Unveiling Empathic Triggers in Online Interactions via Empathy Cause Identification

Calliope Bandera,[1,2]* Gyeongeun Lee,[1] and Natalie Parde[1]
[1]Department of Computer Science, University of Illinois Chicago
[2]Politecnico Di Milano
calliopebandera@gmail.com, {glee87, parde}@uic.edu

Abstract

Effectively learning language patterns that provoke empathetic expression is vital to creating emotionally intelligent technologies; however, this problem has historically been overlooked. We address this gap by proposing the new task of empathy cause identification: a challenging task aimed at pinpointing specific triggers prompting empathetic responses in communicative settings. We correspondingly introduce AcnEmpathize-Cause, a novel dataset consisting of 4K cause-identified sentences, and explore various models to evaluate and demonstrate the dataset's efficacy. This research not only contributes to the understanding of empathy in textual communication but also paves the way for the development of AI systems capable of more nuanced and supportive interactions.

1 Introduction

Empathy enhances interaction quality by conveying understanding of others' emotions and perspectives (Decety and Lamm, 2006), often reducing aggression and improving intergroup relations (Eisenberg et al., 2010). In human-computer interactions, empathy enables conversational agents to automatically recognize users' emotional states and respond sensitively to their needs, in contexts such as customer care or online health support (Sethi and Jain, 2024). Empathetic systems can also support socially isolated individuals and encourage healthier lifestyles (Paiva et al., 2021; Lee and Parde, 2024).

Research on empathy has been far from absent in the natural language processing (NLP) community (Sharma et al., 2020a; Hosseini and Caragea, 2021a; Lahnala et al., 2022; Lee et al., 2024); however, it has focused largely on empathy detection (Rashkin et al., 2019; Sharma et al., 2020b). Comparatively less attention has been given to the triggers underlying empathetic expression, which are critical to allowing a fuller understanding of empathy (Chen et al., 2022; Jiang et al., 2023). In this work, we address this limitation by (1) introducing a manually annotated empathy cause dataset and (2) proposing a new task of empathy cause identification using this dataset.

To foster broad access and easy integration with ongoing studies, we build our dataset as an additional layer to the publicly available AcnEmpathize (Lee and Parde, 2024) dataset.[1] AcnEmpathize contains posts and responses from an acne support community, annotated for the presence of empathy; in our dataset, AcnEmpathize-Cause, we add manual labels identifying sentence-level empathy causes in posts for which the responses are labeled as containing empathy. Alongside the dataset, we define the new task of empathy cause identification and establish performance baselines using machine learning models. Our key contributions include:

- **We introduce AcnEmpathize-Cause**, a novel dataset with sentence-level empathy cause annotations, enabling detailed analysis of empathy triggers in social support dialogues.
- **We define and formalize the new corresponding task of empathy cause identification**, which connects research in emotion cause extraction with empathy modeling.
- **We benchmark multiple approaches for this task**, including a custom attention-based model and prompting-based question answering frameworks, offering insight into the unique challenges of empathy cause detection.

In the remainder of this paper we elaborate on these contributions. We review related research (§2), discuss dataset creation and structure (§3),

*Former affiliation.

[1] Our dataset and code are available at https://github.com/CalliopeBandera/Empathy-Cause-Identification.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 888–899
December 20-24, 2025 ©2025 Association for Computational Linguistics

and detail our modeling approach (§5). Finally, we conclude (§7) with findings and future directions.

2 Related Work

2.1 Empathy Detection

Empathy detection has recently become a popular NLP task supported by datasets from varying sources, including news articles (Buechel et al., 2018) and specialized support networks (Sharma et al., 2020b; Lee et al., 2023). Although this focus is promising for understanding empathy, existing work has notably lacked analysis of the causes of empathy. Understanding these causes is key to gaining deeper contextual insight into empathetic communication and provides a foundation for more nuanced and effective support systems.

EPITOME (Sharma et al., 2020b) contains 10k conversations annotated for three levels of communication rationales and empathy. However, Lee et al. (2023) highlighted limitations to the model, proposing micromodel frameworks that incorporate context in empathy-seeking posts. These works highlight the importance of context in understanding empathy, while failing to address what causes empathy in a conversation.

Similarly, Rashkin et al. (2019) introduced EMPATHETICDIALOGUES, a dataset of 25k emotionally charged conversations, demonstrating its usefulness in modeling empathy. Welivita and Pu (2020) later expanded this dataset with response intents, providing insights into intent and emotion patterns. However, neither dataset annotates the specific triggers that prompt empathetic responses. Buechel et al. (2018) developed a dataset of 2k responses to news articles using Batson's Empathic Concern scale, while Hosseini and Caragea (2021b) introduced IEMPATHIZE, which annotated cancer support messages with the direction of empathy (provided vs. sought). Although these datasets incorporate context, they still do not systematically identify parts of messages that evoke empathetic reactions. Thus, while there are ample data to support the determination of whether text contains empathy, none supports the identification of empathy cause. This highlights the need for a dataset that explicitly captures the triggers of empathy, enabling a deeper understanding of empathetic communication.

2.2 Emotion Cause Extraction

Emotion cause extraction (ECE) is closely related to our proposed task, aiming to identify triggers of emotional experiences. Since empathy itself is emotionally driven, ECE datasets and techniques provide useful inspiration for empathy cause identification. An important dataset for ECE is RECCON (Poria et al., 2021), which includes 10k cause-effect pairs and serves as a benchmark for causal span extraction and entailment tasks. Transformer-based encoders or bi-LSTMs with attention networks are often employed for the task. For example, Gao et al. (2021) and Li et al. (2021) used ECE results to develop empathetic response generators. Additionally, models like those proposed by Minghui et al. (2022), Xu et al. (2019), and Liu et al. (2019) also employ ranking, semi-supervised learning, or syntactic approaches to enhance ECE performance.

Despite the similarities between ECE and empathy cause identification, it is important to note that ECE seeks to find the internal or external triggers of a specific emotion (e.g., anger or sadness) experienced by the speaker and expressed in their own text. In contrast, empathy cause identification seeks to identify aspects of the speaker's text that evoke empathy in the responder, thus involving a relational dynamic between speaker and listener. Beyond this interactional difference, empathy itself is a multi-dimensional, psychologically-grounded phenomenon that encompasses multiple cognitive and emotional processes, rather than emotion alone (Davis, 1980).

Recognizing this distinction, we build on both empathy classification and ECE insights in conceptualizing and addressing empathy cause identification. Our work is based on the AcnEmpathize (Lee and Parde, 2024) dataset and uses contextual modeling and attention-based architectures inspired by ECE approaches to identify empathy causes. By bridging the gap between empathy classification and ECE, our task formulation and approach enable finer-grained understanding of what causes empathy to emerge in a dialogue.

3 Dataset

3.1 AcnEmpathize

AcnEmpathize (Lee and Parde, 2024) is a recently published dataset designed for empathy-related tasks that contains over 12k posts from acne.org, an online support community forum. It comprises three types of posts: (1) *initial posts* that start conversations, (2) *replies* that respond to these posts, and (3) *quotes* that explicitly refer to specific text in another post. Posts were labeled for empathy by three trained annotators from diverse backgrounds.

The dataset includes 1,730 English conversation threads, ranging from single-post threads to threads with up to 23 posts. Among the posts, 2,976 show empathy, while 9,236 do not. AcnEmpathize's focus on a specific domain, with real-world posts reflecting genuine interactions, makes it an ideal foundation for empathy cause identification. We chose to concentrate on this single topic to build a deeper understanding of empathy mechanisms before extending to more diverse domains. For instance, empathetic responses to conditions like acne may differ substantially from those addressing sensitive issues such as domestic violence. However, while suitable for empathy detection, it lacks annotations specifying which sentences in initial posts trigger empathy. We thus used AcnEmpathize as the starting point for annotating empathy cause specifically.

3.2 Annotation Process

We recruited three volunteer annotators to provide empathy cause annotations, mirroring the annotation setup used for AcnEmpathize. The annotators were two authors of this paper as well as a third volunteer; all were graduate computer science students and fluent English speakers. Unlike some prior work that trained crowdworkers without formal NLP or psychology backgrounds (e.g., Sharma et al. (2020a)), our annotators had both relevant academic training in NLP and prior experience annotating empathy texts, supported by guidelines and multiple rounds of discussion. We used the collaborative tool INCEpTION (Klie et al., 2018) (shown in Figure 4 in Appendix B) to collect annotations, choosing it due to its support for span annotation and collaborative workflows.[2]

Annotators, for each thread, were instructed to read through each sentence of each empathetic reply and highlight at least one sentence in the initial post of the thread they deemed to cause empathy in the corresponding reply. Throughout the annotation process, the definition of empathy considered was the one proposed by Davis (1980), utilized also in the AcnEmpathize paper. Annotation at a sentence level was chosen because finer-grained annotations (e.g., phrases) often lack context and are harder to annotate consistently, while coarser units (e.g., entire posts) may contain unrelated content, making it difficult to isolate specific triggers.

Given occasional issues with automatic sentence detection in INCEpTION, the annotators were told to treat spans ending in punctuation as sentences. An initial discussion among annotators clarified definitions of empathy and its triggers, ensuring consistent annotation criteria. During this initial discussion, annotators identified that typical causes of empathy included expressions of emotion (often negative, e.g., "*I feel so sad and hopeless*") and relatable experiences providing context to struggles (e.g., "*I have cystic acne on my body... it's genetic, so it's extra difficult*"). Some cases, such as threads with no replies, posts lacking text, or replies quoting other posts, were excluded to simplify annotation. Annotators also agreed to skip excessively long posts (over 1,000 sentences) since empathy cause was seldom straightforwardly indicated at the sentence level in these posts.

After annotating a pilot round of 100 conversations, the inter-annotator agreement (IAA) was calculated at κ=0.7 using an averaged Cohen's kappa (Cohen, 1960) across all annotator permutations (with pairwise scores of 0.67, 0.70, and 0.73), which is on par with that observed for similar tasks like ECE (e.g., Sharma et al. (2020b) report an IAA of 0.68). Annotators resolved disagreements through discussion, ultimately reaching consensus for all cases. The remaining conversations were equally divided among the annotators.

3.3 Dataset Structure

AcnEmpathize-Cause is structured around pairs of initial posts and replies, emphasizing the interaction between them to capture how empathy triggers are influenced by reply content. Posts were split into sentences based on the same criteria used during annotation.[3] The dataset's primary components are the sentence-separated text of posts and their labels. Labels, stored in a list format, indicate empathy causes ("1" for cause, "0" otherwise) corresponding to the order of sentences in the text. Additional metadata for posts and replies, such as URLs, titles, and user IDs, is included to provide a comprehensive view of the dataset.

[2] Other tools, such as DOCCANO (Nakayama et al., 2018) and Brat (Stenetorp et al., 2012), were considered but lacked this functionality.

[3] Although organizing each sentence of an initial post as a separate row could simplify labeling, this risks losing contextual nuances essential for identifying empathy causes. The current structure, with pairs of posts and replies, preserves this context while allowing flexibility for future tasks.

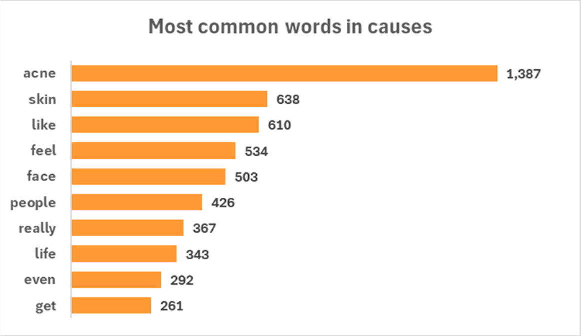


Figure 1: Frequency distribution of the most common words in cause sentences. The values indicate the number of occurrences of each word.

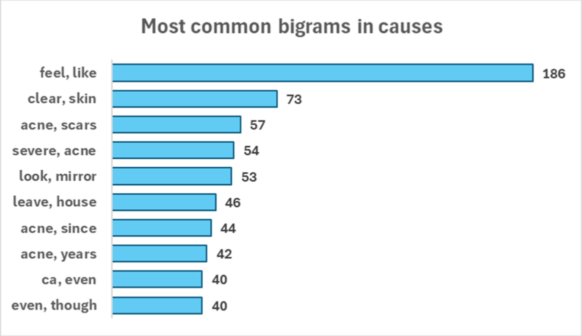


Figure 2: Frequency distribution of the most common bigrams in cause sentences. The values indicate the number of occurrences of each bigram.

Sentence Type	Avg. Sentiment
Cause sentences	-0.0177
Non-cause sentences	0.0032

Table 1: Average sentiment values of cause and non-cause sentences, obtained using TextBlob sentiment polarity analysis. Sentiment polarity ranges from -1 (negative) to 1 (positive).

Emotion	Cause	Non-Cause
Negative	0.9685	0.5642
Positive	0.6322	0.5071
Sadness	0.6299	0.3560
Fear	0.6161	0.3418
Trust	0.4597	0.3493
Anger	0.4482	0.2551
Anticipation	0.4085	0.3162
Disgust	0.3897	0.2473
Joy	0.3396	0.2588
Surprise	0.1910	0.1477

Table 2: Normalized emotion scores for cause and non-cause sentences, computed using the NRC Emotion Lexicon. The values indicate the relative frequency of each emotion within the two categories.

4 AcnEmpathize-Cause Analysis

Overall, AcnEmpathize-Cause contains 3,217 posts, including 1,021 unique initial posts and 2,196 replies, extracted from 1021 conversations. The post count is smaller than in the AcnEmpathize dataset (12k posts), reflecting the removal of non-empathetic replies and posts unrelated to empathy causes. The initial posts contain 45,183 sentences, with 3,931 labeled as causes. On average, each post has 44.25 sentences, and 8.70% of those sentences are cause sentences, indicating a class imbalance. Such an imbalance is typical in ECE datasets, especially in long posts, where meaningful emotion-causing sentences appear less frequently.

Linguistic analysis of the cause sentences, shown in Figure 1 and Figure 2, revealed frequent terms like "acne," "skin," "feel," and "face," reflecting the dataset's focus on the mental struggles associated with acne. These struggles are also evident in common bigrams such as "clear skin" and "acne scars." Additionally, words related to social withdrawal and self-consciousness, such as "people" and "leave house," suggest that shared experiences may trigger empathy.

Sentiment analysis (Table 1) shows that cause sentences have slightly negative sentiment, while non-cause sentences are slightly positive. This aligns with the idea that empathy is often triggered by distressing experiences. Using the NRC Emotion Lexicon (Table 2), we indeed found that cause sentences were predominantly negative, with emotions like sadness, fear, and anger being much more common than in non-cause sentences. Positive emotions, such as trust and joy, also appeared but less frequently. Non-cause sentences showed a more balanced emotional profile, with comparable levels of negative and positive sentiment.

We also applied topic modeling using latent Dirichlet allocation (Blei et al., 2003), identifying five main topics, focusing on emotional struggles, self-reflection, appearance, social comparison, and ongoing acne-related challenges. These topics portray a captivating portrait of empathy in this domain, emphasizing the psychological and social impacts of acne. The detailed topic breakdown is provided in Table 5 in Appendix C. We note that overall, the dataset shows an imbalance influenced by the nature of empathy and its causes—namely, that empathy is triggered by specific, meaningful sentences that resonate with personal experiences. For example, sentences describing personal struggles, such as "*I feel like everyone is staring at my skin, and it makes me not want to leave the house.*" are more likely to evoke empathetic responses than general statements about skincare routines, as supported by both linguistic and emotion analysis. Additionally, because these are real posts, there is inherent noise (i.e., topics less prone to empathy, such as those regarding skincare and medications), further reducing the number of text spans likely to cause empathetic responses. Finally, the dataset's focus on acne-related mental struggles and emotions, as demonstrated by the high frequency of acne-related language, showcases the dataset's specificity.

5 Approach

We frame empathy cause identification as a binary classification task. The goal is to determine which sentences in the initial post are causes of empathy based on the content of the reply, assigning sentence-level labels $y \in \{0, 1\}$, where "1" denotes a cause sentence. While some ECE work attempts to predict text spans, this would be too complex for empathy cause identification in an unbalanced dataset. Sentence-level binary classification simplifies the task by treating each sentence as a fixed span of text, defined as any text ending with a period, question mark, exclamation mark, or ellipsis. We investigated three diverse approaches for empathy cause identification:

- **Attention Network**: Captures contextual information from the initial approach and reply by incorporating embeddings using a co-attention network.
- **Prompting**: Utilizes a fixed prompt to guide the model in predicting the cause sentences.
- **Question Answering**: Frames the task as a question for the model to answer.

These models were selected to examine different broad framings of the task, with the intent that the most effective approach can serve as the focus of finer-grained follow-up studies.

5.1 Attention Network

Attention-based networks have proven effective for ECE, as they can focus on relevant parts of text expressing emotion. For instance, Li et al. (2019) used a multi-attention neural network to link cause and emotion clauses, while Hu et al. (2021) applied a bidirectional hierarchical attention network (BHA) for document-level context. Inspired by the success of these models in ECE, we developed an attention network with the architecture shown in Figure 3. We used a BERT tokenizer (bert-base-uncased) with a maximum sequence length of 64, padding or truncating all sentences to this length. The input IDs and attention masks generated from the tokenizer were passed into the model to generate contextual embeddings.

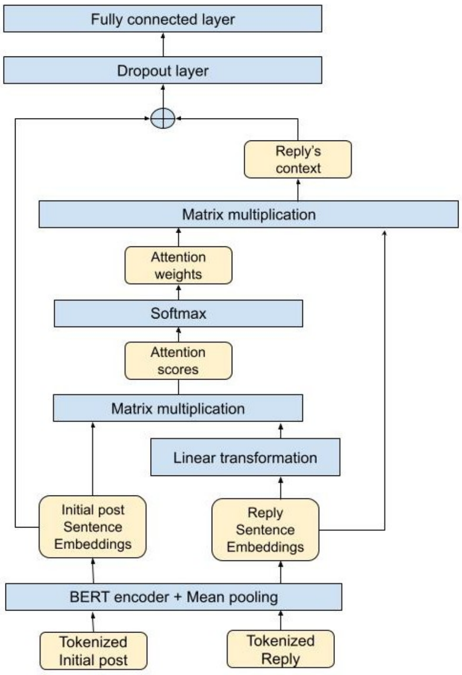


Figure 3: Diagram of the attention network depicting the internal structure of the model. Blue indicates model layers, and yellow indicates the layer's inputs or outputs. Further implementation details and parameters are available in the released code.

We processed the dynamic sentence embeddings, obtained by mean pooling the previous contextual embeddings, through a co-attention network, computing attention scores between the initial post and the reply to assess the relevance of the reply's content in predicting empathy causes in the initial post. Mean pooling was selected over the [CLS] token since it provides a more comprehensive representation and enhances model generalization (Li, 2024). The sentence embeddings were processed as shown in Figure 3.[4]

[4] We also studied different variations of this framework in preliminary experiments but found that they proved less

The combined feature vectors were passed through a dropout layer to prevent overfitting by randomly deactivating neurons during training. Afterward, a fully connected linear layer performed binary classification, producing logits that were transformed into probabilities for determining whether a sentence was an empathy cause.

5.1.1 Training Process

Several techniques were implemented during training to enhance performance and counter overfitting. We used gradient accumulation to simulate larger batches without increasing memory usage; this also made training more stable and improved generalization. During hyperparameter tuning, we found that a batch size of 4 provided the best balance of performance. We used an *AdamW* optimizer, which provides adaptive learning rates and decoupled weight decay, improving generalization. After testing various learning rates and weight decay values, a learning rate of $1 \times e^{-4}$ and a weight decay of 0.01 were selected.

We experimented with two learning rate schedulers: *Cosine Annealing Warm Restarts* and *Reduce LR On Plateau*. We selected the former due to its superior performance. We applied early stopping based on the validation F_1 score with a patience of 4 to prevent overfitting and save computational resources, and saved the model with the best validation F_1 for final testing.

5.2 Prompting and Question Answering

Prompting and question answering (QA) both present viable alternatives to our proposed attention network; thus, we implemented both considering them to be strong baselines for the proposed empathy cause identification task. Prompt-based approaches have recently excelled in a broad range of NLP tasks, including empathy detection (Kong and Moon, 2024), and ECE has recently been framed as a question answering task with evidence that this leads to strong performance (Chandakacherla et al., 2024). For our prompt-based approach, we experimented with manually-defined discrete prompt templates; preliminary experiments suggested that the prompt reported in Appendix A performed best, and thus we used it for our final evaluation.

Following a question answering paradigm allowed us to simplify the task into a structured format; in ECE, this was previously shown to more productively facilitate inference. We reformulated the task as a question answering problem by converting each initial post sentence into a question and using the reply as context. We then fine-tuned *bert-large-uncased* for sequence classification and used a QA inference pipeline to predict whether each sentence was the cause of empathy.

6 Evaluation

6.1 Experimental Setup

We conducted all experiments in a relatively low-resource environment (Google Colab with an L4 GPU and high RAM). When studying different configurations of the co-attention network, we used a *bert-base-uncased* backbone model with a maximum token size of 64 tokens. We held the learning rate constant at $1 \times e^{-4}$ along with a gradient accumulation of 4 and an *AdamW* optimizer with $eps = 1 \times e^{-8}$ and weight decay of 0.01. For all conditions, our scheduler used *Cosine Annealing Warm Restarts* with $T_0 = 5$ and $T_{mult} = 1$. We used a dropout rate of 0.3 with early stopping patience set to 4 epochs based on the validation F_1. We allowed the model to train for up to 20 epochs (training typically stopped earlier due to early stopping), and we randomly split the dataset into 80% training, 10% validation, and 10% test subsets. All reported results are averaged over five independent runs to ensure consistency and reduce variance.

6.2 Co-Attention Network Experiments

We compared performance across various configurations of the co-attention network for empathy cause identification. Many of these configurations were designed to address class imbalance: as previously noted, `AcnEmpathize-Cause` has more non-cause sentences than cause sentences, and imbalances such as this can lead to biased predictions and reduced performance. We explored both the use of class weights in the loss function and focal loss for this purpose. Focal loss (Lin et al., 2018) in particular can be employed to reshape the standard cross-entropy loss and down-weight well-classified examples, offering strong potential to improve performance in imbalanced class settings.

In Table 3, we report the results of our co-attention network experiments. Precision, recall and F_1 all refer to those of the positive class. We

effective. One variation used the [CLS] token instead of mean pooling, while another bypassed the attention mechanism, concatenating the average embeddings of the reply with those of the initial post. A variant of the model processing single sentences at a time was also tested, but it resulted in significant computational overhead without improving performance.

Model	**Test Loss**	**Precision**	**Recall**	$\mathbf{F_1}$	**Accuracy**	**ROC AUC**
Binary Focal Loss	**0.0502**	0.4912	0.4308	**0.4590**	**0.9158**	**0.8266**
Class Weights [1.0, 3.0]	0.8710	0.4063	0.5137	0.4537	0.8895	0.8169
Class Weights [1.0, 11.0]	1.2635	0.3668	**0.5561**	0.4420	0.8746	0.8246
Focal Loss (Simpler Model)	0.0722	0.4674	0.4464	0.4566	0.9051	0.8092
Focal Loss (CLS Token)	0.0713	**0.5210**	0.4015	0.4535	0.9135	0.8025
Random Labeling	-	0.0862	0.4851	0.1465	0.4998	0.4877
Majority Class Labeling	-	0.0000	0.0000	0.0000	0.9115	0.5000

Table 3: Performance comparison of different loss functions and co-attention network configurations for empathy cause identification. The table reports precision, recall, and F_1 for the positive class across five training configurations, along with two baseline methods: random labeling and majority class labeling. The best values are reported in bold.

considered F_1, the harmonic mean of precision and recall, as the most critical metric for evaluating empathy cause identification performance. We compare the following co-attention network conditions:

- **Binary Focal Loss:** Binary focal loss with no weights.
- **Class Weights [1.0, 3.0]:** Cross-entropy loss with class weights of 1 for the majority class and 3 for the minority class.
- **Class Weights [1.0, 11.0]:** Cross-entropy loss with class weights of 1 for the majority class and 11 for the minority class.
- **Focal Loss (Simpler Model):** Binary focal loss and a simpler model without the attention network. All sentence embeddings of the reply are averaged together and concatenated to the initial post's sentence embeddings.
- **Focal Loss (CLS Token):** Binary focal loss and a model for which the [CLS] token is used instead of mean pooling to get the sentence embeddings.
- **Random:** Assigns random labels to each sentence (naïve baseline condition).
- **Majority Class:** Always assigns the majority class (majority=0; naïve baseline condition).

The results demonstrate that binary focal loss leads to the best overall performance (including F_1=0.46, accuracy=0.92, and ROC AUC=0.83). This suggests that focal loss is particularly well-suited for handling the dataset's class imbalance by emphasizing difficult-to-classify samples, and especially those in the minority class (cause sentences). The metrics also reflect the inherent complexity of the task and the challenges posed by the dataset imbalance. Identifying empathy cause requires understanding nuanced contextual relationships between sentences, which makes it difficult for the model to achieve high precision and recall. Despite these challenges, the F_1 achieved by the best model performs far above both naïve baselines, demonstrating the model's ability to learn and perform the task effectively within the constraints of the dataset. This highlights the potential of the proposed co-attention network and provides a solid foundation for future work in empathy cause identification.

6.2.1 Follow-Up Analyses

We also studied the use of threshold adjustment, data augmentation, and undersampling, but found them to be less effective. Adjusting the classification threshold or using undersampling removed critical context, which hurt performance. Data augmentation, such as synonym replacement and backtranslation, introduced noise and failed to improve results. Following this, we experimented with the inclusion of preprocessing techniques including lemmatization and stopword removal, but we observed that these also slightly worsened performance by removing essential contextual information. Finally, to assess the extent to which class imbalance impacted overall empathy cause identification performance, we compared the performance of the model using a balanced version of the dataset (removing a random sample of non-cause sentences) to the dataset in its original form. We observed improved performance in this condition, confirming that class imbalance does increase the complexity of this task.

Method	F_1	P	R	Acc.	ROC AUC
BFL	**0.46**	**0.49**	**0.43**	**0.92**	**0.83**
Prompt	0.08	0.10	0.07	0.86	-
QA	0.03	0.09	0.02	0.90	0.52

Table 4: Performance for the positive class (cause sentences) for the best-performing model for each approach. **P**recision, **R**ecall, $\mathbf{F_1}$, **Acc**uracy, and **ROC AUC** are all rounded to two significant digits for brevity. The table compares the best performing configuration for each approach: Binary Focal Loss (BFL), Few-Shot Prompting on Whole Posts (Prompt), and Question-Answering (QA). The best values are reported in bold.

6.3 Model Comparison

We selected the best-performing co-attention network, trained using binary focal loss, for comparison with our prompting and QA conditions and report these results in Table 4. The results clearly indicate that the attention-based model outperforms the other approaches across key metrics. Using a co-attention mechanism allowed the model to most effectively incorporate context from the reply when predicting empathy causes, making it the best suited for this task among those tested.

In contrast, the prompting and QA approaches struggled to achieve comparable performance. Even when fine-tuned, it appears that their more generic and all-purpose architecture cannot effectively understand the nuanced relationship between the initial post and the reply. Their underperformance highlights the challenges of adapting general-purpose techniques to more specialized tasks. It also conversely highlights that specialized architectures, such as the co-attention network with binary focal loss, can effectively identify causes of empathy. Using LLMs for empathy cause identification may require more careful prompt design or additional fine-tuning to match the performance of a tailored attention network.

7 Conclusion

In this work **we proposed a novel task, empathy cause identification, that seeks to find the specific sentences in a text that evoke empathetic response**. While existing empathy research in NLP focuses on tasks like empathy classification, the identification of more nuanced empathetic conversational dynamics has remained unexplored. Understanding how empathetic responses are triggered can enable the development of more emotionally intelligent AI systems, enhancing their ability to respond compassionately and appropriately in sensitive contexts such as mental health support and customer service. By tackling this problem, this work contributes to advancing the understanding of empathy and the development of empathy-driven AI, paving the way for applications that can better understand and support human interactions.

`AcnEmpathize-Cause` comprises 3217 real-world posts with exhaustive manual labels across 45,183 sentences indicating causes of empathy (positive n=3931), ensuring that a reliable gold standard is provided for this challenging task and offering a robust foundation for advancing research in this area. Through the systematic comparison of a diverse range of models, including our proposed co-attention network, we observed that the class imbalance inherent to empathy cause identification presents modeling challenges, highlighting the importance of specialized techniques like focal loss. Overall, **we observed that the co-attention network performed more effectively than strong prompting and question answering baselines**, showcasing its ability to incorporate contextual information from replies to predict empathy causes in initial posts accurately. Our best-performing model achieved an F_1=0.46, accuracy=0.92, and ROC AUC=0.83.

Future research could aim to explore this task in different domains to evaluate the generalizability of empathy cause detection models as well as domain-specific nuances that may emerge. Additionally, the task could be expanded to include span extraction, allowing the identification of text spans, potentially larger or smaller than a single sentence, that cause individuals to express empathy. Another potential extension could involve annotating the replies in the dataset with the evidence of empathy, enabling an extraction of causes and evidence together. This approach would align with popular ECE methodologies, such as emotion cause pair extraction, and further expand the task's applications.

Overall, by introducing a new dataset, testing multiple models, and addressing inherent challenges such as class imbalance, this paper contributes to advancing our collective computational ability to interpret and respond to the intricacies of human interactions. It lays a foundation for future efforts in the field of empathy cause identification, encouraging further research into more enhanced models and different application fields. Ultimately, the insights gained here showcase and emphasize the potential of intelligent systems to foster empathy and support in human-centered contexts.

Limitations

This paper has some limitations that warrant acknowledgment. The dataset was created through a manual annotation process, which, like any human endeavor, is inherently prone to errors. This is particularly relevant given the potentially subjective nature of empathy, which can lead to different interpretations among annotators. Although we addressed this purposefully through rigorous annotator discussion and calculation of inter-annotator agreement, it is possible that others may disagree with some empathy cause labels.

Additionally, the dataset's domain-specific focus on an acne support community, while beneficial for task prediction within this field, may limit its generalizability to other contexts; although beyond the present scope, future investigation is warranted to probe this further. Moreover, the annotators were women from 20 to 30 years old from different nations. Although it is impossible to guarantee full coverage across demographic groups in absence of an unfeasible number of annotators, it is possible that annotators from other demographic groups may have made different labeling decisions. Finally, as previously discussed, the dataset's significant class imbalance poses limitations in developing and testing high-performing models: we consider this both an obstacle to higher performance and a welcome challenge for future work.

Ethical Considerations

The dataset presented in this paper was created from a public dataset that was reviewed by the Institutional Review Board at its host institution and determined to be exempt from further review. The data was sourced from the acne.org forum; the creators of the original dataset deidentified it manually such that the resulting, publicly available data contains no personally identifiable information. All annotators involved in our `AcnEmpathize-Cause` dataset creation were volunteers, and this new dataset will be made publicly available to facilitate additional research on the topic by others.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. Gyeongeun Lee and Natalie Parde were supported during this work by the National Science Foundation under Grant No. 2125411. Natalie Parde was also supported by the National Institutes of Health under Grants R41NR020667, 1R61DA057629-01A1, and 1R01AG091762-01 during this time. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *Preprint*, arXiv:1808.10399.

Sharad Chandakacherla, Vaibhav Bhargava, and Natalie Parde. 2024. UIC NLP GRADS at SemEval-2024 task 3: Two-step disjoint modeling for emotion-cause pair extraction. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1373–1379, Mexico City, Mexico. Association for Computational Linguistics.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. EmpHi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1074, Seattle, United States. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Mark Davis. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog Sel. Doc. Psychol.*, 10.

Jean Decety and Claus Lamm. 2006. Human empathy through the lens of social neuroscience. *The Scientific World Journal*, 6(1):280363.

Nancy Eisenberg, Natalie D. Eggum, and Laura Di Giunta. 2010. Empathy-related responding: Associations with prosocial behavior, aggression, and intergroup relations. *Social Issues and Policy Review*, 4(1):143–180.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021a. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2021b. It takes two to empathize: One to seek and one to provide. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:13018–13026.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021. Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 558–568, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liting Jiang, Di Wu, Bohui Mao, Yanbing Li, and Wushour Slamu. 2023. Empathy intent drives empathy detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Singapore. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Haein Kong and Seonghyeon Moon. 2024. RU at WASSA 2024 shared task: Task-aligned prompt for predicting empathy and distress. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 380–384, Bangkok, Thailand. Association for Computational Linguistics.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2139–2158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2023. Empathy identification systems are not accurately accounting for context. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1686–1695, Dubrovnik, Croatia. Association for Computational Linguistics.

Gyeongeun Lee and Natalie Parde. 2024. AcnEmpathize: A dataset for understanding empathy in dermatology conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 143–153, Torino, Italia. ELRA and ICCL.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 519–529, Bangkok, Thailand. Association for Computational Linguistics.

Ran Li. 2024. Adapt bert on multiple downstream tasks. Stanford CS224N Default Project.

Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. 2019. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2041–2045. ACM.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection. *Preprint*, arXiv:1708.02002.

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. *Preprint*, arXiv:1907.11692.

Zou Minghui, Pan Rui, Zhang Sai, and Zhang Xiaowang. 2022. Using extracted emotion cause to improve content-relevance for empathetic conversation generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 811–823, Nanchang, China. Chinese Information Processing Society of China.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. 2021. Empathy and prosociality in social agents. In *The handbook on socially interactive agents: 20 Years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 1: methods, behavior, cognition*, pages 385–432.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Preprint*, arXiv:2012.11820.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Surbhi Sethi and Kanishk Jain. 2024. Ai technologies for social emotional learning: recent research and future directions. *Journal of Research in Innovative Teaching & Learning*, 17:213–225.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020a. 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, Online. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. 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, Online. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.

A Appendix: Few-shot Prompt

Here we report the prompt used for the few shot learning method.

```
 alpaca_prompt = """ Below is
 an instruction that describes a
 task, paired with an input that
 provides further context. Write
 a response that appropriately
 completes the request.
{few_shot_examples}
 ### Instruction:
{instruction}
 ### Input:
 Initial post: {initial_post}
 Reply: {reply}
 ### Response:
{response}
 """
```

Where this is the instruction:

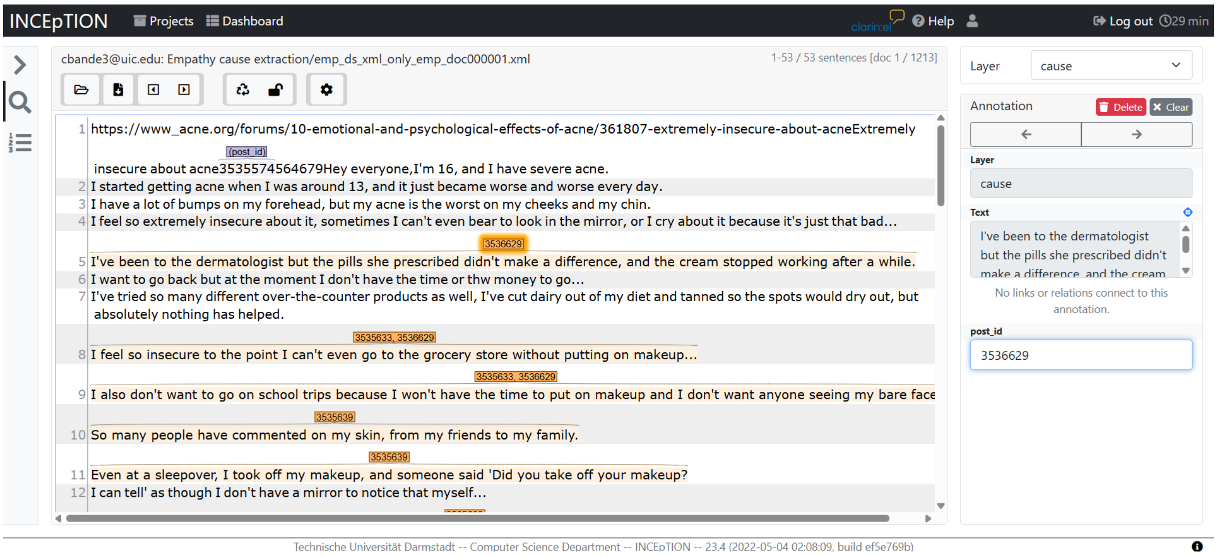
```
fixed_instruction = """Given the
input text, classify whether
the sentences in the initial
post cause empathy based on the
content of the reply.  The
sentences in the initial post
are enclosed in quotation marks
and separated by commas.  The
objective is to identify which
sentences in the initial post
evoke empathy in the reply. Each
reply shows empathy toward the
initial post. For each sentence
in the initial post:
Respond with 1 if it causes
empathy,
Respond with 0 if it does not.
Output a sequence of 0s and 1s
corresponding to each sentence
in the initial post, with the
values separated by commas and no
additional text. """
```

B Appendix: INCEpTION Interface

Figure 4 shows the user interface of the annotation tool INCEpTION used in our study.

C Appendix: LDA Topic Modeling Details

Here reported in Table 5 the complete LDA topic analysis.



cbande3@uic.edu: Empathy cause extraction/emp_ds_xml_only_emp_doc000001.xml

1-53 / 53 sentences [doc 1 / 1213]

1 https://www_acne.org/forums/10-emotional-and-psychological-effects-of-acne/361807-extremely-insecure-about-acneExtremely insecure about acne3535574564679Hey everyone,I'm 16, and I have severe acne.
2 I started getting acne when I was around 13, and it just became worse and worse every day.
3 I have a lot of bumps on my forehead, but my acne is the worst on my cheeks and my chin.
4 I feel so extremely insecure about it, sometimes I can't even bear to look in the mirror, or I cry about it because it's just that bad...
5 I've been to the dermatologist but the pills she prescribed didn't make a difference, and the cream stopped working after a while.
6 I want to go back but at the moment I don't have the time or thw money to go...
7 I've tried so many different over-the-counter products as well, I've cut dairy out of my diet and tanned so the spots would dry out, but absolutely nothing has helped.
8 I feel so insecure to the point I can't even go to the grocery store without putting on makeup...
9 I also don't want to go on school trips because I won't have the time to put on makeup and I don't want anyone seeing my bare face
10 So many people have commented on my skin, from my friends to my family.
11 Even at a sleepover, I took off my makeup, and someone said 'Did you take off your makeup?
12 I can tell' as though I don't have a mirror to notice that myself...

Layer: cause

Annotation

Layer: cause

Text: I've been to the dermatologist but the pills she prescribed didn't make a difference, and the cream

No links or relations connect to this annotation.

post_id: 3536629

Technische Universität Darmstadt -- Computer Science Department -- INCEpTION -- 23.4 (2022-05-04 02:08:09, build ef5e769b)

Figure 4: Interface of the annotation tool software INCEpTION.

Topic	Keywords	Interpretation
1	self, like, feeling, time, going, depressed, feel, life, know, acne	**Self-Reflection and Emotional Struggles**
2	wish, time, look, really, want, acne, im, like, think, feel	**Desire for Change and Reflection**
3	month, like, really, makeup, girl, skin, year, face, acne, want	**Appearance and Social Comparison**
4	life, really, trying, year, face, scar, like, feel, skin, acne	**Ongoing Struggles with Acne and Self-Perception**
5	social, im, make, skin, really, look, people, acne, feel, like	**Social Impact and Self-Consciousness**

Table 5: LDA topic modeling results on cause sentences. Each topic is represented by high-frequency keywords and an interpretation of the main theme.