# Empathy Classification in Text-Based Online Mental Health Support

**Ambika Gupta, Nicole Liu**
University of California, Berkeley
{ambikag, nicoleliu}@berkeley.edu

## Abstract

Mental health support is a critical area of focus both in healthcare and in our everyday experience. Despite a wide array of resources available, there lacks a scalable approach to evaluate the usefulness and appropriateness of text-based mental health support.

In this paper, we adopt a framework that measures the extent of empathy in text-based mental health support in three dimensions. We design a set of models that predict whether a given response is empathetic or not in all three dimensions and recommend distinct approaches of feature engineering depending on the aspect of interest.

## 1 Introduction

While mental health issues are on the rise[1], professional support is struggling to fulfill the demand. In this context, many individuals still default to online forum[2] such as Reddit to share their experiences, ask for advice, and seek help. The range of responses they may receive for a particular post is broad, arbitrary, and unregulated. Some responses show empathy and provide practical advice, while others are unhelpful or even exacerbate the problem by assigning blame.

As these forums have been operating for several years, there is now a vast amount of text-based pairs of posts seeking help and posts responding to them. In this paper, we adopt the framework developed by Sharma et al (2020) to evaluate responses for empathy. EPITOME measures three communication mechanisms of empathy: emotional reaction, exploration, and interpretation. This framework combines the emotional and cognitive components related to the exchange of experiences, feelings, and understanding.

In our work, we use three sub-datasets, each corresponding to one of the three aspects of empathy. Our tasks were binary classifications of whether the responses showed emotional reaction, exploration, or interpretation, respectively for each dataset.

[1] Reinert, M, Fritze, D. & Nguyen, T. (October 2021). "The State of Mental Health in America 2022" Mental Health America, Alexandria VA.

[2] Smith-Merry, J., Goggin, G., Campbell, A., McKenzie, K., Ridout, B., & Baylosis, C. (2019). Social Connection and Online Engagement: Insights From Interviews With Users of a Mental Health Online Forum. *JMIR mental health*, *6*(3), e11084. https://doi.org/10.2196/11084

Our major contributions are twofold.

(1) While the original paper by Sharma et al applies a uniform approach of feature engineering in all three datasets, we instead propose distinct feature engineering techniques depending on the aspect of interest.
(2) Recognizing the challenge of data collection and annotation in the real world, our work utilized a substantially smaller dataset, less than 30% of the original data used by Sharma et al.

With this work, we hope to evaluate the presence of empathy in text-based mental health support in a more efficient way and guide the effort of generating more relevant and impactful mental health resources on a large scale.

## 2 Background

NLP classification and mention detection in mental health represent an area with extensive prior research. Sharma et al (2020) formalized the three-dimensional framework of empathy, which we adopted in this paper. With three labels in their data indicating the empathy level (0, 1, and 2), they achieved between 72% – 92% accuracy scores across the three dimensions. Similar to our goal of binary classification, Upadhyay et al (2023) explored using pre-trained BERT variants to perform binary classification of toxic versus non-toxic content from Twitter and inspirational quotes online. They achieved 85% weighted F1 in their study.

Beyond classification, Rodriguez et al (2021) devised a model to mimic cognitive behavior therapy (CBT) techniques and automatically transform a negative thought into a more positive one. Ji et al (2022) pre-trained two masked language models, MentalBERT and MentalRoBERTa, using mental health data. However, there are limitations in such pre-trained BERT variant models as exposed by Lin et al (2022) and other researchers. Lin et al found that, in masked-token tasks, the models were more likely to fill female-related tokens in sentences related to mental health stigma.

Our research is differentiated from prior research in that it proposes novel feature engineering and de-noising techniques for the classification task.

## 3 Methods

### 3.1 Framework for Empathy

The data includes conversations on online forums such as TalkLife, the largest global peer-to-peer mental health support network, and threads posted on mental health communities within Reddit. It is broken into three datasets that each represent one aspect of empathy from the EPITOME framework. They are made up of (seeker post, response post) pairs with annotated empathy levels as well as rationales for positive labels. The annotations were done by crowdworkers trained by Sharma et al (2020). Here we have a definition of the framework:

- **Emotional Reactions**: Expressing emotions experienced by seeker posts.
- **Interpretations**: Communicating an understanding of feelings and experiences from seeker posts.

- **Explorations**: Exploring feelings and experiences not stated in seeker posts to improve understanding.

### 3.2 Data

The original dataset comprises 3,084 (seeker post, response post) pairs as well as rationales if the labels are positive. The labels indicating empathy have three levels: 0 - no communication, 1 - weak communication, and 2 - strong communication. There is a high class imbalance for all three datasets skewed towards class label 0, though skewness as well as distribution between classes 1 and 2 varied, as shown in Appendix A.

We found that this class imbalance was extremely difficult to address even in BERT models with class weights. We decided to change our scope from multiclass classification to binary classification by combining labels 1 and 2. This is to determine solely if empathy is present in the response and not to which extent it is present. Even with this adjustment, there was noticeable class imbalance between the no empathy class (0) and the new combined class representing empathy (1).

To further account for this imbalance, we augmented the Explorations dataset due to its very low count of empathetic examples using back-translation from English to Spanish to English via Google Spreadsheets. These translations were manually checked to verify similar meanings between the original response and back-translated response and remove duplicates, as shown in Appendix B. We also downsampled the negative class to create an even distribution between the two classes for all datasets.

| **Seeker_post** | **Response_post** | **Labels** |
|---|---|---|
| The sadness will last forever. I feel so lonely and empty and nothing is working | The sadness will not last forever. Tell us more about your predicament and I'll offer some advice and help! | Em: 1<br>In: 0<br>Ex: 1 |
| My diet becomes fucked when i get depressed.. I can't control myself in the grocery store when I'm feeling down. Chips, cookies, soda, cake, you name It. Anyone else a stress eater? | By any chance do you think you're in a loop? Junk food can make you depressed. And being depressed probably makes you eat more junk food? For a while that was my problem too. Even if its just random, ill probably binge and feel bad later. Only to indulge in more later to try to cheer up. | Em: 0<br>In: 1<br>Ex: 1 |

**Table 1: Examples of (seeker post, response post) pairs in our binary classification datasets**

### 3.3 Methodology

Our work consists of three main components: (1) baseline model using Naive Bayes, (2) BERT model with various experiments, (3) BERT variants using the best features from step (2) for each dataset.

Intuitively, both the seeker post and the response post in a sample contribute to the label, especially for the Emotional Reactions and Interpretations datasets. For our baseline, we performed multinomial Naive Bayes using bag of words on both features to evaluate our models.

In step (2), we joined seeker and response posts with a [SEP] token between them, tokenized the text, and used pre-trained BERT-based embeddings. We then fed the CLS token from the transformer layers into a fully connected layer, applied dropout layers and regularization, and finally, an activation function for prediction.

### 3.4 Experiments

We performed a series of experiments to improve the model results. For a comprehensive summary of all experiments as well as their corresponding hypotheses and findings, please refer to Appendix C.

We first tested whether manually encoding token_type_id to "1" for 'response_post' in the combined seeker and response input would give the model more information about the differentiations between the two components.

We then experimented with a variety of seeker and response lengths in the input. For instance, we compared trimming seeker to 60% percentile word count versus no seeker trimming, using responses alone versus seeker + response, using mean of response lengths as max token length versus using mean + 1 standard deviation as max token length, etc.

To further de-noise, we tested using the 'rationales' as inputs when available, instead of the full responses. We tested using the first rationale, longest rationale, and all available rationales. We hypothesized for samples with longer inputs, the model was not able to differentiate between noise and meaningful data. We then split the test sets into two subsets based on input length to validate our hypothesis.

The experiments enabled us to collect the best features that are distinct for each dataset, namely 'seeker_post' + [SEP] + 'response_post' for the Emotional Reactions dataset and truncated responses for the Interpretations and Explorations datasets. We then ran BERT variants, including RoBERTa and ALBERT, as well as GPT-2 using these features.

### 3.5 Metrics

We chose to report on accuracy and F1 scores for all models. Accuray gave us an overall view of the model performance, while F1 combining precision and recall gave a more nuanced view.

## 4 Results and Discussion

| Dataset | Metrics | Naive-Bayes | BERT | | | | | Best Features | |
|---|---|---|---|---|---|---|---|---|---|
| | | Baseline | Seeker and Response | Response: Mean Max Length | Response: Mean+1SD Max Length | Rationale Replacement | Rationale Replacement: Short Test | RoBERTa | ALBERT |
| **Emotional Reactions** | Accuracy | 67.06 | **78.23** | 74.22 | 76.85 | 71.12 | 74.22 | **80.91** | 76.13 |
| | F1-Score | 72.29 | **77.75** | 75.23 | 72.68 | 68.07 | 71.79 | **80.77** | 76.08 |
| **Interpretations** | Accuracy | 66.45 | 76.41 | 80.06 | **81.69** | 81.04 | 78.55 | **84.28** | 82.33 |
| | F1-Score | 67.71 | 73.23 | 76.84 | **80.55** | 78.13 | 66.67 | **82.89** | 78.92 |
| **Explorations** | Accuracy | 67.19 | 80.56 | 81.78 | **83.33** | 80.73 | 90.91 | 88.54 | **92.71** |
| | F1-Score | 67.02 | 78.22 | 81.68 | **82.22** | 75.17 | 92.31 | 88.04 | **92.22** |

**Table 2: Results comparison for all models across all three datasets**

### 4.1 Baseline

For our baseline, our multinomial Naive-Bayes on ‘seeker_post’ and ‘response_post’ performed with an accuracy of 67% across all 3 datasets and an F1 score of 72% for the Emotional Reaction dataset and an f1 score of 67% for the other two.

### 4.2 BERT

Our BERT models did improve from our baseline due to the additional context that BERT provides over Naive-Bayes. While using both ‘seeker_post’ and ‘response_post’, our best results came when chunking ‘seeker_post’ to the average number of tokens . There was a 10-13% increase in accuracy across all datasets as previously, long seeker posts would limit how much of each response post was represented in encodings.

### 4.3 Seeker Removal

We expected all models to perform worse when removing seeker posts as a feature, especially for Interpretations as empathetic responses are meant to communicate understanding of the seeking post. However, this was only true for Emotional Reactions, likely due to the fact that many words related to emotions in seeker posts were reiterated in empathetic responses, emphasizing the emotion. For the other two datasets, the removal of seeker posts actually decreased noise.

Using responses alone, we noticed that with a larger maximum length of encodings, our models continuously performed better, although logarithmically as the maximum length increased. This is due to encodings that represent more of each response, and therefore covering rationales that are located further into responses. We found that performance started to taper off at encoding lengths of one standard deviation above the mean, likely due to increased noise.

### 4.4 Response Replacement with Rationale

In our experimenting with replacing responses with rationales, we actually saw performance stay stagnant or slightly decrease. This is likely due to the lesser amount of noise for empathetic responses in the training data versus the testing data. The best performing replacement method was using the longest rationale. This was surprising in comparison to using all rationales, but likely due to context changes between different rationales of a response. This is supported by the higher number of negative predictions in general across all 3 models, indicating that context was misconstrued.

We hypothesized that in samples with longer inputs, there was too much noise that overshadowed the meaningful content. To test this hypothesis, we compared how the response replacement model performed on different lengths of test inputs. Across all three datasets, it performed better on shorter test inputs than longer test inputs, confirming our hypothesis

### 4.5 BERT Variants

For each dataset, we used the best features to run RoBERTa and ALBERT as higher performing alternatives to BERT. We found that RoBERTa performed best for Emotional Reactions and Interpretations as expected due to RoBERTa’s larger training dataset. Explorations also improved using RoBERTa, but it performed best using ALBERT. This is likely because the Explorations dataset is smaller and its responses are shorter on average, and ALBERT tends to work better with shorter samples.

### 4.6 Other Findings

For our augmented Explorations dataset, we found that it had similar performance to Explorations. In a few cases, it did perform up to 3% better than the non-augmented Explorations, but we can attribute this to overfitting similar responses in train and test.

Using our best features, we also ran GPT-2 as used by Sharm et al. We found our results using RoBERTa and ALBERT to be better as BERT variants tend to perform better on sentiment analysis.

In general, the Explorations dataset had the best results with an accuracy of 93% using RoBERTa, by using 'response_post' as the sole input. This is likely due to its features being more distinct: remarks on exploration are often phrased as a question. The fact that the response posts alone yielded sufficient results regardless of context from the seeking posts begs the question of whether LLMs really understood the contexts while detecting empathy. Alternatively, they only had a narrow understanding of empathy and were merely searching for certain questions or expressions commonly associated with the right labels.

## 5 Conclusion

In this project, we used the data with annotated empathy levels from the EPITOME framework and performed binary classification tasks on the Emotional Reactions, Interpretations, and Explorations datasets. We augmented data with back-translation, experimented with various feature engineering techniques to improve BERT-based models, and tested our highest performing features on BERT variants. Our best models achieved accuracy scores of 81% - 93% and F1 scores of 81% - 92% across the three datasets. Notably, we found that different feature engineering techniques worked best for different datasets. The Explorations dataset had the best results among the three datasets, especially with ALBERT.

Future work includes multiclass classification using the original three labels in the data to further detect nuances in empathy levels. Additionally, mention detection techniques applied to rationale extraction would also complement the classification task.

## References


Ishan Sanjeev Upadhyay, KV Aditya Srivatsa, and Radhika Mamidi. 2022. Towards Toxic Positivity Detection. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 75–82, Seattle, Washington. Association for Computational Linguistics.

Ignacio de Toledo Rodriguez, Giancarlo Salton, and Robert Ross. 2021. Formulating Automated Responses to Cognitive Distortions for CBT Interactions. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 108–116, Trento, Italy. Association for Computational Linguistics.

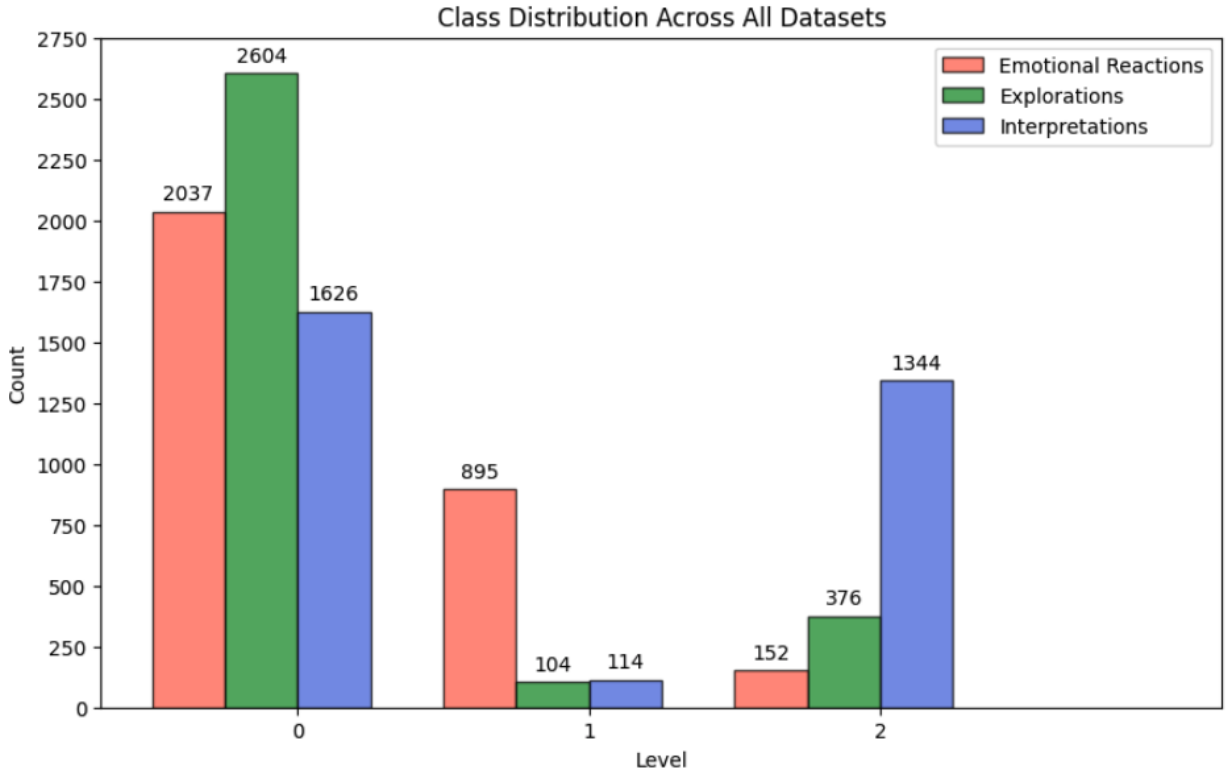
Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered Mental Health Stigma in Masked Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2170, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

**Appendix A.**

Class distribution across all datasets before reclassification. This graph shows the high class imbalance favoring label 0 and varying distributions in levels 1 and 2.



**Appendix B.**

The following table shows examples of back translations (from English to Spanish to English) that are either accepted or rejected.

| Original post | Back-translated post | Decision | Reason |
|---|---|---|---|
| Is that really so bad? Maybe it was the smart decision because you needed that time to recover. You're being kind to yourself when you need it and that's important. Hope you feel better soon. | Is that really so bad? Maybe it was the intelligent decision because you needed that time for recovery. You are being kind with yourself when you need it and that is important. I hope you feel better soon. | Accepted | |
| What were you taking before vs what are you taking now? Are you being prescribed by your general doctor or a psychiatrist? Do you feel like you are in danger right now? | What were you taking before what you are taking now? Are you being prescribed by your general practitioner or a psychiatrist? Do you feel that you are in danger at this time? | Accepted | |
| Hey, what's up? Why do you feel like this? | Hello! How are you? Why do you feel that way? | Accepted | |
| Why do you feel your time is almost up? | Why do you feel that your time is almost awake? | Rejected | Change of meaning |
| I always wondered, why are people so obsessed with prolonging their lives?? Are you just into pain? | I always wondered, why are people so obsessed with prolonging their lives? Does it hurt alone? | Rejected | Too similar |
| What makes you say these things? | What does these things tell you? | Rejected | Change of meaning |

## Appendix C.

Summarization of experiments: the hypothesis and the tests in each step. Bolded text represents better results in each comparison.