The uncivil empathy: Investigating the relation between empathy and toxicity in online mental health support forums

Ming-Bin Chen and **Jey Han Lau** and **Lea Frermann**
{mingbin, laujh, lfrermann}@unimelb.edu.au

Abstract

WARNING:This paper contains content related to suicide and self-harm. We explore the relationship between empathy and toxicity in the context of online mental health forums. Despite the common assumption of a negative correlation between these concepts (Lahnala et al., 2022), it has not been empirically examined. We augment the EPITOME mental health empathy dataset (Sharma et al., 2020) with toxicity labels using two widely employed toxic/harmful content detection APIs: Perspective API and OpenAI moderation API. We find a notable presence of toxic/harmful content (17.77%) within empathetic responses, and only a very weak negative correlation between the two variables. Qualitative analysis revealed contributions labeled as empathetic often contain harmful content such as promotion of suicidal ideas. Our results highlight the need for reevaluating empathy independently from toxicity in future research and encourage a reconsideration of empathy's role in natural language generation and evaluation.

1 Introduction

Natural Language Processing (NLP) technology has been instrumental in both the analysis and enhancement of online discussions, as exemplified by its application in platforms like Reddit (Medvedev et al., 2019). Specifically, the detection of toxicity in online comments has emerged as a widely embraced preventive measure for moderating online discussions (Lees et al., 2022). On the other hand, in recent years, there has been a surge of research interest in NLP on empathy (Raamkumar and Yang, 2022), due to its critical role in human communication and relationship building (Muradova, 2021; Sharma et al., 2020).

In the realm of online public discourse analysis, both toxicity and empathy are frequently studied and discussed within the broader context of civility (Friess and Eilders, 2015). While toxicity is typically characterized as a form of uncivil behavior, empathy is associated with civil interactions that contribute to pro-social outcomes. Some research in the field of NLP has made unexamined implicit assumptions based on this conceptual contrast. One such assumption posits a negative correlation between empathy and toxicity (Lahnala et al., 2022; Oswald, 2023). On the other hand, studies from psychology hold mixed views regarding the relation between the two concepts (Moyers and Miller, 2013; Breithaupt, 2018). While the effect and roles of both toxicity and empathy are complex, developing technology founded on unexamined assumptions entails the risk of unforeseen consequences.

This study analyses the correlation between toxicity and empathy using the human annotated empathy labels of EPITOME, a widely used mental health subreddit empathetic dataset (Sharma et al., 2020), and augmenting it with toxicity labels predicted by two popular APIs. We conduct a qualitative analysis of EPITOME responses which are both empathetic and predicted as toxic. Our key findings and contributions are:

1. 17.77% of human-identified empathetic responses classified as toxic/harmful by APIs.[1]
2. Contrary to intuition, no strong negative correlation found between API predicted toxic/harmful labels and human annotated EPITOME empathetic labels.
3. Qualitative analysis reveals presence of suicidal ideation and the widespread unhelpful responses, suggesting potential risks in finetuning empathetic language generation with EPITOME dataset.

2 Related Work

Toxicity is generally defined as language that is harmful, offensive, or suppressing the expression of others (van Aken et al., 2018). While earlier tox-

[1] We validated the quality of predictions in Appendix C.

Proceedings of The 21st Annual Workshop of the Australasian Language Technology Association, pages 136–147
November 29 - December 1, 2023 ©2023 Association for Computational Linguistics

icity detection tasks primarily focused on binary classification (Dixon et al., 2018), more recent studies have shifted towards incorporating more specific fine-grained labels, such as personal attacks (Wulczyn et al., 2017), hate speech (Hartvigsen et al., 2022) and many more (Price et al., 2020). Recent developments also encompass toxic span detection (Pavlopoulos et al., 2021) and implicit, context-dependent toxicity detection (Hartvigsen et al., 2022; Anuchitanukul et al., 2022).

Some of these advancements have transitioned into production as public APIs, such as the Perspective API (Jigsaw, 2023), and find practical use not only in everyday applications like online forum moderation but also in research fields beyond computer science, such as political science. However, some concerns have been raised regarding the potential inconsistency and oversimplification in the underlying definitions of toxicity within the detection models (Fortuna et al., 2020).

Driven by the interest in developing more engaging and supportive AI agents, empathy has emerged as a prominent theme in recent NLP research (Raamkumar and Yang, 2022). Earlier research on empathy primarily focused on emotional understanding and reactions, whereas recent works delve into the cognitive dimensions of empathy, including perspective-taking (Kim et al., 2021). While numerous studies aim to generate empathetic responses resembling human ones, few concentrate on automated empathy detection. This trend can be attributed, in part, to empathy's diverse definitions, spanning various fields such as cognitive neuroscience and psychology (Singer and Lamm, 2009; Cuff et al., 2016). The EPITOME dataset (Sharma et al., 2020) stands out as the sole dataset to not only label empathy levels but also annotate empathy across three distinct components: Emotion Reaction (ER), Interpretation (IP), and Exploration (EX), encompassing both emotional and cognitive aspects of empathy.

3 Methodology

In this study, we use the sub-reddit version of the EPITOME dataset, which was sourced from 55 mental health focused subreddits. The dataset includes 3081 pairs of support seeker post and peer support response. Each response message is human annotated with the levels (None: 0, Weak: 1, Strong: 2) of the three empathetic components (ER, IP, EX). Appendix A covers the detailed definitions and annotation level criteria.

We use two widely-used APIs for harmful and toxic online content detection, the Perspective API (Jigsaw, 2023) and OpenAI's moderation API (OpenAI, 2023). Perspective API is provided by Google for online content moderation. The underlying models of the API are trained on online comment labels from a variety of sources, like Wikipedia. Given an input message, the API returns continuous scores (0-1) for 6 different toxicity categories. Besides the score, the API also returns the detected toxic spans for each corresponding category.

OpenAI's moderation API was developed primarily for moderating the input and output of their flagship large language model ChatGPT. With less emphasis on toxicity per se, the API is designed to detect harmful and dangerous content. For each input message, it returns an overall binary flag (0,1) and 11 continuous category scores (0-1). Appendix B contains the detailed definitions for the labels of both APIs.

Using both APIs, we (automatically) annotate the *peer support responses* in the EPITOME datasets with toxicity labels. We are primarily interested in empathy/toxicity in EPITOME peer responses, and so feed only the responses into the APIs. To clarify, we do not include the support seeker post as part of the input, and so the classification is done using only the response. Both quantitative and qualitative analysis have been conducted based on the scores from the APIs along with the human annotated EPITOME labels.

To validate the predictions of the two APIs and ensure the validity of this study, we conducted manual annotation on 50 positive (labeled as toxic by at least one API) and 50 negative samples (not labeled as toxic by either API), resulting 0.87 accuracy (details in Appendix C), suggesting that the toxicity predictions are generally reliable. To provide a qualitative understanding on these predictions, we conducted an error analysis and identified that the predominant error cases (12 out of 13) were false positive errors. These errors were largely attributed to the predicted self-harm labels from the OpenAI's moderation API (10 out of the 12), while a smaller subset were related to profanity use (2 out of 12). Upon further qualitative analysis, we identified that the error cases frequently featured lengthy content with mixed intentions. For instance, these cases often began with individuals sharing their own suicidal thoughts or experiences

Label	Frequency	Percentage
IP(E)	1458	47.32
ER(E)	1047	33.98
EX(E)	480	15.58
Profanity(P)	315	10.22
Toxicity(P)	294	9.54
Self-harm(O)	133	4.32
Self-harm/intent(O)	124	4.02
Insult(P)	61	1.98
Harassment(O)	45	1.46
Threat(P)	35	1.14
Violence(O)	25	0.81

Table 1: Frequency and % contining posts of labels from the three label groups with frequency > 20.

Label groups	Count
E(EPITOME)	2381
P (Perspective API)	379
O (OpenAI moderation)	248
$E \cap P$	288
$E \cap O$	203
$E \cap P \cap O$	68
$E \cap (P \cup O)$	423
Total	3081

Table 2: The frequency and intersection frequency of the three labels groups.

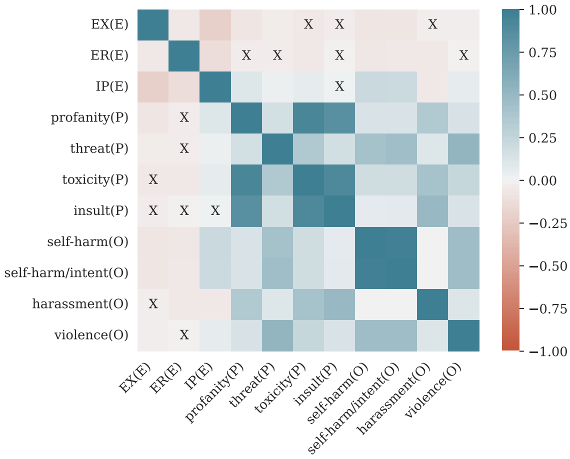

Figure 1: The Pearson correlation between labels from the three label groups. "X" indicates insignificant correlation (p-value > 0.05).

but subsequently shifted towards discouraging suicide. We also observed that the APIs can at times exhibit oversensitivity to the presence of specific keywords like “suicide”, “depression” and “shit”, even when these terms are used with the goal of emphasis rather than offence.

4 Quantitative Analysis

We first analyse the frequency distribution of each toxic/harmful label and its correlation with the EPITOME component levels. In addition, we inspect the difference between the empathetic toxic text and the non-empathetic toxic text. Hereafter, for convenience, we will group the labels into three label groups based on their sources, which are E (EPITOME), P (Perspective API), and O (OpenAI moderation API).

4.1 Frequency Analysis

Table 1 presents the label-level frequency distribution across the three label groups. To calculate these frequencies, we converted the continuous confidence scores within the P and O label groups into binary values, considering any score greater than 0.5 as positive. For labels within group E, we marked both weak (1) and strong level (2) as presence.

Within the Perspective API labels, “profanity” and “toxicity” are the two most frequently occurring labels. Conversely, the OpenAI moderation API primarily identifies “self-harm” and “self-harm/intent” as the most frequent labels. Furthermore, the rarer, more severe forms of toxic or harmful speech labels, such as “hate”, “severe toxicity” and “identity attack”, exhibit frequencies below 20 ($<< 1\%$ of instances) and as such are excluded from our experiments.

Table 2 displays the group-level frequencies for each group individually as well as their intersection. We consider a group label as present if at least one label within the group was labeled as positive. Overall, we observe a notable presence of toxic or harmful labels within empathetic instances.

4.2 Correlation Analysis

Figure 1 illustrates the Pearson correlation among the labels from the three distinct groups. Based on prior work, we would expect (a) positive correlations between labels from the two toxicity APIs and (b) negative correlations between toxicity and empathy labels. Indeed, we note weak to moderate levels of positive correlations observed between the labels from the two APIs (P and O groups). However, the correlations between empathy (E) and

Type	N	% Toxic	Length
All	3084	8.07	47.01
$E \cap (P \cup O)$	423	50.86	58.45
$\neg E \cap (P \cup O)$	109	70.18	34.69

Table 3: Comparing frequency (N), toxic coverage (% toxic), and response length in tokens between empathetic and non-empathetic toxic harmful responses.

Label	Empathetic	Toxic	Helpful
Percentage	88%	74%	26%

Table 4: Summary statistics of the quality analysis with manual annotation.

toxicity (P, O) labels exhibit a mixed pattern, comprising both insignificant correlations (indicated with an 'x' in Figure 1) as well as small but significant positive and negative correlations. While EX and ER show some significant but weak negative correlation with some toxic/harmful labels, IP has weak positive correlation with most of the toxic/harmful labels. Overall, this mixed pattern does not fully align with the common assumption of negative correlation — or, in other words, that the presence of empathy suggests a lack of toxicity and vice versa.

4.3 Toxicity in Empathetic and Non-empathetic Responses

To explore the factors contributing to the toxicity or harm in empathetic responses, we compared between empathetic toxic/harmful ($E \cap (P \cup O)$) and non-empathetic toxic/harmful ($\neg E \cap (P \cup O)$) responses. We used the Perspective API to identify toxic spans and estimate the fraction of toxic language in a response. The results, as shown in Table 3, indicate that empathetic toxic/harmful responses exhibit substantially lower fractions of toxic language, and are generally longer compared to their non-empathetic counterparts.

5 Qualitative Analysis

To better understand the interplay between empathy and toxic/harmful characteristics, we selected a subset of the top 50 samples that exhibited high levels of empathy while also being associated with either of the toxic/harmful group labels ($E \cap (P \cup O)$), and performed another manual annotation. Here we collapsed the fine-grained labels of EPITOME and both APIs categories into two binary labels "empathetic" and "toxic", and included a third class, "helpful" (also binary), to evaluate whether the responses has pragmatic benefit to the seekers. We define "helpful" as *comments or content that have the intention or potential to help/improve the future situation or lessen the negativity of the seeker physically, mentally or emotionally.* Full definitions of all three classes are given in appendix D. The motivation for introducing the "helpful" class is to fill the gap in the current EPINOME annotations, which lack a metric for measuring the desired outcome or utility. In the context of mental health support, we propose the perception of "helpfulness" serves as a proxy for the desired outcome. For this exercise, the first author of this paper annotated all 50 samples.

Table 4 displays the distribution of the three classes in the 50 samples. We see high levels of "empathetic" and "toxic" instances, aligning with the original EPITOME and API annotations (recall that these samples are drawn from $E \cap (P \cup O)$). In contrast, only a smaller proportion of the responses are categorized as "helpful", suggesting that many responses, although labelled as empathetic, are not ultimately helpful in improving the support seeker's situation.

Table 9 in Appendix D provides examples of responses featuring different label combinations and their ratios. In the first example, the response demonstrates an intention to help and convey understanding and uses of profanity for emphasis. In contrast, the second to fourth examples illustrate various instances where both toxicity and empathy are present but there is a lack of any intent to help the seeker. We also see patterns of side-taking and personal tragedy sharing. Notably, the third example contains content indicative of suicidal ideation (despite being emphathetic). Our qualitative analysis also reveals that the predominant contributor to toxic labels is the use of profanity.

6 Conclusion and Limitation

We examined the interplay of empathy and toxicity in responses to support seekers in mental health online discussions.

Our results found a mixed pattern of insignificant or weak (positive/negative) correlations between the EPITOME empathy labels and the toxic/harmful labels obtained from two widely used APIs. We also revealed a significant presence of toxic/harmful content within empathetic instances in the EPITOME dataset, dominated by "profanity" and "self-harm" labels. These outcomes challenge the standard assumption that there is a negative correlation between empathetic and toxic/harmful language.

Interestingly, we found that the majority of empathetic toxic/harmful responses are not helpful for the individuals who are seeking help. We also noticed some well-intent responses being labelled as toxic due to use of profanity. These mislabels could stem from the issues of oversimplification and ambiguity in toxicity definitions, as previously highlighted in relevant studies (Fortuna et al., 2020). As argued by some communication studies (Masullo Chen et al., 2019) (and also seen in our analyses), the utilization of toxic language does not invariably signify malicious intent. Instead, it may function as a tool for emphasis, conveying closeness, or aligning with the conventions of a particular sociolect, or online context. This observation raises further questions about the role of domain- and community-specific conceptualizations of toxicity in the realm of online content moderation.

Furthermore several empathetic instances are identified as containing suicidal ideation. This discovery raises concerns about the potential use of this dataset for empathetic fine-tuning purposes (Lahnala et al., 2022). To address these concerns, we recommend employing fine-grained toxicity detection models or APIs for data filtering along with human manual validation to ensure alignment between the filtered data and the objectives of the fine-tuning task.

We acknowledge a few limitations of this study: First, it only examines a single dataset within the mental health domain, and the predictions do not consider the context of the seeker's post due to API constraints. Second, as demonstrated by both quantitative and qualitative validation of the APIs' performance, the correspondence between the predicted toxic/harmful labels and human judgments is not perfect (though usable given the accuracy). Third, the introduction of the "helpful" label in our analysis is a preliminary endeavor aimed at addressing the absence of a desired outcome metric in EPITOME, and as such is a (gross) simplification of the problem of measuring response utility. More refined measures, like empathic concerns (Zahn-Waxler and Radke-Yarrow, 1990) or self-report surveys, might be worth considering in future studies. And lastly, the final manual annotation (emphathetic, toxic, and helpful) of the responses was done with a single annotator, and more thorough investigation is required to further validate the robustness of our findings.

For future studies, we recommend a re-evaluation and clarification of the role of empathy in text generation and understanding tasks. Given that certain social science studies have indicated potential harm from empathetic behavior (Breithaupt, 2018), further NLP research is needed to identify subcategories of empathy based on context that can either be beneficial or detrimental. Finally, we suggest incorporating a measure of desired or undesired outcomes in future NLP studies, particularly when dealing with complex and sensitive concepts. This approach will facilitate the analysis and validation of the interplay between outcomes and mediating factors, such as empathy.

References

Atijit Anuchitanukul, Julia Ive, and Lucia Specia. 2022. Revisiting contextual toxicity detection in conversations. *ACM Journal of Data and Information Quality*, 15(1):1–22.

Fritz Breithaupt. 2018. The bad things we do because of empathy. *Interdisciplinary Science Reviews*, 43(2):166–174.

Benjamin MP Cuff, Sarah J Brown, Laura Taylor, and Douglas J Howat. 2016. Empathy: A review of the concept. *Emotion review*, 8(2):144–153.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet*, 7(3):319–339.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

Jigsaw. 2023. Perspective api. https://developers.perspectiveapi.com. Accessed: 2023-09-03.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240.

Allison Lahnala, Charles Welch, Béla Neuendorf, and Lucie Flek. 2022. Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4926–4938.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207.

Gina Masullo Chen, Ashley Muddiman, Tamar Wilner, Eli Pariser, and Natalie Jomini Stroud. 2019. We should not get rid of incivility online. *Social Media+ Society*, 5(3):2056305119862641.

Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2019. The anatomy of reddit: An overview of academic research. *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches 10*, pages 183–204.

Theresa B Moyers and William R Miller. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3):878.

Lala Muradova. 2021. Seeing the other side? perspective-taking and reflective political judgements in interpersonal deliberation. *Political Studies*, 69(3):644–664.

OpenAI. 2023. Openai moderation api. https://platform.openai.com/docs/guides/moderation. Accessed: 2023-09-03.

Lisa Oswald. 2023. Effects of preemptive empathy interventions on reply toxicity among highly active social media users.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124.

Aravind Sesagiri Raamkumar and Yinping Yang. 2022. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Tania Singer and Claus Lamm. 2009. The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156(1):81–96.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *EMNLP 2018*, page 33.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Carolyn Zahn-Waxler and Marian Radke-Yarrow. 1990. The origins of empathic concern. *Motivation and emotion*, 14:107–130.

A Definition and Level Criteria of Epitome Components

Component	Definition	Level Criteria
ER(Emotion Reactions)	Expressing emotions such as warmth, compassion, and concern, experienced by peer supporter after reading seekers post.	A weak communication of emotional reactions alludes to these emotions without the emotions being explicitly labeled (e.g., Everything will be fine). On the other hand, strong communication specifies the experienced emotions (e.g., I feel really sad for you).
IP(Interpretations)	Communicating an understanding of feelings and experiences inferred from the seekers post.	A weak communication of interpretations contains a mention of the understanding (e.g., I understand how you feel) while a strong communication specifies the inferred feeling or experience (e.g., This must be terrifying) or communicates understanding through descriptions of similar experiences (e.g., I also have anxiety attacks at times which makes me really terrified).
EX(Explorations)	Improving understanding of the seeker by exploring the feelings and experiences not stated in the post.	A weak exploration is generic (e.g., What happened?) while a strong exploration is specific and labels the seeker's experiences and feelings which the peer supporter wants to explore (e.g., Are you feeling alone right now?).

Table 5: The definition and level criteria of the EPITOME components

B Definition of Perspective API and OpenAI Moderation API Labels

Label	Source	Definition
Toxicity	Perspective API	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Severe toxicity	Perspective API	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Identity attack	Perspective API	Negative or hateful comments targeting someone because of their identity.
Insult	Perspective API	Insulting, inflammatory, or negative comment towards a person or a group of people.
Profanity	Perspective API	Swear words, curse words, or other obscene or profane language.
Threat	Perspective API	Describes an intention to inflict pain, injury, or violence against an individual or group.

Table 6: The definition of Perspective API Labels

Label	Source	Definition
Hate	OpenAI	Content that expresses, incites, or promotes hate based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste. Hateful content aimed at non-protected groups (e.g., chess players) is harrassment.
Hate/Threatening	OpenAI	Hateful content that also includes violence or serious harm towards the targeted group based on race, gender, ethnicity, religion, nationality, sexual orientation, disability status, or caste.
Harassment	OpenAI	Content that expresses, incites, or promotes harassing language towards any target.
Harassment/threatening	OpenAI	Harassment content that also includes violence or serious harm towards any target.
Self-harm	OpenAI	Content that promotes, encourages, or depicts acts of self-harm, such as suicide, cutting, and eating disorders.
Self-harm/intent	OpenAI	Content where the speaker expresses that they are engaging or intend to engage in acts of self-harm, such as suicide, cutting, and eating disorders.
Self-harm/instructions	OpenAI	Content that encourages performing acts of self-harm, such as suicide, cutting, and eating disorders, or that gives instructions or advice on how to commit such acts.
Sexual	OpenAI	Content meant to arouse sexual excitement, such as the description of sexual activity, or that promotes sexual services (excluding sex education and wellness).
Sexual/minors	OpenAI	Sexual content that includes an individual who is under 18 years old.
Violence	OpenAI	Content that depicts death, violence, or physical injury.
Violence/graphic	OpenAI	Content that depicts death, violence, or physical injury in graphic detail.

Table 7: The definition of OpenAI moderation API labels

C Validation of APIs combined performance

To validate the combined performance of the two APIs, we firstly derived an overall toxic flag which is labeled as positive if any APIs returned labels(6 from Perspective API and 11 from Open AI moderation API) is positive. For the manual annotation criteria, we also derived an aggregated definition of overall toxicity by inputting all toxic/harmful labels' definitions into ChatGPT. Subsequently, we employed this unified toxicity criterion to annotate a set of 100 samples, comprising 50 predicted as positive and 50 as negative by the APIs to validate the prediction performance. The outcomes of this validation process are depicted in Figure 2.

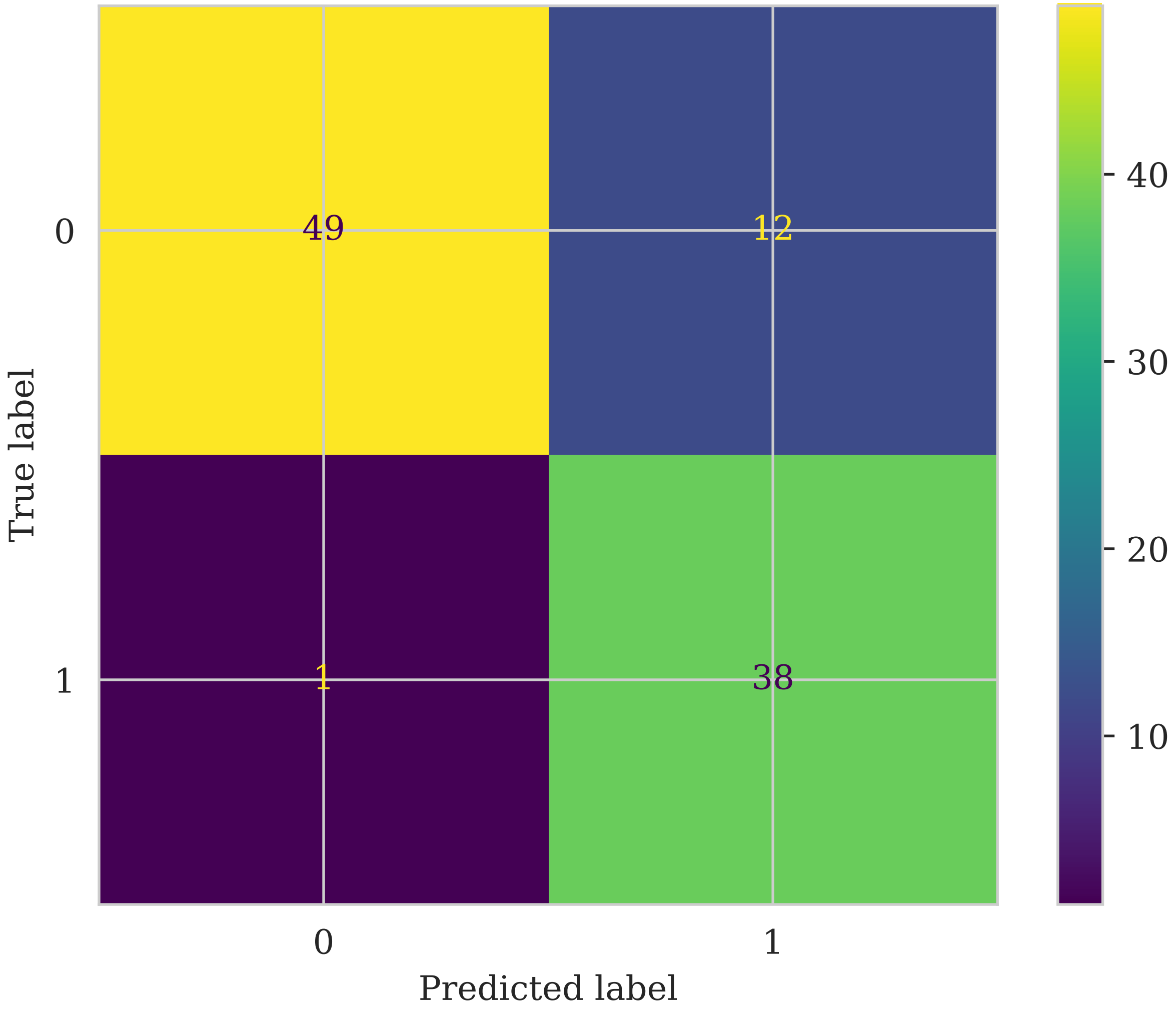

Figure 2: The confusion matrix between the combined predictions of the two APIs(X axis) and the manual annotated binary labels (Y axis).

D Qualitative analysis supplementary

Label	Definition
Empathetic	Comments or content that express emotions like warmth, compassion, and concern in response to seekers' posts, or communicate an understanding of inferred feelings and experiences, or enhance comprehension by exploring unspoken emotions and experiences.
Toxic	Comments or content that are rude, disrespectful, or offensive, likely to drive users away from a discussion, or includes hate speech, harassment, self-harm promotion, sexual content, and graphic depictions of violence or injury, targeting various identity-based and non-protected groups.
Helpful	Comments or content that have the intention or potential to help/improve the future situation or lessen the negativity of the seeker physically, mentally or emotionally.

Table 8: The definition of the labels for the qualitative analysis annotation.

Example seeker post	Example response	Empathetic	Toxic	Helpful	Ratio(%)
I want to die. I fucking hate myself and my miserable existence. I'm a shitty excuse of a human being. I can't even kill myself properly but that probably won't stop me from trying again soon.	What makes you think you're a shitty human being? I don't know who you are and I don't know your story, so pardon me if I'm coming as rude by saying this (that is not my intention, I'm solely trying to help), but maybe some of those reasons aren't founded. For example I often worry that all the people I love will abandon me, and even though it's not really founded and I know in the back of my head that it probably won't happen, it is still one of my largest fears.	1	1	1	16
I almost threw up with anxiety from walking down the high street. I used to be able to rave, go out with friends, work, study. Now I can't walk down the road without having panic attacks. Great	sorry man. If people judge you for it then FUCK THEM	1	1	0	52
I just feel like shit.. I don't really know why. Everything's just so overwhelming all of sudden. I was feeling great this past month, but it's back. Did I do something wrong? I don't know why.	I feel like shit too everyday. I just wanna fade away.	1	1	0	52
What the hell am I doing. I'm failing classes. I'm sabotaging relationships with people I care about. I have little to no motivation for social interaction. Literally what the fuck am I doing. I don't think I've reached a lower point than what I'm currently at.	we on the same boat. havent been going to school, probably already failed some of my classes. theyre just a group of fucking morons anyway. broke up with gf. have no friends. the ones I had were fake. shit.	1	1	0	52
DAE have this? Any happy moment is ruined by this thought: This would be a good moment to be my last.. Every happy experience is accompanied by this. Makes it impossible to enjoy things or get past my issues.	I have tried to kill myself **while happy** for that exact reason.	0	1	0	6

Table 9: Examplar post and response pairs with manual annotation using three labels, empathetic, toxic, and helpful. The last column "ratio" indicates the percentage of samples with the same label combination.