AcnEmpathize: A Dataset for Understanding Empathy in Dermatology Conversations

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

Abstract

Empathy is critical for effective communication and mental health support, and in many online health communities people anonymously engage in conversations to seek and provide empathetic support. The ability to automatically recognize and detect empathy contributes to the understanding of human emotions expressed in text, therefore advancing natural language understanding across various domains. Existing empathy and mental health-related corpora focus on broader contexts and lack domain specificity, but similarly to other tasks (e.g., learning distinct patterns associated with COVID-19 versus skin allergies in clinical notes), observing empathy within different domains is crucial to providing tailored support. To address this need, we introduce AcnEmpathize, a dataset that captures empathy expressed in acne-related discussions from forum posts focused on its emotional and psychological effects. We find that transformer-based models trained on our dataset demonstrate excellent performance at empathy classification. Our dataset is publicly released to facilitate analysis of domain-specific empathy in online conversations and advance research in this challenging and intriguing domain.

Keywords: empathy detection, acne, mental health

1. Introduction

Empathy is a multidimensional construct with emotional and cognitive components by which we feel and understand the experiences of others (Davis et al., 1980). Emotional empathy typically entails a quick, involuntary emotional reaction to the experiences and feelings expressed by others, while cognitive empathy involves a more deliberate process that focuses on interpreting and understanding the emotions of the individuals. Empathic interactions have demonstrated their effectiveness in improving mental health support outcomes (Elliott et al., 2018), and the benefits of employing empathy have been showcased across various settings including conversational agents (Shuster et al., 2019; Roller et al., 2020), customer care dialogue agents (Firdaus et al., 2020; Sanguinetti et al., 2020), and online health support communities (Pérez-Rosas et al., 2017; Khanpour et al., 2017). In these scenarios, empathy plays a pivotal role in delivering supportive interactions that enhance user experiences and overall satisfaction. Moreover, studies have revealed that empathetic conversational agents can improve the mood of socially excluded individuals (De Gennaro et al., 2020) and encourage positive behavior change to promote healthier lifestyles (Lisetti et al., 2013).

The surge of interest in empathetic language has given rise to numerous datasets for empathy recognition and empathetic response generation (Lahnala et al., 2022). Notably, *Empathic Reactions* (Buechel et al., 2018) and *Empathetic Dialogues* (Rashkin et al., 2018) have been widely employed in these tasks, with the former being used for empathy detection shared tasks (Tafreshi et al., 2021; Barriere et al., 2022, 2023). More recently, there has been an increased emphasis on studying empathy in mental health contexts. For instance, the impact of empathy on mental health has been explored by Chen and Xu (2021), with Sharma et al. (2021) striving to facilitate empathic conversations within online mental health support communities. This growing attention to mental health has led to the creation of relevant datasets, as seen in work by Sharma et al. (2020) and Hosseini and Caragea (2021). Nevertheless, these datasets often lack domain specificity, as they encompass a wide range of mental health concerns. Combining various mental health concerns can complicate the recognition and interpretation of empathy since individuals experience challenges in crucially different ways.

Dermatology, and particularly acne, is an underexplored domain in this area, despite its commonality and profound impact on mental health. Acne is a skin condition that affects approximately 85% of young adults at some point in their lives. Multiple studies have highlighted that over 40% of individuals with acne experience depression and anxiety, while 6-7% may face suicidal tendencies (Molla et al., 2021). The National Institute for Health and Care Excellence (NICE) has recognized this connection and issued mental health support guidelines for individuals coping with acne (National Institute for Health and Care Excellence, 2021). This underscores the pressing need for a domain-specific

LREC-COLING 2024, pages 143–153
20-25 May, 2024.

dataset to facilitate better understanding of empathy within the acne community.

To address this limitation, we introduce *AcnEmpathize*[1], a dataset of more than 12K posts with empathy annotations collected from an online acne support community. This dataset is derived from `acne.org`, a platform dedicated to providing information and support to individuals dealing with acne-related issues. More specifically, the data was scraped from the forum titled "Emotional and Psychological Effects of Acne," where people openly discuss their feelings and extend support to one another. Our specific contributions are as follows:

- We systematically curate 12,249 forum posts and associated metadata pertaining to the emotional and psychological effects of acne, and solicit annotations for these posts from three trained annotators.
- We establish performance benchmarks for this data, offering strong baselines through evaluations with different machine learning models.
- We perform comprehensive analyses of this dataset to enhance the collective understanding of empathy usage within the context of emotional and psychological effects of acne.

By addressing the unique challenges faced by individuals in the acne community, our dataset will enable precise and focused analysis, tailored support, and a fresh angle in the study of empathetic language. Specifically, we aim to inspire further development of empathetic chatbots and enhance peer support within online communities across various domains, beyond the acne community.

2. Related Work

Existing empathy detection corpora are typically designed for general purposes and lack domain-specific focus. For instance, Buechel et al. (2018) introduced a dataset comprising 2K messages for empathy prediction in written language. This dataset collected reactions to online news articles that evoke emotions related to suffering and need. It provides gold ratings for both empathy and distress, which were annotated by participants tasked with reading these articles. Additionally, Rashkin et al. (2018) presented a dataset featuring 25k dialogues, which are grounded in situations prompted by specific emotion labels and come with human empathy ratings. Sharma et al. (2020) presented EPITOME, a framework for expressed empathy, and collected 10K conversations from 55 mental health subreddits and TalkLife, a global peer-to-peer mental health support network. The dataset was annotated, and a computational method was proposed for understanding expressed empathy in text-based, asynchronous conversations on mental health platforms. These papers, however, serve general-purpose empathy detection, lacking specificity in their applications.

Hosseini and Caragea (2021) presented a dataset containing 5K messages collected from an online cancer network, the Cancer Survivors Network (CSN). The dataset is annotated with whether each message seeks or provides empathy. Thus, its purpose is tangential to facilitating empathy detection itself, since its focus is instead on differentiating between seeking and providing empathy.

Outside of empathy, datasets related to mental health and skincare domains have been proposed, but they are not suitable for empathy detection tasks. For example, Sharma and De Choudhury (2018) presented a dataset collected from 55 Reddit communities serving various psychological needs, such as Trauma & Abuse, Psychosis & Anxiety, Compulsive Disorders, Coping & Therapy, and Mood Disorders. However, this dataset lacks empathy labels and doesn't distinguish between mental health conditions and their potential causes. Similarly, Fettach and Benhiba (2019) proposed a dataset scraped from Pro-Eating Disorders (Pro-ED) and Pro-Recovery (Pro-Rec) communities on Reddit. Although this dataset explores a specific domain and offers potential for interesting analyses by comparing Pro-ED and Pro-Rec communities, it also lacks empathy labels.

In response to the limitations and gaps observed in existing empathy detection and mental health datasets, we introduce a pioneering empathy detection corpus that focuses on a specific domain centered around acne and its emotional and psychological effects. Our dataset, to our knowledge, is the largest in scale among text-based asynchronous datasets created for empathy detection, featuring over 12K annotated texts.

3. Methods

3.1. Data Collection

Our data was sourced from the website `acne.org`, which serves as a comprehensive resource for individuals dealing with acne. This platform provides information about acne, offers products for which users can leave reviews and ratings, and hosts discussion forums, among other features. We specifically collected our data from the "Emotional and Psychological Effects of Acne" forum, dedicated to discussions about the emotional and psychological aspects of dealing with acne. The majority of con-

[1] https://github.com/gyeongeunlee16/AcnEmpathize

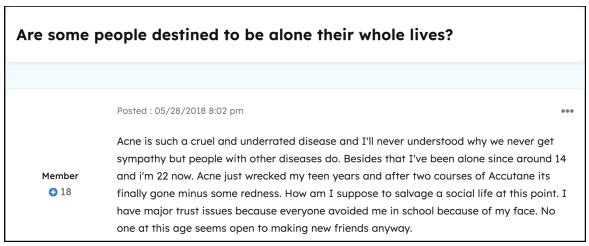

Figure 1: Example of an initial post.

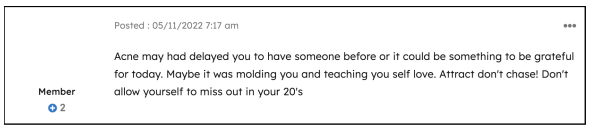

Figure 2: Example of a reply.

versations in this forum revolve around users' experiences of stress, anxiety, and concerns related to their struggles with acne. In response, other users offer empathy and advice. Our research received an exemption approval from the Institutional Review Board (IRB) at the University of Illinois Chicago (UIC) and was determined not to constitute human subjects research.

Within this forum, conversations consist of three types of posts: initial posts, replies, and quotes. An initial post is created with a title by a user who initiates a conversation, typically seeking empathy and support (example shown in Figure 1). The user photo and name have been masked out to de-identify data and messages in public presentations, following guidelines from Benton et al. (2017) to minimize risks for users.

After a user creates a conversation with an initial post, other users can contribute to the conversation by replying to or quoting any post within the conversation. A reply, as depicted in Figure 2, usually involves responding to either the initial post or any other post (indicated with a "@username" prefix if replying to posts other than the initial one). On the other hand, a quote, as shown in Figure 3, is similar to a reply in that it can be made to any post in the same conversation, but it includes a block of original text that the person is quoting, typically with their own opinion expressed under the quoted text.

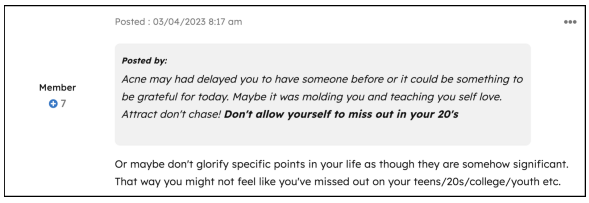

Figure 3: Example of a quote.

At the time of data collection, there were a total of 1,870 conversations available from the "Emotional and Psychological Effects of Acne" forum. However, the number of posts within each conversation varied widely, ranging from 1 (a post created by the initial author with no replies) to 7,740. To address this, we calculated the Interquartile Range (Dekking, 2005, IQR) from the full distribution of posts per conversation across all conversations in the forum. Since the distribution was skewed towards shorter conversations, this suggested that we set the bounds for the number of posts in a conversation to any integer between 1 and 23, inclusive, to remove outlier conversations from our dataset. After filtering the conversations, we were left with 1,740 conversations and a total of 12,249 posts, including replies and quotes.

We collected the URL, title, post ID, user ID, content (text), and, in the case of quotes, the post ID of the quoted text. Additionally, we gathered user information, which encompassed their reputation score,[2] number of product reviews, number of accounts they were following, number of followers, and number of conversations they had created. The extraction of user features was intended to facilitate analyses related to user behaviors and activities and how those qualities might impact the empathy expressed in their posts.

3.2. Preprocessing

Each post obtained in the previous step was preprocessed by removing newline characters from the text and retaining only posts containing at least one alphabetical token.[3] As a result of this preprocessing, blank posts (including those with only whitespace), quotes with no additional text, and posts consisting solely of special characters or emojis, were removed. Posts from the same conversation were grouped together and assigned unique integer conversation IDs. Non-quote posts, encompassing initial posts and replies, were designated with a "-1" label in the "quoted_id" column, which tracks the post ID of the post being quoted. After the preprocessing, a total of 1,730 conversations and 12,212 posts remained.

In some cases, no user information was available. This occurred when either the post was made by an anonymous account or by a deleted profile. For these cases, we modified the values of columns related to reputation score, number of product reviews, number of followed accounts, number of

[2]The method for computing reputation score, a non-negative integer that pertains to the reputation of posters, is not described on `acne.org`.

[3]We set the minimum word count to one alphabetical token to accommodate potentially shorter empathetic expressions, such as "Aw," "That sucks," or "I can relate."

Data	Total
Number of Title Posts	1,730
Number of Quotes	2,591
Number of Replies	7,891
Final Corpus	12,212

Table 1: Data collection statistics.

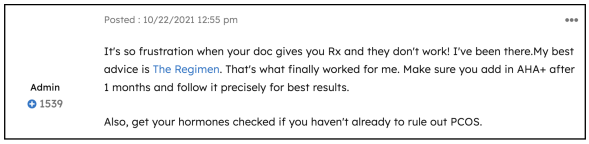

Figure 4: Example of a post that contains empathy.

followers, and number of created conversations from "None" to "-1" to facilitate more straightforward downstream processing. Summary statistics regarding the number of title posts, quote posts, and reply posts along with the overall corpus size are provided in Table 1.

3.3. Annotation Process

The annotation process involved the collaborative efforts of one graduate student (one of the authors) and two undergraduate students who labeled each post as containing empathy (1) or not containing empathy (0). All annotators volunteered for this task. We adopted the annotation guidelines from Sharma et al. (2020), designed specifically for empathy annotation in text-based asynchronous conversations. The annotators were instructed to read Sharma et al. (2020) in order to enhance their understanding of empathy and its framework.

For each post, the annotators were provided with a set of three questions aimed at evaluating whether the post exhibited any communication mechanism of empathy, encompassing emotional reactions, interpretations, and explorations. To further clarify their understanding of empathetic sentences, annotators were provided with concrete examples (refer to Table 2) of each mechanism. If they answered "yes" to at least one of the questions, they were directed to label the post as containing empathy (1). Furthermore, they were guided to mark posts containing solely advice or factual information as not containing empathy (0). Note that our objective was to identify empathy anywhere within posts, regardless of the entire post being empathetic. Figure 4 provides a sample post labeled as "containing empathy," illustrating an instance where a poster communicates an understanding of the seeker's experiences and feelings and then offers advice.

During the initial round of annotation, each annotator labeled 100 randomly sampled posts and took notes highlighting the specific empathetic portions within each post. Following this, they engaged in discussions to address any disagreements, ultimately adjudicating the disagreed-upon labels and attaining a perfect inter-annotator agreement (IAA) following discussion, as measured using the Krippendorff's Alpha (Krippendorff, 1970) coefficient. The second round of annotation involved the labeling of an additional 900 randomly sampled posts, following the same annotation process. An initial IAA of 0.763 was achieved for this sample. The annotators meticulously discussed the areas of discrepancy and again adjusted and adjudicated their labels following discussion to achieve perfect agreement. For the final round of annotation, the remaining posts were equally divided among the three annotators and single-annotated.

3.4. Annotation Discussion

A substantial portion of the discussion time revolved around addressing ambiguous examples. These cases typically featured sentences like "I've been there too" or "I see what you mean," which were frequently used in posts to express an understanding of another person's experience without explicitly sharing their feelings. Some users furthered the conversation by immediately introducing a contrasting viewpoint, elaborating on how "it could be worse," or even offering advice based on their own experiences. As the remainder of the sentences in those posts did not overtly demonstrate empathy, determining whether these posts should be categorized as empathetic or not presented a challenge. Our consensus was that relating to another person's experience often involves an understanding of any associated feelings. Consequently, we decided to label examples expressing any understanding of associated feelings as containing empathy.

There were also vaguely empathetic posts that were difficult to annotate. For example, one user replied to the original author saying "What the hell is wrong with him?" and showing their anger towards the author's attitude. However, without looking at the original text, it was hard to tell whether the user was sharing their own experiences and feelings or being empathetic to the original poster. Making this determination would require us to make assumptions about the original post instead of focusing on detecting empathy in the reply. Thus, in cases such as this, we decided not to make additional assumptions and consider these examples as not containing empathy.

Another topic of discussion revolved around to whom the expressed empathy should be directed. The annotators debated whether it should be directed solely toward the target user (the user to whom they are replying or quoting) or toward any

Types	Questions	Examples
Emotional Reactions	Does the response express or allude to warmth, compassion, concern, or similar feelings of the responder towards the seeker?	• Everything will be fine. • I feel really sad for you.
Interpretations	Does the response communicate an understanding of the seeker's experiences and feelings? In what manner?	• I understand how you feel. • This must be terrifying. • I also have anxiety attacks at times which makes me really terrified.
Explorations	Does the response make an attempt to explore the seeker's experiences and feelings?	• What happened? • Are you feeling alone right now?

Table 2: Examples of texts categorized into three communication mechanisms of empathy.

Ambiguity	Empathy?
Expresses understanding of associated feelings without overt reference to shared feelings	✓
Requires assumptions to be made about a different post	✗
Directed toward anyone in the group including the target user	✓
Expresses kindness or support without other overt empathy	✗
Focuses on personal experience without relating to the experience or emotion of an external target	✗

Table 3: Summary of annotator discussion regarding ambiguities in empathy labeling.

individuals. Some posts, such as "We are all here together" and "I know guys...so tough on us emotionally," appeared empathetic, but it wasn't clear whether this empathy was intended for the target user or if it was extended to a broader, unidentified audience. A consensus was reached among the annotators that empathetic comments should be directed towards anyone within a group that includes the target user. Thus, empathetic posts containing words like "we" and "everyone" were designated as containing empathy.

Mere expressions like "good luck," while seemingly supportive, were not considered to contain empathy because they were simply expressions of kindness and support. Similarly, statements such as "I really do wish the best for every one of you" without other empathetic phrases, did not qualify. An additional challenge arose when annotating examples that acknowledged positive experiences. Instances like "I am so glad to hear the cysts are gone!" and "I'm very happy to see that you're clear" were encountered. However, these posts primarily focused on the feelings of the authors themselves rather than relating to the experiences and emotions of the target user. No similar examples were found in the annotation guidelines provided by Sharma et al. (2020). Therefore, we decided to mark these examples as not containing empathy.

Overall, discussing these ambiguities in the first two rounds of annotations ensured the production of high-quality and consistent annotations in the final round. We summarize the outcomes of our annotator discussion in Table 3.

4. Dataset Analysis

4.1. Dataset Composition

Our completed AcnEmpathize dataset contains a total of 1,730 title posts, 2,591 quotes, and 7,891 replies (see Table 1). On average, each conversation consists of approximately 7.059 posts including the title post, with a median of 6.000 posts and a standard deviation of 5.123 posts. The conversations range from having only a title post to the largest ones containing 23 posts. Larger conversations with more than 23 posts were filtered out during the preprocessing step, as discussed in Section 3.1.

As indicated in Figure 5, our final annotated corpus comprises 2,976 posts containing empathy and 9,236 posts that do not. The imbalanced distribution stems from our strict adherence to the annotation guidelines for determining the presence of empathy, as established by Sharma et al. (2020). This distribution aligns with findings from their work, where similar ratios were reported (2,965:7,178 and 2,406:7,737 for *Empathy* to *No Empathy* for different communication mechanisms of empathy[4]). We

[4]Rather than annotating text instances for empathy holistically, Sharma et al. (2020) subdivided their annotations into different communication mechanisms involved

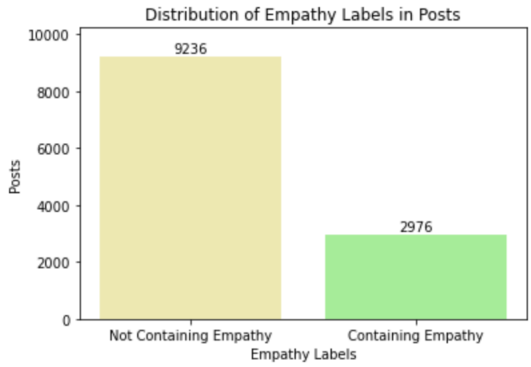

Figure 5: Empathy label distribution.

addressed ambiguities according to the processes described in Section 3.4 and summarized these ambiguities in Table 3; we note that this resulted in a majority of ambiguous examples being annotated as not containing empathy.

In total, AcnEmpathize includes 2,527 unique users, excluding anonymous users who are assigned user IDs of 0. The number of posts made by users ranges from 1 to 257, with an average of 4.398 posts per user (median of 1.000 post and a standard deviation of 11.464 posts). Notably, six "super users" contributed 100 or more posts each. The average reputation score across all posts is 63.271, ranging from 0 to 2,481 across all posts, with a median of 10.000 and a standard deviation of 210.867. Users on average left 1.251 product reviews (with a median of 0.000 and a standard deviation of 4.400), followed 7.516 users (median of 2.000 and a standard deviation of 15.129), and had 9.259 followers (median of 2.000 and a standard deviation of 19.897). Additionally, users created an average of 274.923 forum posts in all available forums on `acne.org` (with a median of 91.000 and a standard deviation of 478.403), covering a variety of other topics beyond our forum of focus ("Emotional and Psychological Effects of Acne").

4.2. Dataset Content

We tokenized each post, removed punctuation and stopwords, performed case normalization, and lemmatized all tokens. We then applied Latent Dirichlet Allocation (Blei et al., 2003, LDA) to the preprocessed data, which resulted in the extraction of 16 coherent topics. The top 10 words for each of these topics are presented in Table 4.

The average number of tokens in each post is approximately 153.884 (with a minimum of 1.000 and a far outlying maximum of 39,464.000), having a median of 92.000 tokens and a standard deviation of 413.778 tokens. This hints at a wide range of token counts, with one outlying post with ID "3566573" containing over 10,000 tokens. To further analyze linguistic patterns in texts containing and not containing empathy, we computed the log odds ratio using an informative Dirichlet prior (Monroe et al., 2008; Hessel, 2016), removing contractions containing stopwords such as "ill" and "youre" since the base tokens of those contractions ("i" and "you") would be considered stopwords. We present the results of this analysis in Figures 6 and 7. Note that common text messaging acronyms (e.g., "lol") were retained since they are not stopwords; however, although "lol" appears in the plot in Figure 6 it

Topic	Words
1	life, acne, thing, let, positive, great, think, get, may, skin
2	skin, month, time, picking, back, started, go, made, way, pick
3	people, think, like, acne, someone, thing, really, know, feel, say
4	acne, diet, food, try, help, work, eat, really, skin, think
5	skin, acne, look, like, people, feel, see, face, think, really
6	girl, woman, guy, attractive, men, make, attraction, shit, beauty, f**k
7	Lol, Yea, independent, hahaha, Choose, lookin, outcome, Looks, Canada, OMG
8	Thanks, Thank, reply, thank, thanks, Wow, sharing, definitely, much, glad
9	wedding, Glad, F**k, refreshing, going, five, recovery, inspirational, haircut, instrument
10	acne, year, life, time, back, could, day, go, thing, still
11	get, scar, help, u, know, skin, thing, good, need, better
12	depression, anxiety, disorder, bipolar, mental, meditation, OCD, diagnosed, form, therapy
13	Great, rash, band, aid, Yep, cent, nope, Screw, Live, Ah
14	like, acne, feel, know, really, want, even, get, go, year
15	taste, tea, input, measure, seed, lemon, Aw, green, Exactly, apple
16	skin, acne, face, pimple, red, week, clear, using, month, got

Table 4: Top 10 words for identified LDA topics.

in the expression of empathy.

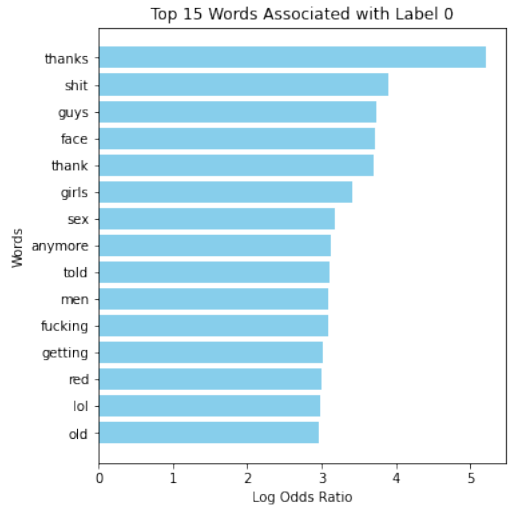

Figure 6: Words most closely associated with not containing empathy class. Content Warning: Contains profanity.

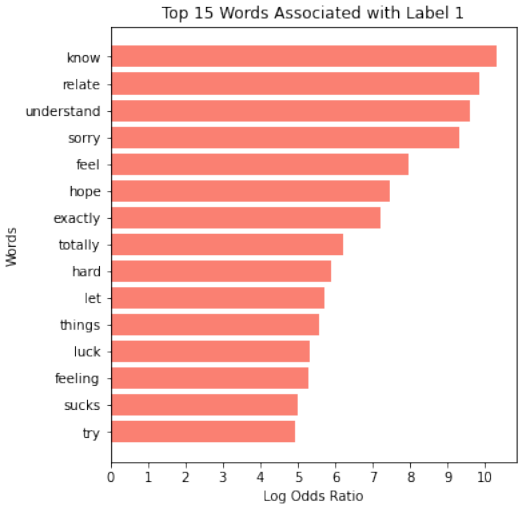

Figure 7: Words most closely associated with containing empathy class.

is more of an expression of lightheartedness rather than a topical word.

Figure 6 showcases the top 15 words closely associated with posts not containing empathy, while Figure 7 presents words closely associated with posts containing empathy. Words in Figure 6 tend to have more negative connotations (e.g., "sh*t" and "f**king"), whereas words in Figure 7 lean towards warmer and compassionate expressions (e.g., "relate," "understand," and "sorry").

5. Proof of Concept

To assess the validity of our annotated dataset, we formulated a text classification problem for empathy labels to evaluate the suitability of AcnEmpathize for domain-specific automated empathy detection. We describe our models, experimental setup, and results in the following subsections.

5.1. Models

To benchmark performance, we considered a variety of baseline and high-performing contemporary text classification models. Specifically, we compared the following conditions:

- **Most Frequent Class:** A baseline model that predicts the most frequent class from the training data, establishing class-specific performance floors with respect to label imbalance.
- **Random:** A baseline model that randomly predicts labels, serving as a performance floor and a comparative proxy for random chance.
- **Naive Bayes:** A well-known probabilistic machine learning algorithm, which leverages Bayes' theorem and the assumption of feature independence for text classification.
- **Logistic Regression:** A linear model that uses the logistic function for binary classification.
- **BERT (Devlin et al., 2018):** A transformer-based model that employs self-attention mechanisms and a bidirectional architecture to capture contextual information in text.
- **RoBERTa (Liu et al., 2019):** An extension of BERT that optimizes the pre-training process for improved performance on NLP tasks.
- **DistilBERT (Sanh et al., 2019):** A distilled version of BERT, designed to be lighter and more computationally efficient while retaining most of its performance.
- **BART (Lewis et al., 2019):** A transformer model trained on the MultiNLI dataset (Williams et al., 2018), tailored for tasks related to natural language inference, combining bidirectional and autoregressive capabilities.

We fine-tuned the transformer-based models and trained the other models entirely on our data. For the transformer-based models, we used the base models from the HuggingFace Transformers library[5] and configured the hyperparameters:

- **max_length:** The maximum length of the input sequences, set to 256.

[5] https://huggingface.co/docs/transformers

		No Empathy			*Empathy*		
Model	**Accuracy**	**Precision**	**Recall**	**F_1**	**Precision**	**Recall**	**F_1**
Most Frequent Class	0.756	0.756	**1.000**	0.861	0.000	0.000	0.000
Random	0.494	0.748	0.499	0.598	0.235	0.479	0.316
Naive Bayes	0.775	0.775	**1.000**	0.873	**0.917**	0.020	0.039
Logistic Regression	0.844	0.864	0.948	0.904	0.737	0.497	0.594
BERT	0.891	0.930	0.928	0.929	0.760	0.766	0.763
RoBERTa	0.885	**0.956**	0.892	0.923	0.703	0.861	**0.774**
DistilBERT	**0.893**	0.926	0.936	**0.931**	0.777	0.746	0.761
BART	0.269	0.809	0.067	0.124	0.231	**0.946**	0.372

Table 5: Results from experiments for each model.

- **lr (learning rate):** The learning rate for the optimizer AdamW, set to 1e-5.
- **num_epochs:** The number of training epochs, set to 3.
- **batch_size:** The batch size for training data, set to 32.

For the Naive Bayes and Logistic Regression models, we used the scikit-learn library[6] and held all parameters at their default values. All models were implemented using Python.

5.2. Experiments

In our experimental setup, we randomly divided the data into an 80% training and 20% testing split for all posts. For BERT, RoBERTa, and DistilBERT, the maximum sequence length was set to 256 (the average token count in preprocessed posts, 154, rounded up to the next power of 2). For Naive Bayes and Logistic Regression, we employed TF-IDF feature vectors and configured the maximum number of features to 5,000. Default values were maintained for all unspecified parameters.

5.3. Results

We evaluated model performance using accuracy, precision, recall, and F_1 for each class label (*No Empathy* and *Empathy*) separately to account for the imbalanced distribution. Accuracy indicates the correctness of predictions, precision assesses the ratio of true positive predictions among all positive predictions, recall measures the proportion of true positive predictions among all actual positives, and the F_1 is a harmonic metric balancing precision and recall. Our findings are summarized in Table 5.

All models generally exhibit stronger performance in the prediction of posts that have *No Empathy* labels. Among these models, the transformer-based BERT, RoBERTa, and DistilBERT stand out with high accuracy (accuracy=0.891, 0.885, and 0.893, respectively) and well-balanced precision, recall, and F_1 for both *No Empathy* and *Empathy* classes. Notably, RoBERTa demonstrates relatively superior performance (precision=0.956, recall=0.892, and F_1=0.923) for predicting the *No Empathy* class versus the *Empathy* class (precision=0.703, recall=0.861, and F_1=0.774).

Logistic Regression also performs well, with high accuracy (accuracy=0.844) and a well-balanced F_1 (F_1=0.904) for the *No Empathy* class. However, the performance for the *Empathy* class is imbalanced, with a higher precision (precision=0.775) but a lower recall (recall=0.497). Naive Bayes demonstrates a similar performance, with reasonably high accuracy (accuracy=0.775) and F_1 (F_1=0.873) for the *No Empathy* class but poor performance on recall (recall=0.020) and F_1 (F_1=0.039). BART performs poorly in terms of accuracy (accuracy=0.269), recall (recall=0.067), and F_1 (F_1=0.124) in the *No Empathy* class and precision (precision=0.231) and F_1 (F_1=0.372) in the *Empathy* class. The subpar performance of BART may be attributed to various factors, such as the imbalanced distribution of posts labeled as containing or not containing empathy, as well as the model's primary focus on natural language inference tasks.

All models far exceeded performance of the naive baselines (Random and Most Frequent Class) with the exception of precision, recall, and F_1 for the class *Empathy* for Most Frequent Class, which was expected given the natural imbalance of the dataset. This clearly establishes validity of the dataset for learning to model empathy within this domain. Despite some variations in model performance, the results underscore the dataset's reasonable performance when used with a range of machine learning models. Furthermore, the dataset offers potential to support empathy detection tasks in the domain of acne and mental health, thereby contributing to the diversity and real-world applicability of natural language processing applications.

[6] https://scikit-learn.org/

6. Limitations

Our work is constrained by several factors. The distribution of posts labeled as either containing or not containing empathy is imbalanced, which could impact the models' performance. While transformer-based models generally performed well, the BART model demonstrated subpar performance. The manual annotation process is inherently subjective and may introduce potential bias and result variability, although we sought to control and correct for this through rigorous annotation standards and follow-up discussion. Lastly, the results and analyses of this study may not generalize to other domains, since they are specifically focused on the `acne.org` community.

7. Conclusion and Future Directions

In this paper, we introduced AcnEmpathize, the first dataset of its kind focusing on empathy detection within the domain of acne and its psychological effects. Comprised of 12K+ conversations, it stands as one of the most extensive corpora annotated for expressed empathy. We make the dataset publicly available to facilitate future research in automated empathy detection and beyond. The dataset's suitability and validity for domain-specific empathy detection have been substantiated through benchmarking experiments, achieving excellent performance for fine-tuned transformer-based models with high overall accuracy and well-balanced precision, recall, and F_1 for both *No Empathy* and *Empathy* classes. In the future, we hope to further develop automated empathy detection methods that allow models to more capably recognize and interpret empathy within this domain, particularly by exploring the use of figurative language and other linguistic elements within the text.

8. Ethical Considerations

As detailed in §3.1, this research obtained an exemption approval from the IRB at UIC for human subjects research. Our primary data source is `acne.org`, a public website that provides open access to forum posts. As explained in §3.3, the annotators participated voluntarily. We make our dataset publicly available, in hope to facilitate research into empathetic language within the dermatology context, specifically relating to acne. It will allow for the exploration of empathetic dialogues, since we preserved available back-and-forth conversations between the users, and may be useful for building chatbots aimed at providing social support in online communities.

Acknowledgements

We thank the student annotators and anonymous reviewers for their helpful feedback.

Bibliographical References

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525.

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

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

Yixin Chen and Yang Xu. 2021. Exploring the effect of social support and empathy on user engagement in online mental health communities. *International Journal of Environmental Research and Public Health*, 18(13):6855.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology*.

Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology*, 10:3061.

F.M. Dekking. 2005. *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer Texts in Statistics. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Yousra Fettach and Lamia Benhiba. 2019. Pro-eating disorders and pro-recovery communities on reddit: text and network comparative analyses. In *Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services*, pages 277–286.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4172–4182.

Jake Hessel. 2016. Implementation: Fightin' words. https://github.com/jmhessel/FightingWords.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Allison Lahnala, Charles Welch, David Jurgens, and Lucie Flek. 2022. A critical reflection and forward perspective on empathy and natural language processing. *arXiv preprint arXiv:2210.16604*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)*, 4(4):1–28.

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

Amr Molla 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*, pages 999–1007.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

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

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Scalerandi Marco, Mana Dario, Simeoni Rossana, et al. 2020. Annotating errors and emotions in human-chatbot interactions in italian. In *The 14th Linguistic Annotation Workshop*, pages 1–12. 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. arxiv 2019. *arXiv preprint arXiv:1910.01108*.

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*, pages 194–205.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13.

Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2019. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *arXiv preprint arXiv:1911.03768*.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.