Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing

Shaoxiong Ji
Aalto University
shaoxiong.ji@aalto.fi

Abstract

Recent applications of natural language processing techniques to suicidal ideation detection and risk assessment frame the detection or assessment task as a text classification problem. Recent advances have developed many models, especially deep learning models, to boost predictive performance. Though the performance (in terms of aggregated evaluation scores) is improving, this position paper urges that better intention understanding is required for reliable suicidal risk assessment with computational methods. This paper reflects the state of natural language processing applied to suicide-associated text classification tasks, differentiates suicidal risk assessment and intention understanding, and points out potential limitations of sentiment features and pretrained language models in suicidal intention understanding. Besides, it urges the necessity for sequential intention understanding and risk assessment, discusses some critical issues in evaluation such as uncertainty, and studies the lack of benchmarks.

1 Introduction

Warning: this paper contains text examples that are negative, depressive, or adverse.

Suicide is a global problem, with most industrialized countries and many emerging markets seeing exceptionally high rates. The latest WHO publication on suicide study,[1] entitled suicide worldwide in 2019, reported that more than 700,000 people die by suicide every year around the world, with one suicide case for every 100 deaths. According to a report on mental health from the World Health Organization,[2] one out of every four persons in the world lives with mental disorders to some extent. And 3 out of 4 people with severe mental disorders do not receive treatment, worsening the problem. Previous research has found a link between mental problems and the likelihood of suicide (Windfuhr and Kapur, 2011). More people live with mental health issues during particular periods, such as the pandemic. Many may not seek care from mental health practitioners due to insufficient mental health services.

Social networking sites offer an essential forum for communication and information sharing online. Online discussions also include much harmful information and can result in issues like cyberstalking or cyberbullying. The result is severe and dangerous because the wrong information is frequently engaged in social cruelty, causing rumors or even mental harm. Research shows a link between cyberbullying and suicide (Hinduja and Patchin, 2010). Victims exposed to too many negative messages or events may become depressed and desperate. Some of them are likely to choose suicide as their option and ask for suicide methods on online social network websites (Starcevic and Aboujaoude, 2015), which is called cybersuicide. Furthermore, some social groups even persuade other individuals to commit suicide together, namely cybersuicide pact. Thus, it is necessary to understand users' intentions by mining their conversations, modeling their profiles, and analyzing their social groups.

On the other hand, social networking also offers a channel for peer support among those living with mental illnesses, allowing social workers to provide proactive social care and early intervention. People experiencing a mental health condition sometimes post their feelings or experience on online discussion forums like Reddit or social networking websites such as Twitter and Weibo. That user-generated content is an essential portal to facilitate automated suicidal risk assessment and allow

[1] Published in 16 June 2021 by the WHO team of Mental Health and Substance Use, available via https://www.who.int/publications/i/item/9789240026643.

[2] Mental health action plan 2013 - 2020, available via http://www.who.int/mental_health/action_plan_2013/mhap_brochure.pdf?ua=1

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4028–4038
December 7-11, 2022

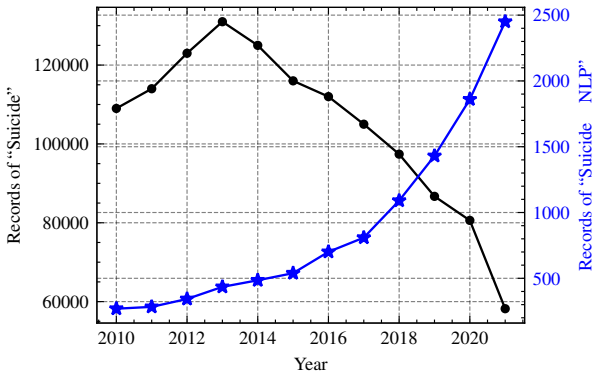


Figure 1: The number of scholarly articles published in the past twelve years. The black line with dot markers presents the records searched with query of "suicide" and the blue line with asterisk markers shows the records with query of "suicide" and "NLP".

social workers to provide effective prevention.

Computational methods using natural language processing techniques have been studied to classify mental disorders and suicide from various data types, mainly social media posts. See the recent review of NLP applied to mental illness detection (Zhang et al., 2022). Suicidal risk assessment in social media is a task that categorizes given social posts into different levels of suicidal risks. Early detection of suicide paves the way for suicide prevention. Research on suicide with NLP is booming. Table 1 shows the number of scholarly articles searched by the Google Scholar search engine in the past twelve years. The number of research papers on suicide declined after 2013, and the number of NLP papers on suicide increased steadily.

Suicide attempt or completed suicide usually starts from suicidal ideation. Thus, early assessment of the intensity of suicidal ideation and effective intention understanding can help predict later suicidal risk better. A recent review (Abdulsalam and Alhothali, 2022) and a comparative analysis of recent techniques (Haque et al., 2022) discussed many publications for suicidal risk classification. For example, some construct sentiment lexicons for posts regarding suicidality and build machine learning classifiers (Sarsam et al., 2021). Recent deep learning-based models build comprehensive neural architectures to improve the classification performance, including attentive relation network (Ji et al., 2022a), enhanced word embedding (Cao et al., 2019), transformer networks (Zhang et al., 2021), and hyperbolic user representation learning (Sawhney et al., 2021c).

However, most existing works that approach suicidal risk assessment as a text classification problem consider less about human intention understanding. Human intention or mental state understanding is a complex cognitive process. This position paper argues that suicidal risk assessment with natural language processing needs a better intention understanding of the social posts. It points out several unresolved issues and open questions, including the limitation of sentiment features and pretrained language models, the importance of sequential intention understanding, the critical ingredients in evaluation, and the need for more benchmarks for visual-grounded intention understanding and multilingualism.

This paper is organized as follows. Section 2.1 highlights the challenges of suicidal intention understanding and differentiates it from suicidal risk assessment. Section 3 proposes the setting of intention understanding in sequence (e.g., the timeline of user's posts). We discuss the obstacles of evaluation and benchmarks in Section 4 and 5. Section 6 reviews additional related work. We conclude this paper in Section 7.

2 Suicidal Intention Understanding

Many factors can lead to suicide, i.e., users' personality (such as hopelessness, severe anxiety, schizophrenia, alcoholism, and impulsivity), social factors (like social isolation, too much exposure to deaths), and adverse life events (including traumatic events, physical illness, affective disorders, and previous suicide attempts). Due to the prevalence of social networks, online users may be exposed to risk factors such as talking about suicide, worsening mood, and cybersuicide pacts. Suicide is also related to other matters, such as access to lethal suicide methods in the physical world.

Adequate measurement of suicidality is vital to assessing people at risk for suicide attempts. The scale of suicidal ideation (SSI) (Beck et al., 1979), one of the classic measures of suicidality, develops a 19-item instrument for quantifying suicidal intention in clinical research. Its variables include race, age, education, civil status, employment status, and psychiatric diagnosis. As an instrument of clinical psychology, SSI was found capable of assessing the subtle changes in levels of depression and hopelessness.

Whether the NLP model can mimic clinical professionals' screening practice and understand the

inherent intention of people living with mental conditions and suicidality becomes a challenging problem. This section discusses intention understanding[3] in suicidal risk assessment with natural language processing.

2.1 Intention Understanding

Human brains infer people's intentions during communication, and decoding human intention involves complex social cognition. Early works on mining people's intentions from social media consider key elements such as intent indicators and intent keywords as features for intention classification (Wang et al., 2015). However, intention understanding in suicidal risk assessment is challenging when it comes to social media scenarios. The communicative nature of risk screening conveys massive sociolinguistic information, gestures, acoustic information, or facial expressions. Unlike clinical screening, social content usually only contains texts when people who live with suicidal ideation do not post selfies or other pictures. Intentions underlying communication in the cyber world are hard to be understood from semantic patterns or syntactic structures in the natural language.

Neuroimaging research reveals that two different neural systems process immediate goals and long-term intentions, and these two systems decode intentions depending on the shared goal of interacting agents (Canessa et al., 2012). Immediate goals are associated with action understanding, while the understanding of long-term intention is associated with the mentalizing system, involving human's mental states (Canessa et al., 2012). Taking the text "I am going to buy a knife" as an example, the immediate goal can be self-harm if we consider that the context is more about mental issues.[4] The long-term intention can be taking one's own life due to chronic suffering from some mental illness.[5] Human intentional communication can also disguise one's actual intention intentionally, making human intention, especially long-term intentions, hard to be decided based on the content of the present message.

[3]The word "understanding" might be over-stating for machine intelligence compared with human-level understanding, especially in current NLP research on suicidal risk assessment. This paper puts this potential debate aside.

[4]The immediate goal can also be eating a meal if the context is more about daily life.

[5]Here is an example, although not all suicide attempts have the self-harm stage.

2.2 Suicidal Intent v.s. Suicidal Risk

Current research classifies suicidal risks into different severities, e.g., high, medium, or low-risk. Suicidal intention is one of the crucial aspects of suicidal risk assessment or stratification. However, current dataset construction considers very little about the annotation of suicidal intent while defining the task of suicidal risk assessment. For example, Ji et al. (2018) and Sinha et al. (2019) built a collection of keywords to filter social posts and annotation guidelines for human annotators to do manual labeling. Such annotation guidelines are usually simple, with several rules, including 1) posts contain a suicide plan and/or previous attempts, or potential suicidal actions; 2) posts express suicidal thoughts or ideation; 3) posts reveal risk factors, e.g., depression and bullying; 4) posts contain somber words. Cao et al. (2019) collected social posts from people who commented on their suicidal thoughts on the last post of a student who died by suicide. These datasets simplify the task as a binary classification problem on whether the post contains suicide-related signals of an individual.

We need a fine-grained understanding of suicidal ideation and intention. The Columbia Suicide Severity Rating Scale (C-SSRS) (Posner et al., 2008) is a clinical measure of suicide severity and is supposed to be administered by a trained individual in suicide screening. Gaur et al. (2019) developed the C-SSRS-based five-label categories (named as supportive, indicator, ideation, behavior, and attempt) and built a suicidal risk severity lexicon covering different levels of suicidal risk severity. Suicidal risk stratification tends to prioritize suicidal ideation in practice (Large et al., 2017). The C-SSRS scale measures the intensity of suicidal ideation by frequency, duration, controllability, deterrents, and reasons for ideation. NLP models for suicide classification should be able to classify fine-grained suicidal risks and detect the genuine intention of an individual. However, many publications in NLP tend not to distinguish the difference between suicidal intent and suicidal risk. There exists nuance between these two concepts:

> Whereas suicidal intent may be regarded as a psychological phenomenon subject to exploration and measurement, suicidal risk is a predictive statement of the probability of the occurrence of a fatal suicide attempt and can be conceived in terms of a complex (although not fully

formulated) equation. (Beck et al., 1979)

Our NLP models for effective suicidal risk assessment are required to "understand" the inherent intention of the text. As the severity of mental issues affects the choice of intervention actions, recognizing genuine intention can facilitate the adoption of corresponding intervention actions. Unfortunately, the ground truth about the user's real intentions behind the social post is usually unavailable in social media data. This further makes the evaluation of intention understanding impossible. A recent study on depression detection leverages clinical questionnaires to improve the out-of-domain generalization (Nguyen et al., 2022). The model constrained to clinical questionnaires with well-designed features for intention evaluation can be a possible solution and is worthy of further investigation.

2.3 Intention & Sentiment

Analyzing users' use of language provides insights helping to detect suicidality or stratify the suicidal risk. Intuitively, emotional words commonly utilized in online posts may vary depending on the focal mental issues of the users. For example, aggressive words may be a salient indicator of anxiety, whereas pessimistic words may be used more frequently for depressed online users. However, Gaur et al. (2021) showed that posts with different severity of suicidal risks on Reddit posts have no significant variation in emotion or sentiment. This paper utilizes the SenticNet suite (Cambria et al., 2020, 2022),[6] which combines symbolic and subsymbolic artificial intelligence for sentiment analysis, to explore the sentiment in a Twitter dataset (Ji et al., 2018) with suicide-associated and control posts. Figure 2 shows the distributions of sentiment attitude, temper, sensitivity, and introspection. The results indicate that the texts with binary classes have no significant differences in sentiment. Methods with sentimental or emotional features, e.g., Sarsam et al. (2021), extract textual features and use machine learning classifiers to fit the data. Neural NLP models that enhance the feature learning with sentiment modules, e.g., Ji et al. (2022a), increase the feature dimension and the model complexity. When carefully optimized, these models usually gain improved classification performance. However, whether the modeling of sentiment information in text can capture the individual's intention is questionable. One worrying guess is that it might only help build a more powerful text classifier.

2.4 Intention & Language Models

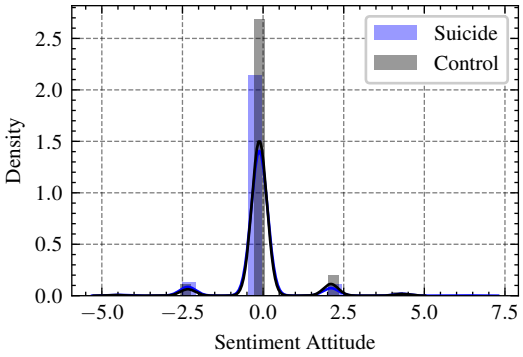
In the era of pretraining, many pretrained large language models have been applied to fine-tune suicide text classifiers. Those pretrained models achieved superior classification performance. One straightforward question in this position paper is whether the pretrained models can "understand" the latent intention to some extent.

Taking the sentence "I am going to buy a knife" as an example,[7] it can be recognized as purchasing intention in a daily context or suicide attempt (i.e., to commit suicide with a knife). Starting with our question, we conduct the fill-mask language modeling task with the sentence "I am going to buy a knife and [MASK]", using several masked language models, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and their domain-specific variants MentalBERT and MentalRoBERTa (Ji et al., 2022b). Figure 3 shows the output of word probabilities. The suicidal ideation does not appear in the prediction of BERT (Figure 3a). RoBERTa in Figure 3b and MentalBERT in Figure 3c predict the suicidal intention ("die") in the fifth place. MentalRoBERTa in Figure 3d recognizes the suicidal intention as the first place.

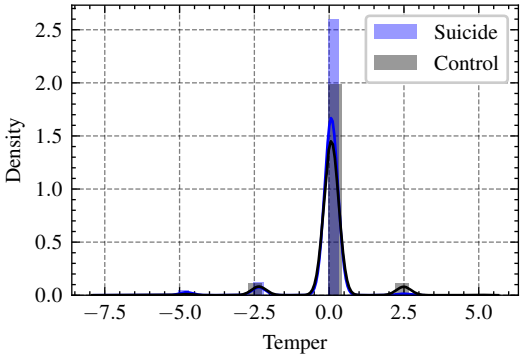
Then, we use another example sentence "This life is not worth living. I am going to buy a knife and [MASK]." that provides a bit more contextual information. The results in Figure 4 show that RoBERTa and MentalBERT tend to predict suicidal intention ("die") with higher probabilities. MentalRoBERTa in Figure 4d predicts "die" with a significantly high probability. These two examples showcase the abilities of masked language models to predict intention as a fill-mask task. Domain-adaptive continued pretraining helps with suicide keyword prediction, although we do not consider the memorization issue of BERT here. While MentalRoBERTa outputs "die" with a high probability, which can be interpreted as suicidal intention, we also see the outputs of RoBERTa ("shoot") can be interpreted as potential anti-social and criminal actions (e.g., shoot at others, although it does not make sense with a knife.).

[6] The API of SenticNet suite is available at http://sentic.net/api/, and the knowledge base can be downloaded from http://sentic.net/downloads/.

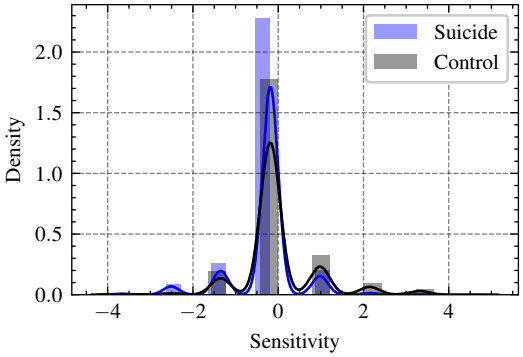
[7] Obviously, this simple example is not associated with the context.



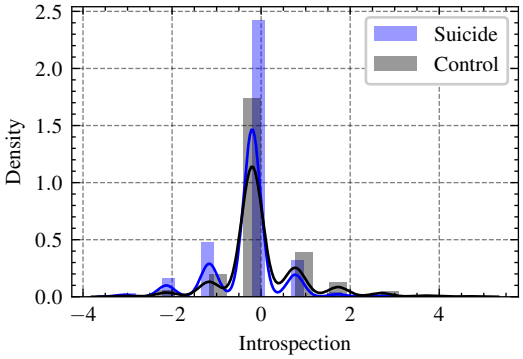
(a) Distribution of Sentiment Attitude



(b) Distribution of Temper

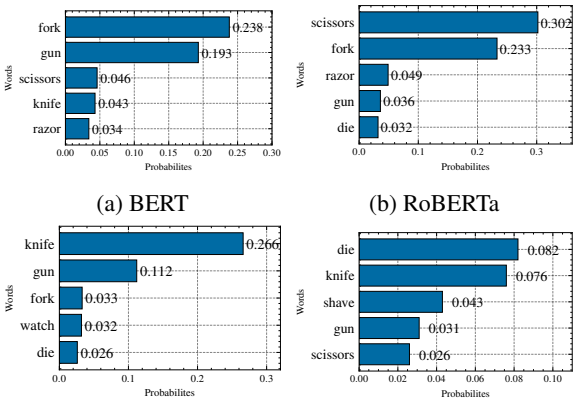


(c) Distribution of Sensitivity



(d) Distribution of Introspection

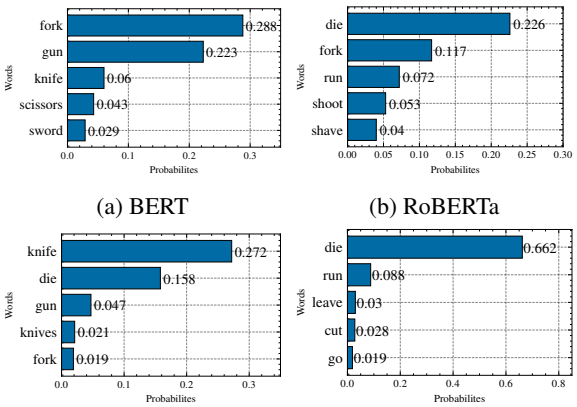
Figure 2: The sentiment distributions of a Twitter dataset with binary classes (Ji et al., 2018), calculated by the SenticNet suite. "Suicide" represents the class whose texts are associated with suicide messages and "Control" means the texts are from the control group without suicide messages.



(a) BERT (b) RoBERTa

(c) MentalBERT (d) MentalRoBERTa

Figure 3: The output of word probabilities in the fill-mask language modeling task using various pretrained masked language models for the sentence "I am going to buy a knife and [MASK]."



(a) BERT (b) RoBERTa

(c) MentalBERT (d) MentalRoBERTa

Figure 4: The output of word probabilities in the fill-mask language modeling task using various pretrained masked language models for the sentence "This life is not worth living. I am going to buy a knife and [MASK]."

We further test how the generative language model interprets the intention inherent in text prompt. We use the previous two examples in the fill-mask task as inputs to generate new texts. The following two boxes show the texts generated by the GPT-2 model hosted in Huggingface.[8]

Text Generated by GPT-2

I am going to buy a knife and kill everyone else for free" and "that is the right thing to do". It's like saying - this is not my position,

[8]The GPT-2 model is available at https://huggingface.co/gpt2

. it is the opinion of all.

Text Generated by GPT-2

This life is not worth living. I am going to buy a knife and I am going to kill everyone I see here, because I will be a part of a great community. It will be worth it to me.

Those produced new texts did not show empathy to the prompt text "This life is not worth living.", but produced hateful text and the tendency of anti-social and criminal behavior ("kill everyone").[9] When applying models to the high-stake domain like suicide prevention, this is worrying.

3 Intention Understanding in Sequence

The suicidal ideation could be chronic, including duration and strength at different stages. More recent works focus on identifying people with suicidal risks based on a single post or a collection of posts, i.e., post or user classification. For example, several works conducted user-level suicidal risk classification (Tsakalidis et al., 2022a). However, people's mood changes over time (Tsakalidis et al., 2022b) and the level of suicidal ideation experiences fluctuations as time goes by (Clum and Curtin, 1993). The communicative nature of social networks and the chronic nature of suicide ideations require NLP models to detect people's suicidal intentions in post sequences or over online conversations.

Recent research developed temporal-aware models to represent the sequence of social posts and predict the social user's suicidal risk. For example, historic tweets are used to improve the prediction of an individual's suicidal risk with time-aware model (Sawhney et al., 2020) and phase-aware model (Sawhney et al., 2021a). The retrospective evaluation designated for a specific period fails to monitor the dynamics of suicidal ideation. We need longitudinal suicidal risk monitoring over social posts' timelines, enabling dynamic social care and timely clinical intervention.

Figure 5 shows a pseudo case of suicide severity assessment in post sequence (similar to sequence labeling for social posts), where an individual post at each time stamp is associated with a severity scale and the severity changes when the user lives with various mental conditions in the timeline or

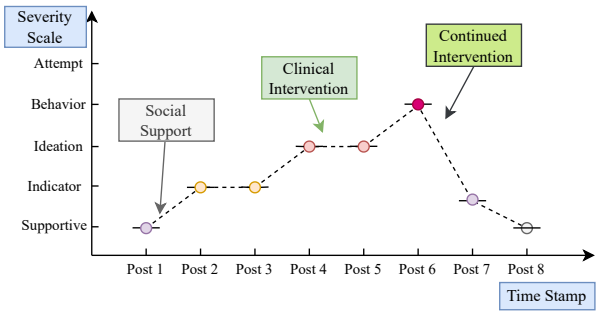


Figure 5: Severity assessment in sequence, where the suicidal risk follows the definition of Gaur et al. (2019) adapted from C-SSRS.

some kind of intervention such as social support or clinical intervention has been taken. Higher levels of suicidal risk are more critical than low risks, and different levels of suicidal risks should be allocated with different clinical resources. Dynamic detection can, in a timely manner, prioritize people needing better mental healthcare resource allocation. It can also integrate with the human-in-the-loop system for more reliable suicidal intention understanding. The ordinal relationship of the severity scale contains critical information for fine-grained classification. Sawhney et al. (2021b) approached user-level suicidal risk assessment as an ordinal regression problem. The ordinal constraints and sequence dependency between time stamps play an essential role in the sequential modeling of suicidal risk monitoring.

4 Evaluation of Suicidal Risk Assessment

An evaluation must be carefully considered when computational approaches are used in suicide research. Most publications on suicidal risk assessment or suicidal ideation detection use aggregated evaluation scores such as accuracy, F1 score, and AUC-ROC to evaluate the predictive performance. The sensitivity or recall, calculated by the ratio of true positives to the total positives, is an essential metric that needs great attention because some low-risk cases might be deprived of social care and die by suicide if a model has a low recall.

For some safety-critical applications like suicidal risk prediction, incorrect predictions may cause ethical costs and even the loss of life. Thus, uncertainty estimation becomes an important task. Dusenberry et al. (2020) analyzed the model uncertainty in medical applications of mortality and diagnosis prediction. No current suicidal ideation detection literature estimates data (aleatoric) or

[9] The whole sentence does not make sense.

model (epistemic) uncertainty. Epistemic uncertainty is required to understand examples different from training data for the life-critical task (Kendall and Gal, 2017). Intelligent intention understanding systems should be able to assign how confident is the model prediction in some erroneous predictions of fine-grained suicidal ideation.

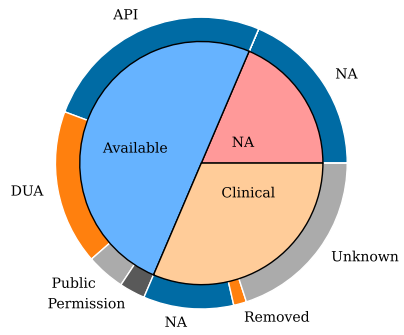
Human uncertainty and disagreement in annotation are ubiquitous. We need to learn with disagreement (Uma et al., 2021). Learning with humans-in-the-loop (Zanzotto, 2019) can also be utilized to select more reliable instances or samples with less ambiguous intentions.

5 Benchmarks

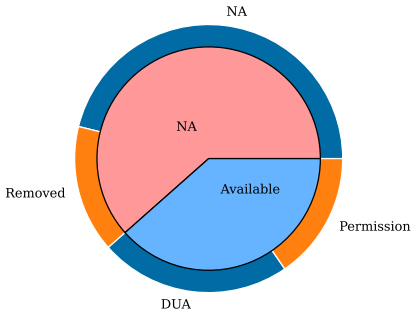
A recent survey on mental health research (Harrigian et al., 2021) reviewed studies published between January 2012 and December 2019 in conferences, workshops, and journals focusing on NLP and healthcare research. This survey investigated the availability of mental health-related text data, as shown in Figure 6a. A review on suicidal ideation detection (Ji et al., 2021) investigated thirteen published datasets, with the availability shown in Figure 6b. Due to the sensitivity of suicide data, all available datasets require data usage agreement or need to be requested with the authors' permission. More are unavailable, removed, or out of maintenance among those published datasets.

Most datasets contain binary classes (i.e., a positive class with suicide messages and a negative class from control groups). These datasets simplify the setting. The severity of suicide and suicidal ideation are fine-grained. The Scale for Suicide Ideation introduces a fine-grained suicidal risk rating scale to measure the intensity of suicide in interview-based screening (Beck et al., 1979). The Self-Monitoring Suicide Ideation Scale (Clum and Curtin, 1993) designed for self-report measures on a daily basis includes the intensity and duration of ideation and the level of control in making a suicide attempt. Check out the review by (Brown, 2001) for more suicide assessment measures. We need more datasets like Gaur et al. (2019) and more annotation about intention and ideation for special social care.

Most datasets contain self-reported posts from social websites like Reddit, where users seek peer support or express personal feelings or experiences. ScAN (Rawat et al., 2022) is a remarkable dataset of suicide events from electronic health records in



(a) Mental Health



(b) Suicide

Figure 6: The availability of benchmarks of NLP on mental health and suicide. DUA means data usage agreement.

the MIMIC III database (Johnson et al., 2016) by filtering the suicide-associated ICD codes and introducing domain experts' annotation. It consists of suicidal behavioral evidence of suicidal attempts and ideations, where suicidal ideation is defined as text mentions about self-harming or taking one's own life in clinical notes. Social posts are not as reliable as clinical notes written by clinicians. Current distantly supervised mental health models lack the generalization ability across different domains (e.g., data platforms and populations) (Harrigian et al., 2020). Moreover, to our knowledge, no datasets have been introduced from the perspective of intention understanding.

Finally, we still lack benchmarks for visual-grounded intention understanding and multilingualism. A picture of cutting the wrist can be suicidal intention or sharing self-harm behavior of other people. With social text, the visual signal can help better understand the intention. Many datasets were collected from Reddit and Weibo, which are used mainly in the US and China. Figure 7 shows that English datasets dominate the field of mental health

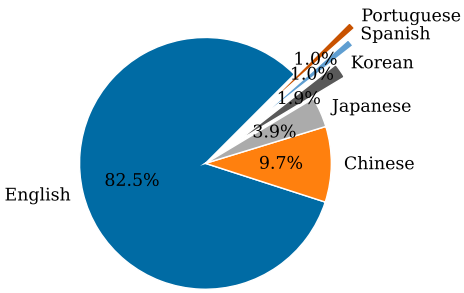


Figure 7: The availability of multilingual texts for mental health research from the survey of Harrigian et al. (2021)

research, which may cause demographic biases.

6 Additional Related Work

Suicidal detection has drawn much research attention since the suicide rate has increased these years. The cause of suicide is complicated and related to many complex interactions of risk factors (O'Connor and Nock, 2014). Feature engineering-based NLP models use N-gram features, knowledge-based features, syntactic features, context features, and class-specific feature (Wang et al., 2012). Machine learning classifiers include regression analysis (Chattopadhyay, 2007), boosting, and SVM classification (Delgado-Gomez et al., 2011). Pacula et al. (2014) proposed a model to identify signs of distress in transcripts of helpline conversations. Pestian et al. (2010) and Delgado-Gomez et al. (2012) compared the effectiveness of several multivariate methods. Other research and methods have been conducted to stop cyber-suicide such as speech pattern recognition, mobile phone network analysis, social media content detection (Larsen et al., 2015), and reply bias assessment for online suicide prevention (Huang and Bashir, 2016). However, there was not much discussion of suicide intent in those works.

Intention classification is helpful for commercial applications such as target advertising in social media (Luong et al., 2016). One of the early studies by Chen et al. (2013) formulates the intention identification from posts in multi-domain discussion forums as a binary classification task (i.e., explicit intent and non-intent posts with a specific focus on buying intention). Similarly, Gupta et al. (2014) classified purchase intention in question-and-answer websites. Wang et al. (2015) studied a multi-class intention classification on social posts on Twitter social networks and proposed the definition of the intent tweet if a tweet follows the following three conditions: 1) containing at least one verb; 2) with an explicit description of the user's intent to perform an activity; 3) in a recognizable way. However, this definition does not fit early detection of suicidal intention because the suicidal intention is usually implicit and early detection requires recognizing suicidal thoughts as early as possible before the suicidal attempt.

7 Conclusion

Intention understanding is an essential aspect of suicidal risk assessment. Robust intention understanding in suicidal risk assessment is needed to provide efficient triage support to social workers and psychiatrists. This position paper discusses suicidal intention understanding and conducts a case study on sentiment analysis and fine-tuning pretrained language models in the context of suicide research. It further points out critical aspects in the task settings and evaluation and the lack of benchmarks.

Social Impact

Early detection of suicide ideation provides a solution to early intervention so that social workers can help people living with mental health issues through proactive conversations. However, no significant evidence shows that suicidal risk assessment can guide decision-making in clinical practice (Large et al., 2017). We suggest that people experiencing a mental health condition seek professional help from psychiatric services. Research on suicidal intention understanding and risk assessment does not aim to replace psychiatrists. It can empower social workers to prioritize social resources for people with mental conditions.

The sensitive nature of suicide-related data requires our research to protect privacy. This study uses social media posts from anonymous users that are manifestly available on the website. Furthermore, these collected posts are stored on password-protected servers. We do not attempt to identify or contact social users.

Acknowledgement

Many thanks to researchers who shared their research publicly, e.g., pretrained language models, datasets, and insights. No funding agency supported this study. The author conducted this study during his free time.

Limitations

This position paper provides a perspective on suicidal intention understanding in suicidal risk assessment with natural language processing. One limitation is that it does not conduct extensive experimental analysis with real-world data. We also need to mind the gap between computational methods (e.g., inductive biases in machine learning and linguistic or distributed representation features) and the theories and findings of psychiatry. For example, a recent study shows that 60% of men who died by suicide in the US have no history of mental illness (Fowler et al., 2022). Existing NLP papers usually regard the history of mental illness as a valuable feature for machine learning models.

References

Asma Abdulsalam and Areej Alhothali. 2022. Suicidal ideation detection on social media: A review of machine learning methods. *arXiv preprint arXiv:2201.10515*.

Aaron T Beck, Maria Kovacs, and Arlene Weissman. 1979. Assessment of suicidal intention: the scale for suicide ideation. *Journal of Consulting and Clinical Psychology*, 47(2):343.

Gregory K Brown. 2001. A review of suicide assessment measures for intervention research with adults and older adults.

Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 105–114.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: a commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *LREC*.

Nicola Canessa, Federica Alemanno, Federica Riva, Alberto Zani, Alice Mado Proverbio, Nicola Mannara, Daniela Perani, and Stefano F Cappa. 2012. The neural bases of social intention understanding: the role of interaction goals. *PLoS One*, 7.

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.

Subhagata Chattopadhyay. 2007. A study on suicidal risk analysis. In *9th International Conference on e-Health Networking, Application and Services*, pages 74–78. IEEE.

Zhiyuan Chen, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Identifying intention posts in discussion forums. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1050.

George A Clum and Lisa Curtin. 1993. Validity and reactivity of a system of self-monitoring suicide ideation. *Journal of Psychopathology and Behavioral Assessment*, 15(4):375–385.

David Delgado-Gomez, Hilario Blasco-Fontecilla, AnaLucia A Alegria, Teresa Legido-Gil, Antonio Artes-Rodriguez, and Enrique Baca-Garcia. 2011. Improving the accuracy of suicide attempter classification. *Artificial Intelligence in Medicine*, 52(3):165–168.

David Delgado-Gomez, Hilario Blasco-Fontecilla, Federico Sukno, Maria Socorro Ramos-Plasencia, and Enrique Baca-Garcia. 2012. Suicide attempters classification: Toward predictive models of suicidal behavior. *Neurocomputing*, 92:3–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. 2020. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 204–213.

Katherine A Fowler, Mark S Kaplan, Deborah M Stone, Hong Zhou, Mark R Stevens, and Thomas R Simon. 2022. Suicide among males across the lifespan: An analysis of differences by known mental health status. *American Journal of Preventive Medicine*.

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

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.

Vineet Gupta, Devesh Varshney, Harsh Jhamtani, Deepam Kedia, and Shweta Karwa. 2014. Identifying purchase intent from social posts. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Rezaul Haque, Naimul Islam, Maidul Islam, and Md Manjurul Ahsan. 2022. A comparative analysis on suicidal ideation detection using nlp, machine, and deep learning. *Technologies*, 10(3):57.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24. ACL.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of Suicide Research*, 14(3):206–221.

Hsiao Ying Huang and Masooda Bashir. 2016. Online community and suicide prevention: Investigating the linguistic cues and reply bias. In *Proceedings of CHI*.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022a. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34:10309–10319.

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

Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022b. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3:160035.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Matthew Michael Large, Christopher James Ryan, Gregory Carter, and Nav Kapur. 2017. Can we usefully stratify patients according to suicide risk? *BMJ*, 359.

Mark E Larsen, Nicholas Cummins, Tjeerd W Boonstra, Bridianne O'Dea, Joe Tighe, Jennifer Nicholas, Fiona Shand, Julien Epps, and Helen Christensen. 2015. The use of technology in suicide prevention. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7316–7319. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Thai-Le Luong, Quoc-Tuan Truong, Hai-Trieu Dang, and Xuan-Hieu Phan. 2016. Domain identification for intention posts on online social media. In *Proceedings of the Seventh Symposium on Information and Communication Technology*, pages 52–57.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459.

Rory C O'Connor and Matthew K Nock. 2014. The psychology of suicidal behaviour. *The Lancet Psychiatry*, 1(1):73–85.

Maciej Pacula, Talya Meltzer, Michael Crystal, Amit Srivastava, and Brian Marx. 2014. Automatic detection of psychological distress indicators and severity assessment in crisis hotline conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4863–4867. IEEE.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*, 2010(3):19.

K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. 2008. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Wilfred R Pigeon, and Hong Yu. 2022. ScAN: Suicide Attempt and Ideation Events Dataset. In *NAACL*.

Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Waleed Alnumay, and Andrew Paul Smith. 2021. A lexicon-based approach to detecting suicide-related messages on twitter. *Biomedical Signal Processing and Control*, 65:102355.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021a. Phase: Learning emotional phase-aware representations for suicide ideation detection

on social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2415–2428.

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)*, pages 7685–7697.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021b. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 22–30.

Ramit Sawhney, Harshit Joshi, Rajiv Shah, and Lucie Flek. 2021c. 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.

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*, pages 941–950.

Vladan Starcevic and Elias Aboujaoude. 2015. Cyberchondria, cyberbullying, cybersuicide, cybersex:"new" psychopathologies for the 21st century? *World Psychiatry*, 14(1):97–100.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022a. Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts. In *Proceedings of The Eighth Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*. Association for Computational Linguistics.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Jinpeng Wang, Gao Cong, Xin Wayne Zhao, and Xiaoming Li. 2015. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Wenbo Wang, Lu Chen, Ming Tan, Shaojun Wang, and Amit P Sheth. 2012. Discovering fine-grained sentiment in suicide notes. *Biomedical Informatics Insights*, 5(Suppl 1):137.

Kirsten Windfuhr and Navneet Kapur. 2011. Suicide and mental illness: a clinical review of 15 years findings from the uk national confidential inquiry into suicide. *British Medical Bulletin*, 100(1):101–121.

Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.

Tianlin Zhang, Annika Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, 5.

Tianlin Zhang, Annika M Schoene, and Sophia Ananiadou. 2021. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet interventions*, 25:100422.