Empathy Speaks in Metaphors: The *Empathy-Metaphor* Corpus of Figurative Language in Empathetic Text

Gyeongeun Lee and Natalie Parde
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois Chicago
{glee87, parde}@uic.edu

Abstract

Metaphorical language is a powerful vehicle for expressing empathy, yet it has received limited attention in computational studies of supportive communication. We introduce *Empathy-Metaphor*, the first corpus that explicitly annotates metaphorical spans in empathetic online peer-support. Building on 2,492 empathetic posts from an acne support forum, the dataset contains over 2,100 manually identified metaphorical spans with strong inter-annotator agreement (κ=0.85). Analyses show that metaphors are frequent, diverse, and strategically positioned, often framing acne as a battle, journey, or shared struggle. Lexical and semantic clustering highlight recurring themes of encouragement and emotional hardship, while lexicon-based category analysis (Empath) emphasizes the prominence of conflict and negative emotion framings. Benchmark experiments demonstrate that transformer models, especially DeBERTa-v3, substantially outperform linear and recurrent baselines, achieving a token-level macro F1 of 0.634 and a span-level macro F1 of 0.440 under relaxed evaluation. These contributions establish a new resource for studying figurative language in empathetic text, providing insights into the creative role of metaphors in online support.

Keywords: Empathy, Metaphor, Figurative language

1. Introduction

Metaphorical expressions such as "*Keep your chin up*" are often used conversationally to convey support, and research has shown that these expressions carry stronger emotional resonance than literal language in the same context (Kövecses, 2000; Goatly, 2007; Citron and Goldberg, 2014; Mohammad et al., 2016). This makes them particularly well-suited for expressing *empathy*, which involves understanding and responding supportively to another person's emotional experience (Davis, 1980). In the context of empathetic communication, metaphors act as linguistic tools to frame another person's experience in relatable and validating terms—acknowledging struggle (e.g., "*carrying a heavy burden*"), encouraging perseverance (e.g., "*keep battling!*"), or offering hope (e.g., "*finding light at the end of the tunnel*"). With an increasing demand for peer support in text-based online platforms (De Choudhury and Kiciman, 2017; Sharma et al., 2021), examining how metaphorical language conveys empathy can deepen our understanding of empathetic communication and help build more effective support systems.

However, resources to facilitate this study are limited. No existing corpus jointly captures metaphor and empathy, leaving a gap in understanding how figurative language operates in supportive contexts. Metaphor resources such as the VU Amsterdam Metaphor Corpus (VUA) (Steen et al., 2010) have propelled computational metaphor research and shared tasks (Leong et al., 2018, 2020), but they focus on expository or argumentative domains (e.g., news articles, academic texts, or political speeches) rather than supportive communication. More recent efforts have explored metaphor in specialized domains, such as religious discourse (Reimann and Scheffler, 2024), medical texts (Shi and Khoo, 2023), or German cross-genre data (Berger et al., 2024), but not in empathetic communication. In contrast, empathy corpora such as EmpatheticDialogues (Rashkin et al., 2019) and online peer-support datasets (Sharma et al., 2021) capture supportive exchanges but do not annotate figurative language. As a result, the role of metaphor in shaping empathetic support remains largely unexplored.

In this paper, we introduce *Empathy-Metaphor*, the first corpus that explicitly annotates metaphorical expressions in online empathetic communication. Our contributions are as follows:

- We present a manually annotated corpus of metaphorical text spans in 2,492 empathetic posts from an online peer-support forum, providing the first language resource at the intersection of figurative expression and empathy.
- We conduct lexical and semantic clustering analyses as well as a lexicon-based category analysis (Empath) to explore how metaphorical language is used in empathetic contexts.
- We benchmark metaphor detection models trained on this dataset using sequence labeling approaches to validate the dataset's utility

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11188–11198
11–16 May 2026.

for modeling and downstream applications.

Together, these contributions build foundations and research capacity for studying figurative language in empathetic communication and for enhancing peer support in online communities.

2. Related Work

2.1. Metaphor Corpora

Existing metaphor corpora differ in their unit of annotation and scope. Early datasets such as MOH-X (Mohammad et al., 2016) and TroFi (Birke and Sarkar, 2006) focus on single target verbs within sentences, where each verb is annotated as metaphor (nonliteral) or literal. MOH-X was manually annotated through crowdsourced human judgments of contextual meaning, while TroFi was labeled via initial unsupervised clustering of verb usages followed by manual verification. The TSV (Tsvetkov et al., 2014) dataset, unlike the previous two, consists of adjective-noun pairs manually labeled by native speakers as metaphorical or literal based on semantic judgment. Larger corpora such as VUA (Steen et al., 2010) and KOMET 1.0 (Antloga, 2020) annotate metaphors at the word level (nouns, verbs, adjectives, and adverbs) in continuous texts, drawn from general, multi-topic sources. Both corpora follow the Metaphor Identification Procedure Vrije Universiteit (MIPVU), introduced in VUA as an extension of the original Metaphor Identification Procedure (MIP) (Group, 2007) that established the first systematic guideline for identifying metaphors.

More recent corpora have expanded metaphor annotation beyond general-domain corpora to specialized contexts such as religious discourse (Reimann and Scheffler, 2024), medical texts (Shi and Khoo, 2023), or cross-genre settings (Berger et al., 2024), often following the MIPVU framework. Despite this progress, the study of metaphor in supportive and empathetic texts remains underexplored and lacks formal resources.

2.2. Empathy Corpora

In empathy-related domains, corpora such as EmpatheticDialogues (Rashkin et al., 2019) and peer-support datasets (Sharma et al., 2021) capture empathetic communication but do not annotate figurative language. However, psycholinguistic literature suggests that metaphors are more emotionally resonant than literal expressions (Kövecses, 2000; Citron and Goldberg, 2014; Mohammad et al., 2016), and recent computational studies back this up by showing that figurative cues can improve empathy detection (Lee et al., 2024b,a) and empathetic response generation (Lee et al., 2025).

To address existing barriers to further research on this topic, we introduce *Empathy–Metaphor*, the first corpus with manually-provided annotations of metaphorical spans in empathetic text. We annotate spans, ranging from one word to multiple words, to capture complete metaphorical expressions in context (e.g., “stuck in a dark hole” or “fighting these scars”). Through this resource, we hope to advance the understanding of empathetic communication through the lens of creative language and ultimately contribute to more effective forms of support.

3. Methods

3.1. Data Collection

Our dataset builds on *AcnEmpathize* (Lee and Parde, 2024), a binary-labeled empathy corpus sourced from *acne.org*’s Emotional and Psychological Effects of Acne forum, where individuals share personal struggles and receive peer support. Each entry consists of either an initial post describing a problem or a reply providing support. We focus on replies, as these contain the empathetic language of interest. We selected this dataset for two reasons. First, prior work (Lee et al., 2024b, 2025) using automatic metaphor, idiom, and hyperbole detection on this dataset has shown its richness in figurative language, particularly with respect to metaphors and idioms, and demonstrated its impact on both empathy detection (Lee et al., 2024b) and generation (Lee et al., 2025). This motivates the need for systematic manual annotation to support further research of figurative language in this setting. Second, focusing on acne—a common dermatological condition among young people with significant mental health consequences (National Institute for Health and Care Excellence, 2021; Molla et al., 2021)—allows for a targeted investigation of empathetic language in an often-overlooked health domain, establishing clear ties to downstream broader impacts.

3.2. Data Preparation and Annotation

After filtering out non-empathetic replies in *AcnEmpathize* using the dataset’s existing binary empathy labels, we obtained a set of 2,492 empathetic replies as the basis for our new metaphor annotation. The annotation task was to identify metaphorical expressions in these replies.

Annotation Process and Guidelines

Three student volunteers participated in annotating the data: a PhD student in Computer Science, a Master’s graduate in Computer Science, and an undergraduate student intern pursuing an interdisciplinary Computer Science degree and working

in an NLP research laboratory. All had relevant formal coursework and research background to quickly onboard to the annotation task. We used the collaborative annotation platform INCEpTION (Klie et al., 2018), which allowed annotators to highlight metaphorical spans directly in the text. The software was previously used to annotate empathy causes in `AcnEmpathize-Cause` (Bandera et al., 2025), a dataset used in our prior generation work (Lee et al., 2025). We adopted the same platform to maintain annotation consistency across our studies.

Our annotation procedure was grounded in MIPVU (Steen et al., 2010), adapted for the targeted empathetic setting through iterative rounds of annotation and discussion. Annotators were instructed to mark the minimal span (one word to several words) communicating a metaphor rather than the entire surrounding sentence(s). A span was annotated if it met three conditions:

1. The word or phrase is used indirectly; that is, that the meaning in context is different from its basic, literal sense.

2. The basic meaning comes from a more concrete or physical domain (e.g., body, nature, or objects) than the implied meaning.

3. The basic meaning helps us understand the abstract or emotional meaning in context.

We included conventional cross-domain mappings (e.g., *carry a burden*) and personifications with explicit noun mappings (e.g., *acne controls me*) while excluding fixed idioms without productive metaphorical meaning (e.g., *I've been there*), purely literal or emotive intensifiers (e.g., *really sad*), and common supportive phrases with no figurative mapping (e.g., *I hear you*). For example, only the phrase *want to disappear* was annotated in "*acne makes me want to disappear*," and *like a pizza* in "*my face looks like a pizza with acne*." An example post and its highlighted metaphorical spans are shown in Figure 1.

To ensure consistency, the annotators first each independently labeled the same sample of 250 replies (approximately 10% of the dataset) in three batches with sizes n=50, n=50, and n=150. After each batch, discussions were held to resolve disagreements and clarify ambiguous cases. Considering the inherent complexity of span annotation, we considered two annotations to be in agreement when their highlighted spans overlapped (e.g., *down* vs. *being down*). Based on this criterion, the inter-annotator agreement (IAA) was measured at κ=0.85 using Fleiss' Kappa (Fleiss, 1971). The remaining 2,242 replies were then divided equally among the annotators.

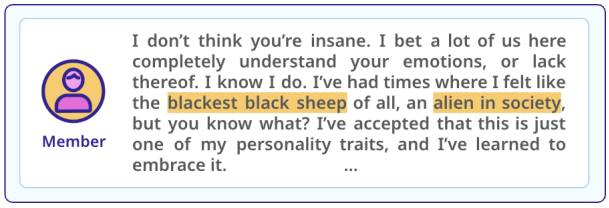

Figure 1: Example post and annotation from *Empathy–Metaphor*. Highlighted spans show how annotators marked metaphorical expressions within empathetic posts.

Posts	Total
Has metaphors	1,312 (52.65%)
No metaphors	1,180 (47.35%)
Total	2,492

Table 1: Distribution of empathetic posts with and without metaphorical spans.

The resulting corpus, *Empathy-Metaphor*, contains manually annotated metaphorical spans in 2,492 empathetic replies from an online peer-support forum. To our knowledge, it is the first resource that explicitly connects metaphor annotation with empathetic communication, opening new opportunities for both linguistic study and computational modeling.[1]

4. Dataset Analysis

4.1. Dataset Overview

Among the 2,492 empathetic posts in the *Empathy-Metaphor* dataset, 1,312 posts (52.65%) contain at least one metaphorical span, while 1,180 posts (47.35%) include only literal or other non-metaphorical expressions (Table 1). Across the 1,312 posts containing metaphorical spans, a total of 2,177 metaphorical spans were identified, of which 1,667 are unique, resulting in a unique-to-total ratio (calculated by dividing the unique metaphorical spans by the total spans) of 0.77 (Table 2). This suggests that while some metaphorical spans recur across posts, a substantial proportion are distinct, showcasing the creativity of metaphorical language in this domain.

To study the distribution of metaphorical language across posts, we calculated the frequency with which metaphor spans appeared per post and show the log-scaled distribution in Figure 2. We observed that most posts contain only a single or

[1]The dataset is publicly available at https://github.com/gyeongeunlee16/AcnEmpathize_Metaphor

Posts	Total
Total metaphor spans (T)	2,177
Unique metaphor spans (U)	1,667
Unique-to-total ratio (U/T)	0.77

Table 2: Counts of metaphor spans and unique metaphorical spans across empathetic posts. The unique-to-total ratio (U/T = 0.77) indicates that most metaphors are distinct rather than repeated.

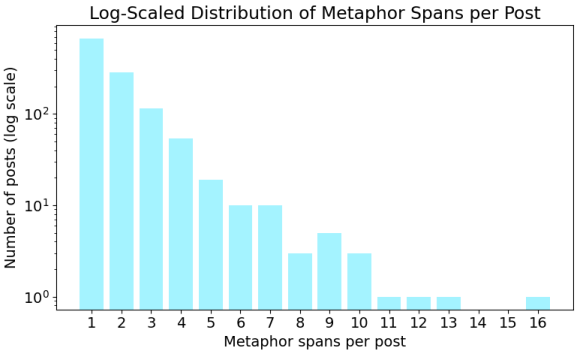

Figure 2: Log-scaled distribution of the number of metaphor spans per post.

a few metaphorical spans, with a small minority having more than ten; the maximum number of metaphorical spans in a single post was 16. Figure 3 presents the frequency distribution of individual metaphorical spans, ranked from most to least common. Confirming the U/T ratio calculated in Table 2, Figure 3 shows that the vast majority of metaphors appear only once or a few times. This further highlights the creative use of metaphors in *Empathy-Metaphor*, suggesting that few metaphors have been conventionalized at scale in an empathetic support context.

4.2. Analysis of Metaphor Usage

We further explore metaphor use in this setting by analyzing the metaphorical spans' placement within posts, their thematic groupings, and the lexicon-based semantic categories (Empath) most closely tied to them.

Positional Distribution

To evaluate metaphor placement within posts, we divided each post into three segments with equal character counts: *opening*, *middle*, and *closing*. We assigned spans to the corresponding segment in which they appeared (e.g., spans appearing within the first third of the post were binned as *opening*). Overall, we found that metaphors were distributed relatively evenly across posts, but were

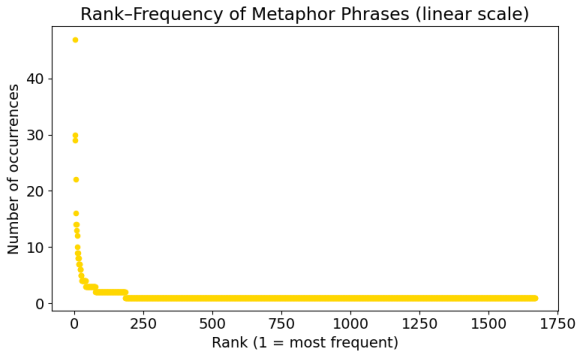

Figure 3: Ranked frequency of metaphorical spans.

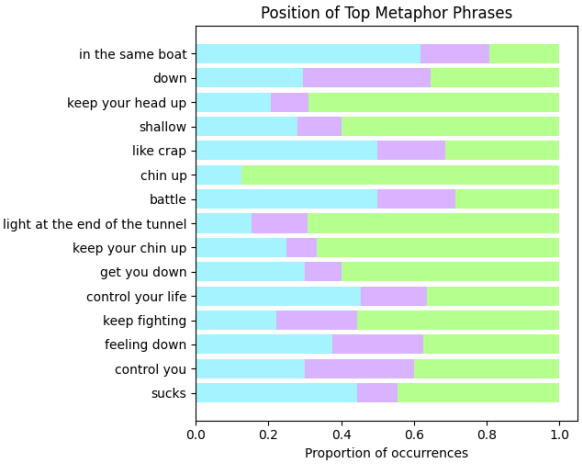

Figure 4: Positional distribution of the most frequent metaphors. Values show the mean proportion of tokens within each span matching the given category. Blue indicates opening spans, purple indicates middle spans, and green indicates closing spans.

found slightly more often in the closing (37.9%) and opening segments (33.2%) than in the middle (29.0%). This indicates that there may be a slight preference toward using metaphors to either initially frame the support or to reinforce empathy at the end. We also examined the positional distribution of the most frequent metaphorical phrases to observe individual patterns (Figure 4). Many of these phrases appeared more often in the opening or closing segments, with top phrases like "*in the same boat*" typically being placed at the start, acknowledging shared struggle, and phrases like "*keep your head up*," and "*chin up*" occurring more commonly in closing positions to reinforce encouragement at the end of posts.

Lexical and Semantic Themes

We then turned to clustering metaphorical phrases, using two complementary approaches to better understand thematic patterns.

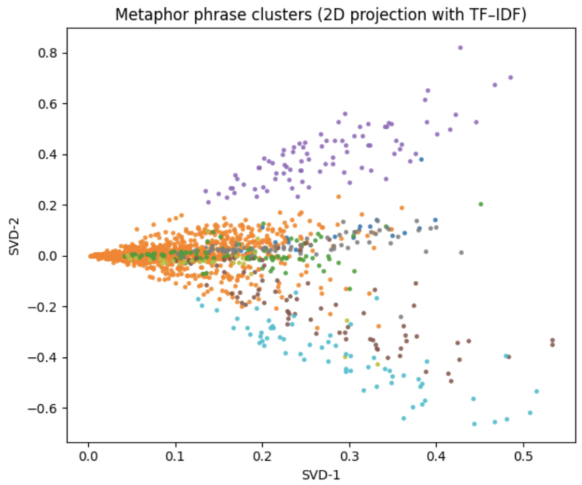

Figure 5: Lexical clusters of metaphorical phrases visualized in 2D using TF–IDF and SVD projection. Each color represents a semantically related group: C0 (head/encouragement), C1 (miscellaneous), C2 (shared experience), C3 (down-related emotions), C4 (fight), C5 (life/control), C6 (journey), and C7 (battle).

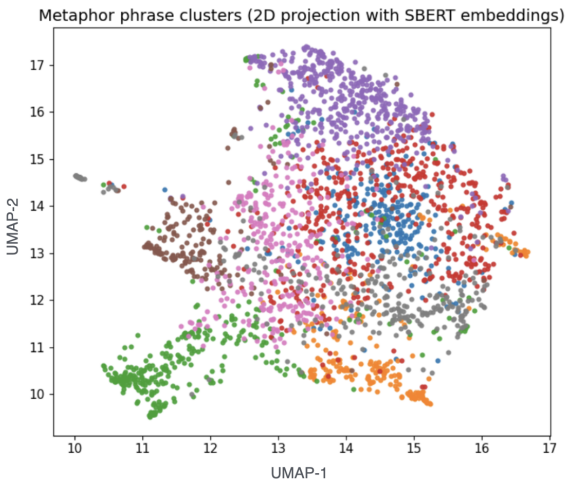

Figure 6: Semantic clusters of metaphorical phrases visualized in 2D using SBERT embeddings and UMAP projection. Each color represents a semantically related group, with representative labels derived using KeyBERT (Grootendorst, 2020) keyword extraction: C0 (struggle/mutual support), C1 (makeup/relationships), C2 (shared feelings), C3 (high school), C4 (accutane/clear skin), C5 (acne scars), C6 (life impacts/anxiety), and C7 (people).

First, we examined lexical themes by representing metaphorical spans using TF–IDF features and clustering these representations using KMeans (Lloyd, 1982). To visualize the resulting clusters, we projected the TF-IDF vectors into two dimensions using Singular Value Decomposition (SVD). This approach groups phrases according to surface lexical similarity, allowing us to identify clusters organized around frequently repeated words or expressions. We selected $k = 8$ clusters after exploratory experiments with values between 5 and 12, balancing interpretability and cluster coherence. The observed major themes are illustrated in Figure 5. C0 focuses on head-related metaphors, often used for encouragement (e.g., "*keep your head up*"). C1 contains miscellaneous words with no cohesive theme (e.g., "*shallow*" or "*man up*"). C2 emphasizes solidarity and shared struggle (e.g., "*in the same boat*"). C3 captures down-related emotional struggle (e.g., "*feeling down*"). C4 has fight-related phrases (e.g., "*fight acne*"), while C5 involves control/life metaphors (e.g., "*control your life*"), showing the dominance of acne over one's life. C6 is journey-related (e.g., "*going through that journey*"), emphasizing endurance and progress. Finally, C7 contains battle-related framings (e.g., "*endless battle*"), depicting acne as an enemy to combat. These clusters highlight recurring lexical patterns centered around topics such as encouragement, solidarity, emotional struggle, and conflict that are used metaphorically to express empathy. The scattered or elongated shapes suggest semantic variation within these themes, reflecting diverse ways metaphorical phrases were used to communicate empathy.

Second, we explored semantic themes by encoding spans with SBERT (Reimers and Gurevych, 2019) embeddings, clustering them with KMeans, and visualizing them using UMAP. Unlike the TF–IDF analysis, which groups phrases by lexical overlap, this method organizes them based on underlying semantic similarity. UMAP parameters were set to *n_neighbors = 20* and *min_dist = 0.08*, which produced stable cluster separation across exploratory runs. The resulting themes are shown in Figure 6. C0 emphasizes acne-related struggle and mutual support. C1 focuses on makeup and romantic relationships, showing how appearance concerns affects dating. C2 captures shared feelings of empathy and C3 centers on high school experiences. C4 includes discussions of treatments such as accutane which is an acne regimen and hopes for getting clear skin. C5 involves acne scarring and C6 reflects the broader life impacts of acne such as anxiety and social life. Finally, C7 discusses people-related concerns (e.g., "When people stare at me, ...").

These two visualizations provide complementary perspectives. Figure 5 reveals commonly occurring themes based on lexical patterns, while Figure 6 uncovers the broader semantic narratives underlying those words. Beyond metaphors of emotional ups and downs or battles in life, we see how these

expressions capture lived experiences: concerns about appearance and the desire to conceal acne with makeup, the impact of acne on dating and relationships, and how the scars extend the struggle even after acne subsides. These patterns illustrate how acne metaphors are used to articulate both immediate emotional states and the longer-term social and psychological dimensions of living with acne.

Lexicon-Based Semantic Category Analysis (Empath)

We further conducted a lexicon-based semantic category analysis of the annotated metaphor spans using Empath (Fast et al., 2016), a tool that estimates normalized word frequencies across more than 200 semantic domains. This allows us to quantify the emotional and topical dimensions of the metaphors, uncovering which categories are most closely tied to empathetic expression. We tokenized all annotated metaphor spans and processed them with the Empath Python library to assign words to predefined categories. Using the normalization option, category counts were divided by the number of tokens in each span and then averaged across all metaphor spans (resulting in small absolute values but informing relative prominence). The results are shown in Figure 7.

Top categories include conflict-related domains like fight, war, and competing, along with negative emotion, strength, and violence-related domains. These results resonate with themes highlighted in the clustering analysis (such as battle, fight, and down-related metaphors) but also bring forward stronger associations with violence and struggle. We then further observe emotion-oriented framing by handpicking a subset of Empath categories related to empathy and affect (Table 3). Among the selected categories, negative emotion emerges most frequently, followed by positive emotion and sadness, with smaller proportions for affection and sympathy. This distribution indicates that metaphorical expressions in empathetic text are more often anchored in negative affect and hardship rather than explicitly signaling warmth or compassion. This pattern is consistent with the broader Empath results (Figure 7) where fight/war and negative emotion categories are most prominent.

Across these analyses, we find that metaphorical expressions in empathetic posts are frequent, diverse, and strategically placed to frame or reinforce support. The clustering analysis captures the breadth of metaphorical language, whereas the Empath analysis quantifies the prominent domains, particularly fight-related and negative emotions. These results illustrate the varied functions of metaphors within empathetic text.

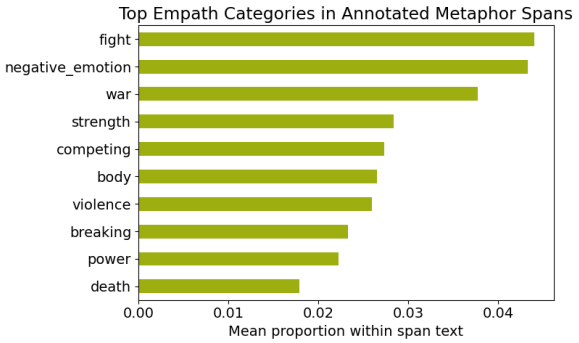

Figure 7: Top Empath categories across annotated metaphor spans. Scores are normalized within each span and averaged across the dataset, showing the relative mean proportion of categories most associated with metaphorical expressions.

Category	Mean Frequency
Negative Emotion	0.043
Positive Emotion	0.017
Sadness	0.015
Affection	0.006
Sympathy	0.005

Table 3: Mean normalized frequency of emotion-related Empath categories within metaphor spans. Values represent the average fraction of tokens per span matching each category.

5. Proof of Concept

In this section, we demonstrate the feasibility of automatically identifying metaphorical expressions in *Empathy-Metaphor*. We frame metaphor detection as a sequence labeling problem, in line with prior work toward identifying metaphors at the token level in continuous text (Leong et al., 2018, 2020). We adopt a BIO (Beginning-Inside-Outside) tagging scheme (Ramshaw and Marcus, 1995) to represent metaphorical spans, allowing us to capture single-word and multiword figurative expressions.

5.1. Models

We evaluated a range of sequence labeling models, from linear and recurrent baselines to pretrained transformers. All models are trained to predict BIO tags over tokenized text, with metaphorical spans labeled as B/I and all other tokens as O. We split the data into 80% training and 20% test sets at the post level. Below we describe the models used, along with their key specifications and training setups.

- **Logistic Regression**: A linear baseline trained on token-level features, with class weighting to handle imbalance.

Model	Token Precision	Token Recall	Token Macro F1	Span F1 (Exact)	Span F1 (IoU $\geq$ 0.5)
Logistic Regression	0.345	**0.669**	0.277	–	–
BiLSTM	0.407	0.533	0.437	0.058	0.115
DistilRoBERTa	0.636	0.491	0.540	0.110	0.244
RoBERTa	0.634	0.592	0.611	0.262	0.402
DeBERTa-v3-base	**0.667**	0.606	**0.634**	0.263	0.411
DeBERTa-v3-large	0.666	0.580	0.616	**0.297**	**0.440**

Table 4: Proof-of-concept results for various sequence labeling models on Empathy-Metaphor. Token-level metrics are macro-averaged. Span-level F1 is reported using both exact boundary matching and relaxed overlap, where IoU (Intersection-over-Union) $\geq$ 0.5 indicates that the predicted and gold spans overlap by at least 50% of their combined length.

- **BiLSTM (Graves and Schmidhuber, 2005)**: A recurrent baseline using bidirectional LSTMs, trained with AdamW (learning rate 1e-3, batch size 64, 4 epochs) and class-weighted loss.
- **DistilRoBERTa-base (Sanh et al., 2019)**: A distilled variant of RoBERTa (Liu et al., 2019) designed for efficiency while retaining most of its performance. Fine-tuned with learning rate 5e-5 (default), batch size 16 (train)/ 32 (test), 2 epochs.
- **RoBERTa-base (Liu et al., 2019)**: A transformer model that improves on BERT (Devlin et al., 2019) with more robust pretraining. Fine-tuned for learning rate 3e-5, batch size 16 (train)/ 32 (test), 4 epochs, weight decay 0.01.
- **DeBERTa-v3-base (He et al., 2023)**: A transformer with disentangled attention that refines BERT/RoBERTa representations for stronger language understanding. Fine-tuned for learning rate 5e-5 (default), batch size 12 (train)/24 (test), 2 epochs.
- **DeBERTa-v3-large (He et al., 2023)**: A larger version of DeBERTa with more parameters and longer context capacity. Fine-tuned for learning rate 5e-5 (default), batch size 8 (train)/ 16 (test), 3 epochs, weight decay 0.01.

We also tested prompting approaches with TinyLlama-1.1B-Chat (Zhang et al., 2024) and Phi-3.5-mini (Abdin et al., 2024), both lightweight instruction-tuned conversational models. They were prompted in zero and few-shot settings but consistently failed to produce reliable span predictions. Given their near-zero performance, we exclude them from quantitative reporting and analysis.

5.2. Experiments

We evaluated all models on the *Empathy-Metaphor* dataset under the BIO tagging setup introduced in Section 5.1. Linear and recurrent baselines were trained from scratch on TF-IDF token-level features, while transformer-based models were fine-tuned using the Hugging Face Transformers library (Wolf et al., 2020). Results are averaged over three runs and reported at two granularities: token-level precision, recall, and macro F1-score on BIO tags, and span-level precision, recall, and macro F1-score requiring exact boundary matches. Macro averaging is used to account for class imbalance by weighting rare metaphor labels equally with the majority class.

In addition, we also report a relaxed span F1 using Intersection-over-Union (IoU) $\geq$ 0.5, which counts a prediction as correct if it overlaps with a gold span by at least half of its length. This provides a more realistic measure of model performance and reflects the inherent subjectivity of figurative language boundaries, since metaphorical expressions often admit multiple reasonable annotator interpretations (e.g., one annotator may mark "*in the same boat*" while another marks "*same boat*"). These evaluation metrics allow us to assess both token-level recognition of metaphorical expressions and span-level boundary accuracy.

5.3. Results

Table 4 presents results of our experiments. At the token level, transformer models consistently outperformed linear and recurrent baselines. Logistic Regression achieved 0.277 macro F1, relying only on shallow token features, while BiLSTM improved to 0.437 macro F1 by capturing sequential dependencies. Pretrained transformers achieved larger gains for macro F1, with DistilRoBERTa reaching 0.540, RoBERTa 0.611, and DeBERTa-v3 variants yielding the strongest results: DeBERTa-v3-base at 0.634 and DeBERTa-v3-large at 0.616. This performance gap demonstrates how large-scale contextual pretraining can enable transformer-based models to better capture metaphorical expressions in empathetic text compared to models trained only on our dataset.

Performance dropped under span-level evaluation, reflecting the challenge of predicting exact boundaries of multiword metaphors. For instance, in a span like "*in the same boat*," all tokens including "in" and "the" are marked as metaphorical under BIO tagging. This can also lower token-level scores, and even more span-level scores since they require exact match of span boundaries. As a result, BiLSTM scored only 0.058 for exact span F1, while transformer-based models improved performance, with DeBERTa-v3-large achieving the best score of 0.297, followed by DeBERTa-v3-base (0.263), RoBERTa (0.262), and DistilRoBERTa (0.110). Unlike token-level results, where DeBERTa-v3-base was strongest, DeBERTa-v3-large performed best in span-level prediction, suggesting that larger models may better capture phrase boundaries even if token-level accuracy is slightly lower.

Relaxed span evaluation with IoU $\geq$ 0.5 shows clear improvements over exact matching. DeBERTa-v3-large still achieved the highest F1 of 0.440, with DeBERTa-v3-base and RoBERTa following closely at 0.411 and 0.402. This pattern is consistent with the exact-span results and further suggests that DeBERTa-v3-large is particularly effective at capturing approximate phrase boundaries. Moreover, the relaxed span evaluation shows that models frequently identify overlapping regions of metaphorical text even when exact boundaries differ. The inclusion of IoU-based metrics therefore provides a more realistic view of model performance by accounting for partial overlaps that still capture the same metaphorical expression.

From our experiments, results show that transformer-based models can reliably identify metaphorical spans in *Empathy-Metaphor*, while span-level performance shows the added complexity of multiword metaphorical spans. This gap motivates future work on boundary-sensitive modeling and more nuanced evaluation methods.

6. Discussion

Our results systematically and comprehensively confirm that metaphorical language is central to empathetic expression in acne peer support. Metaphors were used frequently, appearing in over half of the empathetic posts. They were also strategically positioned, often in the closing segment to reaffirm encouragement or persistence and at the opening to set the supportive tone. Another key finding is that metaphorical expressions in this support community are quite creative. Although metaphors appeared frequently, most spans occurred only once or a few times, and their frequency distribution was highly skewed. This indicates that metaphors were used in highly individualized ways, rather than relying on a fixed set of conventionalized phrases. Such diversity makes it difficult to isolate a single overarching pattern, which is why we complement the analysis with lexical/semantic clustering and Empath-based categories.

Clustering analyses revealed a range of topics in metaphor use. The lexical clusters (TF–IDF) highlighted recurring themes of conflict and struggle, such as wanting to have control over acne, feeling down, or fighting and persisting through the journey more than pure encouragement. The semantic clusters (SBERT) brought out more nuanced concerns, including appearance and makeup, relationships, and the lasting impact of scars. The Empath analysis reinforced this picture, showing that negative emotion and conflict-related categories were especially salient. Note that these were not hostile expressions, but empathetic framings that validated hardship, highlighted shared struggle, and conveyed resilience.

The detection experiments extend these findings by testing whether computational models can recognize the metaphorical patterns we observed. Transformer models outperformed linear and recurrent baselines, showing that large-scale general-domain pretraining can be valuable for detecting metaphorical expressions. At the same time, the gap between token and span-level performance suggests they still struggle with capturing the exact start and end tokens of the expressions. Relaxed span evaluation (IoU $\geq$ 0.5) indicated that models still identified overlapping regions of the target expressions. Therefore, future work may require more complex models that account for boundary sensitivity or evaluation frameworks that better align with human interpretation of metaphorical expressions.

7. Conclusion

In this work, we introduced *Empathy-Metaphor*, the first dataset of metaphor span annotations in empathetic text. Extending an existing empathy corpus with figurative language annotations, this resource offers a fresh angle for studying both metaphor and empathetic language. Our analyses revealed that metaphors in acne peer support are frequent, diverse, and strategically placed, often portraying acne as something that can be fought or controlled rather than passively endured. Through systematic benchmarking experiments, we empirically established that this dataset also supports the development of computational models for span-level metaphor detection. More broadly, it also provides insights that may help inform support strategies in conversational agents or peer-support training across different health communities. We release the dataset and our benchmark models publicly to promote further work on this topic.

8. Limitations

Several limitations remain in this study. Although we developed detailed annotation guidelines and conducted multiple rounds of discussion and adjudication, some subjectivity is inevitable in interpreting metaphorical expressions. Phrase-level detection is also more intuitive for humans than for automated systems, and existing methods may struggle to capture the full range of figurative language with precise span boundaries. Lastly, our findings are grounded in English-language acne support communities. This focused scope allows for in-depth analysis within a specific domain, but may not generalize to other languages or contexts.

9. Ethical Considerations

This study uses data (*AcnEmpathize*) from a publicly available online acne support forum (*acne.org*). Usernames were anonymous and no personal identifiers were provided. All content focused on discussions of skin health and emotional support, and annotators participated voluntarily. We encourage responsible use of this dataset in ways that respect users' expressions of vulnerability and support research on empathy and online peer support.

10. Acknowledgements

We thank our annotators for enabling the creation of this resource. Both authors 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.

11. Bibliographical References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Špela Antloga. 2020. Metaphor corpus komet 1.0. Accessed via European Language Grid.

Calliope Chloe Bandera, Gyeongeun Lee, and Natalie Parde. 2025. Unveiling empathic triggers in online interactions via empathy cause identification. In *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.

Maria Berger, Nieke Kiwitt, and Sebastian Reimann. 2024. Applying transfer learning to German metaphor prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1383–1392, Torino, Italia. ELRA and ICCL.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, Trento, Italy. Association for Computational Linguistics.

Francesca M. M. Citron and Adele E. Goldberg. 2014. Metaphorical sentences are more emo-

tionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Mark Davis. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*, 10:85–103.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 32–41.

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

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA. Association for Computing Machinery.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378—382.

Andrew Goatly. 2007. *Washing the brain: Metaphor and hidden ideology*. John Benjamins Publishing Company, Amsterdam, Netherlands. ID: 2007-03412-000.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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, Santa Fe, New Mexico. Association for Computational Linguistics.

Zoltán Kövecses. 2000. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press, New York, NY, US. ID: 2001-16940-000.

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, Zhu Wang, Sathya N. Ravi, and Natalie Parde. 2024a. EmpatheticFIG at WASSA 2024 empathy and personality shared task: Predicting empathy and emotion in conversations with figurative language. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 441–447, Bangkok, Thailand. Association for Computational Linguistics.

Gyeongeun Lee, Zhu Wang, Sathya N. Ravi, and Natalie Parde. 2025. From heart to words: Generating empathetic responses via integrated figurative language and semantic context signals. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4490–4502, Vienna, Austria. Association for Computational Linguistics.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024b. 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.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the

2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Stuart P. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Amr Molla, Hassan Alrizqi, Emtinan Alharbi, Arwa Alsubhi, Saad Alrizqi, and Omar Shahada. 2021. Assessment of anxiety and depression in patients with acne vulgaris in medina: A case-control study. *Clinical, Cosmetic and Investigational Dermatology*, 14:999–1007.

National Institute for Health and Care Excellence. 2021. Acne vulgaris: management.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

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.

Sebastian Reimann and Tatjana Scheffler. 2024. Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th EMC2 - Energy Efficient Machine Learning and Cognitive Computing*, Vancouver, British Columbia.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Jiayi Shi and Zhaowei Khoo. 2023. Words for the hearts: a corpus study of metaphors in online depression communities. *Frontiers in Psychology*, Volume 14 - 2023.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. A method for linguistic metaphor identification.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model.