Pouring Your Heart Out: Investigating the Role of Figurative Language in Online Expressions of Empathy

Gyeongeun Lee and **Christina Wong** and **Meghan Guo** and **Natalie Parde**
Department of Computer Science
University of Illinois at Chicago
{glee87, cwong49, mguo13, parde}@uic.edu

Abstract

Empathy is a social mechanism used to support and strengthen emotional connection with others, including in online communities. However, little is currently known about the nature of these online expressions, nor the particular factors that may lead to their improved detection. In this work, we study the role of a specific and complex subcategory of linguistic phenomena, figurative language, in online expressions of empathy. Our extensive experiments reveal that incorporating features regarding the use of metaphor, idiom, and hyperbole into empathy detection models improves their performance, resulting in impressive maximum F_1 scores of 0.942 and 0.809 for identifying posts without and with empathy, respectively.

1 Introduction

Empathy is a complex multidimensional communicative tool that involves the capacity to recognize and respond to the emotional experiences of individuals seeking help (Davis et al., 1980). It is employed in therapeutic interactions (Elliott et al., 2018), including in the digital realm where millions seek solace and support within online communities (Eysenbach et al., 2004). Interest in online empathy has surged recently, leading to the creation of datasets in generalized domains like mental health (Sharma et al., 2020; Hosseini and Caragea, 2021b). However, these broad investigations of empathy across diverse conditions make it difficult to draw nuanced conclusions regarding the nature of empathetic language. Most investigations arising from large-scale initiatives (e.g., shared tasks organized by Barriere et al. (2022, 2023)) have predominantly relied on black-box approaches and focused on emotions and user demographics, hindering the development of a more comprehensive understanding of empathy in online environments and underscoring the need for further research.

We conduct a two-pronged investigation to address these gaps. We (1) focus our study on domain-specific detection of expressed empathy (Barrett-Lennard, 1981) related to the emotional and psychological effects associated with acne. By delving into this specific condition known to affect social and mental well-being (Molla et al., 2021), our work paves the way for more specialized empathy understanding and support mechanisms. We also (2) investigate the role of figurative language in expressions of empathy within this dataset, addressing an unexplored aspect of contemporary empathy detection research. We further demonstrate that integrating features describing the presence of metaphor, idiom, and hyperbole enhances empathy detection performance across feature-based and pretrained language model (PLM) classifiers, offering a deeper understanding of empathetic expressions beyond emotional and demographic indicators. Our key contributions include:

- We detect empathy in a new dataset, AcnEmpathize, with over 12K posts from an acne-focused online forum.
- We analyze the role of figurative language in empathetic expressions, focusing on idioms, metaphors, and hyperboles.
- We demonstrate enhanced empathy detection performance by integrating figurative language features, motivating the need for more focused study of the linguistic phenomena giving rise to empathy.

We hope that the outcomes from this research inspire further work towards improved AI-driven support systems, including empathetic chatbots tailored to specific needs. Additionally, our research may provide avenues for enhancing empathy training and feedback mechanisms for peer supporters in online communities, ultimately elevating the quality of support available to those seeking help.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 519–529
August 11-16, 2024 ©2024 Association for Computational Linguistics

2 Background

2.1 Datasets for Empathy Detection

Despite the growth of online support communities (generally understood to be rich sources of empathy) in recent years, there are few publicly available empathy detection datasets. *Empathic Reactions* (Buechel et al., 2018) was one of the earliest, composed of reactions to 2K online news articles covering general topics related to suffering. While utilized in empathy detection tasks for the Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) (Tafreshi et al., 2021; Barriere et al., 2022), its relatively small size and lack of domain specificity pose challenges in discerning nuanced traits of empathy. *EPITOME* (Sharma et al., 2020) consists of 10k pairs of posts and responses gathered from 55 mental health-related subreddits and the TalkLife[1] support network. Similar to *Empathic Reactions*, *EPITOME* aggregates various mental health conditions and is not well-suited for conducting in-depth empathy analysis targeted at specialized concerns.

On the other hand, recent datasets by Hosseini and Caragea (2021a,b) include 5K sentences from an online cancer network and 3K tweets related to cancer. Their focus on distinguishing between seeking and providing empathy, however, makes the work peripheral to ours. Additionally, their datasets are considerably smaller than AcnEmphathize (see in §3), which encompasses 121K sentences across 12K posts. Finally, the dataset Omitaomu et al. (2022) used for WASSA 2023 extends *Empathic Reactions*, with further annotations on self-report empathy, second-person perceived empathy, and turn-level empathy, among other features. Despite these advancements, the dataset still lacks comprehensive exploration of domain-specific empathy.

2.2 Text-based Empathy Detection Methods

Many methods for detecting empathy leverage PLMs, often incorporating features related to demographic information. For instance, Kulkarni et al. (2021) proposed a multi-task framework integrating RoBERTa-base (Liu et al., 2019) with demographic and personality information. Similarly, Chen et al. (2022) fine-tuned RoBERTa-large by incorporating fine-grained demographic attributes. Hosseini and Caragea (2021a) alternatively incorporated emotion and sentiment knowledge into their multi-task training system. In a slightly different approach, Lin et al. (2023) proposed a unified ensemble network of sentiment-enhanced RoBERTa-based models without additional features. Omitaomu et al. (2022) also developed a neural architecture with an attention layer and fine-tuned RoBERTa-base. Most of these approaches focus narrowly on sentiment or demographic features, although effective empathy requires appropriately responding to emotions beyond merely understanding them (Davis et al., 1980). In text-based empathy, this entails the ability to effectively convey understanding in written language. This may be facilitated by the use of more complex linguistic phenomena, such as figurative language.

2.3 Figurative Language in Empathy

Despite the pervasiveness of figurative expressions (Citron et al., 2016), research on emotion-conveying language has predominantly dealt with literal language. Expressions are considered figurative if they deliver meaning beyond their literal interpretation (Bowdle and Gentner, 2005). For example, saying *She had a* ***rough day*** communicates that the day was *difficult*, rather than referring to the rough texture (Citron et al., 2016). On an emotional level, figurative language, expressed through what are often referred to as *figures of speech*, is often employed to evoke stronger emotions.

Metaphors, in particular, have been extensively studied for their ability to convey emotional intensity (Fainsilber and Ortony, 1987; Gibbs Jr et al., 2002; Fussell and Moss, 2014; Dankers et al., 2019). They are not only more emotionally charged than their literal counterparts (Citron and Goldberg, 2014) but also enhance the performance of text-based emotion prediction (Dankers et al., 2019). Yet, little is known of when and how people use figurative language in specialized emotional settings. This creates rich opportunity for the investigation of figurative language use in domain-specific empathy. We further define the types of figurative language included in our study in §4.1.

3 Data

The scarcity of domain-specific empathy datasets has constrained the depth to which it may be studied in specialized settings. Our new dataset, AcnEmpathize (Lee and Parde, 2024), focuses entirely on acne-related conversations. Our objective in developing this dataset was to foster enhanced analy-

[1] https://www.talklife.com

sis of empathetic interactions in a well-motivated (Eysenbach et al., 2004; Molla et al., 2021) domain-specific setting. This dataset is publicly available.[2] Here, we provide a concise overview of the dataset for context. We encourage readers to refer to Lee and Parde (2024) for further details.

3.1 Data Collection and Annotation

AcnEmpathize consists of 12,212 forum posts annotated for the presence of empathy, gathered from the "Emotional and Psychological Effects of Acne" forum on acne.org.[3] Our data collection efforts and subsequent study of the data were reviewed and granted an exemption from further review by the Institutional Review Board (IRB) at the University of Illinois at Chicago (UIC). We collected forum conversations, each including an initial post and reply posts, and filtered them based on post count (1 to 23) using the Interquartile Range (IQR) (Dekking, 2005) to exclude outliers. This process yielded 1,740 conversations with a total of 12,249 posts. After preprocessing our text by removing newline characters and posts with fewer than one alphabetical token,[4] we ended up with a final count of 1,730 conversations and 12,212 posts.

These posts were then annotated as containing empathy (1) or not containing empathy (0) by three graduate and undergraduate student volunteers (all authors of this paper, and all studying computer science with formal training in natural language processing) at UIC. The annotators, all females aged between 21 and 25, followed annotation guidelines for general-domain empathy annotation provided by Sharma et al. (2020). In the initial round of annotation, 100 posts were randomly sampled and labeled by each annotator. Upon completion, they engaged in discussions and made adjustments to achieve a perfect inter-annotator agreement (IAA) measured using the Krippendorff's Alpha coefficient (Krippendorff, 1970). An additional 900 randomly sampled posts were annotated in the same manner, with an initial IAA of 0.763, followed by discussions and adjustments until perfect agreement was reached. The remaining posts were then divided equally to be single-annotated among the three annotators. The final dataset includes 2,976 posts labeled as *containing empathy* (*Empathy*) and 9,236 posts labeled as *not containing empathy* (*No Empathy*).

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 1: Top 10 words for identified LDA topics derived from entire posts (both *Empathy* and *No Empathy*).

3.2 Dataset Analysis

AcnEmpathize has an average of 7.059 posts per conversation, with a median of 6.000 posts and a standard deviation of 5.123 posts. Posts average 153.884 tokens, with a median of 92.000 tokens and a standard deviation of 413.778 tokens. The label ratio (2,976 *Empathy* and 9,236 *No Empathy* posts) is similar to that reported by Sharma et al. (2020), who indicated 2,965 *Empathy* and 7,178 *No Empathy* posts for their empathy communication mechanism *Interpretations*.

Table 1 demonstrates the top 10 most preva-

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

[3] The website acne.org is an online platform offering acne-related support and information.

[4] We did not set a minimum word count for utterances to preserve potentially empathetic expressions in shorter responses such as "That sucks" or "I can relate."

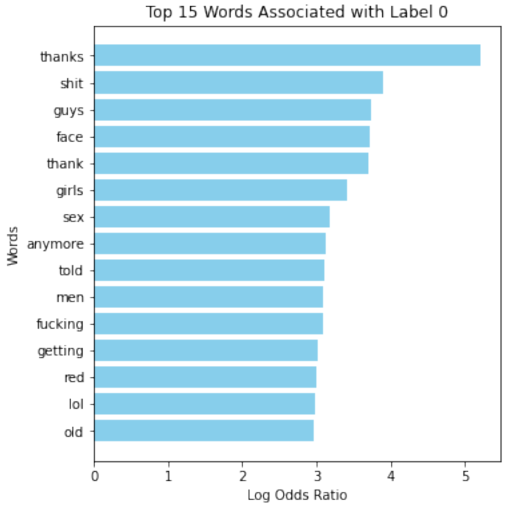

Figure 1: Words most closely associated with *No Empathy*. **Content Warning:** Contains profanity.

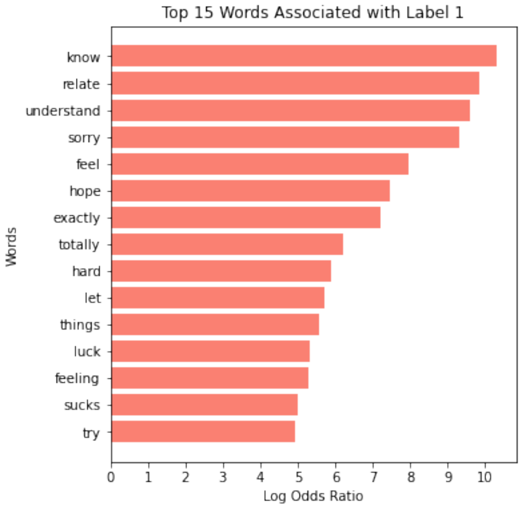

Figure 2: Words most closely associated with *Empathy*.

lent words from different topics represented in the dataset. Topics were generated by applying Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to the full dataset. The discussions primarily revolve around acne and skin conditions, including thoughts and feelings (e.g., "feel" and "think"), relevant lifestyle factors (e.g., "diet" and "food"), societal perceptions of attractiveness (e.g., "attraction" and "beauty"), and mental health concerns (e.g., "depression" and "anxiety").

To further examine the distinct characteristics of *Empathy* and *No Empathy* posts, we computed the log odds ratio separately for each class using an informative Dirichlet prior (Monroe et al., 2008; Hessel, 2016). The top 15 words associated with *No Empathy* and *Empathy* classes are shown in Figures 1 and 2, respectively. The *No Empathy* class contains slang and provocative words (e.g., "sh*t" and "f**king") while the *Empathy* class contains affirmative and supportive words (e.g., "relate," "understand," and "luck").

4 Figurative Language in AcnEmpathize

The use of figurative language in empathy is unexplored, although Citron et al. (2016) found connections between affective properties and German idioms. We focus on three forms of figurative language—idioms, metaphors, and hyperboles—and systematically analyze how these figures of speech are used within AcnEmpathize.

4.1 Figurative Language Detection

We employed Lai et al. (2023)'s publicly available multi-figurative language detection approach to facilitate our analysis. This approach identifies idioms, metaphors, and hyperboles by introducing a multitask framework that incorporates template-based prompt learning using mT5 (Xue et al., 2020) across five existing figurative language datasets. Their prompt, `Which figure of speech does this text contain? (A) Literal (B) [Task] | Text: [Text]`, assigns a specific task corresponding to each figurative language type. The method showcases accuracy for hyperbole at 0.823, idiom at 0.815, and metaphor at 0.813 when trained on English data. We reproduced the work and applied it to posts in AcnEmpathize. For each post, we iteratively applied prompting for each type of figurative language to every sentence, while keeping track of whether any of the sentences within the post were marked as containing at least one type. The figurative language types included in our work are defined as follows.

Idiom. Idiom derives meaning not from literal interpretations of constituent words but from cultural and contextual understanding. It is often used to describe concrete ideas or situations (Nunberg et al., 1994). For instance, in our dataset, the expression *"Just **keep your chin up** buddy"* encourages a poster to stay optimistic about their skin condition, rather than to literally raise their chin.

Metaphor. Metaphor helps explain ideas by assigning new meanings to conventional terms, such that one concept is framed in the more accessible terms of another (Lakoff and Johnson, 1980). Unlike idiom, metaphor can be interpreted based on

Language Category	**Total**
Figurative	7,407 (61%)
Literal	4,805 (39%)

Table 2: Total posts in each language category.

an understanding of the underlying concepts, but it should not be interpreted based on the literal definitions of those concepts. The focus of metaphoric expressions lies in conveying abstract ideas more effectively. For example, the sentence *"I totally agree* ***acne is a curse****!"* from our dataset directly compares acne to a curse, portraying it as a misfortune that may bring negative consequences.

Hyperbole. Often referred to as "exaggeration" (Claridge, 2010), hyperbole is a rhetorical device that magnifies aspects of a situation to evoke strong emotions. In our dataset, the sentence *"yea i can relate, my mood changes* ***every second****"* utilizes hyperbole to vividly depict frequent mood swings, highlighting the intensity of unstable emotions.

4.2 Analysis on AcnEmpathize

After applying Lai et al. (2023)'s model to the posts in AcnEmpathize,[5] we observe that the dataset demonstrates a prolific use of figurative language, with 7,407 posts containing one or more figures of speech compared to 4,805 posts with only literal language (see Table 2). A further breakdown of figurative language prevalence for posts with and without empathy is provided in Table 3. The "Total Figurative" row counts posts with at least one form of figurative language, not the sum of idioms, metaphors, and hyperboles, since a single post may contain multiple types of figurative language.

In both *Empathy* and *No Empathy* groups, idiom was the most commonly used figurative language, followed by metaphor and hyperbole. Our chi-square tests with 1 degree of freedom, as shown in Table 4, reveal that idioms and metaphors are significant indicators of empathy in posts. Building upon this finding, we analyze the use of these two language types within *Empathy* posts.

Tables 5 and 6 display the top 10 words for topics identified by BERTopic (Grootendorst, 2022) for posts containing idioms and metaphors. We set the number of topics to 15 and kept the default n-gram range of (1,1). Commonly appearing words like "feel" and "know" in both groups suggest that people employ figurative language to extend empathy regarding acne-related struggles by sharing their own experiences.

	Empathy	**No Empathy**	**Total**
# Posts	2,976	9,236	12,212
# Idiom	1,476 (50%)	3,822 (41%)	5,298 (43%)
# Metaphor	1,048 (35%)	2,120 (23%)	3,168 (26%)
# Hyperbole	665 (22%)	1,990 (22%)	2,655 (22%)
Total Figurative	2,066 (69%)	5,341 (58%)	7,407 (61%)

Table 3: Figurative language distribution across posts.

Language	χ^2	**p**	**Significance**
Idiom	61.51	4.40e-15	✓
Metaphor	175.49	4.67e-40	✓
Hyperbole	0.80	3.71e-01	✗

Table 4: Statistical analysis of figurative language in posts containing empathy.

We also calculated the average emotional intensity scores within these groups using the NRC Emotion/Affect Intensity Lexicon (Mohammad, 2017) (see Table 7). To do so, we created frequency dictionaries for words listed in Tables 5 and 6 and correspondingly computed the weighted scores for joy, anger, sadness, and fear. Words not present in the lexicon were excluded from the analysis. In general, we found that posts containing empathy and metaphors exhibit a slightly more intense tone compared to those containing empathy and idioms. The metaphor group has particularly higher scores for fear, with a score of 0.589 (on a scale from 0–1) compared to the idiom group's score of 0.273. Within the metaphor group, words with the highest emotional intensity include "hell," "nightmare," and "fear." The metaphor group also exhibited higher anger scores, with prominent words like "vicious," "fighting," and "hell."

In general, we found that "fight" is a prevalent word in the metaphor group, occurring often in expressions like *"Keep* ***fighting*** *acne."* This validates findings from conceptual metaphor theory (Lakoff and Johnson, 1980) and prior corpus linguistics studies (Stefanowitsch and Gries, 2007) indicating the prevalence of "fight" metaphors in large-scale

[5] We manually assessed accuracy by sampling approximately 10% of the posts and evaluating them based on the definitions and usages of idioms, metaphors, and hyperboles from the sources outlined in §4.1. Our evaluation showed an approximate figurative language detection accuracy of 85%.

Topic	Words
1	like, people, life, im, know, get, feel, things, time, go
2	acne, skin, people, like, get, clear, face, im, know, even
3	feel, im, know, like, sorry, alone, get, better, life, hope
4	it, know, control, think, try, hard, give, you, worth, done
5	mirror, makeup, look, see, mirrors, ugly, people, feel, like, wear
6	rude, people, matter, hate, sucks, judge, deserve, it, much, like
7	luck, hope, best, helps, helped, you, updated, works, mate, good
8	head, hang, keep, there, up, high, man, say, try, here
9	boat, im, were, similar, exactly, unplugged, man, you, system, bro
10	tunnel, journey, light, road, end, better, forever, get, coming, battle

Table 5: Topics associated with posts containing empathy and idioms, with topics selected by BERTopic.

Topic	Words
1	acne, like, im, life, feel, know, people, face, get, think
2	acne, skin, clear, life, let, im, people, much, like, know
3	life, depression, anxiety, stress, really, things, im, get, feel, go
4	confidence, people, shallow, self, personality, flaws, selfesteem, insecurities, appearance, esteem
5	hope, lose, faith, lost, give, like, know, going, good, always
6	cure, disease, heal, drug, time, it, said, illness, certainly, healing
7	thoughts, brain, mind, head, autopilot, mindset, useful, them, thinking, thought
8	fear, fears, nightmare, enough, panic, could, biggest, wish, shatter, noticeable
9	boat, im, feel, exactly, were, similar, unplugged, bro, pretty, three
10	battle, fight, strength, fighting, conquer, way, keep, this, strong, continue
11	diet, gut, sugar, eating, eat, cut, check, food, leaky, intake
12	heart, sucks, goes, breaks, know, you, soul, hard, go, really
13	cycle, picking, vicious, pick, mentally, energy, cardio, drained, harmful, load
14	mirror, reflection, mirrors, looking, waking, triggers, checking, morning, avoid, know
15	college, earth, destined, particularly, wondering, forever, bad, breakout, hell, university

Table 6: Topics associated with posts containing empathy and metaphors, with topics selected by BERTopic.

datasets such as the VU Amsterdam Metaphor Corpus (Steen et al., 2010). These metaphors are often used to underscore the concept of determination in overcoming obstacles—in the context of acne, the use of "fight" implies a shared meaning of perseverance and resilience. Words used in the idiom group, however, show higher intensity scores related to sadness and joy. "Hope," "good," and "luck," have the highest intensity scores associated with joy, indicating support for individuals dealing with acne. Expressions like *"There is **light at the end of the tunnel**"* and *"I am **in the same boat**"* within the idiom group further emphasize its optimistic nature.

Finally, to investigate empathetic posts containing idioms and metaphors at a finer-grained level, we generated trigram collocations using the likelihood ratio from localized sentences identified within the posts as containing idioms and metaphors. We summarize these collocations in Table 8. The *idiom-only*[6] group carries a similar sentiment as the idiom-containing group in Table 5, which can be attributed to the predominance of idiomatic expressions in our dataset. ***"Easier said than done"*** is a dominant idiomatic expression used in AcnEmpathize to convey warmth and understanding. Many trigrams, such as (itll, get, better), hint at the optimistic and encouraging tone of *idiom-only* replies.

On the other hand, trigrams from the *metaphor-only* group primarily address topics related to self-esteem, with less emphasis on mental health concerns or the war-like nature of acne seen in Table 6. Discussions in these posts often revolve around how acne lowered one's self-esteem, exemplified by trigrams like (tore, self, esteem), and are characterized by messages of encouragement and hope. These discussions sometimes include advice on rebuilding self-esteem, as reflected in *metaphor-only* replies such as *"anything to **boost the self-esteem** and promote relaxation will help too."*

[6] *Idiom* and *idiom-only* are two distinct terms, where *idiom* refers to posts containing idioms inclusively, potentially along with hyperboles and/or metaphors, while *idiom-only* refers to posts that exclusively contain idioms without any other types of figurative language. Later, we define *metaphor-only* similarly to *idiom-only*.

	Joy	Anger	Sad	Fear	Avg.
Idiom	0.467	0.567	0.544	0.273	0.521
Metaphor	0.436	0.592	0.539	0.589	**0.530**

Table 7: Weighted emotional intensity scores for words identified by BERTopic for Idiom and Metaphor groups.

Figurative Language	**Trigram Collocations**
Idiom	(easier, said, done), (thing, get, better), (make, feel, like), (know, feel, like), (get, better, soon), (hope, get, better), (youll, get, better), (feel, get, better), (let, get, better), (itll, get, better)
Metaphor	(low, self, esteem), (self, esteem, boost), (lowered, self, esteem), (rebuilding, self, esteem), (ruining, self, esteem), (self, esteem, completely), (self, esteem, grew), (self, esteem, practically), (self, esteem, surprisingly), (tore, self, esteem)

Table 8: Top 10 trigrams for idioms and metaphors in posts containing empathy.

5 Empathy Detection in AcnEmpathize

Following our analysis of figurative language use in AcnEmpathize, we investigated its role in empathy detection. We aimed to enhance performance by incorporating figurative language features more directly into our models. To ensure broad understanding of this phenomenon, we experimented with both feature-based and PLM paradigms. For all experimental conditions, we utilized an 80:20 random split for training and testing data.

5.1 Models

Baseline Models

We employed two baseline models, **Most Frequent** and **Random**, to establish performance floors. Most frequent predicts the most frequent class from the training data (*No Empathy*), while Random randomly predicts labels.

Feature-Based Models

Our feature-based models included **SVM**, **Naive Bayes**, and **Logistic Regression** models trained using LIWC (Tausczik and Pennebaker, 2010) psycholinguistic features. We selected these models due to their widely documented success on a range of feature-based text classification tasks. We used the LIWC 2022 edition[7] to extract 119 varied psycholinguistic features from each post. Then, we employed *SelectKBest*, a univariate feature selection approach from the scikit-learn library (Pedregosa et al., 2011), with the default parameter *f_classif*. This method calculates the F-score for each feature with respect to the target variable (the empathy label in this case) and selects the top $k = 5$ scoring features. The LIWC features selected through this process include *Analytic*, *Linguistic*, *Function*, *Insight*, and *Feeling*. *Analytic* refers to analytical thinking, a metric of logical and formal thinking. *Linguistic* encompasses various linguistic dimensions such as pronouns and articles. *Function* refers to total functional words such as "the," "to," and "I." *Insight* captures words such as "know," "how," "think," and "feel." *Feeling* encompasses words such as "feel," "hard," "cool," and "felt."

Pre-Trained Language Models

We also experimented with three diverse PLM-based methods, as well as an approach ensembling all three. We included two fine-tuned PLM approaches: **RoBERTa-large-mnli** (Liu et al., 2019), a RoBERTa-large model specifically fine-tuned on the MNLI (Multi-Genre Natural Language Inference) corpus, and the most up-to-date version of **RoBERTa-twitter-sentiment** (Loureiro et al., 2022), a RoBERTa-base model fine-tuned for sentiment analysis on Twitter data. This allowed us to examine the performance of fine-tuned PLMs for domain-specific empathy detection in cases when the pretraining data was primarily general-purpose data (RoBERTa-large-mnli) and in cases when the pretraining data was closer to the target domain (RoBERTa-twitter-sentiment). We used versions of these models from the HuggingFace libraries[8] with the following hyperparameters: `max_length=256` (the average token count in preprocessed posts, 154, rounded up to the next power of 2), `learning rate` = 1e-5 for the optimizer AdamW, `num_epochs=3`, and `batch_size` for training set to 8 for RoBERTa-large-mnli and 16 for RoBERTa-twitter-sentiment.

Our third PLM method was a prompt-based learning condition. We used **T5** (Raffel et al., 2020), an advanced encoder-decoder model pretrained on a mix of unsupervised and supervised tasks. We performed zero-shot prompting with frozen language model weights for this setting, relying entirely on language parameters learned during the pretraining process. We manually speci-

[7] `https://www.liwc.app/`

[8] RoBERTa-twitter-sentiment: `https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest`, and RoBERTa-large-mnli: `https://huggingface.co/FacebookAI/roberta-large-mnli`

		No Empathy			*Empathy*			
Model	**Accuracy**	**Precision**	**Recall**	**F$_1$**	**Precision**	**Recall**	**F$_1$**	**Macro F$_1$**
Most Frequent	0.756	1.000	0.861	0.000	0.000	0.000	0.000	0.000
Random	0.494	0.748	0.499	0.598	0.235	0.479	0.316	0.457
SVM$_{\text{LIWC}}$	0.771	0.771	1.000	0.871	0.000	0.000	0.000	0.436
Naive Bayes$_{\text{LIWC}}$	0.770	0.772	0.996	0.870	0.417	0.009	0.018	0.444
Logistic Regression$_{\text{LIWC}}$	0.767	0.772	0.989	0.868	0.333	0.018	0.034	0.451
SVM$_{\text{LIWC+FIG}}$	0.813	0.838	0.939	0.886	0.656	0.388	0.488	0.687
Naive Bayes$_{\text{LIWC+FIG}}$	0.801	0.845	0.908	0.875	0.586	0.438	0.502	0.689
Logistic Regression$_{\text{LIWC+FIG}}$	0.808	0.841	0.926	0.882	0.623	0.411	0.496	0.689
RoBERTa-large-mnli	0.900	0.957	0.898	0.926	0.715	0.864	0.782	0.854
RoBERTa-twitter-sentiment	0.896	0.953	0.911	0.931	0.738	0.848	0.789	0.860
T5	0.896	0.942	0.922	0.932	0.857	0.809	0.781	0.864
Ensemble	0.896	0.952	0.911	0.931	0.739	0.844	0.788	0.860
RoBERTa-large-mnli$_{\text{FIG}}$	0.902	0.946	0.926	0.936	0.766	0.821	0.793	0.865
RoBERTa-twitter-sentiment$_{\text{FIG}}$	**0.910**	0.950	0.933	0.941	0.786	0.834	**0.809**	**0.875**
T5$_{\text{FIG}}$	0.909	0.935	0.948	**0.942**	0.816	0.778	0.797	0.870
Ensemble$_{\text{FIG}}$	0.900	0.950	0.919	0.934	0.754	0.837	0.793	0.864

Table 9: Results from experiments for different models.

fied the following discrete prompt template: `Does the following text contain empathy? [X]` with each post [X] in our test set being used to fill the prompt. T5 was imported from the HuggingFace library[9] with the same parameters defined for RoBERTa-large-mnli above. Finally, our ensemble approach (Dietterich, 2000) incorporated all of our PLM conditions, incorporating dynamic weighting based on each model's confidence scores calculated as the maximum probability obtained from the softmax output of each model's predictions. All PLMs were run on a T4 GPU in under 2 hours.

Figurative Language Features

We constructed one-hot encoded labels for detected idioms, metaphors, and hyperboles. During the fine-tuning or prompting phase for PLM conditions, we appended these to the text of each post through an additional embedding layer. For the feature-based conditions (LIWC), we appended these one-hot encodings directly to the feature vectors. This allowed us to evaluate the impact of these linguistic elements in numerous settings by comparing model performance with and without them.

5.2 Results

The results were obtained using the scikit-learn library (Pedregosa et al., 2011) and averaged over three runs (see Table 9). We observe a marked improvement across all models with the addition of figurative language features (FIG). The highest overall accuracy and F$_1$ were achieved by the PLM approaches for both *Empathy* and *No Empathy* labels. RoBERTa-twitter-sentiment$_{\text{FIG}}$ achieved the best overall accuracy at 0.910 and an F$_1$=0.809 for the *Empathy* label, an increase from 0.896 accuracy and F$_1$=0.789 without FIG features. Similarly, T5$_{\text{FIG}}$ attained the highest F$_1$ score for *No Empathy*, up from F$_1$=0.932 (T5 without FIG). These enhancements were consistent across all PLMs, including the ensemble approach. The impact was even more pronounced in the feature-based models. For instance, SVM trained solely with LIWC features initially showed zero precision and recall for predicting the *Empathy* label, which jumped to an F$_1$=0.488 with the incorporation of FIG features. Comparable improvements were observed across the board, as evidenced by increased overall accuracy and F$_1$ scores.

5.3 Discussion

To further understand the contribution of idiom, metaphor, and hyperbole to empathy detection performance, we performed follow-up analyses focusing separately on each figurative language feature. We took RoBERTa-twitter-sentiment, the best-performing PLM model (Macro F$_1$=0.875 and accuracy=0.910), and SVM, a feature-based model that demonstrated relatively strong performance (Macro F$_1$=0.687 and accuracy=0.813) from Table 9. We then trained models incorporating each figurative language type individually, and report these results in Table 10. For SVM, idiom features returned

[9] https://huggingface.co/google-t5/t5-base

		No Empathy			*Empathy*			
Model	**Accuracy**	**Precision**	**Recall**	**F$_1$**	**Precision**	**Recall**	**F$_1$**	**Macro F$_1$**
SVM$_{\text{LIWC+IDIOM}}$	0.815	0.839	0.941	0.887	0.663	0.394	0.494	0.691
SVM$_{\text{LIWC+METAPHOR}}$	0.814	0.839	0.938	0.886	0.654	0.394	0.492	0.689
SVM$_{\text{LIWC+HYPERBOLE}}$	0.814	0.839	0.938	0.886	0.655	0.394	0.492	0.689
RoBERTa-twitter-sen$_{\text{IDIOM}}$	**0.895**	0.959	0.903	**0.930**	0.727	0.869	**0.792**	**0.861**
RoBERTa-twitter-sen$_{\text{METAPHOR}}$	0.892	0.956	0.901	0.928	0.721	0.859	0.784	0.856
RoBERTa-twitter-sen$_{\text{HYPERBOLE}}$	0.887	0.955	0.897	0.925	0.711	0.857	0.777	0.851

Table 10: SVM$_{\text{LIWC}}$ and RoBERTa-twitter-sentiment performance with different figurative language features.

minimally higher performance than metaphor and hyperbole features. Comparing these results to those in Table 9 (SVM$_{\text{LIWC}}$ and SVM$_{\text{LIWC+FIG}}$), we observe that models with individual figurative language features perform slightly better than those with combined features. For RoBERTa-twitter-sentiment, idiom features also yield the highest performance, with slightly more pronounced improvements over metaphor and hyperbole features compared to SVM. However, this is a noticeable drop from the performance of RoBERTa-twitter-sentiment$_{\text{FIG}}$ in Table 9 when all features are combined. This analysis confirms that the relationship between figurative language and empathy expression is complex and some of its types may be interdependent. In the future, it may be beneficial to explore additional figurative language types such as sarcasm, simile, and paradox for further insights.

6 Conclusion

In this paper, we investigate figurative language use in the new domain-specific empathy detection corpus, AcnEmpathize. We find that incorporating figurative language features into domain-specific empathy detection models improves their performance, and we achieve an impressive maximum F$_1$=0.942 and F$_1$=0.809 when identifying posts with and without empathy, respectively. We release all models[10] and data[2] publicly to encourage follow-up research by others.

In our systematic analysis of figurative language use in this dataset, we find confirmatory and intriguing associations between empathy, idiom, and metaphor. Insights resulting from this study hold promise for improving peer-to-peer support and paving the way for the development of empathetic chatbots that cater to the concerns of different online communities. Promising future directions include investigating the interplay between various forms of figurative language and the implications of their combined use, and broadening our scope to include additional forms of figurative language.

Limitations

Our study has several limitations. The annotation process for AcnEmpathize may have involved subjectivity, despite our extensive discussions to reduce potential biases. Imbalances exist in the labels concerning the presence of empathy and figurative languages, potentially impacting model performance and analyses. Future researchers are advised to carefully examine each category or apply appropriate weighting mechanisms when utilizing our data for their studies. We also did not investigate the interdependence between different types of figurative language; we underscore that this presents an intriguing direction for future research.

Ethical Considerations

Our study was granted an exemption by the Institutional Review Board at UIC, determined as not involving direct human subjects research. The primary data source for our research, acne.org, consists of publicly available, anonymous posts, which do not include personal information about users. All annotators for the AcnEmpathize dataset participated voluntarily and are recognized as co-authors of this paper. We recommend that our dataset be used for research purposes.

Acknowledgements

N. Parde was partially supported by the National Science Foundation under Grant No. 2125411 and the National Institutes of Health under Grant No. 1R41NR020667-01. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the NSF or the NIH.

[10] https://github.com/gyeongeunlee16/detect_AcnEmpathize

References

Godfrey T Barrett-Lennard. 1981. The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology*, 28(2):91.

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, Toronto, Canada. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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

Brian F Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological review*, 112(1):193.

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*.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. Iucl at wassa 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*.

Francesca MM Citron, Cristina Cacciari, Michael Kucharski, Luna Beck, Markus Conrad, and Arthur M Jacobs. 2016. When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for german (panig). *Behavior research methods*, 48:91–111.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of cognitive neuroscience*, 26(11):2585–2595.

Claudia Claridge. 2010. *Hyperbole in English: A corpus-based study of exaggeration*. Cambridge University Press.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. 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 2218–2229.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *JSAS: catalog of selected documents in psychology*, 10.

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

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

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.

Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. 2004. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *Bmj*, 328(7449):1166.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2(4):239–250.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Raymond W Gibbs Jr, John S Leggitt, and Elizabeth A Turner. 2002. What's special about figurative language in emotional communication? In *The verbal communication of emotions*, pages 133–158. Psychology Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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

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

Mahshid Hosseini and Cornelia Caragea. 2021b. 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.

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

Atharva Kulkarni, Sunanda Somwase, Shivam Rajput, and Manisha Marathe. 2021. Pvg at wassa 2021: A multi-input, multi-task, transformer-based architecture for empathy and distress prediction. *arXiv preprint arXiv:2103.03296*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. *arXiv preprint arXiv:2306.00121*.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL*.

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.

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. Ncuee-nlp at wassa 2023 shared task 1: Empathy and emotion prediction using sentiment-enhanced roberta transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 548–552.

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*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Saif M Mohammad. 2017. Word affect intensities. *arXiv preprint arXiv:1704.08798*.

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*, 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.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *arXiv preprint arXiv:2205.12698*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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*.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, et al. 2010. A method for linguistic metaphor identification. *Amsterdam: Benjamins*.

Anatol Stefanowitsch and Stefan Th. Gries, editors. 2007. *Corpus-Based Approaches to Metaphor and Metonymy*. De Gruyter Mouton, Berlin, New York.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, João Sedoc, Sven Buechel, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 92–104. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.