2022

A Risk-Averse Mechanism for Suicidality Assessment on Social Media

Ramit Sawhney
University of South Carolina - Columbia

Atula Tejaswi Neerkaje
Manipal Institute of Technology, Manipal University

Manas Gaur
University of South Carolina - Columbia

Publication Info

Preprint version *Association for Computational Linguistics 2022 (ACL 2022)*, 2022.

A Risk-Averse Mechanism for Suicidality Assessment on Social Media

Ramit Sawhney[1]*, **Atula Tejaswi Neerkaje**[2]*, **Manas Gaur**[1]
[1]AI Institute, University of South Carolina, SC, USA
mgaur@email.sc.edu
[2]Manipal Institute of Technology, Manipal, India
atula.neerkaje@learner.manipal.edu

Abstract

Recent studies have shown that social media has increasingly become a platform for users to express suicidal thoughts outside traditional clinical settings. With advances in Natural Language Processing strategies, it is now possible to design automated systems to assess suicide risk. However, such systems may generate uncertain predictions, leading to severe consequences. We hence reformulate suicide risk assessment as a selective prioritized prediction problem over the Columbia Suicide Severity Risk Scale (C-SSRS). We propose SASI, a risk-averse and self-aware transformer-based hierarchical attention classifier, augmented to refrain from making uncertain predictions. We show that SASI is able to refrain from 83% of incorrect predictions on real-world Reddit data. Furthermore, we discuss the qualitative, practical, and ethical aspects of SASI for suicide risk assessment as a human-in-the-loop framework.

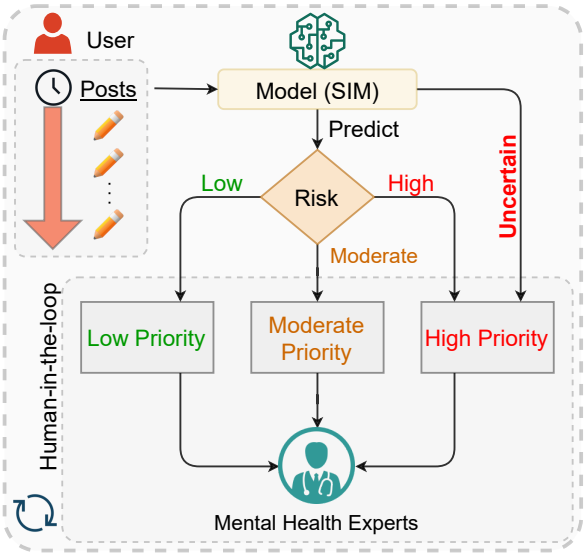

Figure 1: End-to-end pipeline for suicide risk assessment. When SASI assesses the posts, it returns the predicted risk level along with a certainty score. With a human-in-the-loop framework, these predictions can be sorted into various risk levels. SASI assigns high priority to uncertain predictions, for an immediate review by mental health experts.

1 Introduction

Suicide is a global phenomenon responsible for 1.3% of deaths worldwide (WHO, 2019). While it is the leading cause of death among 14-35 year olds in the US (Hedegaard et al., 2021), suicide rates have increased by 13% in Japan between July to September 2020 (Tanaka and Okamoto, 2021). It hence becomes critical to extend clinical and psychiatric care, which relies heavily on identifying those at risk. While 80% of patients do not undergo clinical treatment, 60% of those who succumbed to suicide denied having suicidal thoughts to mental health experts (McHugh et al., 2019). However, studies show eight out of ten people shared suicidal thoughts on social media (Golden et al., 2009).

The advent of Natural Language Processing (NLP) shows promise for suicide risk assessment based on online user behavior (Ji et al., 2021b; Choudhury et al., 2016), with automatic risk assessment algorithms outperforming traditional clinical methods (Coppersmith et al., 2018; Linthicum et al., 2019). Numerous deep learning methods already exist, which include leveraging suicide-related word-embeddings (Cao et al., 2019), social graphs (Mishra et al., 2019; Sinha et al., 2019; Cao et al., 2022; Sawhney et al., 2021b) and historical context (Matero et al., 2019; Gaur et al., 2019).

However, mental health is a safety-critical realm, where technological failure could lead to severe harm to users on social media (Sittig and Singh, 2015). One such case was covered by Register (2020), wherein a medical bot suggested a mock patient kill themselves, demonstrating that unintended harmful behavior can emerge from AI systems (Amodei et al., 2016; Chandler et al., 2020).

*Authors contributed equally

Despite the significant power of traditional NLP methods, such models are inherently designed to make a prediction even when not confident. This poses a challenge when working with critical tasks like suicide risk assessment, for which it may be hard to make a prediction due to various reasons such as task hardness or contained ambiguity. Such a system may associate a lower risk level to a user who needs urgent help. A resulting delayed response from mental health experts may lead to adverse consequences. We hence need systems that assign high priority to uncertain predictions, for immediate review and response.

Contributions: We reformulate suicide risk assessment as a prioritized prediction task which factors in uncertainty, and propose **SASI**: A Risk-Averse Mechanism for **S**uicidality **A**ssessment on **S**ocial Med**I**a. SASI is risk-averse in the sense that it is self-aware, as it incorporates a selection function to measure uncertainty. Based on a set threshold value, SASI refrains from making a prediction when it is uncertain. We show that SASI can act as a tool to efficiently prioritize users who need immediate attention. Through a human-in-the-loop framework that involves a domain expert, SASI assigns high priority to uncertain predictions to avoid critical failure (Figure 1). We demonstrate the effectiveness of SASI using a real-world gold standard Reddit dataset. Through a series of experiments, we show SASI refrains from making 83% of incorrect predictions. We further demonstrate its effectiveness through a qualitative study and discuss the ethical implications.

2 Methodology

2.1 Columbia Suicide Severity Risk Scale

The Columbia Suicide Severity Rating Scale (C-SSRS) is an authoritative questionnaire employed by psychiatrists to measure suicide risk severity (Posner et al., 2011). There are 3 items in the scale: Suicide Ideation, Suicide Behavior, and Suicide Attempt. Each C-SSRS severity class is composed of a conceptually organized set of questions that characterize the respective category. Responses to the questions across the C-SSRS classes eventually determine the risk of suicidality of an individual (Interian et al., 2018; McCall et al., 2021). One of the challenges researchers face when it comes to dealing with social media content is the disparity in the level of emotions expressed (Gaur et al., 2019). Since the C-SSRS was originally designed for use in clinical settings, adapting the same metric to a social media platform would require changes to address the varying nature of emotions expressed. For instance, while in a clinical setting, it is typically suicidal candidates that see a clinician; on social media, non-suicidal users may participate to offer support to others deemed suicidal (Gaur et al., 2021). To address these factors, two additional classes were defined (Gaur et al., 2019) to the existing C-SSRS scale with three classes: Suicide Indicator and Supportive (Negative class).

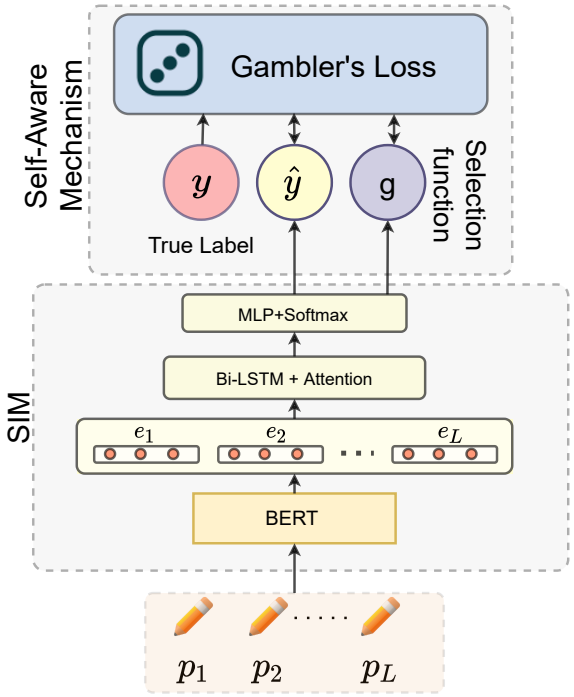

Figure 2: An overview of SASI: SASI incorporates a risk-averse, self-aware mechanism to any given suicide ideation model (SIM) by training using Gambler's Loss. It refrains from predicting when uncertain.

2.2 Problem Formulation

Following existing work (Gaur et al., 2019; Sawhney et al., 2021a), we formulate the problem as a classification task to predict the suicidal risk of the user $u_i \in \{u_1, u_2, \cdots, u_N\}$, whose posts $P_i = \{p_1^i, p_2^i, \cdots, p_T^i\}$ are authored over time in a chronological order, with the latest post being p_T^i. We denote the label set $\mathbf{Y}$ = {Support (SU), Indicator (IN), Ideation (ID), Behaviour (BR), Attempt (AT)} in increasing order of severity risk, defined based on the C-SSRS. For a given Suicide Ideation Model, our goal is to expand the cardinality of the label space to $|\mathbf{Y}| + 1$ so as to enable an option to refrain when the model is uncertain.

2.3 Suicide Ideation Model (SIM)

Each post made by a user could provide detailed context of suicidal thought manifestation over time (Oliffe et al., 2012). To capture this property, we draw inspiration from existing state-of-the-art (SOTA) models (Gaur et al., 2019; Matero et al., 2019; Sawhney et al., 2021a; Ji et al., 2021a) which use LSTM based backbones. To encode each post p_k^i, we use the 768-dimensional representation of the `[CLS]` token obtained from BERT (Devlin et al., 2019) as e_k^i=BERT(p_k^i). As shown in Figure 2, we then pass each post embedding sequentially through a bi-directional LSTM, given as h_k^i = Bi-LSTM(e_k^i). We thus obtain the sequence of hidden states, $\boldsymbol{x} = [h_1^i, h_2^i, \cdots, h_T^i]$, where $h_k^i \in \mathbb{R}^H$, and H is the hidden dimension. To filter out relevant signals from the potentially vast user history (Shing et al., 2020), we pass the hidden state sequence through an attention layer. The final layer is a multilayer perceptron (MLP) to obtain the prediction vector $\hat{\boldsymbol{y}}$, given as:

$$\begin{aligned} &\hat{\boldsymbol{y}} = f(\boldsymbol{x}), \quad \text{where} \\ &f(\boldsymbol{x}) = \text{Softmax}(\text{MLP}(\text{Attention}(\boldsymbol{x}))) \end{aligned} \tag{1}$$

2.4 Self-Aware Mechanism

To make the model self-aware, we transform the model such that it makes a prediction only when certain (Liu et al., 2019). As shown in Figure 2, the model $f : \mathbb{R}^{T\times H} \rightarrow \mathbf{Y}$ is augmented with a selection function $g : \mathbb{R}^{T\times H} \rightarrow (0, 1)$, which is an extra logit. The augmented model is described as a piece-wise function, given by:

$$(f, g)(\boldsymbol{x}) := \begin{cases} \text{Refrain}, & \text{if } g \geq \tau \\ \text{argmax}(\hat{\boldsymbol{y}}), & \text{otherwise} \end{cases} \tag{2}$$

Where the threshold $\tau \in (0, 1)$, argmax($\hat{\boldsymbol{y}}$) $\in \mathbf{Y}$. Let $p = (f, g)(\boldsymbol{x})$, where $p \in \mathbf{Y} \cup \{\text{Refrain}\}$ denote the final prediction by the model for a user u_i. Human moderators can then define the level of granularity of these predictions, and sort them into priority levels as desired. As an example, moderators may choose to have only three levels of priority, where the user is high priority if $p \in \{\text{AT, BR, Refrain}\}$, moderate if $p \in \{\text{ID, IN}\}$ and low if $p \in \{\text{SU}\}$. With the addition of the Refrain option, uncertain predictions will have highest priority, alleviating the possibility of high-risk users being neglected.

It is essential to note that the confidence threshold τ is not utilized during training, rather as a threshold variable to calibrate data coverage (cov) during evaluation. The cov fraction of total samples is what SASI predicts on, leaving out $(1-cov)$ samples for which SASI is most uncertain. Specifically, we can choose some value τ such that there will be $(1 - cov)$ samples for which $g \geq \tau$. The idea behind this approach is to trade-off $(1 - cov)$ samples for immediate review by mental health experts in exchange for higher model performance on the cov samples about which it is confident.

2.5 Network Optimization

In any m-class classification problem, if the model assigns a high probability score to the wrong class, then learning becomes difficult due to vanishing gradients (Ziyin et al., 2020). To account for the additional refrain option in the augmented label space, we train SASI using Gambler's Loss (Liu et al., 2019). Gambler's loss allows the gradients to propagate through g instead, by abstaining from assigning weights to any of the m classes. Thus, the model learns a distribution of noisy/uncertain data points characterized by the selection function g. The loss function is given as:

$$\mathcal{L} = -\sum_{j}^{|\mathbf{Y}|} y_j \cdot \log(\hat{y}_j \cdot r + g) \tag{3}$$

where y_j is the true label, and the reward r is a hyperparameter. A higher value of r discourages restraint. Since the loss function directly learns g, it does not depend on the coverage (Liu et al., 2019), and can be manually set to any value during evaluation.

3 Experimental Setup

3.1 Dataset

We use the dataset released by Gaur et al. (2019), which contains Reddit posts of 500 users filtered from an initial set of 270,000 users across several mental health and suicide-related subreddits, such as r/StopSelfHarm (SSH), r/selfharm (SLF), r/bipolar (BPL), r/BipolarReddit (BPR), r/BipolarSOs, r/opiates (OPT), r/Anxiety (ANX), r/addiction (ADD), r/BPD, r/SuicideWatch (SW), r/schizophrenia (SCZ), r/autism (AUT), r/depression (DPR), r/cripplingalcoholism (CRP), and r/aspergers (ASP). The posts were annotated by practicing psychiatrists into five increasing risk levels based on the Columbia Suicide Severity Risk Scale (Posner et al., 2011), leading to an acceptable

average pairwise agreement of 0.79 and a group-wise agreement of 0.73. The class distribution of each category with increasing risk level is: Supportive (20%), Indicator (20%), Ideation (34%), Behaviour (15%), Attempt (9%). On average, the number of posts made by a user is 18.25±27.45 with a maximum of 292 posts. The average number of tokens in each post is 73.4±97.7.

3.2 Evaluation Metrics

We first describe the evaluation metrics that measure how well the model performs on the *cov* samples. Following Gaur et al. (2019), we use graded variants of F1 score, Precision, and Recall, where we alter the formulation of False Negatives (FN) and False Positives (FP). FN is modified as the ratio of the number of times predicted severity of suicide risk level (k^p) is less than the actual risk level (k^a) over N number of samples. FP is the ratio of the number of times the predicted risk (k^p) is greater than the actual risk (k^a), given as:

$$FN = \frac{\sum_{i=1}^{N} I(k_i^a > k_i^p)}{N}$$
$$FP = \frac{\sum_{i=1}^{N} I(k_i^p > k_i^a)}{N} \quad (4)$$

Let P_{T} denote the total number of test samples, $P_{\text{corr+refrain}}$ the sum of samples that have either been correctly predicted or have been refrained, P_{refrain} the total number of refrained samples, and P_{in} the number of incorrect predictions among the refrained samples. We additionally introduce two metrics, *Robustness* and *Fail-Safe Rejects*, as:

$$\text{Robustness} = \frac{P_{\text{corr+refrain}}}{P_{\text{T}}}$$
$$\text{Fail-Safe Rejects} = \frac{P_{\text{in}}}{P_{\text{refrain}}} \quad (5)$$

Robustness captures the fraction of samples which are correctly classified or instead sent for immediate review. *Fail-Safe Rejects* captures the fraction of refrained samples which were indeed erroneous. A higher Fail-Safe Rejects score hence implies that human moderators will be subjected to a lesser amounts of redundant work.

4 Results

4.1 Performance Comparison

We compare the performance of SASI with various state-of-the-art baselines in Table 1. Sequential models like Suicide Detection Model (SDM) (Cao et al., 2019) and ContextBERT (Matero et al., 2019) generally outperform ContextualCNN (Gaur et al., 2019), which uses a bag-of-posts approach. SISMO (Sawhney et al., 2021a) shows further improvements by modeling the ordinal nature of risk labels. SASI significantly outperforms ($p < 0.005$) these methods for various values of coverage (*cov*), demonstrating its ability to avoid committing to erroneous predictions by characterizing its confidence (Liu et al., 2019).

Model	Gr. Prec.	Gr. Recall	FScore	Robustness	Fail-Safe Rejects
Contextual CNN	0.65	0.52	0.59	-	-
SDM	0.61	0.54	0.57	-	-
ContextBERT	0.63	0.57	0.60	-	-
SISMO	0.66	0.61	0.64	-	-
SASI (Cov 100%)	0.67*	0.62	0.66*	0.48	-
SASI (Cov 85%)	*0.69**	*0.65**	*0.67**	*0.61*	**0.83**
SASI (Cov 50%)	**0.71***	**0.69***	**0.70***	**0.73**	*0.65*

Table 1: We report the median of results over 10 random seeds. * indicates the result is statistically significant with respect to SISMO ($p < 0.005$) under Wilcoxon's signed-rank test. **Bold** denotes best performance while *Italics* denotes second best.

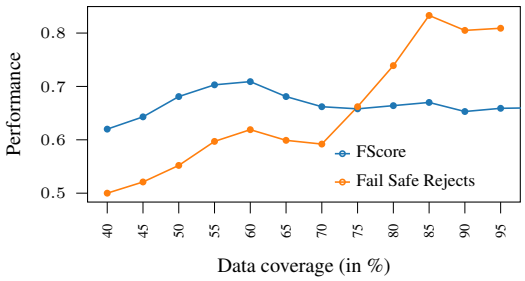

Figure 3: Changes in performance metrics with increasing coverage, averaged over 10 random seeds.

4.2 Coverage and Performance Trade-off

We further evaluate SASI for various values of target coverage (*cov*) by calibrating the threshold τ. As shown in Figure 3, lower coverage leads to an increase in Graded Recall, Precision, and FScore (Table 1), as the model only keeps *cov* predictions which it is highly certain about. However, we observe a decrease in Fail-Safe Rejects due to an increasingly cautious approach employed by the model, which implies an increased fraction of originally correct predictions that need to be manually reviewed. We hence observe a trade-off, wherein we must seek to achieve competitive performance on the *cov* samples, while at the same time not overburden moderators with the $(1 - cov)$ samples. For lower coverage values (say 50%), human modera-

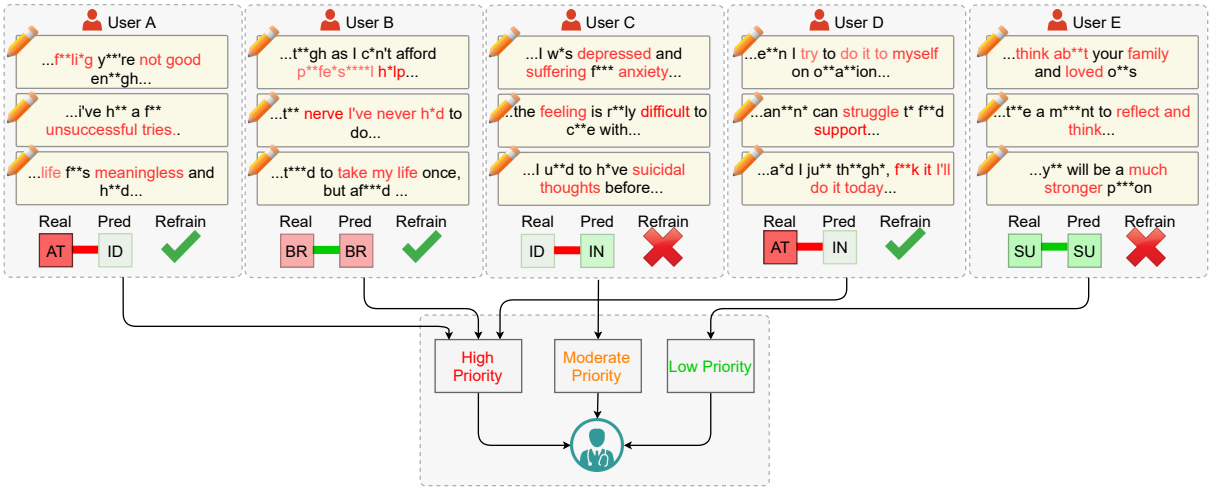

Figure 4: We show SASI can be used for efficient prioritization of users during suicide risk assessment. For each user, we show the real labels next to predicted labels, while also indicating whether SASI refrained from making that prediction. We further demonstrate how SASI sorts the users into priority levels. All examples in this paper have been paraphrased as per the moderate disguise scheme (Bruckman, 2002) to protect user privacy.

tors may be overburdened by having to review a lot of redundant samples. On the other hand, we note that SASI (85%) provides more utility, as it statistically outperforms SOTA models like SISMO, while maintaining a fail-safe rejection score of 83% and a competitive robustness score of 61%.

4.3 Qualitative Analysis

The essence of SASI lies behind its ability to refrain from making misleading predictions over high-risk samples. We study five users with snippets of their posts, as shown in Figure 4. We observe the model makes erroneous predictions on high-risk users A and D. However, SASI refrains from committing to these predictions, assigning these users a high priority for immediate review and response. SASI chooses to refrain despite predicting the risk level of user B correctly, possibly because it employs a cautious approach due to phrases such as '*take my life*' scattered in the user's timeline. This user, who is already of relatively high risk, is hence assigned a high priority. User E shows a very low sign of risk, which is confidently captured by SASI without needing to refrain. User C is an erroneous case wherein SASI is confident, yet makes a wrong prediction. However, the user is not high risk and gets assigned to the same priority level as the true risk label. While this example is not a cause for concern, certain situations may arise where SASI also confidently assigns a low-risk score to a high-risk user, opening avenues for future work that involves integrating and reformulating ordinal regression over the principles of Gambler's loss.

5 Conclusion

With a motivation to provide a robust solution to fine-grained suicide risk assessment on social media, we present SASI, a framework that integrates the concept of selective prioritization to existing deep learning based risk-assessment techniques. SASI is self-aware, wherein it refrains from making a prediction when uncertain, and instead assigns high priority to such data samples for immediate review by mental health experts. We demonstrated the effectiveness of SASI through quantitative evaluations on real-world data, wherein SASI avoided high-risk situations by refraining from making 83% of incorrect predictions. Through a qualitative analysis, we described how SASI can be used as a part of a human-in-the-loop framework, facilitating efficient responses from mental health experts.

Acknowledgements

We thank Prof. Amit Sheth for reviewing the paper and providing valuable feedback and support. We would also like to thank the anonymous reviewers for their insightful suggestions on various aspects of this work.

Ethical Considerations

We work within the scope of acceptable privacy practices suggested by Chancellor et al. (2019) and considerations presented by Fiesler and Proferes (2018) to avoid coercion and intrusive treat-

ment. The primary source of the dataset used in this study is Reddit. Although Reddit is intended for anonymous posting, we take further precautions by performing automatic de-identification of the dataset using named entity recognition (Zirikly et al., 2019). All examples used in this paper are further been anonymized, obfuscated, and paraphrased for user privacy (Benton et al., 2017) and to prevent misuse as per the moderate disguise scheme suggested by Bruckman (2002). Taking inspiration from Benton et al. (2017), we also keep the annotation of user data separate from raw user data on protected servers linked only through anonymous IDs. Our work focuses on building an assistive tool for screening suicidal users and providing judgments purely based on observational capacity. We acknowledge that it is almost impossible to prevent abuse of released technology even when developed with good intentions (Hovy and Spruit, 2016). Hence, we ensure that this analysis is shared only selectively to avoid misuse such as Samaritan's Radar (Hsin et al., 2016).

We further acknowledge that the studied data may be susceptible to demographic, expert annotator, and medium-specific biases (Hovy and Spruit, 2016). While the essence of our work is to aid in the early detection of at-risk users and early intervention, any interventions must be well-thought, failing which may lead to counter-helpful outcomes, such as users moving to fringe platforms, making it harder to provide assistance (Kumar et al., 2015). Care should be taken to not to create stigma, and interventions must hence be carefully planned by consulting relevant stakeholders, such as clinicians, designers, and researchers (Chancellor et al., 2016), to maintain social media as a safe space for individuals looking to express themselves (Chancellor et al., 2019). It is also essential that clinicians and human moderators are not overburdened (Chancellor et al., 2019). For instance, "Alarm fatigue" is when alarms are so excessive, many of which are false positives, that healthcare providers become desensitized from alarms (Drew et al., 2014).

We also agree that suicidality is subjective (Keilp et al., 2012), wherein the interpretation may vary across individuals on social media (Puschman, 2017). We do not make any diagnostic claims, rather help prioritize the users that should be evaluated by the medical professionals first, as part of a distributed human-in-the-loop framework (de Andrade et al., 2018).

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *ArXiv preprint*, abs/1606.06565.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected inhuman subjects research on the internet. *Ethics and Inf. Technol.*, 4(3):217–231.

Lei Cao, Huijun Zhang, and Ling Feng. 2022. Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*, 24:87–102.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 79–88, New York, NY, USA. Association for Computing Machinery.

Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 1171–1184, New York, NY, USA. Association for Computing Machinery.

Chelsea Chandler, Peter W Foltz, and Brita Elvevåg. 2020. Using machine learning in psychiatry: the need to establish a framework that nurtures trustworthiness. *Schizophrenia bulletin*, 46(1):11–14.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 2098–2110. ACM.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology*, 31(4):669–684.

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

Barbara J Drew, Patricia Harris, Jessica K Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloS one*, 9(10):e110274.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit P. Sheth, Randy S. Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 514–525. ACM.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PloS one*, 16(5):e0250448.

Robert N Golden, Carla Weiland, and Fred Peterson. 2009. *The truth about illness and disease*. Infobase Publishing.

Holly Hedegaard, Sally C Curtin, and Margaret Warner. 2021. Suicide mortality in the united states, 1999-2019. *NCHS data brief*, (398):1–8.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Honor Hsin, John Torous, and Laura Roberts. 2016. An Adjuvant Role for Mobile Health in Psychiatry. *JAMA Psychiatry*, 73(2):103–104.

Alejandro Interian, Megan Chesin, Anna Kline, Rachael Miller, Lauren St. Hill, Miriam Latorre, Anton Shcherbakov, Arlene King, and Barbara Stanley. 2018. Use of the columbia-suicide severity rating scale (c-ssrs) to classify suicidal behaviors. *Archives of suicide research*, 22(2):278–294.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2021a. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021b. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

John G Keilp, Michael F Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K Burke, Hanga Galfalvy, Maria A Oquendo, and J John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of affective disorders*, 140(1):75–81.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, page 85–94, New York, NY, USA. Association for Computing Machinery.

Kathryn P Linthicum, Katherine Musacchio Schafer, and Jessica D Ribeiro. 2019. Machine learning in suicide science: Applications and ethics. *Behavioral sciences & the law*, 37(3):214–222.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2019. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10622–10632.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

William V McCall, Ben Porter, Ashley R Pate, Courtney J Bolstad, Christopher W Drapeau, Andrew D Krystal, Ruth M Benca, Meredith E Rumble, and

Michael R Nadorff. 2021. Examining suicide assessment measures for research use: Using item response theory to optimize psychometric assessment for research on suicidal ideation in major depressive disorder. *Suicide and Life-Threatening Behavior*, 51(6):1086–1094.

Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open*, 5(2).

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

John L Oliffe, John S Ogrodniczuk, Joan L Bottorff, Joy L Johnson, and Kristy Hoyak. 2012. "you feel like you can't live anymore": Suicide from the perspectives of canadian men who experience depression. *Social science & medicine*, 74(4):506–514.

Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, 168(12):1266–1277.

Cornelius Puschman. 2017. Bad judgment, bad ethics? *Internet Research Ethics for the Social Age*, page 95.

The Register. 2020. Researchers made an openai gpt-3 medical chatbot as an experiment. it told a mock patient to kill themselves.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021a. Towards ordinal suicide ideation detection on social media. WSDM '21, page 22–30, New York, NY, USA. Association for Computing Machinery.

Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021b. 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*, pages 2176–2190, Online. Association for Computational Linguistics.

Han-Chin Shing, Philip Resnik, and Douglas Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, Online. Association for Computational Linguistics.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. #suicidal - A multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 941–950. ACM.

Dean F Sittig and Hardeep Singh. 2015. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. In *Cognitive informatics for biomedicine*, pages 59–80. Springer.

Takanao Tanaka and Shohei Okamoto. 2021. Increase in suicide following an initial decline during the covid-19 pandemic in japan. *Nature human behaviour*, 5(2):229–238.

WHO. 2019. Suicide data.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Liu Ziyin, Blair Chen, Ru Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2020. Learning not to learn in the presence of noisy labels. *ArXiv*, abs/2002.06541.