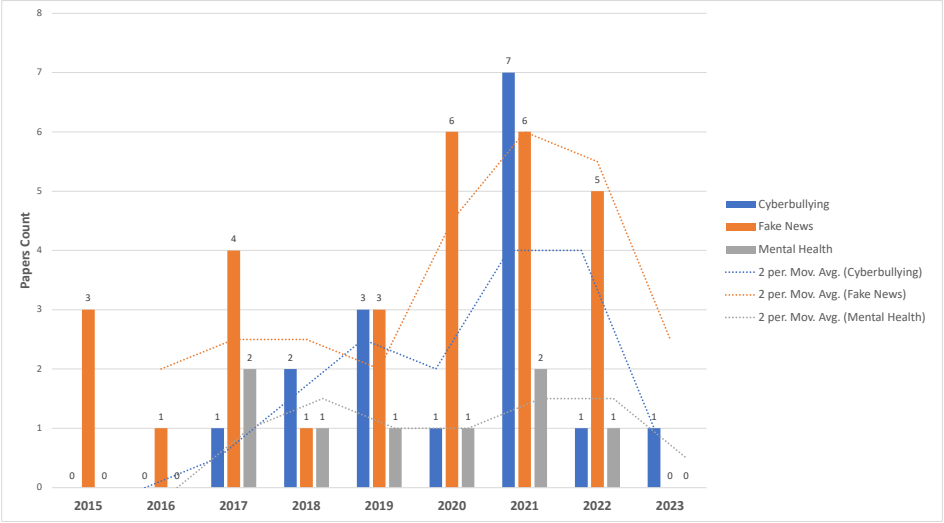

Figure 1: Number of Publications by Risk Type Over Time.

(4) The paper presented a real-time approach using a computational or algorithmic approach such as Natural Language Processing (NLP) or Machine Learning (ML) that consists of both aspects of real-time models' definition: 1) sequential input and 2) timeliness [18].
(5) The paper provided a new computational approach or an enhancement of an existing approach, rather than only training or fine-tuning off-the-shelf computational models that are mainly designed for general use cases and may not provide the same level of precision when applied to specific domains like real-time risk detecting in social media.

We coded a paper as relevant if it met all the criteria above, which resulted in 45 papers. To identify additional relevant papers that were not yielded in our initial search, we cross-referenced the citations of the relevant papers. This cross-referencing resulted in 33 unique papers that were potentially relevant, of which 8 papers met our inclusion criteria. After one more iteration of this process, no additional relevant papers were identified, which suggests that we reached a saturation point. Therefore, our final search resulted in **53** relevant papers for our review.

3.2 Data Analysis Approach

To answer RQ1, we used a thematic analysis approach [16] to code papers related to how real-time risk detection was defined. To answer RQ2 and RQ3, we utilized the human-centered framework presented by Razi et al. [120] to review papers based on the 1) ecological validity of the dataset for detecting the risks, 2) investigating if the models are grounded in human theory, knowledge, and understanding, 3) performance of algorithms in terms of meeting end users' needs, 4) their outcomes when deployed in real-world settings. Razi's et al. [120] framework was created based on computational sexual risk detection; therefore, while coding the papers, we identified and added emerging codes that were not covered in this framework and suited our *real-time* detection for a generalized view of risks. We iteratively created new codes to suit the real-time aspect of the detection process. Codes were allowed to overlap for double-coding. Two coders labeled the same 10% of articles, and we calculated Cohen's Kappa IRR [31] to ensure the robustness of our coding process, which was 0.87. The researchers met to discuss the articles to resolve conflicts. Once a consensus was reached, the codes were updated. The remaining articles were then divided and coded by the two coders. The first author identified key themes by reviewing and conceptually grouping the final codes. The definitions of our codes and grounded codes that emerged from our data are shown in Table 1.

4 RESULTS

We present the characteristics of our dataset, followed by our results organized by our over-arching research questions.

4.1 Defining and Operationalizing "Real-time" Risk Detection on Social Media

4.1.1 Types of Risks Detected. As illustrated by Figure 1, the research papers considered three main types of risk: fake news (55%), cyberbullying (30%), and mental illness (15%). There was a pronounced surge in articles on the real-time detection of fake news in the years 2020 and 2021. This time span coincided with the COVID-19 pandemic and the concurrent escalation in both rumors and fake news dissemination [55]. This alignment potentially contributed to the escalated scholarly output during this period to combat all kinds of fake news, including rumors and misinformation. The historical significance of 2017 for the increase in the number of publications on real-time risk detection for mental health could be aligned with the launch of the ERISK workshop, [2] as discussed in our Background section. A common theme among these papers was the incorporation of the time-evolving aspect of the risk when building risk detection algorithms. For instance, cyberbullying implies a repetitive behavior over time [82], while mental illness symptoms,

[2] https://erisk.irlab.org/2017/index.html

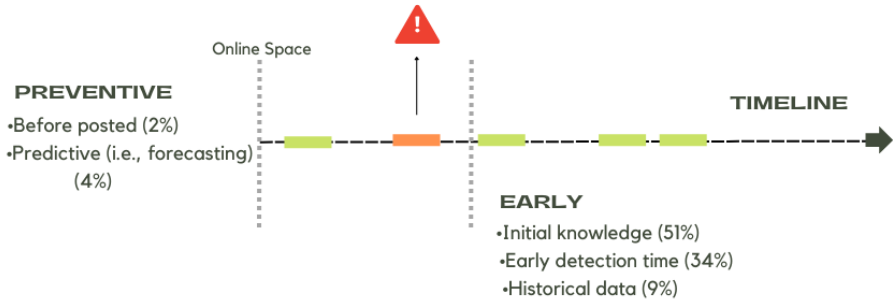

Figure 2: Timing in real-time risk detection approaches.

Table 2: Real-Time Approches

Real-Time	Types	References
Early (94%)	Initial Knowledge (51%)	[19, 25, 27, 38, 59, 65, 69, 70, 73, 80–82, 88, 89, 93, 97, 98, 100, 123, 125, 126, 137, 156, 157, 160, 165, 166]
	Early Detection Time (34%)	[28, 29, 45, 49, 77, 79, 87, 113, 116, 117, 147, 151, 155, 158, 161, 162, 164, 167]
	Historical Data (9%)	[68, 127, 128, 154, 168]
Preventive (6%)	Before Posted (2%)	[36]
	Predictive (i.e., forecasting) (4%)	[78, 91]

such as eating disorders and depression appear for a longer period of time to ensure correct diagnosis of mental illness [117, 161]. Additionally, papers that addressed fake news and rumor detection presented the timing as the spread or propagation of such content throughout the network.

4.1.2 Definitions and Operationalization of 'Real-Time' Risk Detection. For RQ1, we first set out to understand how "real-time" in real-time risk detection models were defined and operationalized in the reviewed papers. All papers in this review accounted for and discussed the timing of risk detection when presenting their approach. However, our review revealed that real-time or timely detection had different definitions, ranging from preventative risk detection to early risk detection. As shown in Table 2 and Figure 2, most of the papers (94%) presented early risk detection approaches to retrospectively detect the risks after it was posted online; yet, differed mainly in terms of the focus. Over half of the papers (51%), which form the majority of early approaches, focused on detecting the risk using *early stages* of information propagation or interactions (e.g., when only the initial retweets or comments became available). These approaches were trained to learn the risk using partial information of online content, such as the comments within the first 1-, 3-, and 5-day period [65]. The rest of the papers considered the optimization of the least number of observations needed to make an accurate decision as part of models' learning. A common challenge among these papers was how the model could achieve high accuracy with the lack of sufficient cues.

Another trend for the definition of "real-time" in the real-time risk detection papers revolved around the *detection time* to be as "early-as-possible" after the risk already occurred, without considering limited online content when training the models. Therefore, the latency of risk detection, the time gap between when a risk is detected and when it's posted online, became a crucial metric. This measure was compared to the model's accuracy over time to showcase the trade-off between early detection and accuracy. The improvement in model accuracy came at the expense of early detection, implying that the more data the models used to give accurate predictions, the more time the model would take to provide predictions. The third group of early detection methods expanded their input scope by incorporating *historical data* when detecting the risks early. As mentioned earlier, mental health indicators often depend on the presence of symptoms over an extended period. Consequently, we observed that the majority of papers utilized historical online content, such as changes in emotions [128], to the detection of mental health issues.

In this review, we found that sometimes, real-time risk detection was defined as a "preventative" approach that attempts to prohibit the risk from reaching online platforms (6%). The preventative approaches could be divided into two types: 1) prevent the risk offline prior to being posted online (2%) [91] or 2) predicting the possibility of risk occurrence in the future (4%) [36, 78]. In both preventative approaches, the main goal was to prevent the risk from reaching online spaces, attempting to reduce its possible harm. For the first type of preventative approach, Masud et al. [91] presented a normalization real-time model for hate speech that intervenes when users type hateful keywords (i.e., an auto-complete fashion) and suggests normalized texts as an alternative before toxic words are posted online. This approach aimed to encourage individuals to post less toxic opinions online by proactively sensitizing them.

The second type of preventative approach (i.e., forecasting) formulated the risk detection problem for a given post and its initial history of comments, the model should forecast the risk, in the upcoming comments. Unlike early approaches that relied on delays

to measure the models' earliness, forecasting approaches relied on measuring the leading time for the model to accurately predict future risk incidents. Specifically, the effectiveness of these proactive approaches was measured by how accurate the model was in predicting the risk within N number of future comments. For instance, Dahiya et al. [36] evaluated the foreseeability of the hate intensity model by illustrating how far the model can predict hateful comments for $n = 1, 3, 5, 7$ and showed that the model performed well even for 7 future comments with Mean Absolute Percentage Error (MAPE) of < 40. Forecasting or predicting social media risks can be useful to prevent risks from reaching online spaces and reduce damage. However, the applicability of such predictions was found to be more challenging than the early approaches as it is difficult to determine the exact occurrence and impact of risks beforehand, making it challenging to evaluate the effectiveness of forecasting models objectively.

4.2 Applying a Human-Centered Lens to Computational Trends for Real-Time Risk Detection

We organize our results by the dimensions of Razi et al.[120]'s framework to highlight trends and best practices in HCML. Then, we present our findings from the 53 articles analyzed in parallel structure to our codebook in Table 1.

4.2.1 Characteristics of the Datasets. The HCI and HCML communities care a great deal about leveraging ecologically valid datasets that are representative of the target populations they set out to study [120, 135]. Given that data size and type are the foundation of algorithmic development, the HCML lens emphasizes making sure that the datasets match the real-world users' context [56]. From a human-centered perspective, collecting ground truth annotations from humans, specifically from actual victims ensures that the training risk detection models reflect real-world experiences and accurately represent the risks users face online [66]. Furthermore, leveraging human insights and theories when building real-time risk detection for social media is crucial for building human-centered models [12]. Below, we summarize the trends we saw in the real-time risk detection literature compared to these best practices from the HCML literature.

Data Source, Privacy-Level, and Selection Criteria: All of the articles in this review relied on utilizing publicly available datasets. Over half of the papers (68%) used a scraped dataset from Twitter, followed by Weibo (26%). Thirteen papers (25%) used datasets from image and video-based social media platforms (i.e., Instagram (15%) and Vine (9%)). Reddit discussions (9%) were also utilized for real-time risk detection, while only two papers (4%) [91, 157] used a dataset from Facebook. This made the Twitter platform (Now 'X') the most dominant platform for datasets that were used to train and evaluate real-time risk detection models. Since most of the datasets were scraped from public social media posts (94%), data were scraped from unidentifiable users of the platform, without specifying users' target characteristics (e.g., profile characteristics). Only three (6%) papers [68, 117, 161] studied specific user groups instead of using a general query to collect social media data from any user, such as obtaining data from both depressed and non-depressed users. While they focused on collecting data from targeted groups, the selection criteria were still based on keywords and hashtags within the posts for identification. Table 4 in the appendix reports the datasets that the papers examined in this literature.

Dataset Size and Data Types: We found that 47% of the papers used a large social media dataset that consisted of more than 1 million instances with a maximum of 40 million tweets [73]. In addition, another set of papers (25%) utilized medium-sized datasets with thousands of instances, which ranged from equal to or more than 10K instances (13%) to more than 100K instances (11%). We also found (28%) of the papers used equal to or more than a thousand instances, with a maximum of around 5K posts [36, 80, 87, 91, 151]. All the papers reviewed relied on text and metadata (e.g., mainly the time of the post) datasets. Only 15% of the papers used datasets that included media data, such as images or videos derived from Instagram or Vine. However, the authors only used the textual features extracted from captions of media, media content, comments posted on the media, or meta information such as the number of likes and shares. For instance, Chelmis and Zois [25] only used the text of the extracted emotional cues from the Instagram pictures to train their early cyberbullying risk detection model while López-Vizcaíno [82] used the extracted textual features from the videos, including the nature of the video content and emotions.

Ground Truth: We found a noteworthy proportion of papers (72%) were based on existing labeled datasets that were ready for other researchers to use, which was illustrated in Table 5 in the appendix. Given the large scale of the collected data, most of these datasets were labeled using automatic approaches such as keyword or event searching (30%) or fact-checking websites (8%). A few of the ready-to-use datasets were labeled through researchers (4%), experts (6%), or crowdsourcing (8%). Of the research papers that we reviewed, 38% of the papers collected and labeled their dataset, most of which used human annotations (28%), including researchers (9%), psychology students or professionals (6%), crowdsourcing (6%), and experts (e.g., platform monitoring team) (6%). The rest of the papers relied on automated or rule-based approaches to label the dataset (9%). For example, Petrescu et al. [113] used the "Hateful Symbols or Hateful People" dataset to label by checking if there was a hateful symbol and speech term in the tweet, it was labeled as harmful.

Class Distribution: Half of the papers (58%) leveraged unbalanced datasets while the rest (42%) used balanced datasets for their risk detection models (as illustrated in Table 6 in appendix). While the unbalanced distribution of classes reflects the realistic distribution of the risks in social media platforms as the online risk interactions are significantly less than safe interactions, this unbalanced distribution could yield severe consequences based on learner's prediction bias toward the majority class [132]. Only 11% of the papers presented an approach or discussed how to ensure the model fairness when using unbalanced datasets. These papers introduced improvements to the models' equation to ensure the reliability of the model results such as using a modified misclassification costs ratios [98], adding Class-Balanced loss [35], and Focal Loss [74], which apply "a class-wise re-weighting scheme", that were presented by Sawhney et al. [127]. Liu et al. [81] took another direction by using the PU-learning approach [71], to learn from positive (P) and unlabeled (U) instances. This PU-learning framework identifies a sample of pseudo-negative instances from the unlabeled

dataset and the classifier was then trained using these samples. The authors showed that the accuracy performance remained the same across the fully labeled balanced dataset and the unbalanced dataset, presenting promising results for future research to adopt.

4.2.2 Pre-Processing and Model Development. The human-centered perspective highlights that it is crucial for the computational models to be grounded in human knowledge and human theories [12, 120]. This grounding ensures that these models can better understand and serve the social and psychological needs of individuals and society at large [21]. In addition, human-centered models should provide local interpretability for individual detection decisions, making it easier for users to understand the model's reasoning behind those detection decisions [76]. In this section, we describe the computational trends of data pre-processing and model development for real-time risk detection.

Data Pre-Processing: Most of the articles (83%) used a "streaming-like" approach by segmenting the dataset into equal sizes of chunks of data that mostly were predefined either by a fixed time window such as [69, 87, 161] or a fixed number of posts such as [19, 68, 165] as listed in Table 7 in appendix. These chunks of data were fed to the training models sequentially to produce real-time risk detection decisions. Scholars in this literature have identified that segmenting the datasets is a limitation as these chunks are not representative of real-world peoples' interactions. Therefore, a few papers (17%) implemented sequential training by incrementally adding data (i.e., posts and comments) in chronological order as they were available online to mimic how the data was typically available in the real world, without any segmentation. In fact, Leiva and Freire [68] compared the "first n", which is the first n of messages that were concatenated to make predictions, and the "dynamic" setting, which was messages that were used as they became available to make the predictions when a confidence value reached a certain threshold. They found the dynamic approach with a 0.5 confidence value threshold outperforming the first n chunks of data, with a 0.05 early risk detection measure and 0.77 recall, illustrating that the dynamic approach could be considered the best practice in this field.

Features Selection: Upon reviewing real-time risk detection for social media literature, we found all papers (100%) relied on machine learning features, such as textual, network, user-based, or temporal features; among them, significant emphasis was given to leveraging textual (66%) features for detecting social media risks (Table 8). The textual features were found to be drawn mainly from the posts, comments, image captions, or video descriptions, which included text embedding, Linguistic Inquiry and Word Count (LIWC), term-frequency-inverse document frequency (TF-IDF), and Bag-of-Words (BoW). Meanwhile, 51% of the papers emphasized the importance of considering the social network contextual clues such as the social network (i.e., derived from users' relations) and conversational network (i.e., formed through users' retweets or comments for a given post), which found to improve the detection models comparing with comment streams [128]. Domain-specific or theory-driven features were found in 32% of the reviewed articles. For instance, the real-time risk detection approaches to detecting mental illness mainly relied on prior psychological research such as suicide ideation on emotional reactivity [145], intensity [75], and instability [105] to identify a user's emotional spectrum over time. Theories such as the wisdom of the crowd and domain-specific measures such as degree of skepticism and susceptibility were leveraged in papers [69, 77, 79, 97] to harness the fake news and rumors detection algorithms. For interpersonal risks, we found a trend among the papers that mainly focused on using domain-specific features such as hate lexicons (e.g., Hatebase[3] and Luis von Ahn's Offensive/Profane Word List [4]) or counting the number of hateful instances [36, 73, 78]. A few papers (11%) leveraged domain-specific user behavioral features that described user behaviors associated with certain risks such as their influence and role in rumors propagation [59] or extracting the digital "user footprint" of their abusive behaviors across multiple platforms [157].

Features Computation: Features in real-time risk detection models were learned in a sequential and incremental fashion. With the massive scalability of social media, computing these features was one of the challenges discussed in the real-time risk detection literature to provide scalable and timely risk detection solutions while maintaining sufficient accuracy. Most of the papers we reviewed (92%) used a straightforward approach: computing the features over time by doing a full rerun on the data as they become available [69]. A few papers (8%) introduced approaches to reduce the feature computation timing by applying an attention mechanism to differentiate the importance of calculating the features from risk comments [73] or sorting features in increasing order based on their importance to make an early accurate decision based on the most important features [25]. Dahiya et al. [36] took another direction by leveraging majorization-minimization algorithms [30], where the model computed the parameters only on the recently observed data, which led to faster processing. Another approach presented in the papers was that once the features were computed for the first set of input (e.g., comments) when new input was available, the features for the new input were calculated while reusing the previously calculated features, leveraging the incremental computation [52, 53]. Unlike the models that perform a full run as each new data is available, this approach resulted in less re-computation overhead and would capture the naturalistic way of users' online interactions to provide timely risk detection.

Input Prioritization: In this review, we identified that most of the papers (96%) applied the real-time risk detection models by considering equal importance to classify all public conversations (posts and comments) without having a procedure to prioritize these data for detection. The significance of this procedure is mainly related to increasing the responsiveness of the detection approach to protecting people when needed. In addition, having less number of conversations or messages to classify or schedule the examined input for risk detection would lower the computational overhead for feature calculation, which in turn would produce faster risk decisions. Only two papers (4%) [116, 164] in our review identified and addressed this gap. Under the assumptions that "most media sessions are not bullying in nature, so not all media sessions need to be monitored equally", Rafiq et al. [116] applied the resources on public media sessions (i.e., posts and associated comments) that were most likely to result in cyberbullying by presenting a Dynamic Priority Scheduler (DPS), which dynamically assigned high

[3] www.hatebase.org
[4] www.cs.cmu.edu/~biglou/resources/bad-words.txt

priority to sessions to be examined by the detection model and low priority to the ones that were postponed until new comments were available. This scheduler showed maximum responsiveness when it was compared with other traditional approaches. Zang et al. [164] applied a machine learning algorithm to calculate the False/True probabilities based on the initial features of the events (multiple tweets and replies about a certain event). The events with a high probability were assigned a True/False label and the false information events were moved to another step to be tracked for the final decision. By doing so, the final detection algorithm had a smaller set of possible false events to track them rather than inefficiently tracking all events.

Algorithms: Most of the real-time risk detection models addressed in the literature implemented either deep learning (60%) or statistical approaches (40%). Papers that relied on statistical theories mainly leveraged Markov models (11%), Bayesian model (9%), posterior probability (8%), and State Space models (4%) as illustrated in Table 9 in the appendix. Unlike the aforementioned traditional models that do not account for the sequentiality of data, deep learning models used commonly within real-time risk detection papers were expected and proved to be effective. Yet, off-the-shelf deep learning models were found to suffer when implemented in real-time risk detection, mainly because they can not account for the uneven time interval between responses or comments. Therefore, Sawhney et al. [127] proposed an approach that utilizes a monotonically decreasing function of elapsed time to transfer time into appropriate weights. Convolutional Neural Network (CNN)–based classifiers often generate complex and less interpretable representations of text. Therefore, works such as Liu et al. [81] improved CNN for fake news detection by adding a position-aware attention mechanism, which is an extension of the basic attention mechanism [94], which was used to learn how much attention should be given to the data points in the sequence.

For statistical models such as the Markov model, the risk detection problem was formulated as a sequential or time-series text data that was represented as a chain of posts/comments. The papers we reviewed presented their improvements to tailor these models for the types of risk that were tackled. For example, Li et al. [69] improved the standard Kalman Filter, which is a mathematical algorithm used for state estimation to achieve progressive detection through learning the temporal information of time-series data that arrive irregularly. A few papers (11%) discussed approaches where deep learning and statistical models were combined to foster the deep learning models' capabilities to capture irregular conversations in an evolving nature. For instance, Dahiya et al. [36] utilized state-space models that were combined with deep-learning models, known as deep-state models, where there was a sequence of unknown states that were considered as learned features to represent the data, and then, at each time step, the model provided a probabilistic estimate of the future hidden states conditioned to all available data up to time.

4.2.3 Real-time Risk Detection Evaluation. HCML framework highlights that when evaluating computational risk detection models, it is imperative to look into the models with human-centered perspectives to understand whether the models could make accountable and fair decisions [92, 120]. This assessment could be achieved by incorporating qualitative explanations that go beyond the known quantitative performance metrics. In addition, leveraging the human-in-the-loop approach is one of the important standards that should be incorporated in building human-centered algorithmic results [37]. In the following section, we will provide a detailed explanation of the quantitative and qualitative assessments of the models' performance, as derived from the literature we have examined.

Detection Performance: Timeliness and accuracy are associated metrics in real-time risk detection literature. All papers in this review leveraged commonly known accuracy metrics, including F1-score (70%), Accuracy (53%), recall (58%), precision (51%), Area Under Curve (11%), and Root Mean Square Error (RMSE) (4%). As we explained previously, most of the papers reviewed in this paper focused primarily on training the models using predefined fixed chunks of data. As a result, the models' evaluation was also done using these chunks of data. The majority of articles evaluated the timeliness of the models (i.e., accuracy performance over time) based on chunks of a fixed number of posts (53%) and fixed time windows (21%). These chunks were fed to the models in chronological order to measure the models' accuracy performance across fixed chunks of posts or time windows. A few papers (21%) used the detection time to evaluate how timely is the model. These papers mostly built models that learned when to stop using a widely known problem in statistics called the Markov Optimal Stopping problem [134]. Meanwhile, two papers [158, 166] took another direction by leveraging reinforcement learning to identify the optimal number of observations needed to make the detection decision, which is the most promising approach that could provide assessments about the models' performance when it is deployed in real-world applications.

Explainability: Our review revealed that (57%) of the papers presented explanations for the model performance beyond the timeliness and known accuracy measures (Table 10). These papers presented qualitative explanations of real-time risk detection, including qualitative analysis (32%), error analysis (13%), case studies (9%), and human evaluations (2%). Qualitative or error analysis was discussed in papers to further explain their models' performance such as [78, 81, 100]. Due to the goal of implementing the models in social media, the papers have mainly focused on minimizing and discussing the false negatives [78, 81]. For instance, Liu and Guberman et al. [78] found that the false negatives of their hostility forecasting mainly occurred when there was no indication of escalation with many consecutive similar innocuous messages sent. Case studies were leveraged to illustrate and explain how the model performed over time using a case. None of the reviewed papers integrated user studies or human evaluations, with the exception of Masud et al. [91], who surveyed 25 participants to assess their model that altered hateful texts and found that their model outperformed other hate normalization models in terms of generating reduced hateful comments and more fluent sentences. This human evaluation demonstrated that the effective performance of their hate normalization model extended beyond the dataset that was used during training.

4.2.4 Applications and Interventions of Real-Time Social Medial Risk Detection. HCML places a strong emphasis on building systems-based artifacts to foster machine learning transparency that allow

humans to explore machine learning used features and decision results to build trust, making these models less of a "black box" and enhancing their usability and societal impact [144]. The real-world artifacts should be designed to empower users to interact with, question, and comprehend the algorithms' inner workings to elevate stakeholders' oversight such as victims, clinical practitioners, and social media platform owners or moderation teams [120]. The HCML community also promotes the development of risk intervention to ensure that machine learning models are designed to minimize harm and adverse consequences for individuals and society [62]. Below, we identify the artifacts and risk interventions that were presented in the literature.

Applications: The majority of the papers (92%) focused on presenting the algorithmic approaches that enhanced real-time risk detection, yet presenting system artifacts or APIs that could be integrated with social media platforms was rarely done in the literature. Two papers (4%) presented an online-offline detection approach where the model is fully trained offline and the trained model was deployed in a social media platform or server hosting services [36, 157]. Another two papers [91, 167] developed an interface to demonstrate the performance of their models in the real world. For instance, Zou et al. [167] developed a web interface in which users can search for an event, then an alert would appear if the event was likely to be a rumor along with three visualizations that illustrated 1) the event's timeline to show the event evolution along the time deployed, 2) the propagation structure on social media, and 3) user information graph. Only one paper by Rafiq et al. [116] conducted a simulation for their model in Amazon AWS virtual machine instances with 1GB memory to evaluate the scalability of their model by replicating the 100,000 media sessions' traffic up to the scale of 39 million sessions.

Interventions: The majority of the papers we reviewed (89%) focused on the detection algorithms and their performances. However, a few papers (11%) presented an intervention strategy such as alerts, alternating the risk language in posts, or immunizing certain users from receiving risk content. Three papers (6%) presented alerts that were raised when cyberbullying instances were detected [25, 116, 160]. Intuitively, these alerts should be raised after a classifier produced a decision; however, these papers discussed waiting until certain positive decisions have been made to avoid false positives, which was identified as a trade-off between responsiveness and precision. For example, Yao et al. [160] introduced an approach that reviewed Instagram comments as they became available over time and raised an alert only when the total number of comment–level detection decisions topped a certain threshold. Two papers (4%) [113, 155] leveraged the network immunization approach with the goal of minimizing the spread of risk information such as hate speech or rumors. This approach is mainly derived from network science and graph theory to identify these nodes or users effectively. For instance, Petrescu et al. [113] utilized preventive immunization, which worked on the network without knowing the source of risk content and was applied after detecting hateful content, by lowering the rank of that particular post in the feed. As such, we have identified prior efforts to advance the state-of-the-art of real-time risk detection approaches. Next, we will briefly describe the human-centered gaps and recommendations to direct future research to the best practices in this field.

5 DISCUSSION

In this section, we describe the identified gaps along with the recommendations to address the identified gaps and advance the real-time risk detection approaches computationally and from a human-centered perceptive.

5.1 Identifying the Gaps in Real-Time Risk Detection Research from the Human-Centered Perspectives

First, our analysis provided an opportunity to extend Razi et al.'s framework for systematic reviews of computational risk detection literature by adding unique dimensions for human-centered perspectives of real-time risk detection. The new dimensions and codes that emerged included characteristics of the **dataset** (i.e., selection criteria, dataset size, class distribution), **pre-processing and model development** (i.e., data processing, feature computation, input prioritizing), and **evaluation** (i.e., timeliness). This methodological contribution is valuable for future systematic and human-centered reviews of computational risk detection literature that involve real-time approaches. In this section, we describe the gaps in the social media real-time risk detection literature (illustrated in Table 3) and how to address them from a human-centered perspective moving forward.

5.1.1 Datasets Gap: The Absence of Ecologically Valid Datasets. We raise several questions regarding the ecological validity of datasets for real-time online risk detection. The current approach heavily relies on publicly available text datasets scraped from platforms, excluding input from humans, victims, or survivors of these risks at any stage. Depending solely on such data could hinder the effectiveness of real-time online risk detection. Moreover, collecting the data and ground truth annotations from humans, specifically from actual victims ensures that the training risk detection models reflect real-world experiences and accurately represent the users and risks they face online [8]. Further, in section 4.1.1, we described the *time-evolving nature of risks*; relying solely on static public datasets might overlook the nuanced dynamics inherent in how these phenomena unfold over time. We note the need for capturing temporal patterns, such as escalation during cyberbullying, the gradual unfolding of mental health symptoms, or the trust-building stages that have been well-documented for sexual grooming [20], which require longitudinal data for timely and accurate identification. Additionally, in fake news and rumor detection, acknowledging the progress of content spread could be crucial for collecting datasets that represent the dynamic nature of these risks and associated human behaviors. Therefore, we recommend considering data collection methods and advanced systems designed to gather real-time and continuous data streams, or at least robustly simulate interactions that occur over time. This approach should be tailored for specific populations (e.g., risk victims or survivors), the actual contexts of risks, as well as the dynamic aspects of risk escalation and human communication.

5.1.2 Models Gap: The Need for Grounding Models with Human Behaviors. Our analysis revealed that the existing models were grounded on primarily computational efficiency considerations, without considering human understanding or theories. Most papers used a streaming-like data processing approach with data chunks

Table 3: State-of-the-art in real-time risk detection computational approaches and the identified human-centered gaps.

	State-of-the-art Computational Approaches	Gaps from the Human-Centered Lens
Dataset	Utilized large-scale, public datasets with established ground truth.	The absence of ecologically valid datasets that are representative of targeted population and the contexts of their online risk experiences.
Models	Trained models using streaming-like data, textual and social network features, and improved deep learning.	Lack of grounding pre-processing, features, and models with human behaviors.
Evaluation	Evaluated the models' using chunks of data for timeliness and qualitative and error analysis to interpret the models' performances.	Lack of human evaluations of the models' performance.
Applications	Presented novel algorithmic approaches for real-time risk detection.	Lack of artifacts to deploy the models in real-world settings and personalized interventions to intervene after detection.

lacking conversation context, which could lead to the model misinterpreting or missing potential risks. Only 32% of the papers developed features based on human theories and domain-specific knowledge to capture nuanced context. Therefore, we highlight the significance of acknowledging the dialectical nature of human communication and the dynamic changes in behavior within risk contexts when designing features to enhance the effectiveness of real-time risk detection algorithms. This acknowledgment emphasizes the need for designing online features that capture sequential conversational data rather than traditional (i.e., all conversation at once) or chunk-based features [150]. In addition, well-established methodological approaches like discourse analysis [17], which provide a foundation for in-depth exploration of the structural aspects of human communication, could be useful to craft these features by identifying the time-evolving nature of human communication such as shifts in tone, frequency of aggressive language, shifts in mood or self-disclosure, or changes in narrative. Incorporating such approaches into the design of algorithms enables a more nuanced interpretation of online interactions that aligns closely with human understanding. In this review, we also found heavy reliance on the high capability of deep learning models; yet, these models were not inherently human-centered since these models often operated as "black boxes." This can hinder stakeholders, including the victim, from understanding the models' output [143]. Therefore, there is still a need to adopt human theories widely and human-centered real-time risk detection that effectively identifies social media risks.

5.1.3 Evaluation Gap: Involving Humans in the Evaluation Process is Needed. We found most of the papers relied on purely computational metrics (e.g., accuracy and timeliness) without incorporating user studies or human insights into the evaluation of developed risk detection models and identifying the effectiveness of the models' timeliness in protecting people. We even conducted additional searches for subsequent user studies related to the reviewed papers, but we only found one publication by Chang et al. [22] in which Liu's et al [78] hostility forecasting model was embedded into a tool assessed by end users. Their data collection included a survey on participants' experiences with incivility, responses to tool warnings, and overall impressions, alongside real-time recording of drafting behavior via usage log data. They found that the proactive incivility warnings enhanced participants' awareness of their interactions by reflecting more on conversation tension, spending more time drafting comments, and revising replies to mitigate any tension. Similar to this research, future real-time risk detection models could consider incorporating human evaluations to ensure that these models align with human values, ethics, the complexities of online communication, and aligned with evolving risk dynamics, ultimately leading to more effective, trustworthy, and responsible models. However, we recognize the complexities involved in carrying out these evaluation studies concerning ethical considerations, especially those related to algorithmic bias [22]. These issues present difficulties in mitigating potential negative impacts like the reinforcement of stereotypes or the marginalization of vulnerable groups, as noted by Xu et al. [157]. Despite these challenges, in the user study, Chang et al. [22] established a pathway for future researchers to navigate and potentially tackle these ethical and technical concerns in conducting user studies to evaluate risk detection algorithms. Consequently, we emphasize the need for collaborative initiatives to engage in ethical discussions, aiming to identify the best practices for conducting such important user studies [5].

5.1.4 Applications and Interventions Gap: Need for More Real-Time Interventions and Real-World Applications. Most papers focused more on presenting effective detection algorithms without presenting system-based artifacts and interventions using real-time risk detection algorithms. The existence of such systems is a necessary prerequisite for research on real-world algorithm deployment, system design, and user experience resulting from the use of such systems. These studies are important to improve our understanding of how users engage with and react to applications designed for risk detection. In fact, deploying risk detection models in real-world applications has become more of an industrial problem than an inherent expectation in research presenting these detection algorithms [66]. Moreover, the availability of open-source risk detection systems or algorithms is limited, often confined to proprietary platforms or academic papers [5]. Therefore, HCI scholars could bridge this gap by redirecting the fields' attention and resources toward

developing interfaces and interventions. Additionally, fostering interdisciplinary collaboration between experts in ML and HCI fields could lead to the development of such systems and algorithms with the goal of aligning them with user expectations. Building artifacts should be designed to empower users, including victims, clinical practitioners, and social media platform owners, to interact with and understand how these algorithms work. When stakeholders can explore predictions, understand decision factors, and question the algorithm's outputs, they can intervene if needed to align with human values and privacy considerations, improving the algorithm together. For instance, employing personalized interventions could play an important role in offering targeted support based on individuals' preferences, needs, and behaviors, while empowering users with a sense of control and autonomy [26]. As a result, these artifacts would foster real-time risk detection models' transparency, making the algorithms less of a "black box" and enhancing their usability and societal impact.

5.2 Establishing a Research Agenda for Real-Time Risk Detection on Social Media

We make several recommendations for advancing real-time risk detection approaches based on our review. Figure 3 illustrates our conceptualized and comprehensive framework to direct future research with recommendations for the best technical and human-centered practices for real-time risk detection algorithms.

5.2.1 Towards Leveraging Streaming Mechanisms for Ecologically Valid Datasets. We propose that real-time risk detection training and testing could eventually be accomplished using private and multimedia streaming data or online processing of continuous data, instead of relying on predefined chunks of data. Social media environments are characterized by rapidly changing data patterns, influenced by user behaviors, trends, and external events; therefore, continuous data streaming systems could capture this dynamism [61]. Developers of real-time risk detection are encouraged to construct personalized data stream processing systems utilizing open-source software and tools such as Kafka, rather than using commercial or proprietary systems that may not adhere to the users' privacy [58, 60]. Future research in real-time risk detection could also leverage informative reviews on data streaming systems such as [60, 61] that provide insightful information about the usability, features, and real-world use case scenarios for different data streaming systems.

From a human-centered perspective, to ensure that training models reflect the real-world experiences of users, these data streams could be obtained directly from victims or survivors of online risks in human-subject studies with their consent. We acknowledge that data collection from such vulnerable populations is uniquely challenging as it requires researchers to ask them to share and label their intimate online conversations while ensuring that participating in research does not harm them [99, 119]. We suggest that researchers set guidelines beforehand and make sure to follow established recommendations to enhance the ethical implementation of research involving survivors and victims of traumatic online risks, for example by putting in place formal consultation procedures for participants [34]. In addition, scholars in the HCI field have initiated efforts such as MOSafely[5] (Modus Operandi Safely) with the objective of establishing a multidisciplinary collaborative community that concentrates on safeguarding young individuals in online environments. This innovative approach may serve as a potential avenue for resource-sharing, encompassing datasets and algorithms, to effectively address the online security concerns pertaining to at-risk youth. It is advisable for future research endeavors to actively participate in such collaborative initiatives with a strong ethical foundation, prioritizing the protection of the rights and well-being of people.

5.2.2 Toward an Optimized Real-Time Models' Efficiency Grounded in Human Understanding. In this review, most papers leveraged an aggregate view of features over time or time windows. By aggregating the features, the models consider them as independently distributed, meaning that the features calculated for one set of data are unrelated to newly available data, which fails to capture valuable information from adjacent time periods (i.e., evolutionary data) [57]. Due to these reasons, we suggest that an optimal solution for calculating the features could account for the computation overhead. We suggest that the best practice identified to address this issue in this review revolved around calculating the features for a set of data, and then when new data became available, the features of this new data were only calculated while the previously calculated features were reused, which has mainly been implemented through incremental computation [116]. This approach proved to reduce the computation complexity of feature calculation and provide faster classification than classical approaches. Therefore, setting a benchmark for the performance of the incremental computation in features engineering and balancing between the models' computation efficiency and accuracy is a task for future research in real-time to further investigate.

Unlike ML scholars who mainly leveraged the data-driven features for real-time risk detection, socio-psychological researchers often employ survey-based or interview methods to capture contextual information directly from human subjects. However, in both cases, the data may not provide a holistic understanding of human experiences or behaviors that could be helpful for real-time risk detection. Therefore, we call ML and socio-psychological researchers to collaborate on designing a complementary approach to utilize data-driven features that provide objective insights about users' experiences and subjective data collected through well-designed surveys to provide a holistic understanding of human experiences, behaviors, or perspectives of risk. Additionally, prior research has shown that context-based features improve the detection accuracy performance [140], and this is particularly relevant to real-time risk detection on social media where users' interactions could transform within minutes. Therefore, the dynamic updates of user behaviors ensure that the prediction model remains reflective of the shifting patterns of these behaviors [142]. Yet, capturing this time-evolving context in terms of scale becomes challenging and has been identified as a crucial avenue for future development [4]. Therefore, future research in real-time risk detection approaches could further investigate the applicability of incorporating such features, monitoring how this will affect the models' scalability.

[5] https://www.mosafely.org/

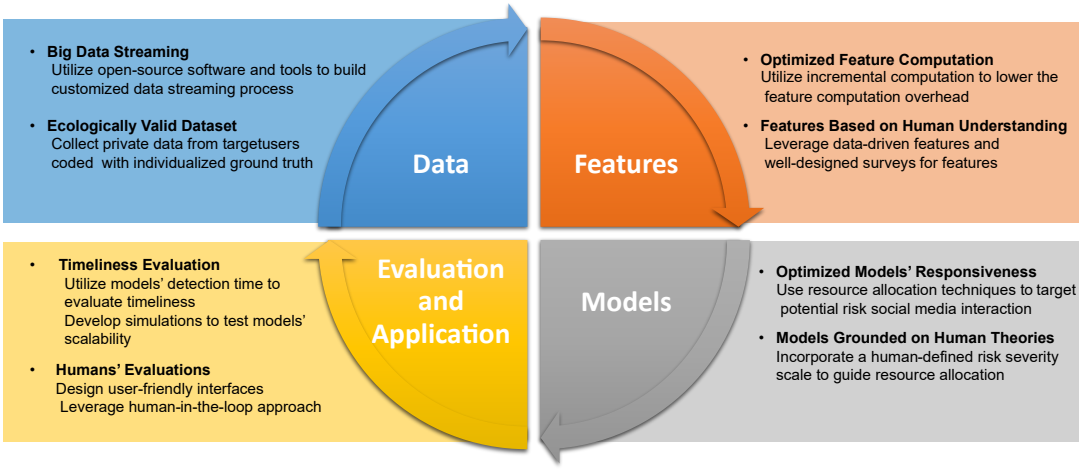

Figure 3: Recommended Computational and Human-Centered Framework For Real-Time Risk Detection Approaches for Social Media

5.2.3 Towards Advancing Real-Time Risk Detection Responsiveness and Interpretability. We identified two papers that adopted a procedure to enforce the risk detection models targeting potentially risky social media interaction (i.e., priority schedulers and machine learning-based ranking) [116, 164]. These papers pave the way for future real-time risk detection algorithms to be more responsive by allocating resources to focus on conversations that are more likely to require immediate intervention. However, we also suggest exploring and using other resource allocation techniques such as adaptive allocation to continuously monitor the workload of the risk detection system and allocate resources based on the volume of conversations and the urgency of risk detection, or predictive allocation to anticipate periods of high-risk activity based on historical patterns to allocate resources during these periods. Reinforcement scheduling [11] leverages reinforcement learning that could be used to learn when potentially risky conversations or at-risk populations would need the detection algorithms. Therefore, future research is encouraged to adopt these proactive allocations of resources to effectively target conversations or threads that potentially contain risky content and ensure faster responses to emerging risks, creating a safer online environment for social media users.

An ultimate approach to ground these recommended resource allocation techniques with human understanding could be by incorporating a human-defined risk severity scale; therefore, more efforts to understand the severity of online risk from the perspectives of users (e.g., [152]) are needed. One approach could be identifying profiles of at-risk individuals that might need more attention from the risk detection algorithms, using unsupervised clustering techniques [139] or by leveraging survey-based data to feed well-established statistical techniques (e.g., Mixture Factor Analysis [96]). Future research on improving the real-time risk detection algorithms is warranted to leverage such human-centered practices when optimizing the models' responsiveness. To achieve this goal, we urge HCI scholars to collaborate with ML and Data Scientists to guide the resource allocation process based on human understanding. In addition, Our findings inform that combining deep learning models with state models can help incorporate domain-specific constraints and handle uncertainty more effectively. Therefore, we recommend future researchers to investigate combining deep learning and statistical models in an ensemble approach, and how it could impact the models' interpretability. Since ensemble methods aggregate predictions from multiple models, they can capture complex patterns while benefiting from the transparent insights of statistical models.

5.2.4 Towards Designing Applications to Incorporate Human Evaluation and Personalized Interventions. In our review, we found that reinforcement learning (RL) such as Q-learning or deep reinforcement learning has the most promising potential to provide information when the detection decision was made instead of relying on pre-defined chunks of data [72, 138]. These techniques advance the detection models to know what level of cues is enough for the model to review the input and provide the detection decision. Besides the detection time, another performance measure should be considered: how well the detection models perform as data volumes increase using a nearly realistic environment. Incorporating scalability simulations into the evaluation process is crucial for ensuring that real-time risk detection models can effectively handle the dynamic nature of social media data streams. During these simulations, it is important to identify potential bottlenecks, limitations, and performance degradation in detection latency and compute times that are essential to enhance the overall responsiveness. Therefore, we recommend that future research on real-time risk detection provide performance metrics that are more useful when deploying the models in real-world settings.

To fill the gap between technical solutions and human expectations, a growing body of work has highlighted the importance of human insights into algorithmic performances to facilitate HCML by informing developers entrusted with designing ethical machine learning algorithms and decision-makers tasked with implementing such systems in social contexts [64]. Given explainability and fairness perceptions are highly context-dependent and can vary

substantially across domains, tasks, and algorithmic designs [141], human involvement in evaluation processes is essential. One way to reflect human perceptions in the evaluation of machine learning systems is through interactive machine learning system design [7] in which human end users are iteratively involved in the model development process. Participatory design strategies allow the users to learn about how the machine-learning model works by instantly testing various inputs and examining the impact of the models [42, 149, 153]. More importantly, these user-led cycles of trial-and-error discovery processes can help developers steer the model to improve model outcomes in ways that satisfy those who are affected by the models the most. Therefore, we call for more collaborative approaches among multi-stakeholders including developers, designers, and users to work together through co-design sessions [1], or even more long-term efforts such as the advisory board of users [95]. This way, we can make sure that real-time risk detection models are working in ways that meet users' expectations and benefit those who are affected by online risks.

Designing user-centric or personalized interventions could involve multiple steps to ensure the effectiveness of these interventions. First, researchers are recommended to gather data to identify the target users and create profiles of users, which include scraping social network data and self-reported to understand individuals' needs, behaviors, and preferences. Scholars have also called for going beyond individual characteristics to explore the effectiveness of contextual characteristics such as culture [121]. Additionally, these interventions should be adaptable to the evolving nature of users' behaviors and needs by continuously monitoring user interactions and feedback to ensure that the support provided remains relevant and engaging. Nudges or gamification could be integrated with these interventions to improve the overall user experience [6, 15]. The design of personalized interventions should possess visual attractiveness, simplicity, and personal relevance in order to resonate with any particular population [102]. Future researchers are encouraged to collaborate with interdisciplinary teams such as psychologists, user experience designers, and data scientists to ensure that intervention designs consider psychological, technical, and ethical dimensions.

5.3 Reconsidering 'Timing' in Real-Time Risk Detection

We found that 94% of the papers operationalized real-time risk detection tasks as an early detection approach that worked on the data retrospectively to detect the risk as early as possible. Social media-rich data have been proved in prior studies to be successfully used to predict the future (i.e., forecasting) across different domains and contexts such as marketing, finance, and sociopolitical events [124]. However, despite our adoption of a comprehensive view of online risks in contrast to prior reviews [66, 120], we observed a similar trend in terms of a conspicuous dearth of preventive methodologies with respect to risk prediction and mitigation. A possible explanation of this rare implementation could fundamentally arise from challenges such as data noise, biases inherent in social media data, limited generalizability, and the inherent difficulty in integrating domain-specific knowledge and theoretical frameworks [114]. In addition to these identified challenges, in our review, we observed that the rapid dissemination of information on social media frequently resulted in temporal intervals that are insufficiently extensive for models to anticipate and proactively address emerging risks before they materialize or escalate. With that being said, we presented three research papers [36, 78, 91] as exemplars that future scholars may consider when seeking to apply and explore the efficacy of their novel preventive methodologies across diverse datasets and various risks.

Another interesting finding in our review was the trade-offs identified within early risk approaches between accuracy and latency in ways that the more data the models use to give accurate predictions, the more time the models take to provide predictions. Trade-offs in ML-based computational systems have been well-documented in the literature, especially between fairness and accuracy, which is a value-sensitive and open question for further discourse [108–110]. We highlight that striking a balance between accuracy and timely detection is indeed an important and challenging aspect of real-time risk detection, especially given that real-time risk detection models are designed to provide "just-in-time" intervention to support those who are (potentially) at risk. Hence, careful consideration of how to balance the two is essential for future work toward designing value-sensitive and effective computational systems to support individuals and society. One way to accomplish this balance might be by defining acceptable trade-off thresholds between accuracy and latency. For example, accepting a certain drop in accuracy if it significantly reduces latency. Thus, the optimal balance between accuracy and latency will vary based on the specific use cases and requirements of the detection task.

In fact, preventive and early approaches aim to safeguard people on social media platforms from potential harm, yet they differ in terms of timing and focus. On one hand, early detection is rooted in real-world data, which could lead to more accurate risk assessments than preventive approaches. Yet, the time required for detection, analysis, and response may result in a delay between risk emergence and effective intervention, reducing its efficacy within a rapidly evolving environment such as social media platforms. On the other hand, the preventive approach utilizes the predictive indicators to take action "before" the risk incident occurs or the victim suffers from the risk. However, applicability concerns have been posed about this approach as explained previously that may lead to unnecessary content removal or user restrictions if these models were not trained very well [24]. As such, each approach (i.e., preventive and early risk detection) has pros and cons that are warranted to be balanced in future research. We suggest that ultimately, a combination of both strategies along with late risk mitigation, tailored to the specific context and nature of risks, can be the most effective way forward in building a safe online landscape, as illustrated in Figure 4. This means that preventive, early, and late risk mitigation strategies could be developed hand in hand to provide a comprehensive risk detection approach that detects the risk as early as possible in case the predictive indicators fail to forecast and mitigate risks beforehand. Late risk mitigation could serve as an analysis stage of the risks that were missed by the preventive and early approaches or the risks' long-term impact (e.g., cyberbullying and following mental health indicators). Adopting this approach forms a full cycle of real-time risk detection algorithms to effectively ensure individuals' safety.

Figure 4: Comprehensive real-time risk detection approaches.

5.4 Limitations and Future Research

There are several limitations of our review that are worth mentioning. First, while our review was comprehensive, it is possible that we did not include all published work that met our inclusion criteria. Additionally, we limited our inclusion criteria to papers that developed a novel computational real-time risk detection model for social media-related risks. Considering the computational complexities involved in developing and assessing these algorithms, it is probable that the human-centered evaluations of these systems were left for subsequent work, although we did not find many in this vein. Consequently, we strongly encourage more research focused on HCI aspects of real-time risk detection on social media, including intervention-based approaches, interface design, user experiments, and real-world system deployment. Further, there may have been some papers that met our inclusion criteria that were held out-of-scope because it was difficult for us to evaluate relevancy due to inconsistent reporting standards. Therefore, we urge the HCI and ML research communities to converge on local norms for reporting important metrics uniformly across fields to increase the communities' ability to synthesize the results in a way that moves the fields cohesively forward. Furthermore, this review primarily concentrates on peer-reviewed research, yet it is worth noting that numerous social media companies are independently developing proprietary algorithms for real-time risk detection [46]. To advance the field more effectively, fostering collaboration between academic and industry researchers could prove to be highly advantageous. Finally, all human-centered research is nuanced, complicated, and context-dependent. As such, insights regarding specific risk types may not be directly applicable to other risks, especially when comparing interpersonal risks, such as cyberbullying or mental health to community-level risks, such as fake news. Therefore, future researchers should use their discretion, as well as domain experts' opinions, as to what recommendations make sense in the context of their work.

6 CONCLUSION

In an increasingly digitalized society, individuals face growing complexity due to the diverse range of social media risks, impacting both individuals and society as a whole. Detecting these risks accurately and timely has become a pressing necessity to facilitate effective interventions for various stakeholders, including governments, online platforms, societies, and academic communities. While previous studies have made great progress in advancing real-time risk detection approaches for social media, our review revealed a lack of integration with human understanding and behaviors in these approaches. Therefore, we strongly recommend that future research prioritize placing humans at the center of designing, developing, and testing real-time risk detection systems to ensure their effectiveness in real-world settings. As our review highlights, as HCI researchers, it is imperative for us to join forces with ML developers and researchers to bridge the gap between theoretical socio-psychological knowledge and the hands-on implementation of computational solutions for real-time risk.

ACKNOWLEDGMENTS

This research is supported in part by the U.S. National Science Foundation under grants #IIP-2329976, #IIS-2333207, #CNS-1942610 and by the William T. Grant Foundation grant #187941. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the research sponsors. We would also like to thank all the participants who donated their data and contributed to the research.

REFERENCES

[1] Zainab Agha, Zinan Zhang, Oluwatomisin Obajemu, Luke Shirley, and Pamela J. Wisniewski. 2022. A case study on user experience bootcamps with teens to co-design real-time online safety interventions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, USA, 1–8.

[2] Yunis Ali Ahmed, Mohammad Nazir Ahmad, Norasnita Ahmad, and Nor Hidayati Zakaria. 2019. Social media for knowledge-sharing: A systematic literature review. *Telematics and informatics* 37 (2019), 72–112.

[3] Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, 1311–1323.

[4] Mohammed Ali Al-Garadi, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. 2019. Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access* 7 (2019), 70701–70718.

[5] Ashwaq Alsoubai, Xavier V Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J Wisniewski. 2022. 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*. ACM, USA, 197–200.

[6] Ashwaq Alsoubai, AFSANEH RAZI, ZAINAB AGHA, SHIZA ALI, GIANLUCA STRINGHINI, MUNMUN DE CHODHURY, and PAMELA J WISNIEWSKI. 2024.

Profiling the Offline and Online Risk Experiences of Youth to Develop Targeted Interventions for Online Safety. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW2 (2024), 36 pages.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. ACM, USA, 1–13.

[8] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-centered data science: an introduction*. MIT Press, USA.

[9] Muhammad Arif. 2021. A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges. *Journal of Information Security and Cybercrimes Research* 4, 1 (2021), 01–26.

[10] Brent Barnhart. 2022. 41 of the most important social media marketing statistics for 2022. https://sproutsocial.com/insights/social-media-statistics/

[11] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Candace Thille, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, and John C Mitchell. 2020. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, USA, 1–12.

[12] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (2017), 2053951717718854.

[13] Gema Bello-Orgaz, Jason J Jung, and David Camacho. 2016. Social big data: Recent achievements and new challenges. *Information Fusion* 28 (2016), 45–59.

[14] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (2019), 421–422.

[15] Sarah C Boyle and Joseph W LaBrie. 2021. A Gamified, Social Media–Inspired, Web-Based Personalized Normative Feedback Alcohol Intervention for Lesbian, Bisexual, and Queer-Identified Women: Protocol for a Hybrid Trial. *JMIR Research Protocols* 10, 4 (2021), e24647.

[16] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association, USA.

[17] Gillian Brown and George Yule. 1983. *Discourse analysis*. Cambridge university press, UK.

[18] Stefan D Bruda and Selim G Akl. 2003. Real-time computation: A formal definition and its applications. *International Journal of Computers and Applications* 25, 4 (2003), 247–257.

[19] Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications* 133 (2019), 182–197.

[20] Amparo Elizabeth Cano, Miriam Fernandez, and Harith Alani. 2014. Detecting child grooming behaviour patterns on social media. In *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings 6*. Springer, USA, 412–427.

[21] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32.

[22] Jonathan P Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. 2022. Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–37.

[23] Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)* 13, 3 (2019), 1–51.

[24] Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941* 1 (2021), 12 pages.

[25] Charalampos Chelmis and Daphney-Stavroula Zois. 2021. Dynamic, incremental, and continuous detection of cyberbullying in online social media. *ACM Transactions on the Web (TWEB)* 15, 3 (2021), 1–33.

[26] Jie Chen, C Daniel Mullins, Priscilla Novak, and Stephen B Thomas. 2016. Personalized strategies to activate and empower patients in health care and reduce health disparities. *Health Education & Behavior* 43, 1 (2016), 25–34.

[27] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers 22*. Springer, USA, 40–52.

[28] Lu Cheng, Ruocheng Guo, Yasin N Silva, Deborah Hall, and Huan Liu. 2021. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Transactions on Data Science* 2, 2 (2021), 1–23.

[29] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the twelfth acm international conference on web search and data mining*. ACM, USA, 339–347.

[30] Emilie Chouzenoux and Jean-Christophe Pesquet. 2017. A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation. *IEEE Transactions on Signal Processing* 65, 18 (2017), 4770–4783.

[31] Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.

[32] Congress. 2023. https://www.congress.gov/bill/118th-congress/senate-bill/1409/text

[33] Congress. 2023. https://www.congress.gov/bill/115th-congress/house-bill/1865

[34] Lisa D Cromer and Elana Newman. 2011. Research ethics in victimization studies: Widening the lens. *Violence against women* 17, 12 (2011), 1536–1548.

[35] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, USA, 9268–9277.

[36] Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would your tweet invoke hate on the fly? forecasting hate intensity of reply threads on twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, USA, 2732–2742.

[37] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management* 60, 2 (2023), 103219. https://doi.org/10.1016/j.ipm.2022.103219

[38] Prasannakumaran Dhanasekaran, Harish Srinivasan, S Sowmiya Sree, I Sri Gayathri Devi, Saikrishnan Sankar, and Vineeth Vijayaraghavan. 2021. SOMPS-Net: Attention Based Social Graph Framework for Early Detection of Fake Health News. In *Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings*. Springer, USA, 165–179.

[39] Rajendra T Dodhiawala, NS Sridharan, Peter Raulefs, and Cynthia Pickering. 1989. Real-Time AI Systems: A Definition and An Architecture.. In *IJCAI*. Citeseer, USA, 256–264.

[40] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint* 1702 (2017), 13 pages.

[41] Suman Dowlagar and Radhika Mamidi. 2021. A survey of recent neural network models on code-mixed indian hate speech data. In *Forum for Information Retrieval Evaluation*. ACM, USA, 67–74.

[42] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. ACM, USA, 211–223.

[43] V Elliott, N Christopher, A Deck, and L Schwartz. 2021. The Facebook papers reveal staggering failures in the global South. Rest of world.

[44] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 12. AAAI, USA, 10 pages.

[45] Hugo Jair Escalante, Esaú Villatoro-Tello, Sara E Garza, A Pastor López-Monroy, Manuel Montes-y Gómez, and Luis Villaseñor-Pineda. 2017. Early detection of deception and aggressiveness using profile-based representations. *Expert Systems with Applications* 89 (2017), 99–111.

[46] Facebook. 2023. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/

[47] Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*. ACM, USA, 105–114.

[48] Bharath Ganesh and Jonathan Bright. 2020. Countering extremists on social media: Challenges for strategic communication and content moderation. , 6–19 pages.

[49] Suyu Ge, Lu Cheng, and Huan Liu. 2021. Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*. ACM, USA, 496–506.

[50] Gary W Giumetti and Robin M Kowalski. 2022. Cyberbullying via social media and well-being. *Current Opinion in Psychology* 45 (2022), 101314.

[51] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[52] Matthew A Hammer, Jana Dunfield, Kyle Headley, Nicholas Labich, Jeffrey S Foster, Michael Hicks, and David Van Horn. 2015. Incremental computation with names. In *Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*. ACM, USA, 748–766.

[53] Matthew A Hammer, Khoo Yit Phang, Michael Hicks, and Jeffrey S Foster. 2014. Adapton: Composable, demand-driven incremental computation. *ACM SIGPLAN Notices* 49, 6 (2014), 156–166.

[54] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, USA, 501–512.

[55] Kris Hartley and Minh Khuong Vu. 2020. Fighting fake news in the COVID-19 era: policy insights from an equilibrium model. *Policy sciences* 53, 4 (2020), 735–758.

[56] Rex Hartson and Pardha S Pyla. 2018. *The UX book: Agile UX design for a quality user experience.* Morgan Kaufmann, USA.

[57] Yu He, Jianxin Li, Yangqiu Song, Mutian He, Hao Peng, et al. 2018. Time-evolving Text Classification with Deep Neural Networks.. In *IJCAI*, Vol. 18. IJCAI, USA, 2241–2247.

[58] Bhole Rahul Hiraman et al. 2018. A study of apache kafka in big data stream processing. In *2018 International Conference on Information, Communication, Engineering and Technology (ICICET)*. IEEE, USA, 1–3.

[59] Hao Huang, LiHua Zhou, and YiTing Jiang. 2021. Early detection of fake news based on multiple information features. In *2021 4th International Conference on Data Science and Information Technology*. ACM, USA, 414–419.

[60] Haruna Isah, Tariq Abughofa, Sazia Mahfuz, Dharmitha Ajerla, Farhana Zulkernine, and Shahzad Khan. 2019. A survey of distributed data stream processing frameworks. *IEEE Access* 7 (2019), 154300–154316.

[61] Supun Kamburugamuve, Geoffrey Fox, David Leake, and Judy Qiu. 2013. Survey of distributed stream processing for large stream sources. *Grids Ucs Indiana Edu* 2 (2013), 1–16.

[62] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and PM Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, USA, 45–55.

[63] Rebecca Kern. 2022. Push to rein in social media sweeps the states. https://www.politico.com/news/2022/07/01/social-media-sweeps-the-states-00043229

[64] Kimon Kieslich, Birte Keller, and Christopher Starke. 2022. Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society* 9, 1 (2022), 20539517221092956.

[65] Alex Kim and Sangwon Yoon. 2022. Detecting Rumor Veracity with Only Textual Information by Double-Channel Structure. *10th International Workshop on SocialNLP* 1 (2022), 10 pages.

[66] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A human-centered systematic literature review of cyberbullying detection algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

[67] Thomas J Laffey, Preston A Cox, James L Schmidt, Simon M Kao, and Jackson Y Readk. 1988. Real-time knowledge-based systems. *AI magazine* 9, 1 (1988), 27–27.

[68] Victor Leiva and Ana Freire. 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*. Springer, USA, 428–436.

[69] Ke Li, Bin Guo, Jiaqi Liu, Jiangtao Wang, Haoyang Ren, Fei Yi, and Zhiwen Yu. 2022. Dynamic probabilistic graphical model for progressive fake news detection on social media platform. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 5 (2022), 1–24.

[70] Ke Li, Bin Guo, Siyuan Ren, and Zhiwen Yu. 2022. AdaDebunk: An Efficient and Reliable Deep State Space Model for Adaptive Fake News Early Detection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM, USA, 1156–1165.

[71] Xiao-Li Li and Bing Liu. 2005. Learning from positive and unlabeled examples with different data distributions. In *Machine Learning: ECML 2005: 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005. Proceedings 16*. Springer, USA, 218–229.

[72] Yuxi Li. 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274* 1 (2017), 58 pages.

[73] Ken-Yu Lin, Roy Ka-Wei Lee, Wei Gao, and Wen-Chih Peng. 2021. Early prediction of hate speech propagation. In *2021 International Conference on Data Mining Workshops (ICDMW)*. IEEE, USA, 967–974.

[74] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. IEEE, USA, 2980–2988.

[75] Paul S Links, Rahel Eynan, Marnin J Heisel, and Rosane Nisenbaum. 2008. Elements of affective instability associated with suicidal behaviour in patients with borderline personality disorder. *The Canadian Journal of Psychiatry* 53, 2 (2008), 112–116.

[76] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[77] Iouliana Litou, Vana Kalogeraki, Ioannis Katakis, and Dimitrios Gunopulos. 2016. Real-time and cost-effective limitation of misinformation propagation. In *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, Vol. 1. IEEE, USA, 158–163.

[78] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In *Twelfth international aaai conference on web and social media*. AAAI, USA, 10 pages.

[79] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM, USA, 1867–1870.

[80] Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. AAAI, USA, 8 pages.

[81] Yang Liu and Yi-Fang Brook Wu. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–33.

[82] Manuel F López-Vizcaíno, Francisco J Nóvoa, Victor Carneiro, and Fidel Cacheda. 2021. Early detection of cyberbullying on social media networks. *Future Generation Computer Systems* 118 (2021), 219–229.

[83] David E Losada, Fabio Crestani, and Javier Parapar. 2017. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*. Springer, USA, 346–360.

[84] David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk: early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France*, Vol. 9. Springer, USA, 343–361 pages.

[85] David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk at CLEF 2019: Early Risk Prediction on the Internet (extended overview). *CLEF (Working Notes)* 4 (2019), 21 pages.

[86] David E Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)* 1 (2020), 14 pages.

[87] Menglong Lu, Zhen Huang, Binyang Li, Yunxiang Zhao, Zheng Qin, and Dongsheng Li. 2022. Sifter: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 429–442.

[88] Zhengliang Luo, Tiening Sun, Xiaoxu Zhu, Zhong Qian, and Peifeng Li. 2021. Early Rumor Detection with Prior Information on Social Media. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part V 28*. Springer, USA, 282–289.

[89] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM, USA, 1751–1754.

[90] Anshu Malhotra and Rajni Jindal. 2022. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing* 1 (2022), 109713.

[91] Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, USA, 3524–3534.

[92] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[93] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*. ACM, USA, 1148–1158.

[94] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. *Advances in neural information processing systems* 27 (2014), 9 pages.

[95] Megan A Moreno, Anna Jolliff, and Brad Kerr. 2021. Youth advisory boards: Perspectives and processes. *Journal of Adolescent Health* 69, 2 (2021), 192–194.

[96] Bengt Muthén and Bengt O Muthén. 2009. *Statistical analysis with latent variables*. Wiley New York, NY, USA.

[97] Christof Naumzik and Stefan Feuerriegel. 2022. Detecting false rumors from retweet dynamics on social media. In *Proceedings of the ACM Web Conference 2022*. ACM, USA, 2798–2809.

[98] Imara Nazar, Daphney-Stavroula Zois, and Mengfan Yao. 2019. A hierarchical approach for timely cyberbullying detection. In *2019 IEEE Data Science Workshop (DSW)*. IEEE, USA, 190–195.

[99] Elana Newman, Elizabeth Risch, and Nancy Kassam-Adams. 2006. Ethical issues in trauma-related research: A review. *Journal of Empirical Research on Human Research Ethics* 1, 3 (2006), 29–46.

[100] Shirin Nilizadeh, François Labrèche, Alireza Sedighian, Ali Zand, José Fernandez, Christopher Kruegel, Gianluca Stringhini, and Giovanni Vigna. 2017. Poised: Spotting twitter spam off the beaten paths. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, USA, 1159–1174.

[101] Robert Nishihara, Philipp Moritz, Stephanie Wang, Alexey Tumanov, William Paul, Johann Schleier-Smith, Richard Liaw, Mehrdad Niknami, Michael I Jordan,

and Ion Stoica. 2017. Real-time machine learning: The missing pieces. In *Proceedings of the 16th Workshop on Hot Topics in Operating Systems*. ACM, USA, 106–110.

[102] Monica Marina Nour, Anika Saiyara Rouf, and Margaret Allman-Farinelli. 2018. Exploring young adult perspectives on the use of gamification and social media in a smartphone platform for improving vegetable intake. *Appetite* 120 (2018), 547–556.

[103] Fayika Farhat Nova, MD Rashidujjaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. 2019. Online sexual harassment over anonymous social media in Bangladesh. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*. ACM, USA, 1–12.

[104] Cindy A O'Reilly and Andrew S Cromarty. 1985. " Fast" Is Not" Real-Time": Designing Effective Real-Time AI Systems. In *Applications of Artificial Intelligence II*, Vol. 548. SPIE, USA, 249–257.

[105] JE Palmier-Claus, Peter J Taylor, F Varese, and D Pratt. 2012. Does unstable mood increase risk of suicide? Theory, research and practice. *Journal of affective disorders* 143, 1-3 (2012), 5–15.

[106] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)* 1 (2021), 864–887.

[107] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 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*, Vol. 13390. Springer, USA, 233–256.

[108] Jinkyung Park, Ramanathan Arunachalam, Vincent Silenzio, Vivek K Singh, et al. 2022. Fairness in Mobile Phone–Based Mental Health Assessment Algorithms: Exploratory Study. *JMIR formative research* 6, 6 (2022), e34366.

[109] Jinkyung Park, Rahul Ellezhuthil, Ramanathan Arunachalam, Lauren Feldman, and Vivek Singh. 2022. Toward Fairness in Misinformation Detection Algorithms. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, Vol. 16. AAAI, USA, 11 pages.

[110] Jinkyung Park, Rahul Dev Ellezhuthil, Joseph Isaac, Christoph Mergerson, Lauren Feldman, and Vivek Singh. 2023. Misinformation Detection Algorithms and Fairness across Political Ideologies: The Impact of Article Level Labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM, USA, 107–116.

[111] Jinkyung Park, Joshua Gracie, Ashwaq Alsoubai, Gianluca Stringhini, Vivek Singh, and Pamela Wisniewski. 2023. Towards Automated Detection of Risky Images Shared by Youth on Social Media. In *Companion Proceedings of the ACM Web Conference 2023*. ACM, USA, 1348–1357.

[112] Jinkyung Park, Irina Lediaeva, Maria Lopez, Amy Godfrey, Kapil Chalil Madathil, Heidi Zinzow, and Pamela Wisniewski. 2023. How affordances and social norms shape the discussion of harmful social media challenges on reddit. *Human Factors in Healthcare* 3 (2023), 100042.

[113] Alexandru Petrescu, Ciprian-Octavian Truică, Elena-Simona Apostol, and Panagiotis Karras. 2021. Sparse Shield: Social Network Immunization vs. Harmful Speech. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM, USA, 1426–1436.

[114] Lawrence Phillips, Chase Dowling, Kyle Shaffer, Nathan Hodas, and Svitlana Volkova. 2017. Using social media to predict the future: a systematic literature review. *arXiv preprint arXiv:1706.06134* 1 (2017), 55 pages.

[115] Nektaria Potha and Manolis Maragoudakis. 2014. Cyberbullying detection using time series modeling. In *2014 IEEE International Conference on Data Mining Workshop*. IEEE, USA, 373–382.

[116] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. 2018. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. ACM, USA, 1738–1747.

[117] Diana Ramírez-Cifuentes, Marc Mayans, and Ana Freire. 2018. Early risk detection of anorexia on social media. In *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*. Springer, USA, 3–14.

[118] Shubhangi Rastogi and Divya Bansal. 2022. A review on fake news detection 3T's: typology, time of detection, taxonomies. *International Journal of Information Security* 22 (2022), 1–36.

[119] Afsaneh Razi, Ashwaq AlSoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J Wisniewski. 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*. ACM, USA, 1–9.

[120] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J Wisniewski. 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 (2021), 1–38.

[121] Luiz Rodrigues, Armando M Toda, Paula T Palomino, Wilk Oliveira, and Seiji Isotani. 2020. Personalized gamification: A literature review of outcomes, experiments, and approaches. In *Eighth international conference on technological ecosystems for enhancing multiculturality*. ACM, USA, 699–706.

[122] David Rogers, Alun Preece, Martin Innes, and Irena Spasić. 2022. Real-time text classification of user-generated content on social media: Systematic review. *IEEE Transactions on Computational Social Systems* 9, 4 (2022), 1154–1166.

[123] Nir Rosenfeld, Aron Szanto, and David C Parkes. 2020. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*. ACM, USA, 1018–1028.

[124] Dimitrios Rousidis, Paraskevas Koukaras, and Christos Tjortjis. 2020. Social media prediction: a literature review. *Multimedia Tools and Applications* 79, 9-10 (2020), 6279–6311.

[125] S Santhoshkumar and LD Dhinesh Babu. 2020. Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. *Social network analysis and mining* 10 (2020), 1–17.

[126] Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022. Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. ACM, USA, 1716–1727.

[127] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, USA, 7685–7697.

[128] Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies*. Association for Computational Linguistics, USA, 2176–2190.

[129] Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the US child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, USA, 1–15.

[130] Rikki Schlott. 2023. China is hurting our kids with TikTok but protecting its own youth with Douyin. https://nypost.com/2023/02/25/china-is-hurting-us-kids-with-tiktok-but-protecting-its-own/

[131] Victor P Seidel, Timothy R Hannigan, and Nelson Phillips. 2020. Rumor communities, social media, and forthcoming innovations: The shaping of technological frames in product market evolution. *Academy of Management Review* 45, 2 (2020), 304–324.

[132] Naeem Seliya, Taghi M Khoshgoftaar, and Jason Van Hulse. 2009. A study on the relationships of classifier performance metrics. In *2009 21st IEEE international conference on tools with artificial intelligence*. IEEE, USA, 59–66.

[133] Kang G Shin and Parameswaran Ramanathan. 1994. Real-time computing: A new discipline of computer science and engineering. *Proc. IEEE* 82, 1 (1994), 6–24.

[134] Albert N Shiryaev. 2007. *Optimal stopping rules*. Vol. 8. Springer Science & Business Media, USA.

[135] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[136] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[137] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 58, 5 (2021), 102618.

[138] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[139] Kristina P Sinaga and Miin-Shen Yang. 2020. Unsupervised K-means clustering algorithm. *IEEE access* 8 (2020), 80716–80727.

[140] Sara Owsley Sood, Judd Antin, and Elizabeth Churchill. 2012. Using crowdsourcing to improve profanity detection. In *2012 AAAI Spring Symposium Series*. AAAI, USA, 6 pages.

[141] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9, 2 (2022), 20539517221115189.

[142] VS Subrahmanian and Srijan Kumar. 2017. Predicting human behavior: The next frontiers. *Science* 355, 6324 (2017), 489–489.

[143] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, USA, 1–16.

[144] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 2019. A review of the gaps and opportunities of nudity and skin detection algorithmic research for the purpose of combating adolescent sexting behaviors. In *Human-Computer Interaction. Design Practice in Contemporary Societies: Thematic Area, HCI 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part III 21*. Springer, USA, 90–108.

[145] Nicholas Tarrier, Patricia Gooding, Lynsey Gregg, Judith Johnson, Richard Drake, Socrates Trial Group, et al. 2007. Suicide schema in schizophrenia: The effect of emotional reactivity, negative symptoms and schema elaboration. *Behaviour research and therapy* 45, 9 (2007), 2090–2097.

[146] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[147] Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. 2017. Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM transactions on knowledge discovery from data (TKDD)* 11, 4 (2017), 1–36.

[148] Piyush Vyas, Gitika Vyas, Akhilesh Chauhan, Romil Rawat, Shrikant Telang, and Madhu Gottumukkala. 2022. Anonymous Trading on the Dark Online Marketplace: An Exploratory Study. In *Using Computational Intelligence for the Dark Web and Illicit Behavior Detection*. IGI Global, USA, 272–289.

[149] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM, USA, 1–15.

[150] Jialei Wang, Peilin Zhao, Steven CH Hoi, and Rong Jin. 2013. Online feature selection and its applications. *IEEE Transactions on knowledge and data engineering* 26, 3 (2013), 698–710.

[151] Honghao Wei, Xiaohan Kang, Weina Wang, and Lei Ying. 2019. QuickStop: A Markov optimal stopping approach for quickest misinformation detection. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 2 (2019), 1–25.

[152] Pamela Wisniewski, Heng Xu, Mary Beth Rosson, Daniel F Perkins, and John M Carroll. 2016. Dear diary: Teens reflect on their weekly online risk experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, USA, 3919–3930.

[153] Christine T Wolf. 2019. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, USA, 252–257.

[154] Liang Wu, Jundong Li, Xia Hu, and Huan Liu. 2017. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, USA, 99–107.

[155] Liang Wu and Huan Liu. 2019. Debunking rumors in social networks: A timely approach. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, USA, 323–331.

[156] Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A state-independent and time-evolving network for early rumor detection in social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, USA, 9042–9051.

[157] Teng Xu, Gerard Goossen, Huseyin Kerem Cevahir, Sara Khodeir, Yingyezhe Jin, Frank Li, Shawn Shan, Sagar Patel, David Freeman, and Paul Pearce. 2021. Deep entity classification: Abusive account detection for online social networks. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*. USENIX, USA, 18 pages.

[158] Xiaofei Xu, Ke Deng, and Xiuzhen Zhang. 2022. Identifying cost-effective debunkers for multi-stage fake news mitigation campaigns. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. ACM, USA, 1206–1214.

[159] Jia Xue, Bolun Zhang, Qiaoru Zhang, Ran Hu, Jielin Jiang, Nian Liu, Yingdong Peng, Ziqian Li, and Judith Logan. 2023. Using Twitter-Based Data for Sexual Violence Research: Scoping Review. *Journal of Medical Internet Research* 25 (2023), e46084.

[160] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*. ACM, USA, 3427–3433.

[161] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*. ACM, USA, 1191–1198.

[162] Peiling Yi and Arkaitz Zubiaga. 2023. Learning like human annotators: Cyberbullying detection in lengthy social media sessions. In *Proceedings of the ACM Web Conference 2023*. ACM, USA, 4095–4103.

[163] Peiling Yi and Arkaitz Zubiaga. 2023. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media* 36 (2023), 100250.

[164] Jianwei Zhang, Jinto Yamanaka, and Lin Li. 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*. ACM, USA, 282–289.

[165] Tong Zhao, Bo Ni, Wenhao Yu, Zhichun Guo, Neil Shah, and Meng Jiang. 2021. Action sequence augmentation for early graph-based anomaly detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM, USA, 2668–2678.

[166] Kaimin Zhou, Chang Shu, Binyang Li, and Jey Han Lau. 2019. Early rumour detection. 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)*. Association for Computational Linguistics, USA, 1614–1623.

[167] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. 2015. Real-time news certification system on sina weibo. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, USA, 983–988.

[168] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. DepressionNet: Learning Multi-modalities with User Post Summarization for Depression Detection on Social Media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Virtual Event, Canada,) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/3404835.3462938

Table 4: Social Media Platforms.

Social Media Platforms	Counts (Percent)	References
Twitter	36 (68%)	[19, 27, 36, 38, 45, 69, 70, 73, 77, 79–81, 87–89, 91, 93, 97, 100, 113, 123, 125–128, 137, 147, 154–156, 158, 161, 164, 166, 168]
Weibo	14 (26%)	[27, 59, 69, 70, 80, 81, 87–89, 151, 156, 165–167]
Instagram	8 (15%)	[25, 28, 29, 49, 78, 98, 160, 162]
Vine	5 (9%)	[29, 49, 82, 116, 162]
Reddit	5 (9%)	[65, 68, 91, 117, 165]
Facebook	2 (4%)	[91, 157]

Table 5: Ground Truth Annotations

Annotations	Counts (Percent)	References
Existing	38 (72%)	[19, 25, 27–29, 38, 45, 49, 59, 65, 68, 70, 73, 77, 80, 81, 87–89, 91, 93, 97, 98, 116, 117, 123, 125, 137, 151, 155, 156, 158, 160, 162, 164, 166–168]
Outsiders	15 (28%)	[36, 69, 78, 79, 82, 91, 100, 126–128, 147, 154, 157, 161, 165]
Auto	5 (9%)	[79, 100, 113, 154, 157]

Table 6: Class Distribution

Class Distribution	Counts (Percent)	References
Unbalanced	31 (58%)	[19, 25, 28, 29, 38, 45, 49, 65, 68, 73, 78, 81, 82, 87, 88, 93, 98, 113, 117, 123, 125–128, 137, 154, 155, 157, 162, 165, 168]
Balanced	22 (42%)	[27, 36, 59, 69, 70, 77, 79, 80, 89, 91, 97, 100, 116, 147, 151, 156, 158, 160, 161, 164, 166, 167]

Table 7: Dataset Processing

Dataset Type	Counts (Percent)	References
Chunks of data	44 (83%)	[19, 27, 29, 36, 38, 45, 49, 59, 65, 69, 70, 73, 78–82, 88, 89, 91, 93, 97, 100, 113, 116, 117, 123, 125–128, 137, 147, 151, 154–158, 161, 162, 164–168]
Dynamical	9 (17%)	[19, 25, 28, 68, 77, 79, 87, 98, 160]

Table 8: Features

Features	Counts (Percent)	Types	References
ML-Based	53 (100%)	Textual (66%)	[19, 25, 27–29, 45, 49, 59, 65, 68–70, 73, 78, 80, 81, 87, 88, 91, 98, 100, 117, 126–128, 147, 156, 160–162, 164–168]
		Network (51%)	[29, 38, 49, 59, 73, 77, 79, 80, 82, 89, 93, 97, 100, 113, 123, 125, 126, 128, 137, 147, 151, 155, 158, 164, 165, 167, 168]
		User (30%)	[29, 38, 45, 49, 59, 70, 80–82, 88, 89, 97, 116, 125, 157, 167]
		Temporal (21%)	[27–29, 73, 80, 97, 125, 128, 137, 155, 156]
		Sentiment (19%)	[25, 68, 69, 82, 89, 91, 116, 164, 167, 168]
Domain-Specific	17 (32%)		[36, 65, 70, 77–79, 87, 89, 116, 117, 151, 154, 157, 160, 161, 164, 168]

Table 9: Machine Learning Models

Approach	Counts (Percent)	Model	References
Statistical	21 (40%)	Bayes	[19, 69, 70, 82, 97, 98, 100, 155, 160]
		Markov Models	[25, 36, 68, 69, 97, 147, 151, 156]
		Hawkes process	[97, 128, 158]
Deep Learning	31 (60%)	LSTM	[27, 38, 70, 73, 91, 113, 126, 127, 147, 158]
		Graph Neural Network	[77, 79, 93, 123, 128, 137, 154, 165]
		Transformers-Based	[65, 88, 91, 113, 168]
		CNN	[59, 81, 125]
		Neural Network	[45, 87, 167]
		Gated recurrent units	[80, 165]

Table 10: Models' Explainable Approaches

Models' Explanibility	Counts (Percent)	References
Qualitative analysis	16 (32%)	[19, 28, 29, 78, 81, 87, 97, 98, 123, 126, 127, 137, 147, 162, 166, 168]
Error analysis	7 (13%)	[19, 91, 100, 126, 128, 151, 154]
Case study	5 (9%)	[49, 69, 70, 73, 165]
Human evaluation	1 (2%)	[91]