An Annotated Dataset for Explainable Interpersonal Risk Factors of Mental Disturbance in Social Media Posts

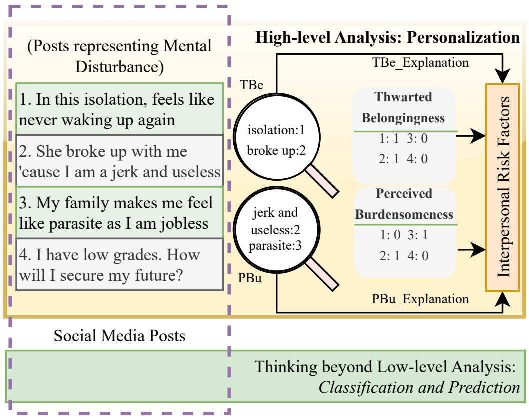
*Muskan Garg, **Amirmohammad Shahbandegan, †Amrit Chadha, **Vijay Mago
*Mayo Clinic, Rochester, MN 55901, USA
**Lakehead University, Thunder Bay, ON P7B 5E1, Canada
†Thapar Institute of Engineering & Technology, Patiala, PB 147005, India

Abstract

With a surge in identifying suicidal risk and its severity in social media posts, we argue that a more consequential and *explainable* research is required for optimal impact on clinical psychology practice and personalized mental healthcare. The success of computational intelligence techniques for inferring mental illness from social media resources, points to natural language processing as a *lens* for determining Interpersonal Risk Factors (IRF) in human writings. Motivated with limited availability of datasets for social NLP research community, we construct and release a new annotated dataset with human-labelled explanations and classification of *IRF* affecting mental disturbance on social media: (i) Thwarted Belongingness (TBe), and (ii) Perceived Burdensomeness (PBu). We establish baseline models on our dataset facilitating future research directions to develop real-time personalized AI models by detecting patterns of TBe and PBu in emotional spectrum of user's historical social media profile.

1 Introduction

The World Health Organization (WHO) emphasizes the importance of significantly accelerating suicide prevention efforts to fulfill the United Nations' Sustainable Development Goal (SDG) objective by 2030 (Saxena and Kline, 2021). Reports released in August 2021[1] indicate that 1.6 million people in England were on waiting lists for mental health care. An estimated 8 million people were unable to obtain assistance from a specialist, as they were not considered *sick enough* to qualify. As suicide remains one of the leading causes of the death worldwide[2], this situation underscores the need of mental health interpretations from social media data where people express themselves and their thoughts, beliefs/emotions with ease (Wongkoblap et al., 2022). The individuals dying by suicide hinder the psychological assessments where a self-reported text or personal writings might be a valuable asset in attempting to assess an individual's specific personality status and mind rationale (Garg, 2023). With strong motivation of thinking beyond low-level analysis, Figure 1 suggests *personalization* through higher-level analysis of human writings. As, the social media platforms are frequently relied upon as open fora for honest disclosure (Resnik et al., 2021), we examine mental disturbance in Reddit posts aiming to discover Interpersonal Risk Factors (IRF) in text.



(Posts representing Mental Disturbance)
1. In this isolation, feels like never waking up again
2. She broke up with me 'cause I am a jerk and useless
3. My family makes me feel like parasite as I am jobless
4. I have low grades. How will I secure my future?
High-level Analysis: Personalization
TBe_Explanation
TBe
isolation:1 broke up:2
jerk and useless:2 parasite:3
PBu
Thwarted Belongingness
1: 1 3: 0
2: 1 4: 0
Perceived Burdensomeness
1: 0 3: 1
2: 1 4: 0
Interpersonal Risk Factors
PBu_Explanation
Social Media Posts
Thinking beyond Low-level Analysis: Classification and Prediction

Figure 1: Overview of the problem formulation depicting the need of identifying interpersonal risk factor in texts. The texts [1-4] are annotated as 0: absence or 1: presence of the interpersonal risk factors TBe and PBu.

Interpersonal relationships are the strong connections that a person with their closest social circle (peers, intimate-partners and family members) which can shape an individual's behavior and range of experience (Puzia et al., 2014). Affecting such interpersonal relationships influences the associated risk factors resulting in mental disturbance. According to *interpersonal-psychological theory of suicidal behavior* (Joiner et al., 2005), suicidal

[1] https://www.theguardian.com/society/2021/aug/29/strain-on-mental-health-care-leaves-8m-people-without-help-say-nhs-leaders

[2] https://news.un.org/en/story/2021/06/1094212

Findings of the Association for Computational Linguistics: ACL 2023, pages 11960–11969
July 9-14, 2023 ©2023 Association for Computational Linguistics

Dataset	Media	Size	Exp.	Task	Avail.
(Kivran-Swaine et al., 2014)	Twitter	4454	×	Responses to expressions of loneliness	No
(Badal et al., 2021)	Interviews	97 adults	×	Isolation and loneliness in older adults	No
(Mahoney et al., 2019)	Twitter	22477	×	Loneliness disclosures throughout the day	No
(Ghosh et al., 2022)	Suicide Notes	350 notes	×	TBE and PBU in Suicide Notes	OR
Ours	Reddit	3522	YES	Explainable TBE and PBU in Social Media Posts	YES

Table 1: Historical evolution of language resources for classifying lonesomeness in texts. OR: On Request, TBE: Thwarted Belongingness and PBU: Perceived Burdensomeness

desire arises when a person experience persistent emotions of (i) Thwarted Belongingness (TBE)[3], and (ii) Perceived Burdensomeness (PBU)[4]. As a starting point for our research, this cross-sectional study facilitates the language resource for discovery of underlying users with prospective self-harm/suicidal tendencies to support and compliment existing literature (Bialer et al., 2022; Tsakalidis et al., 2022; Gaur et al., 2018) as intrinsic classification task.

Computational approaches may better understand the technological advancements in psychology research, aiding the early detection, prediction and evaluation, management and follow-up of those experiencing suicidal thoughts and behaviors. Most automated systems require available datasets for computational advancements. Past studies show that the availability of relevant datasets in mental healthcare domain is scarce for IRF due to sensitive nature of data as shown in Table 1 (Su et al., 2020; Garg, 2023). To this end, we introduce an annotated Reddit dataset for classifying TBE and PBU. The explanatory power of this dataset lies in supporting the motivational interviewing and mental health triaging where early detection of potential risk may trigger an alarm for the need of a mental health practitioner. We adhere to ethical considerations for constructing and releasing our dataset publicly on Github[5].

2 Dataset

2.1 Corpus Construction

Haque et al. (2021) used two subreddits $r/depression$ and $r/suicidewatch$ to scrape the SDCNL data and to validate a label correction methodology through manual annotation of this dataset for *depression* versus *suicide*. They addressed the then existing ethical issues impacting *dataset availability* with public release of their dataset. In addition to 1896 posts of SDCNL dataset, we collected 3362 additional instances from Reddit on $r/depression$ and $r/SuicideWatch$ through PRAW API[6] from 02 December 2021 to 04 January 2022 with about 100 data points per day (to maintain variation in the dataset). On initial screening, we found (i) posts with no self-advocacy, (ii) empty/irrelevant posts. We manually filter them to deduce self-advocacy in texts leveraging 3155 additional samples, which results in a total of 5051 data points (Garg et al., 2022). We removed 694 of the data points depicting no assessment of mental disturbance. Moreover, people write prolonged texts when they indicate IRF which is inline with the conventional arguments where prolonged remarks get better responses from others in comparison of the transient remarks (Park et al., 2015). The length of real-time Reddit posts varies from a few characters to thousands of words. We limit the maximum length of every post to 300 words resulting in 3522 posts as a final corpus.

2.2 Annotation Scheme

Classification of IRF, being a complex and highly subjective task, may induce errors with naive judgment. To mitigate this problem, we build a team of three experts: (i) *a clinical psychologist* for training annotators and validating annotations with psychological viewpoint, (ii) *a rehabilitation counselor* for comprehending human mind to understand users' IRF, and (iii) *a social NLP expert* suggesting text based markings in Reddit posts. To negotiate and mitigate the trade-off between three different perspectives, our experts build annotation guidelines[7] to mark (i) TBE, and (ii) PBU. The experts annotated 40 samples of the corpus in isolation using these annotation guidelines to avoid

[3] An unpleasant emotional response to distinguished isolation through mind and character

[4] Characterized by apperceptions that others would 'be better off if I were gone,' underlying unwelcoming society

[5] https://github.com/drmuskangarg/Irf

[6] https://praw.readthedocs.io/en/stable/

[7] Please see the Annotation Guidelines in Appendix B.

biases and discover possible dilemmas due to the subjective nature of tasks. Therefore, we accommodate perplexity guidelines to simplify the task and facilitate unbiased future annotations.

1. **TBe or PBu in the Past**: To check if the condition of a person with disconnected past is still alarming prospect of self-harm or suicidal risk. For instance, '*I was so upset being lonely before Christmas and today I am celebrating New Year with friends*'. We frame rules to handle risk indicators about the past because a person attends celebration and overcome the preceding mental disturbance which means filling void with external event. With neutral opinion by NLP expert about double negation, our clinical psychologist argues presence of risk in their perception which may again evolve after some time and thus, marks this post with presence of the TBe.

2. **Ambiguity with *Social Experiences***: Relationships point to the importance of the ability to take a societal pulse on a regular basis, especially in these unprecedented times of pandemic-induced distancing and shut-downs. People mention major societal events such as breakups, marriage, best friend related issues in various contexts suggesting different user perceptions. We mitigate this problem with two statements: (i) Any feeling of void/missing/regrets/or even mentioning such events with negative words should be marked as presence of TBe such as consider this post: *'But I just miss her SO. much. It's like she set the bar so high that all I can do is just stare at it.'*, (ii) Anything associated with fights/quarrels/general stories should be marked with absence of TBe such as consider the post: *'My husband and I just had a huge argument and he stormed out. I should be crying or stopping him or something. But I decided to take a handful of benzos instead.'*

2.3 Annotation Task

Three postgraduate students underwent eight hours of professional training by a senior clinical psychologist leveraging annotation and perplexity guidelines. After three successive trial sessions to annotate 40 samples in each round, we ensured their alignment on interpreting task requirements and deployed them for annotating all data points in the corpus. We obtain final annotations based on the majority voting mechanism for binary classification task <TBe, PBu>.[8] We validate three annotated files using Fliess' Kappa inter-observer agreement study on classifying TBe and PBu where kappa is calculated as 78.83% and 82.39%, respectively.

Criteria	Absent	Present
Thwarted Belongingness		
Number of Posts	1595	1927
Avg. #(Words)	134.68	132.58
Avg. #(Sentences)	7.73	7.61
Max. number of Sentences	49	49
Avg. #(Words) in Explanations	-	3.45
Perceived Burdensomeness		
Number of Posts	2375	1147
Avg. #(Words)	132.98	136.54
Avg. #(Sentences)	7.65	7.79
Max. number of Sentences	49	32
Avg. #(Words) in Explanations	-	4.04

Table 2: The statistics of Reddit dataset to determine presence or absence of TBe and PBu and its explanation.

Furthermore, we carry out an inter-annotator agreement study with group annotations[9] for text-spans extraction in positive data points. The results for agreement study in two-fold manner: (i) 2 categories (agree, disagree) and (ii) 4 categories (strongly agree, weakly agree, weakly disagree, strongly disagree), are obtained as 82.2% and 76.4% for agreement study of <TBe_exp>, and 89.3% and 81.3% for agreement study of <PBu_exp>, respectively.

2.4 Dataset Statistics

On observing the statistics of our dataset in Table 2, we found 54.71% and 32.56% of positive data points with underlying 255489 and 156620 words for TBe and PBu, respectively. It is interesting to note that although the average number of sentences to express PBu is less than TBe, the observations are different for average number of words. We calculate the Pearson Correlation Coefficient (PCC) for our cross-sectional study on TBe and PBu as 0.0577 which shows slight correlation between the two. Our dataset paves the way for longitudinal studies which is expected to witness increased PCC due to wide spread emotional spectrum (Kolnogorova et al., 2021; Harrigian et al., 2020). On

[8] Sample of dataset is given in Appendix A.

[9] A group of three student annotators extracting explanations and generating a final lists of explanations for TBe as <TBe_exp> and for PBu as <PBu_exp>

Table 3: Comparison of SOTA baseline models' performance

Model	Thwarted Belongingness				Perceived Burdensomeness			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
LSTM	61.40	92.77	72.00	63.67	44.65	80.90	54.69	62.35
GRU	63.57	91.26	**73.06**	**66.70**	60.87	74.77	**63.75**	**78.90**
BERT	69.70	76.97	72.30	68.97	56.47	53.00	52.20	72.56
RoBERTa	71.23	73.54	71.35	68.97	67.27	37.52	45.51	74.93
DistilBERT	70.24	74.08	71.15	68.50	51.15	31.89	36.93	71.71
MentalBERT	77.97	77.40	**76.73**	**75.12**	64.22	65.75	**62.77**	**78.33**
OpenAI+LR	79.00	83.59	**81.23**	78.62	82.66	63.08	71.55	84.58
OpenAI+RF	79.06	80.68	79.86	77.48	83.33	49.23	61.90	81.36
OpenAI+SVM	81.31	80.34	80.83	**78.90**	79.15	74.77	**76.90**	**86.19**
OpenAI+MLP	81.40	75.56	78.37	76.92	72.08	77.85	74.85	83.92
OpenAI+XGB	81.22	79.83	80.52	78.62	80.36	68.00	73.67	85.05

Model	Task	P	R	F1
LIME	TBe	14.24	53.05	20.88
	PBu	18.47	46.83	25.18
SHAP	TBe	15.74	50.16	22.27
	PBu	20.77	49.89	27.92

Table 4: Performance Evaluation of explanations of MentalBERT model through LIME and SHAP.

changing TBe from absence to presence, we observe high rate of increase in positive data points of PBu (((675 - 472)/472) which is 43.00%) as compared to the absence of PBu (((1252-1123)/1123) which is 11.48%) suggesting the probability of high correlation in the presence of TBe and PBu, respectively which are given in Table 5.

	PBu: 0	PBu: 1
TBe: 0	1123	472
TBe: 1	1252	675
%Δ	129/1123 = 0.1148	203/472 = 0.4301

Table 5: Dataset statistics for Thwarted Belongingness and Perceived Burdensomeness.

The most frequent words for identifying (i) TBe are *alone, lonely, nobody to talk, someone, isolated, lost*, and (ii) PBu are *die, suicide, suicidal, kill, burden, cut myself.*[10] Our approach for identifying TBe and PBu goes beyond a simple keyword detector. Instead, we utilize a more sophisticated method that considers the context and relationships between words. For instance, consider a following sample:

> Massive party at a friend's house- one of my closest friends is there, loads of my close friends are there, i wasn't invited. wasn't told. only found out on snapchat from their stories. spending new years eve on teamspeak muting my mic every time i break down :)

Despite the absence of trigger words, our approach flags this post as positive for TBu based on its indicators 'friend', 'teamspeak', 'friends', 'invited', 'snapchat', to name a few.

3 Experiments and Evaluation

3.1 Baselines

We perform extensive analysis to build baselines with three different conventional methods. We first apply **Recurrent neural networks** where a given text, embedded with GloVe 840B-300[11], is sent to a 2-layer RNN model (LSTM, GRU) with 64 hidden neurons and the output is forwarded to two separate fully connected heads: (i) TBe and (ii) PBu. Each of the fully connected blocks have one hidden layer with 16 neurons and ReLU activation function, and an output layer with sigmoid activation. The loss function is *Binary_CrossEntropy* and optimizer is *adam* with $lr = 0.001$. Next, we apply **pretrained transformer-based models**. The input is tokenized using a pre-trained transformers' tokenizer to obtain a 768-dimensional vector which is then fed to a similar fully connected network as the previous architecture with hidden layer size as 48. We experimented with *roberta-base, bert-base-uncased, distilbert-base-uncased, and mental/mental-bert-base-uncased* models. Finally, we use the **Ope-**

[10] WordCloud is given in Appendix C.

[11] https://nlp.stanford.edu/projects/glove/

nAI embeddings API[12] to convert the input text into 1536-dimensional embeddings through '*text-embedding-ada-002*' engine which are used to train a classifier. We test the robustness of this approach over: (i) Logistic Regression, (ii) Random Forest, (iii) Support Vector Machine (iv) Multi Layer Perceptron, and (v) XGBoost. We further use two explainable methods: (i) **LIME** and (ii) **SHAP** on one of the best performing transformer-based models, MentalBERT (Ji et al., 2022), to obtain the top keywords (Danilevsky et al., 2020; Zirikly and Dredze, 2022). We compare them with the ground truth ROUGE scores for – Precision (P), Recall (R), and F1-score (F).

4 Experimental Settings

For consistency, we used the same experimental settings for all models and split the dataset into the train, validation, and test sets. All results are reported on the test set, which makes up 30% of the whole dataset. We used the grid search optimization technique to optimize the parameters. To tune the number of layers (n), we empirically experimented with the values: learning rate (lr): lr $\in \{0.001, 0.0001, 0.00001\}$ and optimization (O): O $\in$ {'Adam', 'Adamax', 'AdamW'} with a batch-size of 16, 32 were used. We used base version pre-trained language models (LMs) using HuggingFace[13], an open-source Python library. We used optimized parameters for each baseline to find precision, recall, F1-score, and Accuracy. Varying lengths of posts are padded to 256 tokens with truncation. Each model was trained for 20 epochs, and the best-performing model based on the average accuracy score was saved. Thus, we set hyperparameter for our experiments as *Optimizer* = Adam, learning rate = 1e-3, batch size= 16, and epochs=20.

4.1 Experimental Results

Table 3 shows the performance of state-of-the-art methods in terms of precision, recall, F1-score, and accuracy. The current models have moderately low performance in this task, possibly due to a lack of ability to capture contextual information in the text. MentalBERT, a transformer-based language model, initialized with BERT-Base and trained with mental health-related posts collected from Reddit, had the best performance among BERT-based models, with an F1-score of 76.73% and 62.77% for TBe and PBu, respectively. This is likely due to the fact that it was trained on the same context as the task, namely health-related posts on Reddit. The combination of OpenAI embeddings and a classifier outperforms RNN and transformer-based models. The highest F1-Score of 81.23% was achieved by logistic regression for TBe, while the best performing model for PBu was SVM with an F1-score of 76.90%. We also analyzed the explainability of the model using LIME and SHAP methods of explainable AI for NLP on the best performing transformer model (MentalBERT) for TBe and PBu. We obtain results for all positive data points in the testing dataset and observe high recall of text-spans with reference to the ground truth as shown in Table 4. We find the scope of improvement by limiting the superfluous text-spans found in the resulting set of words. The consistency in results suggests the need of contextual/domain-specific knowledge and infusing commonsense to improve explainable classifiers for a given task.

5 Conclusion and Future Work

We present a new annotated dataset for discovering interpersonal risk factors through human-annotated extractive explanations in the form of text-spans and binary labels in 3522 English Reddit posts. In future work, we plan to enhance the dataset with more samples and develop new models tailored explicitly to TBe and PBu. The implications of this work include the potential to improve public health surveillance and other mental healthcare applications that rely on automatically identifying posts in which users describe their mental health issues. We keep the implementation of explainable AI models for multi-task text classification, as an open research direction for Open AI and other newly developed responsible AI models. We pose the discovery of new research directions for future, through longitudinal study on users' historical social media profile to examine interpersonal risk factors and potential risk of self-harm or suicidal ideation. As we focus on Reddit data as a starting point of our study, exploring other forums could be an interesting research direction.

Acknowledgement

We express our gratitude to Veena Krishnan, a senior clinical psychologist, and Ruchi Joshi, a reha-

[12] https://beta.openai.com/docs/guides/embeddings/embedding-models

[13] https://huggingface.co/models

bilitation counselor, for their unwavering support throughout the project. Additionally, we extend our heartfelt appreciation to Prof. Sunghwan Sohn for his consistent guidance and support. This project was partially supported by NIH R01 AG068007. This project is funded by NSERC Discovery Grant (RGPIN-2017-05377), held by Vijay Mago, Department of Computer Science, Lakehead University, Canada.

Limitations

There might be linguistic discrepancies between Reddit users and Twitter users who post about their mental disturbance on social media. Social media users may intentionally post such thoughts to gain attention of other social media users but for simplicity, we assume the social media posts to be credible. Thus, we assume that the social media posts are not misleading. We acknowledge that our work is subjective in nature and thus, interpretation about wellness dimensions in a given post may vary from person to person.

Ethical Considerations

The dataset we use is from Reddit, a forum intended for anonymous posting, users' IDs are anonymized. In addition, all sample posts shown throughout this work are anonymized, obfuscated, and paraphrased for user privacy and to prevent misuse. Thus, this study does not require ethical approval. Due to the subjective nature of annotation, we expect some biases in our gold-labeled data and the distribution of labels in our dataset. Examples from a wide range of users and groups are collected, as well as clearly defined instructions, in order to address these concerns. Due to high inter-annotator agreement (κ score), we are confident that the annotation instructions are correctly assigned in most of the data points. It is reproducible with the dataset and the source code to reproduce the baseline results which is available on Github.

To address concerns around potential harms, we believe that the tool should be used by professionals who are trained to handle and interpret the results. We recognize the huge impact of false negatives in practical use of applications such as mental health triaging, and we shall continue working towards improving its accuracy and reducing the likelihood of false negatives. We further acknowledge that our work is empirical in nature and we do not claim to provide any solution for clinical diagnosis at this stage.

References

Varsha D Badal, Camille Nebeker, Kaoru Shinkawa, Yasunori Yamada, Kelly E Rentscher, Ho-Cheol Kim, and Ellen E Lee. 2021. Do words matter? detecting social isolation and loneliness in older adults using natural language processing. *Frontiers in Psychiatry*, 12.

Amir Bialer, Daniel Izmaylov, Avi Segal, Oren Tsur, Yossi Levi-Belz, and Kobi Gal. 2022. Detecting suicide risk in online counseling services: A study in a low-resource language. In *Proceedings of the 29th International Conference on Computational Linguistics COLING*, pages 4241–4250.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, pages 1–24.

Muskan Garg, Chandni Saxena, Veena Krishnan, Ruchi Joshi, Sriparna Saha, Vijay Mago, and Bonnie J Dorr. 2022. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. In *Language Resources Evaluation Conference (LREC)*.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. " let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 753–762.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Am i no good? towards detecting perceived burdensomeness and thwarted belongingness from suicide notes. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5073–5079. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *International Conference on Artificial Neural Networks*, pages 436–447. Springer.

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.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190. European Language Resources Association (ELRA).

Thomas E Joiner et al. 2005. *Why people die by suicide*. Harvard University Press.

Funda Kivran-Swaine, Jeremy Ting, Jed Brubaker, Rannie Teodoro, and Mor Naaman. 2014. Understanding loneliness in social awareness streams: Expressions and responses. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 256–265.

Kateryna Kolnogorova, Nicholas P Allan, Shahrzad Moradi, and Tracy Stecker. 2021. Perceived burdensomeness, but not thwarted belongingness, mediates the impact of ptsd symptom clusters on suicidal ideation modeled longitudinally. *Journal of Affective Disorders*, 282:133–140.

Jamie Mahoney, Effie Le Moignan, Kiel Long, Mike Wilson, Julie Barnett, John Vines, and Shaun Lawson. 2019. Feeling alone among 317 million others: Disclosures of loneliness on twitter. *Computers in Human Behavior*, 98:20–30.

Sungkyu Park, Inyeop Kim, Sang Won Lee, Jaehyun Yoo, Bumseok Jeong, and Meeyoung Cha. 2015. Manifestation of depression and loneliness on social networks: a case study of young adults on facebook. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 557–570.

Megan E Puzia, Morganne A Kraines, Richard T Liu, and Evan M Kleiman. 2014. Early life stressors and suicidal ideation: Mediation by interpersonal risk factors. *Personality and Individual Differences*, 56:68–72.

Philip Resnik, April Foreman, Michelle Kuchuk, Katherine Musacchio Schafer, and Beau Pinkham. 2021. Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior*, 51(1):88–96.

Shekhar Saxena and Sarah Kline. 2021. Countdown global mental health 2030: data to drive action and accountability. *The Lancet Psychiatry*, 8(11):941–942.

Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. 2020. Deep learning in mental health outcome research: a scoping review. *Translational Psychiatry*, 10(1):1–26.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2022. Social media big data analysis for mental health research. In *Mental Health in a Digital World*, pages 109–143. Elsevier.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39.

A Sample Dataset

The sample dataset is given in Table 6.

B Annotation Guidelines

We follow The Interpersonal Needs Questionnaire (INQ) in association with our experts to set required guidelines. According to the Baumeister and Leary (1995) theory of the *need to belong*, **thwarted belongingness** (TBE) is a psychologically-painful mental state that results from inadequacy of connectedness. It contains detailed set of instructions to mark latent feeling of disconnectedness, missing someone, major event such as death, or being ignored/ostracized/alienated, as TBE.

Marking:
0: No Thwarted Belongingness
1: Thwarted Belongingness present

Perceived burdensomeness (PBU) is a mental state characterized by making fully conscious perception that others would "be better off if I were gone," which manifests when the need for social competence. The Self-Determination Theory (Ryan & Deci, 2000) proposes the association of family discord, unemployment, and functional impairment with suicide across the lifespan. Detailed set of instructions were given to mark the major feeling of *being a burden on other people and/or society*, as PBU.

Marking:
0: No Perceived Burdensomeness
1: Perceived Burdensomeness present

TBE and PBU are the most proximal mental states that precede the development of thoughts of suicide—stressful life events, mental disorders,

Table 6: A sample of dataset to examine interpersonal risk factors and their explanations for mental health problems

Text	TBe	TBe_exp	PBu	PBu_exp
To be rather blunt, I'm single, stuck living with parents and working shitty hours. I don't have any friends, I've never been in a proper, loving relationship and I'm a socially awkward loser. Other people see me as a burden, people hate talking to me, and I'm tired of continuing on with this. It's been 10 years since this mess started, do I not deserve a life worth living?	1	Social awkward	1	See me as a burden
I have lost around 8 friends over the past two years. They leave without even saying goodbye. It's literally just my personality. I'm a "downer" apparently. I'm scared that I'll be alone forever. Should I change so that someone will like me?	1	Alone forever	0	-
I'm having thoughts about killing myself to escape all of this. Its the most dumb thing to do but i feel like im running out of choices. We're not financially stable. I'm a student. I should have wore a condom. What should i do.	0	-	1	killing myself
I only take Lexapro. I was watching some videos on these guy that call themselves "Preppers" and they prep for the end of the world. They say that people on any types of drugs will become unstable and focused on getting their fix or whatever. Is that us?	0	-	0	-

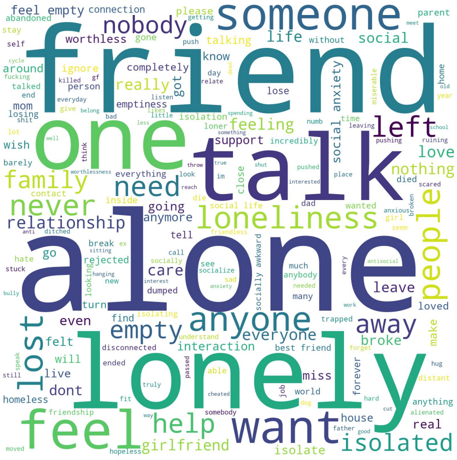


Figure 2: Wordcloud for Thwarted Belongingness

and other risk factors for suicide are relatively more distal in the causal chain of risk factors for suicide. These IRF are posited to be dynamic and amenable to therapeutic change.

C Word Frequency in Explanations

The wordcloud for explanations are shown in Figures 2 and 3.

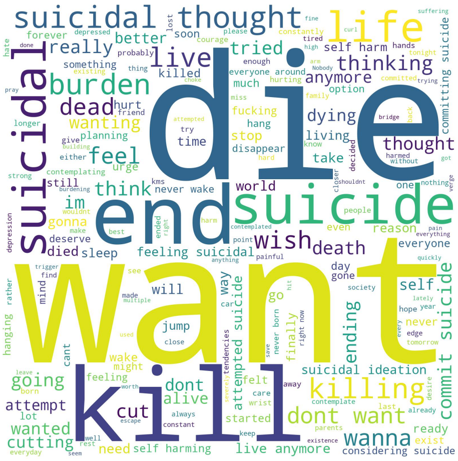


Figure 3: Wordcloud for Perceived Burdensomeness

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.