Dartmouth College
Dartmouth Digital Commons

Computer Science Senior Theses Computer Science

Spring 5-31-2023

Connecting Linguistic Expressions and Pain Relief Through Transformer Model Construction and Analysis

Sarah M. Chacko
sarah.m.chacko.23@dartmouth.edu

Follow this and additional works at: https://digitalcommons.dartmouth.edu/cs_senior_theses

Part of the Computer Sciences Commons, and the Psychological Phenomena and Processes Commons

Recommended Citation

Chacko, Sarah M., "Connecting Linguistic Expressions and Pain Relief Through Transformer Model Construction and Analysis" (2023). *Computer Science Senior Theses*. 10.
https://digitalcommons.dartmouth.edu/cs_senior_theses/10

This Thesis (Undergraduate) is brought to you for free and open access by the Computer Science at Dartmouth Digital Commons. It has been accepted for inclusion in Computer Science Senior Theses by an authorized administrator of Dartmouth Digital Commons. For more information, please contact dartmouthdigitalcommons@groups.dartmouth.edu.

CONNECTING LINGUISTIC EXPRESSIONS AND PAIN RELIEF THROUGH TRANSFORMER MODEL CONSTRUCTION AND ANALYSIS

A Thesis
Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Bachelor of Arts

in

Computer Science

by Sarah Chacko

June 2023

Abstract

Chronic pain is a widespread problem that significantly impacts quality of life. Overprescription and abuse of pain medication continues to be a major public health issue and can further burden patients due to a fragmented health care system. Previous research has suggested a possible psychological basis to pain and the potential for safer, non-pharmacological alternatives for pain relief. This project leverages language models to study chronic pain development and relief through psychological treatments, which will be assessed through responses to post-treatment interviews. A transformer-based natural language processing model is employed to identify connections between language expressions and pain on a dataset of back pain questionnaires. The features of the text are further analyzed through SHAP analysis to explain the model's predictions. From the results, we discovered no significant correlation between the predicted and observed values of the general regression and classification models. We also found a slightly stronger correlation for the regression model for the placebo treatment, and no transfer in performance from generated data-trained regression and classification models. Further study of this topic could lead to more reliable prediction of pain relief by linguistic features.

Preface

I would like to thank Dr. Soroush Vosoughi, who was kind enough to let me into his lab and work on a project which I am truly passionate about. I also express my heartfelt gratitude to Weicheng Ma, who guided me throughout this process and played a fundamental role in my research.

Lastly, I would like to thank my wonderful parents, Jacob and Mary, and my sister, Rachael, for providing endless love and support throughout this journey. None of this would have been possible without you.

Contents

Abstract ii

Preface iii

1 Introduction **1**

1.1 Background 1

1.1.1 Chronic pain and its treatment 1

1.2 Problem statement 4

1.3 Related work 5

2 Methodology and Experiments **7**

2.1 Data 7

2.1.1 Original Dataset 7

2.1.2 LIWC Analysis 9

2.1.3 Generated Dataset 10

2.2 Model Creation 12

2.2.1 What is BERT? 12

2.2.2 Goal 1: Predict pain relief level 12

2.2.3 Goal 2: Predict pain relief level by treatment group 13

2.3 Model Analysis 14

2.3.1 SHAPley Analysis of classification tasks 14

2.3.2 Sentiment analysis using SST-2 14

3 Results and Discussion 16

3.1 Dataset . . . 16

3.1.1 LIWC Analysis . . . 18

3.2 Model Creation . . . 19

3.2.1 Goal 1: Predict pain relief level . . . 19

3.2.2 Goal 2: Predict pain relief level by treatment group . . . 21

3.3 Model Analysis . . . 23

3.3.1 SHAP Analysis . . . 23

3.3.2 Sentiment analysis using SST-2 . . . 27

3.4 Limitations . . . 31

4 Future Work 34

5 Conclusion 36

Chapter 1

Introduction

Section 1.1

Background

1.1.1. Chronic pain and its treatment

The impact of chronic pain. Chronic pain is a debilitating disease that can severely affect an individual's ability to perform regular activities. Recent studies estimate that chronic pain is found in at least 1 in 5 adults in the United States [1]. It is one of the most common reasons that adults seek medical care in the United States, placing further strain on the existing fractured healthcare system. Not only does chronic pain restrict movement and participation in daily life, it has been associated with increased anxiety, depression, and dependence on narcotic pain medication [2]. Furthermore, chronic pain not only results in physical symptoms but also financial burden at both the individual and collective scales. In addition to individuals facing the high cost of treatment, the impact of persistent pain can be large scale. An older estimate of the costs of chronic pain ranges from 560 billion to 635 billion dollars, which is composed of direct costs of health care, lost workdays, and decreased wages [3]. With the direct and indirect costs of chronic pain in mind, this problem poses a formidable challenge to those seeking to reduce treatment costs and improve quality of life.

Chronic pain management. Treating persistent pain is paramount; if left untreated, the pituitary-adrenal axis will be activated and cause suppression of the immune system, leading to a host of other health issues. Patients are also more inclined to develop anxiety and depression disorders, as they experience feelings of helplessness and hopelessness without a regimented treatment plan [4]. Chronic pain management encompasses a diverse range of different treatments and effectiveness. These treatments can include evidence-based non-pharmacological treatment, such as physical and psychological therapies. These treatments are typically safer than their pharmacological counterparts, as there is no risk of relapse or dependence. This pathway is commonly recommended to patients regardless of whether or not they are taking medication. Alternatively, medication-based treatments are another option for the alleviation of pain. Analgesics can encompass different classes, including antidepressants, nonsteroidal anti-inflammatory drugs (NSAIDs), and opioids. [5]. Each class poses different benefits and risks for the patient which a treatment team must review when considering the use of a medication.

Overprescription of opioids as pain relievers. Though opioids are potent pain relievers, their inclusion in a treatment plan must be preceded by a careful analysis of the necessity for a patient's condition. There are major risks involved with choosing opioid therapy. Opioids are highly addictive as they result in increased endorphin release within the reward centers of the brain. This mechanism of action is responsible for one of the most lethal drug epidemics in the United States, with over three million people dependent on opioids. Of this total, an estimated two million people abuse prescription opioid-based pain medication [6]. Though opioids are typically not recommended for long courses of treatments due to its severe adverse effects, patients will often refuse diminution or cessation in order to avoid withdrawal symptoms and continue relief [7]. This epidemic exposes the negligence on the part of healthcare professionals and an inappropriate prescription of strong pain medication. A better understanding of addiction by the healthcare team and a more controlled approach to prescription and cessation must be employed to treat these patients and prevent future misuse.

Biases with pain detection. The prescription of pain medication can also be analyzed critically from racial, gender-based, and class-based lenses. Certain groups, such as black Americans, receive poorer treatment from white healthcare providers due to these providers' underlying perceptual contributions and false beliefs about biological differences between races. Studies have revealed the subconscious link between white healthcare professionals who endorse false beliefs about increased pain tolerance in black patients and under-prescription of pain medication [8].

Socioeconomic status (SES) can also play a subliminal role in the development and management of chronic pain. Individuals with a lower SES experience a disproportionately high incidence of chronic pain and a greater likelihood of receiving a prescription for opioid medication. In addition, individuals with lower levels of education are more likely to hold inaccurate conceptions of pain and are more likely to develop ineffective methods to deal with their pain rather than seeking treatment [9]. As healthcare professionals can hold implicit biases which can impact their decision-making treatment, there is a need for an objective tool that can supplement this process.

The rise of non-pharmacological treatment options. Both the opioid epidemic and the increasingly evident biases in pain prescription have spurred the efforts of researchers in the field of alternative, non-pharmacological therapeutic options. Supplementary treatments, such as physical therapy and cognitive behavioral therapy, provide safe alternatives to pharmacological treatment and are recommended in most treatment plans. These advances are accompanied by a shift in the understanding of pain from a more traditional signal of nerve damage to a much more complex idea with a greater psychoanalytic focus. The expression and perception of pain is influenced by psychological conditions, such as the existence of anxiety or depression. It is also dictated by other variables, such as beliefs, attitudes, and mood [10]. Thus, non-medication-dependent treatments should be seriously considered as a viable alternative to pharmacologic options when crafting a treatment plan for a patient.

Section 1.2

Problem statement

Chronic pain is a debilitating, widespread condition with serious physical, mental, social, and financial effects. The goal of this project is to employ language models to accurately predict pain relief among individuals with chronic pain. To achieve this goal, we utilize a dataset consisting of interviews by participants with chronic back pain who selectively underwent either psychological treatment, a saline injection, or no treatment. More specifically, we fine-tune a transformer-based model to predict the level of pain relief that a patient experienced after treatment. We then examine the possible relevant features of the models and the reasoning it uses to make its predictions. Ultimately, we gain further insight into whether linguistic markers in speech can be predictive of pain. By analyzing the patient's speech directly, we can bypass the biases present in human healthcare professionals.

Section 1.3

Related work

The psychology behind pain perception. Previous research has suggested a possible psychological basis to pain relief, and alternative methods to pain relief could potentially be safer for patients. One case study by Linton et al. encourages physical therapists to augment treatment of chronic pain by identifying underlying psychosocial factors and developing a personalized treatment plan, including methods of cognitive-behavioral or fear-avoidance therapy, to address these factors [11]. The implementation of non-pharmacological strategies could be a promising avenue in reducing the number of prescribed opiates. This study also suggests that screening tools that assess psychosocial risk factors can be useful, which could be another way to incorporate machine learning into the realm of pain relief.

Machine learning techniques in the discussion of pain. The clinical use of machine learning in pain medicine is quickly growing in recent years, with approaches ranging from diagnosis to management of pain [12]. Most of the work, however, focuses on extracting information about pain from clinical notes and radiology scans. There is untapped potential to analyze pain directly from the patient's speech, which can reduce bias. A similar study aimed to classify headache pain into cluster or migraine headaches from self-reported narratives from patients using common classifiers [13]. The researchers also conducted a lexicon-based sentiment analysis to determine patterns of word choices made by the users when talking about pain, and they discovered substantial negative sentiment surrounding mentions of pain. This study demonstrates the effectiveness of using machine learning in classification for patients with chronic pain. However, these techniques were accomplished without the use of deep learning, unlike the transformer-based model we employ in this paper.

Overall, classification of pain using linguistic data and techniques in natural language processing (NLP) is not extensively studied, and there is much work to be done.

Chapter 2

Methodology and Experiments

The following chapter details the successful and unsuccessful approaches that we used with the goal of finetuning and analyzing pretrained language models for predicting pain relief.

Section 2.1

Data

2.1.1. Original Dataset

The main data source that we worked with was obtained from our collaborators of the Wager lab from the Psychological and Brain Sciences lab at Dartmouth College [14]. The data originated from a study which aimed at measuring the utility of psychological treatment for primary chronic back pain. Though 138 patients in the original study completed all the stages of the experiment, only 80 interviews were available for analysis. It consists of a set of 80 interviews from patients of varied demographics. These anonymized interviews were accompanied by physiologic and demographic information, such as pain in different areas of the body, education level, and alcohol/drug use frequency.

The patients were randomized to one of three treatments: pain reprocessing therapy (PRT), open-label placebo treatments, or control usual care. PRT is a therapy based on the theory that the brain constructs primary chronic pain when there is no tissue damage. By re-evaluating the causes and level of pain, PRT has potential to help patients afflicted with pain using solely a psychological basis. 26 patients were assigned to PRT, 25 patients were assigned to the placebo treatment, and the remaining 29 were asked to continue their typical treatment. The participants were interviewed before and after the treatment, and both evaluation sessions included an assessment with fMRI. Information about the participant included their level of pain before and after the treatment. After analysis of the data, PRT was proven to yield significant decreases in chronic back pain, deeming it a viable treatment option for those with chronic back pain.

In the dataset, each participant has rated their pain level before and after the treatment on a scale from 0 to 10, with 0 representing no pain and 10 representing extremely high pain. The numerical score of pain relief was calculated by the difference in these two ratings. Of the entire cohort, the highest pain relief difference is 6.5, representing a high initial pain rating and low final pain rating, resulting in significant pain relief after the treatment. The lowest pain relief difference is -1.5, indicating that the participant experienced greater pain after the treatment. These values were then scaled between 0 and 1.

We read through each interview and isolated responses to six standardized questions present in the majority of interviews that would elucidate information about the patient's pain condition and the effect of the treatment. These questions are listed as follows.

(a) How long have you had chronic back pain?

(b) What have you previously tried to alleviate your back pain?

(c) What was your initial expectation at the start of the study? Did you expect the treatment to help you?

(d) Has this study changed your relationship to your chronic back pain?

(e) If the treatment helped, what about the treatment do you think was the most helpful?

(f) How much pain are you in right now?

The goal of separating the interviews by questions was to determine if certain questions are more influential in predicting back pain than others.

We then manually classified each interview as 'No pain relief,' 'Slight pain relief,' or 'Significant pain relief' based on the content of the interview. The pain relief was counted regardless of whether or not the pain relief was experienced from the treatment. Of the 80 patients, 30 patients experienced no pain relief, 25 experienced slight pain relief, and 25 experienced significant pain relief. Thus, the dataset was somewhat equally split between the three classes.

2.1.2. LIWC Analysis

We used the Linguistic Inquiry and Word Count (LIWC) dictionary in order to quantify sentiment within the dataset. This dictionary is a list of words associated with labels, such as 'posemo' (positive emotion) and 'negemo' (negative emotion). This lexicon was used to assess positive and negative sentiment in the interviews by calculating the percentage of posemo/negemo labels relative to the number of words in an interview. The percent sentiment was calculated for the entirety of the dataset and per treatment group.

2.1.3. Generated Dataset

Through the process of using the original dataset for model creation, we found issues relating to the low number of interviews available. In addition, the answer to each of the six questions was not found in every interview. Thus, we made the decision to generate supplementary data using prompt engineering and GPT-4.

Prompt Engineering. GPT-4 is a large, multimodal language model created by OpenAI that utilizes deep learning to accomplish conversational tasks. In this case, we used a sample prompt as an input in order to receive a response from the model from the perspective of a patient. For every question, five prompts of varying pain relief classes were used to generate 60 unique prompts. In total, 1800 responses were generated. All repeats of responses were removed.

In order to adequately simulate patients encompassing all classes of pain relief, we engineered the prompts in such a way that the model would predict a response given a certain pain relief level. Many iterations of prompts were conducted consistently guide GPT-4 towards a human-like response that would provide insight to the predictive model. There were two types of prompts that were used for generation.

The first type of prompt included the question and a starter response to indicate pain relief, but it did not include the hypothetical patient. For example, for the fifth question, a prompt would be modeled as:

Answer the question by using the prompt, and write a paragraph of three or more sentences. <s> What was your initial expectation at the start of the study? Did you expect the treatment for your chronic back pain to help you? <s> My initial expectation was that the treatment wouldn't alleviate my back pain, but I was proved wrong.

This prompt guided GPT-4 to produce responses that answered the question with an emphasis on pain relief. An appropriate sample response is:

"I think the most helpful aspect of the treatment was that it was tailored to my individual needs. My doctor took the time to understand my specific situation and develop a treatment plan that addressed my particular pain. Additionally, the combination of physical therapy, medications, and lifestyle changes I was prescribed worked together to reduce my pain and improve my quality of life. Finally, the ongoing support and guidance I received from my doctor and other healthcare providers was invaluable in helping me stay on track with my treatment plan."

In the second type of prompt, a sample prompt included a hypothetical patient with a certain pain relief label, the question, and a starter response to guide the model. For example, for the first question, a prompt would be modeled as:

"Suppose you are a patient with chronic back pain whose pain has been relieved a little bit after treatment. Answer the question by using the prompt, and write a paragraph of three or more sentences. <s> How long have you had chronic back pain? </s> I still have pain now, but it was worse earlier"

An appropriate sample response is:

"I first started to have chronic back pain a few months ago. I have been undergoing treatment for it since then, and I can feel a difference in my pain levels. It's not completely gone, but the treatment has helped to relieve it to a certain degree. I'm hopeful that with continuing treatment I can eventually find complete relief."

The initial 1800 responses were generated with the first type of prompt, but the second type of prompt was preferred as it mitigated the time needed to label these responses. However, due to time constraints with sentence generation, a mix of the two prompt types were used for the questions.

Due to prohibitive time constraints, only automatically generated responses for question 3 and question 5 were able to be manually labeled with labels of either 'No pain relief', 'Slight pain relief', or 'Significant pain relief'.

Section 2.2

Model Creation

We determined that a pretrained language model would be sufficient in predicting pain relief by language analysis. The concept of transfer learning is employed when deep learning models are trained on other large corpora to accomplish a certain task, and small changes are made to narrow the scope of the model. For this task, we used the BERT base uncased model. We finetuned different models with the intention of examining the potential of prediction for two different goals: predict pain relief level, and predict pain relief level by treatment group.

2.2.1. What is BERT?

BERT is a transformers-based model that was pretrained on a large corpora of raw texts in the English language. Transformers rely on the encoder-decoder model, which essentially transforms a sequence of words into a vector representation encapsulating the semantics of a word and its position in the phrase. Transformers also employ an attention mechanism to calculate weights for each token in the input, and then it subsequently uses these weights to generate a prediction. The core intention of this model is to be fine-tuned for a downstream task. Finetuning involves adding an additional layer on top of this model and retraining the entire model on a different set of data for a more specific task, which is the method that we have employed.

2.2.2. Goal 1: Predict pain relief level

Our initial goal primarily focused on generally predicting pain relief level.

Original dataset. First, we finetuned a model on every interview and its manually-annotated label from the original dataset and used it for the classification task of labelling a test set of interviews with either 'No pain relief,' 'Slight pain relief,' or 'Significant pain relief.'

We then finetuned a model on the entire dataset and the numerical scores associated with the scaled pain relief difference provided by the participant, and used the finetuned model for the regression task of assigning each interview in a test set with a numerical score between 0 and 1.

As additional metrics for evaluation of the regression models, we calculated the Spearman and Pearson correlation coefficients for the data. Both the Spearman and Pearson coefficients represent the statistical strength between two sets of data. Spearman measures the monotonic relationship, while Pearson measures the linear relationship between two variables. We also calculated the Root Mean Square Error, which is a commonly used statistic for measuring the quality of predictions from a regression model. Each of these statistical measures aids in evaluating the performance of regression models.

Generated Dataset. After working with the original dataset for this goal, we turned to the generated dataset to determine if it could accomplish similar or higher levels of accuracy, as we were able to train on a significantly larger dataset with this method. As questions 3 and 5 were manually labeled, we finetuned one model on both of these questions. These models were then evaluated on a test dataset of generated and gold standard data to determine if these models could be utilized in prediction.

2.2.3. Goal 2: Predict pain relief level by treatment group

Our next goal focuses on predicting the level of pain alleviation based on the treatment group of the participant. First, models were trained on the interviews and corresponding multi-class labels to accomplish the classification task. Next, models were trained on the interviews with their numerical pain rating to complete the regression task. The performance of the classification and regression tasks was then compared.

Similarly to the regression models from Goal 1, we have calculated the RMSE and Spearman/Pearson correlation coefficients to measure and compare the efficacy of these regression models.

Section 2.3

Model Analysis

Following the finetuning of these models, we analyze the model weights and attention layers to understand more about why the models make certain predictions.

2.3.1. SHAPley Analysis of classification tasks

SHAP operates by removing each token in a given textual example and measuring the difference of the model outputs when said feature is not present. This helps us understand how and to what degree each word in an interview affects the prediction of the model. The numerical difference is represented visually on a sample of text by shading a token by color and intensity. A higher intensity of color illustrates a higher importance of a feature in prediction. Red indicates features that pushed the probability of making a certain prediction higher, while blue indicates features that pushed the probability lower. By calculating these SHAP values per token, we can recognize certain features that are more important for pain prediction than others.

2.3.2. Sentiment analysis using SST-2

Lastly, we conducted an analysis in order to understand which models and which specific layers are best for sentiment analysis. We accomplished this task using the Stanford Sentiment Treebank (SST-2) dataset, available through the Huggingface AI community. This corpus contains phrases extracted from movie reviews, and each sentence is accompanied by a binary label of negative or positive. We first started by finetuning a clean BERT model on this data and evaluating it on a test dataset from this corpus. This was used as a baseline for sentiment analysis to compare the performance of the pain models. Each finetuned model was evaluated on the SST-2 dataset, and the performance by layer of each model was recorded. Using this method, we were able to understand which layers were most important for sentiment analysis within our finetuned models, and whether sentiment was important for our models' predictions.

Chapter 3

Results and Discussion

In this section, we provide numerical and visual representations of the results and the conclusions that have been formed from this data.

Section 3.1

Dataset

Original Dataset. We conducted an analysis on the dataset to gain an initial understanding of the spread of pain relief and sentiment within the interviews. First, we counted the number of pain relief labels within the full dataset to understand if there is any class imbalance, which is displayed in table 3.1.

Table 3.1: Distribution of pain reliefs labels in full dataset

Class Labels	**Full Dataset**
No pain relief	30
Slight pain relief	25
Significant pain relief	25

We then calculated the number and percentage of patients experience each pain relief class to understand the distribution of pain relief per treatment group, as displayed in table 3.2.

Table 3.2: Distribution counts of pain relief labels per treatment group

	Group 1		**Group 2**		**Group 3**	
	Count	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
No Pain Relief	2	0.125	6	0.24	22	0.76
Slight Pain Relief	5	0.3125	15	0.6	5	0.17
Significant Pain Relief	9	0.5625	4	0.16	2	0.069

From this table, we can see that there is a clear class imbalance within the treatment groups. The pain relief labels are not distributed evenly within each group. In the treatment group, over half of the samples experienced significant pain relief. In the saline injection group, 60% of the samples experienced slight pain relief. Lastly, in the control group, 75% of the participants had no pain relief, which was to be expected.

Generated Dataset. We conducted a similar analysis on the generated dataset to understand the class balance within the dataset. Though responses were generated for all six questions, the resulting dataset consists of only questions 3 and 5, which were manually labeled.

Table 3.3: Distribution of pain reliefs labels in generated dataset

Class Labels	**Generated Dataset**
No pain relief	120
Slight pain relief	208
Significant pain relief	171

Though the prompts were written in such a way to split the distribution as evenly as possible, there is a greater amount of slight pain relief labels than no pain relief and significant pain relief.

3.1.1. LIWC Analysis

Original Dataset. In order to understand the positive and negative sentiment present in the interview texts, we conducted an LIWC analysis by counting the number of words with 'posemo' tags (positive emotion) and 'negemo' tags throughout the entire dataset.

We then conducted an LIWC analysis by each treatment group in order to view the sentiment types per treatment, considering that we have determined there is a class imbalance of pain relief between the treatment groups.

Table 3.4: Distribution counts of sentiment-related token in the dataset using the LIWC dictionary.

	Group 1		**Group 2**		**Group 3**		**Total**	
	Count	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
Total 'Posemo'	1723	2.87%	1206	3.18%	1225	2.98%	4154	2.99%
Total 'Negemo'	1368	2.28%	656	1.73%	841	2.05%	2865	2.06%
Total 'Anx'	228	0.38%	61	0.16%	92	0.22%	381	0.27%
Total 'Anger'	70	0.12%	25	0.07%	56	0.14%	151	0.11%
Total 'Sad'	151	0.25%	63	0.17%	86	0.21%	300	0.22%

We discovered that the number of positive and negative sentiment labels are evenly distributed throughout the treatment groups, insinuating that there is no significant difference in sentiment linguistically between all treatment groups.

Generated Dataset. We then conducted a similar analysis for the generated dataset.

Table 3.5: Distribution counts of sentiment-related token in the generated dataset using the LIWC dictionary.

	Generated Dataset	
Category Names	**Count**	**Percentage**
Total 'Posemo'	1263	3.88%
Total 'Negemo'	1600	4.91%
Total 'Anx'	126	0.39%
Total 'Anger'	31	0.10%
Total 'Sad'	111	0.34%

Compared to the percentage distribution of sentiment within the original dataset, the generated dataset is very close. There is no significant imbalance of a negative or positive class within the generated dataset, similarly to the original dataset.

Section 3.2

Model Creation

Next, we finetuned models with our two goals in mind.

3.2.1. Goal 1: Predict pain relief level

Original Dataset. We first began by finetuning a clean BERT model over the observed pain ratings to accomplish a regression task. Since the original dataset was smaller, the test dataset was comprised of only 24 samples. Though the original ratings of the test dataset had a variable range of 0 to .875, the predicted pain ratings from the model had a more narrow, median-centered scope of .327 to .500. In order to evaluate the performance of this model, we calculate the root mean squared error value (RMSE). This calculation is typically performed when observing the difference between a model's predicted value and its corresponding observed value. We also calculated Spearman and Pearson's correlation coefficient to measure the degree of correlation between the observed numerical values and the predicted values.

Table 3.6: Performance results for regression model trained on full dataset

	Full Dataset
Spearman Coefficient	0.0074
Pearson Coefficient	-0.011
RMSE	0.2331

The Spearman coefficient indicates that there is a very weak correlation between the observed and predicted values, as the coefficient is within the range of 0 to 0.19. The Pearson coeffcient is close to 0, corroborating this finding. Thus, we can conclude that this regression model is not able to predict pain relief ratings accurately.

We then finetuned a clean BERT model over the manual labels to accomplish a classification task. The test dataset consisted of 24 samples. The classification accuracy for this task was 37.5%, which indicates that this task is unable to be completed using classification, as the performance is worse than random chance.

Generated Dataset. In order to address the issue of a small sample size, we finetuned a model on a GPT-3-generated dataset, and then evaluated this model on both a subset of the generated dataset and the entirety of the original dataset. Due to time constraints, we were able to manually label questions three and five, so the model was trained on this combined dataset of 583 samples.

The test dataset for the generated dataset consisted of 175 samples, and the test dataset for the original dataset consisted of 74 samples.

Table 3.7: Classification performance of model trained on generated data

	Generated Data	**Original Data**
% Accuracy	91.43%	33.78%

The model trained on generated data performed well on the dataset of generated data, but poorly on the dataset of the original dataset. This may have been a result of the lexical similarities of the generated data. Though none of the responses were identical, they contained similar phrases which may have been important in its prediction. Thus, we can conclude that the current methods of generating data are not sufficient for prediction of true human responses relating to pain.

3.2.2. Goal 2: Predict pain relief level by treatment group

We first fine-tuned separate models on each treatment group for the classification task, and the performances are reported in Table 3.8.

Table 3.8: Performance of classification model trained by treatment group

	Group 1	**Group 2**	**Group 3**
% Accuracy	87.50%	50%	77.80%
Number of samples	8	8	9

The results demonstrate increased accuracy for prediction of the PRT and control groups. However, the results of the model are heavily skewed by the makeup of the data, since each group has demonstrated a severe class imbalance. Since the number of samples in the training set are low, the model has a tendency to predict solely the majority class.

Next, we trained models for the regression task by treatment group. We used the Spearman correlation coefficient, Pearson correlation coefficient, and RMSE to judge the fit of these models.

Table 3.9: Performance results for regression models trained on individual treatment groups

	Group 1	**Group 2**	**Group 3**
Spearman Coefficient	-0.05	0.26	0
Pearson Coefficient	-0.21	0.37	0.3
RMSE	0.25	0.16	0.19

For groups 1 and 3, the results indicate a very weak correlation, as both of the Spearman scores are between 0.0 and 0.19. Both of their Pearson scores are between -.3 to .3, which again indicates a low strength of association and corroborates the earlier finding.

There is a slightly greater correlation between the observed and predicted scores for group 2 as evidenced by the higher Spearman score of .26, but it is still weak. Its Pearson coefficient strengthens this finding, as it is higher and indicates a medium strength of association.

As lower RMSE indicates a better model, group 2 once again performs better than group 1 and group 3.

The class imbalance present in the treatment group split was also reflected in the predicted values of the regression models. The Group 1 model predicted values within the range of .599 to .80, Group 2 predicted within the range of .376 to .497, and Group 3 predicted within the range of .180 to .306. Clearly, the regression models detected the bias towards certain classes present in each group and predicted accordingly.

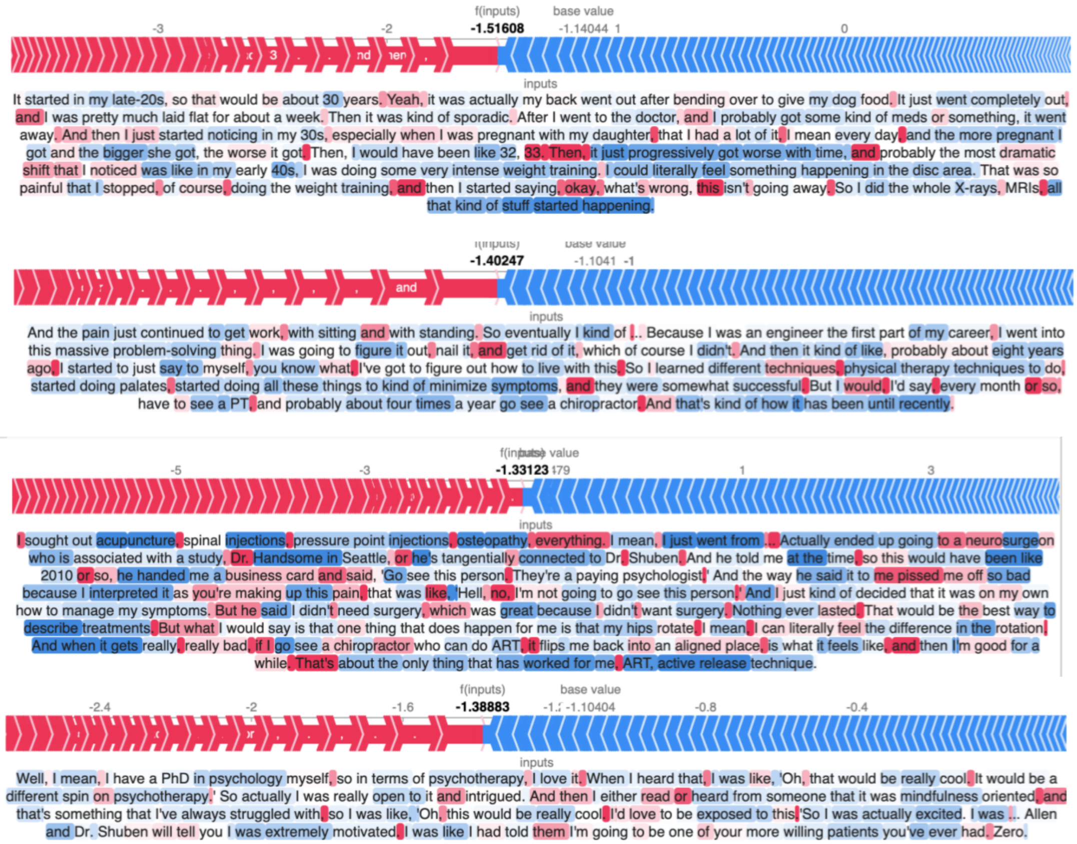
Taken together, these findings demonstrate the increased accuracy of prediction for group 2, and it outperforms models for group 1 and 3. However, this correlation is medium strength at best.

Section 3.3

Model Analysis

3.3.1. SHAP Analysis

Due to time constraints, we were only able to analyze the classification model weights for Goal 2. We looked at the three classification models trained on the treatment group data, and calculated the SHAP scores by token for a representative correctly predicted and incorrectly predicted interview. The scores in blue represent words that strongly predict towards higher pain relief. The scores in red represent words that strongly predict towards lower pain relief. A darker hue of each color represents the most important features, while lighter hues contribute less to the prediction of the model.



It started in my late-20s, so that would be about 30 years. Yeah, it was actually my back went out after bending over to give my dog food. It just went completely out, and I was pretty much laid flat for about a week. Then it was kind of sporadic. After I went to the doctor, and I probably got some kind of meds or something, it went away. And then I just started noticing in my 30s, especially when I was pregnant with my daughter, that I had a lot of it, I mean every day, and the more pregnant I got and the bigger she got, the worse it got. Then, I would have been like 32, 33. Then, it just progressively got worse with time, and probably the most dramatic shift that I noticed was like in my early 40s, I was doing some very intense weight training. I could literally feel something happening in the disc area. That was so painful that I stopped, of course, doing the weight training, and then I started saying, okay, what's wrong, this isn't going away. So I did the whole X-rays, MRIs, all that kind of stuff started happening.

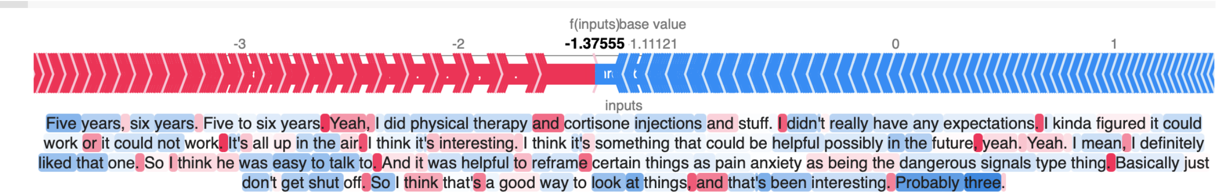
And the pain just continued to get work, with sitting and with standing. So eventually I kind of I... Because I was an engineer the first part of my career, I went into this massive problem-solving thing. I was going to figure it out, nail it, and get rid of it, which of course I didn't. And then it kind of like, probably about eight years ago, I started to just say to myself, you know what, I've got to figure out how to live with this. So I learned different techniques, physical therapy techniques to do, started doing palates, started doing all these things to kind of minimize symptoms, and they were somewhat successful. But I would, I'd say, every month or so, have to see a PT, and probably about four times a year go see a chiropractor. And that's kind of how it has been until recently.

I sought out acupuncture, spinal injections, pressure point injections, osteopathy, everything. I mean, I just went from ... Actually ended up going to a neurosurgeon who is associated with a study, Dr. Handsome in Seattle, or he's tangentially connected to Dr. Shuben. And he told me at the time, so this would have been like 2010 or so, he handed me a business card and said, 'Go see this person. They're a paying psychologist.' And the way he said it to me pissed me off so bad because I interpreted it as you're making up this pain, that was like, 'Hell, no, I'm not going to go see this person.' And I just kind of decided that it was on my own how to manage my symptoms. But he said I didn't need surgery, which was great because I didn't want surgery. Nothing ever lasted. That would be the best way to describe treatments. But what I would say is that one thing that does happen for me is that my hips rotate. I mean, I can literally feel the difference in the rotation. And when it gets really, really bad, if I go see a chiropractor who can do ART, it flips me back into an aligned place, is what it feels like, and then I'm good for a while. That's about the only thing that has worked for me, ART, active release technique.

Well, I mean, I have a PhD in psychology myself, so in terms of psychotherapy, I love it. When I heard that, I was like, 'Oh, that would be really cool. It would be a different spin on psychotherapy.' So actually I was really open to it and intrigued. And then I either read or heard from someone that it was mindfulness oriented, and that's something that I've always struggled with, so I was like, 'Oh, this would be really cool. I'd love to be exposed to this.' So I was actually excited. I was ... Allen and Dr. Shuben will tell you I was extremely motivated. I was like I had told them I'm going to be one of your more willing patients you've ever had. Zero.

Figure 3.1: Correctly predicted interview from Group 1 - PRT. This interview was correctly predicted as Significant pain relief.

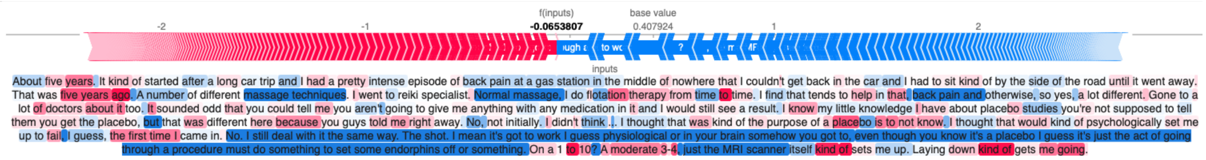
As the model for Group 1 strongly predicted Significant pain relief, it is evident that the designated important features for significant relief is much more predictive than the features for slight or no pain relief. The most important features for high pain relief involve words about treatment, such as 'acupuncture' and 'injections'. However, the model is not as strong when isolating tokens that are not predictive of pain relief. This is evidenced by the strong red markings on unimportant tokens, such as punctuation and coordinating conjunctions.



Five years, six years. Five to six years. Yeah, I did physical therapy and cortisone injections and stuff. I didn't really have any expectations. I kinda figured it could work or it could not work. It's all up in the air. I think it's interesting. I think it's something that could be helpful possibly in the future, yeah. Yeah. I mean, I definitely liked that one. So I think he was easy to talk to. And it was helpful to reframe certain things as pain anxiety as being the dangerous signals type thing. Basically just don't get shut off. So I think that's a good way to look at things, and that's been interesting. Probably three.

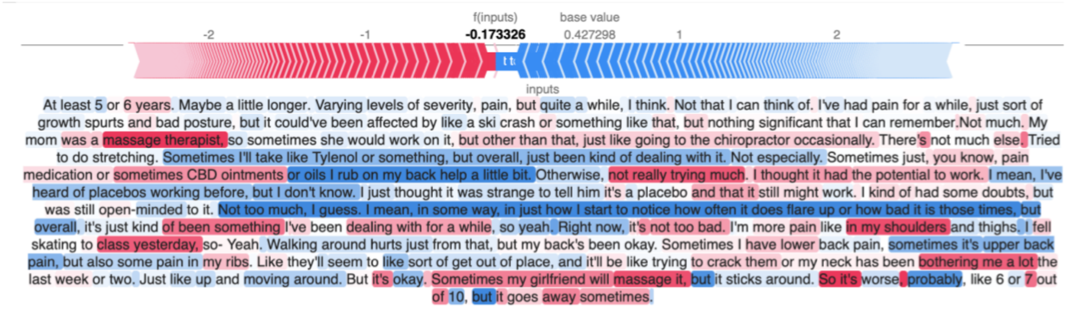
Figure 3.2: Incorrectly predicted interview from Group 1 - PRT. This interview was incorrectly predicted as Significant pain relief, but represents No pain relief.

As discussed earlier, this model has poor performance when predicting labels of slight and no pain relief. Again, the red indicators highlight filler words and punctuation. This interview also contains few mentions of pain relief, which has contributed to the incorrect prediction. The prediction was influenced by the mentions of treatments, which skewed the prediction towards Significant pain relief. The blue labels, which are indicative of high pain relief, highlight more topical words, like 'injections' and 'three' (in mentioning current pain).



About five years. It kind of started after a long car trip and I had a pretty intense episode of back pain at a gas station in the middle of nowhere that I couldn't get back in the car and I had to sit kind of by the side of the road until it went away. That was five years ago. A number of different massage techniques. I went to reiki specialist. Normal massage, I do flotation therapy from time to time. I find that tends to help in that, back pain and otherwise, so yes, a lot different. Gone to a lot of doctors about it too. It sounded odd that you could tell me you aren't going to give me anything with any medication in it and I would still see a result. I know my little knowledge I have about placebo studies you're not supposed to tell them you get the placebo, but that was different here because you guys told me right away. No, not initially. I didn't think ... I thought that was kind of the purpose of a placebo is to not know. I thought that would kind of psychologically set me up to fail, I guess, the first time I came in. No. I still deal with it the same way. The shot. I mean it's got to work I guess physiological or in your brain somehow you got to, even though you know it's a placebo I guess it's just the act of going through a procedure must do something to set some endorphins off or something. On a 1 to 10? A moderate 3-4, just the MRI scanner itself kind of sets me up. Laying down kind of gets me going.

Figure 3.3: Correctly predicted interview from Group 2 - Saline Injection. This interview was correctly predicted as Slight pain relief.

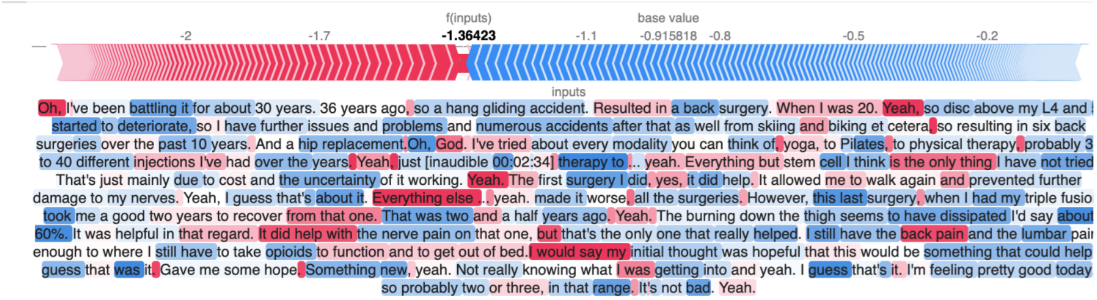


At least 5 or 6 years. Maybe a little longer. Varying levels of severity, pain, but quite a while, I think. Not that I can think of. I've had pain for a while, just sort of growth spurts and bad posture, but it could've been affected by like a ski crash or something like that, but nothing significant that I can remember.Not much. My mom was a massage therapist, so sometimes she would work on it, but other than that, just like going to the chiropractor occasionally. There's not much else. Tried to do stretching. Sometimes I'll take like Tylenol or something, but overall, just been kind of dealing with it. Not especially. Sometimes just, you know, pain medication or sometimes CBD ointments or oils I rub on my back help a little bit. Otherwise, not really trying much. I thought it had the potential to work. I mean, I've heard of placebos working before, but I don't know. I just thought it was strange to tell him it's a placebo and that it still might work. I kind of had some doubts, but was still open-minded to it. Not too much, I guess. I mean, in some way, in just how I start to notice how often it does flare up or how bad it is those times, but overall, it's just kind of been something I've been dealing with for a while, so yeah. Right now, it's not too bad. I'm more pain like in my shoulders and thighs. I fell skating to class yesterday, so- Yeah. Walking around hurts just from that, but my back's been okay. Sometimes I have lower back pain, sometimes it's upper back pain, but also some pain in my ribs. Like they'll seem to like sort of get out of place, and it'll be like trying to crack them or my neck has been bothering me a lot the last week or two. Just like up and moving around. But it's okay. Sometimes my girlfriend will massage it, but it sticks around. So it's worse, probably, like 6 or 7 out of 10, but it goes away sometimes.

Figure 3.4: Incorrectly predicted interview from Group 2 - Saline Injection. This interview was incorrectly predicted as Slight pain relief, but represents No pain relief.

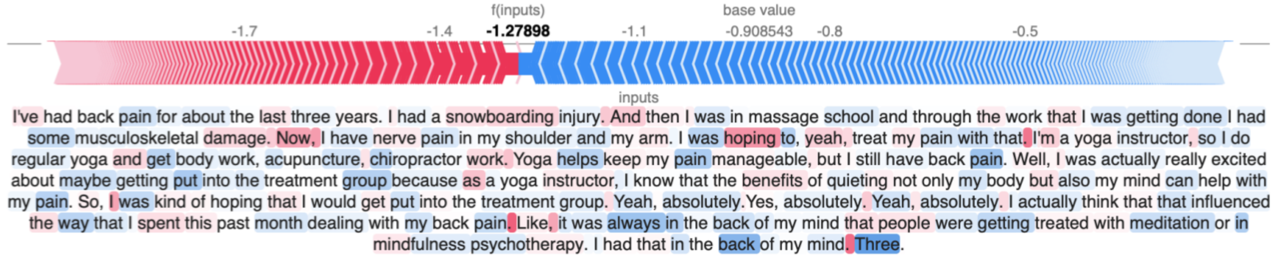
Both of the blue and red labels on the correctly and incorrectly predicted interviews are more accurately placed on the text and seems to avoid filler words more than the previous model. They address words related to pain, therapies, and locations of pain. The more equal amounts of blue and red amounts and intensities may have contributed to the Slight pain relief prediction.

Overall, Group 2 appears to mark the more important words for pain in both directions of the SHAP values than Group 1, whose red direction marked more unimportant features.



Oh, I've been battling it for about 30 years. 36 years ago, so a hang gliding accident. Resulted in a back surgery. When I was 20. Yeah, so disc above my L4 and started to deteriorate, so I have further issues and problems and numerous accidents after that as well from skiing and biking et cetera, so resulting in six back surgeries over the past 10 years. And a hip replacement.Oh, God. I've tried about every modality you can think of, yoga, to Pilates, to physical therapy, probably 3 to 40 different injections I've had over the years. Yeah, just [inaudible 00:02:34] therapy to ... yeah. Everything but stem cell I think is the only thing I have not tried That's just mainly due to cost and the uncertainty of it working. Yeah. The first surgery I did, yes, it did help. It allowed me to walk again and prevented further damage to my nerves. Yeah, I guess that's about it. Everything else ... yeah. made it worse, all the surgeries. However, this last surgery, when I had my triple fusio took me a good two years to recover from that one. That was two and a half years ago. Yeah. The burning down the thigh seems to have dissipated I'd say about 60%. It was helpful in that regard. It did help with the nerve pain on that one, but that's the only one that really helped. I still have the back pain and the lumbar pai enough to where I still have to take opioids to function and to get out of bed. I would say my initial thought was hopeful that this would be something that could help guess that was it. Gave me some hope. Something new, yeah. Not really knowing what I was getting into and yeah. I guess that's it. I'm feeling pretty good today so probably two or three, in that range. It's not bad. Yeah.

Figure 3.5: Correctly predicted interview from Group 3 - Control. This interview was correctly predicted as No pain relief.



I've had back pain for about the last three years. I had a snowboarding injury. And then I was in massage school and through the work that I was getting done I had some musculoskeletal damage. Now, I have nerve pain in my shoulder and my arm. I was hoping to, yeah, treat my pain with that. I'm a yoga instructor, so I do regular yoga and get body work, acupuncture, chiropractor work. Yoga helps keep my pain manageable, but I still have back pain. Well, I was actually really excited about maybe getting put into the treatment group because as a yoga instructor, I know that the benefits of quieting not only my body but also my mind can help with my pain. So, I was kind of hoping that I would get put into the treatment group. Yeah, absolutely.Yes, absolutely. Yeah, absolutely. I actually think that that influenced the way that I spent this past month dealing with my back pain. Like, it was always in the back of my mind that people were getting treated with meditation or in mindfulness psychotherapy. I had that in the back of my mind. Three.

Figure 3.6: Incorrectly predicted interview from Group 3 - Control. This interview was incorrectly predicted No pain relief, but represents Significant pain relief.

Similar to Group 1, the blue direction for Group 3 is more accurate in detecting key features for pain prediction than the red direction, which corroborates the strong bias towards prediction for No pain relief. The strong red labels are placed on unimportant phrases, such as "Yeah" and "Everything else."

Overall, Group 2 appears to mark important features well for both the No pain relief direction and the Significant pain relief direction, and this manifests by the bias towards Slight pain relief predictions. Groups 1 and 3 display increased performance towards Significant pain relief and No pain relief respectively, which explains the strong lean in performance towards the extremes.

3.3.2. Sentiment analysis using SST-2

In order to understand if sentiment was being analyzed by our models to make predictions, we employed the use of the SST-2 dataset.

These results were compared to the ordered layers for the regression and classification models of each treatment group.

Table 3.10: Ordered layers of regression models evaluated on SST-2 data

Group 1		**Group 2**		**Group 3**	
Layer Ranking	**% Accuracy**	**Layer Ranking**	**% Accuracy**	**Layer Ranking**	**% Accuracy**
8	54.78%	8	57.64%	8	57.58%
5	50.11%	12	50.11%	12	50.11%
12	50.11%	2	50.05%	2	50.05%
6	50.05%	5	50.05%	6	50.05%
2	50.00%	6	50.05%	5	50.00%
1	49.95%	1	49.95%	1	49.95%
7	49.95%	4	49.95%	7	49.95%
3	49.89%	7	49.95%	3	49.89%
4	49.89%	3	49.89%	4	49.89%
9	49.84%	9	49.45%	9	49.40%
10	47.97%	11	48.19%	10	48.46%
11	47.75%	10	47.47%	11	47.09%

Table 3.11: Ordered layers of classification models evaluated on SST-2 data

Group 1		**Group 2**		**Group 3**	
Layer Ranking	**% Accuracy**	**Layer Ranking**	**% Accuracy**	**Layer Ranking**	**% Accuracy**
8	54.78%	8	57.64%	8	57.58%
5	50.11%	12	50.11%	12	50.11%
12	50.11%	2	50.05%	2	50.05%
6	50.05%	5	50.05%	6	50.05%
2	50.00%	6	50.05%	5	50.00%
1	49.95%	1	49.95%	1	49.95%
7	49.95%	4	49.95%	7	49.95%
3	49.89%	7	49.95%	3	49.89%
4	49.89%	3	49.89%	4	49.89%
9	49.84%	9	49.45%	9	49.40%
10	47.97%	11	48.19%	10	48.46%
11	47.75%	10	47.47%	11	47.09%

From the data, it is apparent that there is no significant difference between the three regression models when considering the accuracies associated with the layer rankings. All the regression models predict sentiment equally well, which is around random choice.

The same is true for the classification models, as there is no clear difference when comparing these models to each other.

There is a slight performance difference of the layers within both the regression and classification models in that layer 8 is favored in sentiment analysis, and it is at the top of the rankings for every model. At the same time, layers 10 and 11 consistently predict sentiment at a lower accuracy than the other layers.

The model results were compared to the control model, which is a clean BERT-base-uncased model finetuned and evaluated on the SST-2 dataset.

Table 3.12: Ordered layers of clean BERT model evaluated and finetuned on SST-2 data

Clean BERT Model	
Layer Ranking	**% Accuracy**
11	86.91%
10	51.55%
6	49.71%
4	49.25%
8	49.25%
1	49.02%
2	49.02%
3	49.02%
7	49.02%
0	48.91%
5	48.91%
9	48.22%

The drop in performance in the finetuned model indicates that finetuning the model on the back pain dataset hinders the performance of sentiment analysis. Thus, the model is not learning sentiment-related features, and the tasks of predicting pain relief with the given dataset and sentiment analysis are only loosely related, if at all.

This is consistent with the equal levels of sentiment between the three groups that was discovered in the LIWC analysis. Since the negative and positive sentiment is equally distributed between the three treatment groups, there is little distinction between the groups for the model to learn from when making a prediction.

Section 3.4

Limitations

We faced certain limitations and obstacles when examining the dataset, which possibly inhibited the performance of these models.

The pain ratings before and after the interviews were based off of subjective numbers by the patients. However, a significant amount of time had passed between the interviews, and it is possible that the patient had forgotten their original score in order to make a viable comparison between the two scores. As ratings are subjective between participants, it is difficult to compare these scales to one another, and these ratings may have been artificially inflated or deflated by the patient's memory.

Patients may have experienced pain relief unrelated to the treatments, which could have incorrectly influenced the prediction of the model. For example, patients in the control group were asked to continue their normal treatment, and one patient attributes their pain relief to a medication they have been taking, not any treatment specifically for the study. The interviewee proceeds to state,

"Right now I don't have any pain at all. But there are external conditions that are happening...So I was diagnosed with autoimmune disease about a month ago. And started on Prednisone. And Prednisone is an anti-inflammatory drug...so that's helped the inflammation in my back. So I haven't had much pain. But it's not the shot, it's the Prednisone."

The inclusion of such cases, whether explicitly stated by the patient or not, may have affected the results.

The data was initially prepared with labels for the interviews. However, these labels were based off of the numeric pain ratings, not the textual content of the interviews. The original researchers predefined cutoffs within the pain difference range to classify interviews into 'Significant,' 'Slight,' or 'No' pain relief labels. In spite of this, there were multiple instances of interviews that did not adequately fit into these categories. Due to this mis-categorization, the interviews were re-classified by one individual. It is important to have a team audit the labels to provide supplementary feedback and verify the accuracy.

Furthermore, the interviews were manually split up by question. Though there was some standardization in the questions being asked, not every interview was asked every question. This led to gaps in the data, with questions 4 and 5 being asked less frequently than questions 1 and 6.

In addition, some of the data was less helpful to predict on, such as one-answers ("No", "Yeah").

The incompleteness of the interviews amplified most significant limitation of this project, which was the size of the dataset. This was especially noticeable splitting up the interviews by treatment groups, as the test datasets when evaluating the fine-tuned models consisted of 7-9 interviews. As there was also a class imbalance within the treatment groups, this has led to the severe bias in predicting towards one class. Though fine-tuning is a better method for small datasets than training a model from scratch, the limited amount of data can severely limit the number and generalizability of conclusions we can make. Though we attempted to account for this by generating more sentences with GPT-4, these AI-generated sentences are no match for the variability of human responses. This conclusion was made from the discrepancy in performance when training the model on the generated data and evaluating it on the gold standard participant responses. Many of the challenges associated with the generated dataset related to prompt creation and labeling. Though none of the responses were identical, there were similar phrases that the model may have perceived that would lead to a significantly higher performance among evaluation on the generated test set. More work must be done in the prompt creation stage to encourage GPT-4 to produce responses of a higher diversity.

Chapter 4

Future Work

In this section, we will detail approaches that will be examined in the future to better understand the general problem.

We aspire to expand the generation of sentences through number of sentences and variety of prompt generation. Rather than generating prompts by pain relief label, we will either provide a pain relief rating or ask GPT-4 to generate a pain relief rating based on the response it provides. Additionally, we could try generating data by treatment group to strengthen the findings for Goal 2. Example prompts could be modeled as,

"Suppose you are patient whose chronic back pain has decreased from a 10 to 4 after {insert treatment and explanation} on a scale from 0 to 10, with 0 representing no back pain and 10 representing extreme back pain. How would you respond to the following question: {insert question}"

Similar to the other method of generating classification data, we will train a model on this generated regression data and evaluate it both on the actual and generated dataset. In order to assess the performance of the regression models, we will again use the Spearman and Pearson coefficients alongside the RMSE score.

We also will try to generate prompts with gender, race, and/or ethnicity in mind to develop a greater range of responses that will incorporate more viewpoints. However, the responses generated will be highly dependent on the data that was used to train GPT-4, and researchers must be careful to exclude bias that may be present in the responses. A sample prompt could be modeled as,

"Answer the following question from the perspective of an Asian woman who has been experiencing chronic back pain for ten years. {insert question}"

We hope to analyze the SHAP contributions of each token for the regression models, as we have done for the classification models for Goal 2. This analysis will give us a better understanding of the importance for each feature for the model predictions. We also plan to examine the SHAP contribution per sentence, rather than solely per token.

Finally, we intend to append one more goal to the existing two. While we have previously examined the dataset with respect to predicting pain relief generally and by treatment group, we will investigate the possibility of predicting the treatment group by the interview and the pain relief label. In order to handle multiple text inputs, one potential solution could involve appending the label to the end of the interview, separated by a special token such as <CLS>. Being able to identify treatment groups by linguistic markers could have greater implications for predicting the outcomes of certain treatments for an individual.

Chapter 5

Conclusion

In this work, we have examined the potential to predict pain relief from linguistic features. This study has implications in treating the costly problem of pain management in the United States. We have employed a variety of different methods within natural language processing to explore two different goals: 1) to predict pain relief, and 2) to predict pain relief based on treatment group. Through finetuning general classification and regression models, we have discovered no significant correlation between predicted values and observed values. We also trained models by treatment group, which resulted in greater accuracy for the classification models and a slightly stronger correlation in the placebo regression model. From SHAPley analysis of the models, we discovered certain features that contributed to making predictions. We also analyzed the weak role of sentiment analysis in model prediction. There was a significant limitation in the size of the dataset, which future work will address by further refining prompts for GPT-4 response generation. Our hope is that this research serves as a foundation for future analysis of pain relief with machine learning tools.

References

[1] R. Jason Yong, Peter M. Mullins, and Neil Bhattacharyya. "Prevalence of chronic pain among adults in the United States". en-US. In: *PAIN* 163.2 (Feb. 2022), e328. ISSN: 0304-3959. DOI: `10.1097/j.pain.0000000000002291`. URL: `https://journals.lww.com/pain/fulltext/2022/02000/prevalence_of_chronic_pain_among_adults_in_the.31.aspx` (visited on 05/17/2023).

[2] James Dahlhamer. "Prevalence of Chronic Pain and High-Impact Chronic Pain Among Adults — United States, 2016". en-us. In: *MMWR. Morbidity and Mortality Weekly Report* 67 (2018). ISSN: 0149-21951545-861X. DOI: `10.15585/mmwr.mm6736a2`. URL: `https://www.cdc.gov/mmwr/volumes/67/wr/mm6736a2.htm` (visited on 05/16/2023).

[3] Darrell J. Gaskin and Patrick Richard. "The economic costs of pain in the United States". eng. In: *The Journal of Pain* 13.8 (Aug. 2012), pp. 715–724. ISSN: 1528-8447. DOI: `10.1016/j.jpain.2012.03.009`.

[4] Nancy Wells, Chris Pasero, and Margo McCaffery. "Improving the Quality of Care Through Pain Assessment and Management". eng. In: *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Ed. by Ronda G. Hughes. Advances in Patient Safety. Rockville (MD): Agency for Healthcare Research and Quality (US), 2008. URL: `http://www.ncbi.nlm.nih.gov/books/NBK2658/` (visited on 05/17/2023).

[5] Center for Substance Abuse Treatment. *Managing Chronic Pain in Adults With or in Recovery From Substance Use Disorders*. eng. SAMHSA/CSAT Treatment Improvement Protocols. Rockville (MD): Substance Abuse and Mental Health Services Administration (US), 2012. URL: `http://www.ncbi.nlm.nih.gov/books/NBK92048/` (visited on 06/03/2023).

[6] Alexander M. Dydyk, Nitesh K. Jain, and Mohit Gupta. "Opioid Use Disorder". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: `http://www.ncbi.nlm.nih.gov/books/NBK553166/` (visited on 05/17/2023).

[7] Mohammadreza Azadfard, Martin R. Huecker, and James M. Leaming. "Opioid Addiction". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. URL: `http://www.ncbi.nlm.nih.gov/books/NBK448203/` (visited on 05/17/2023).

[8] Kelly M. Hoffman et al. "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites". In: *Proceedings of the National Academy of Sciences* 113.16 (Apr. 2016). Publisher: Proceedings of the National Academy of Sciences, pp. 4296–4301. DOI: `10.1073/pnas.1516047113`. URL: `https://www.pnas.org/doi/10.1073/pnas.1516047113` (visited on 05/17/2023).

[9] Nicole Atkins and Karim Mukhida. "The relationship between patients' income and education and their access to pharmacological chronic pain management: A scoping review". In: *Canadian Journal of Pain* 6.1 (), pp. 142–170. ISSN: 2474-0527. DOI: `10.1080/24740527.2022.2104699`. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9450907/` (visited on 05/17/2023).

[10] Stanley I Innes. "Psychosocial factors and their role in chronic pain: A brief review of development and current status". In: *Chiropractic & Osteopathy* 13 (Apr. 2005), p. 6. ISSN: 1746-1340. DOI: `10.1186/1746-1340-13-6`. URL:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1151654/ (visited on 06/03/2023).

[11] Steven J Linton, Ida K Flink, and Johan W S Vlaeyen. "Understanding the Etiology of Chronic Pain From a Psychological Perspective". In: *Physical Therapy* 98.5 (May 2018), pp. 315–324. ISSN: 0031-9023. DOI: 10.1093/ptj/pzy027. URL: https://doi.org/10.1093/ptj/pzy027 (visited on 05/18/2023).

[12] Maria Matsangidou et al. "Machine Learning in Pain Medicine: An Up-To-Date Systematic Review". en. In: *Pain and Therapy* 10.2 (Dec. 2021), pp. 1067–1084. ISSN: 2193-651X. DOI: 10.1007/s40122-021-00324-2. URL: https://doi.org/10.1007/s40122-021-00324-2 (visited on 09/05/2022).

[13] Nicolas Vandenbussche et al. "Using natural language processing to automatically classify written self-reported narratives by patients with migraine or cluster headache". In: *The Journal of Headache and Pain* 23.1 (Sept. 2022), p. 129. ISSN: 1129-2377. DOI: 10.1186/s10194-022-01490-0. URL: https://doi.org/10.1186/s10194-022-01490-0 (visited on 05/18/2023).

[14] Yoni K. Ashar et al. "Effect of Pain Reprocessing Therapy vs Placebo and Usual Care for Patients With Chronic Back Pain: A Randomized Clinical Trial". In: *JAMA Psychiatry* 79.1 (Jan. 2022), pp. 13–23. ISSN: 2168-622X. DOI: 10.1001/jamapsychiatry.2021.2669. URL: https://doi.org/10.1001/jamapsychiatry.2021.2669 (visited on 05/23/2023).