Examining Medical Narratives of Eating Disorder Recovery on Reddit

Anthony Campbell
anthonycampb@umass.edu

Jeng-Yu Chou
jchou@umass.edu

Marisa Hudspeth
mhudspeth@umass.edu

Trang Nguyen
tramnguyen@umass.edu

Chau Pham
ctpham@umass.edu

1 Introduction

Eating disorders (EDs) are complex, psychiatric disorders characterized by persistently abnormal eating behaviors. Over time, these disorders can have disruptive effects on physical health and psychosocial functioning. Recovery from an eating disorder can be a long and arduous process, often involving medical and psychological interventions, as well as support from friends and family. Despite being a major public health concern, ED recoveries receive little open discussion due to prevalent social stigma. In recent years, individuals suffering from EDs have increasingly turned to online communities such as Reddit to share their experiences of recovery, seek support, and raise awareness.

In this project, we looked into the Reddit posts of those with eating disorders. Using natural language processing (NLP) techniques (prompt engineering, topic modeling, sentiment lexicon induction, and connotation frames), we examined the following questions:

1. How do the topics mentioned by users who suffer from EDs change over time? How does the community-specific sentiment change across multiple posts by the same users?
2. What are the most helpful factors for people with EDs?
3. What are the most common triggers or negative factors that affect people with EDs?
4. Which entities (ED, family, medical professionals, food) have the most power compared to the writer?

By answering these questions, we hoped to not only uncover general trends but also provide comparative insights across time and between different subreddits. Our code is available at https://github.com/nguyentr17/cs685-narrative-analysis

2 What you proposed vs. what you accomplished

- ~~Acquire and pre-process data~~
  - *Initially we aimed to collect three years of data, from 01/01/2020 - 01/31/2023, to get ED-related narratives/stories from before, during, and after the peak of the COVID-19 pandemic. Due to restrictions with the PushshiftAPI, some posts during that time were unable to be collected from certain subreddits. We decided to widen the period from 2015 to 2023 to make up for the number of submissions unable to be collected.*
- ~~Implementing code for character, sentiment, event, and relation extraction tasks~~
  - *We did not use BERT to do clinical extraction, instead we used ChatGPT. We also decided to focus less on clinical extractions and more on factors that may help or harm someone's eating disorder recovery process.*
- ~~Analyze the extracted output and do an error analysis~~
  - *We initially planned to examine the posts for the temporal trends of sentiments with the VADER sentiment tool. Instead, we used the SentProp (Hamilton et al., 2016) algorithm to induce domain-specific sentiment lexicons from unlabeled text and observe average polarity scores of words across a variety of subreddits and submissions (posts).*

- ~~Work on the final report~~

3 Related work

3.1 Online Health Narratives

Online platforms have provided the opportunity for people to share their medical stories, often in forums dedicated to specific diseases, disorders, or medications. Recent research has employed topic modeling, sentiment analysis, and other methods to analyze these online narratives. Giorgi et al. used topic modeling to compare language between Reddit users who engage in non-suicidal self-injury versus those who have substance use disorders. Murray et al. used topic modeling and sentiment analysis to temporally track the symptoms of online posters who tested COVID-19 positive.

In a more relevant work, Antoniak et al. analyzed the common sequences of events and power dynamics in online birth stories. They inferred the sequence of events using an LDA topic model (Blei et al., 2003a). In addition, they extracted characters using fuzzy matching against a predefined lexicon and then analyzed the power of these personas with connotation frames (Sap et al., 2017). These approaches directly inspired the topic and persona analyses in our project.

3.2 LLMs for Information Extraction

More recently, more powerful models such as ChatGPT have been used to extraction information. Yang et al. investigated the potential use of ChatGPT for mental health analysis and emotional reasoning. They performed comprehensive evaluations using 11 datasets across five tasks: binary and multi-class mental health condition detection, cause/factor detection of mental health conditions, emotion recognition in conversations, and causal emotion entailment. The results suggest that ChatGPT surpasses traditional neural network methods but lags behind advanced task-specific methods. Gao et al. leveraged large language models (LLMs) in event extraction, a fundamental yet challenging task in NLP. The researchers conducted experiments to assess ChatGPT's feasibility and performance for this task, in comparison to task-specific models like EEQA. The findings indicate that ChatGPT's performance, on average, is only 51.04% of that of the task-specific model in long-tail and complex scenarios. The model was found to be not robust enough and was highly sensitive to different prompt styles. Wei et al. focuses on zero-shot information extraction (IE), a method that aims to build IE systems from unannotated text. The researchers propose a two-stage framework (ChatIE), which transforms the zero-shot IE task into a multi-turn question-answering problem, leveraging the capabilities of large language models (LLMs) like ChatGPT. They extensively evaluate this framework on three IE tasks: entity-relation triple extract, named entity recognition, and event extraction, using six datasets across two languages. The empirical results demonstrate that ChatIE performs impressively, even surpassing some full-shot models on several datasets, such as NYT11-HRL.

3.3 LLMs for Classification

Researchers have also begun to examine the use of LLMs such as ChatGPT for classification tasks. This has some advantages over traditional methods such as fine-tuning BERT: less or no training data (if doing zero-shot) and no compute required. Recently, researchers have tested ChatGPT specifically on Computational Social Science tasks, finding that it generally does not outperform fine-tuned models but still agrees with human evaluations (Ziems et al., 2023).

4 Your dataset

Text data (submissions only, no comments or replies) was collected from subreddits, including the following —
"anorexiaflareuphelp",
"AnorexiaNervosa",
"AnorexiaRecovery",
"bingeeating",
"BingeEatingDisorder",
"BingeEatingRecovery",
"bulimia",
"BulimiaAndAnaSupport",
"BulimiaRecovery",
"eating_disorders",
"EatingDisorderHope",
"EatingDisorders",
"EDAnonymous",
"EdAnonymousAdults",
"EDRecovery_public",
"edsupport",
"NotOtherwiseSpecified",
"PurgingDisorder
– as listed by Donati and Strapparava along with a couple of additional eating disorder-related sub-

Task	Number of posts after pre-processing	Median number of words per post after pre-processing	Preprocessing methods
Narrative Detection	16,672	406	Text from all collected posts (non null, non deleted) were used.
Domain-specific Sentiment Lexicon	16,672	406	Text from all collected posts (non null, non deleted) were used.
Topic Modeling	2,841	178	Narrative posts by users with at least 2 and at most 50 posts.
Factor & Trigger extraction	2,841	178	Narrative posts by users with at least 2 and at most 50 posts (same sample as Topic Modeling).
Agency and Power of Personas	7676	182	Default spaCy pipeline[1] and spaCy pipeline with coreference resolution[2]

Table 1: Data statistics and pre-processing pipeline for each task.

reddits. These subreddits contain submissions of firsthand eating disorder experiences and recovery stories. We used the Pushshift API to collect submissions from the listed subreddits from January 2015 through January 2023. We did not annotate the collected text ourselves; however, some subreddits do use Reddit Flair flags, special submission labels placed by the submitter, which we made use of to give context to submissions.

4.1 Data Preprocessing

Statistics and preprocessing pipelines for all tasks can be found in Table 1.

For the task of Narrative Detection, we used all collected text. The resulting sample is used for the tasks of Topic Modeling, Factor & Trigger extraction, and Power and Agency analysis.

For the task of Domain-Specific Sentiment Lexicons, we took all of the submission text ("selftext") collected without filtering for the length of the submission or any other properties. This text was then lemmatized, tokenized, and had stopwords removed using NLTK (Loper and Bird, 2002) so that we could calculate the polarity scores for the roots of similar words (e.g. "gain", "gained", "gaining").

For the task of Topic Modeling, we wanted to examine possible diachronic topical changes in posts written by the same user. Therefore, we had two main criteria for data preprocessing in this task. First, we looked at posts that were classified as narrative, a result of the technique in section 5.1. Second, we only retained users who post at least twice so that we can examine possible changes in topics across time. We decided to focus on individual changes, and thus we made sure to filter out users who post more than 50 times. This strategy helped us avoid anonymous or admin users whose posts are submitted by a large number of users. Our resulting sample for this task has 2,841 posts, with an average of approximately 178 words in each post. This sample was also used for the task of Factor & Trigger extraction.

4.2 Data annotation

For the trigger extraction task, we collected and manually annotated a sample of 216 posts to evaluate different instruction prompting methods. More details on how the sample is selected and annotated can be found in Section 5.3.

5 Your approach

5.1 Narrative Detection

We followed Ganti et al. and finetuned a BERT-based classification model to detect whether a Reddit post is a narrative or not. We used the posts with flair flags (Table 2) to construct a set of labeled data ($N = 1700$, with 1388 narrative posts and 312 non-narrative posts). We then split the labeled dataset into training, validation, and test sets (0.6/0.2/0.2) while making sure that the ratio of the two classes was similar between the sets. Our best model had a validation accuracy of 0.86 and test accuracy of 0.85 (fine-tuned with 2 epochs and

a learning rate of $5e-5$). Since this task's purpose is only to filter out non-narrative posts to have a smaller subset of posts to gain social insights from and it is not our primary task, we decided that the performance suffices and did not do further hyperparameter tuning.

Using the fine-tuned classifier, we are able to filter out to 7676 narrative-only posts for further analyses as described in the next sections. In this set of narrative posts, the minimum number of words is 100, the median is 182, and the max number of words is $2,823$.

Class	Flair Flags Used
Narrative	"story", "progress", "recovery", "support", "rant", "vent", "success"
Non-Narrative	"announcement", "educational", "research request", "resources", "advertise", "link", "information", "discussion"

Table 2: Flair Flags Used for Supervised Narrative Classification Task

5.2 Domain-Specific Sentiment Lexicons

We used the SentProp algorithm from the SocialSent (Hamilton et al., 2016) code package which takes unlabeled text data and induces domain-specific sentiment lexicons from it. We decided to use this method because we wanted to evaluate how the sentiment of individual words and submissions differed across subreddits depending on the eating disorder associated with a subreddit and the purpose a subreddit serves in the Reddit eating disorder community.

First, we grouped the preprocessed submission text by the subreddit each submission belonged to and obtained the word types per subreddit. Then for each of these groups, we trained word embeddings (128 dimensions) for each individual subreddit using Word2vec (Mikolov et al., 2013) to be used by the SentProp algorithm along with the word types per subreddit. Other than applying the SentProp [3] algorithm on our own data, we did not make any changes to the algorithm itself. The files in our uploaded code that are associated with this code include word2vec_word_embeddings.ipynb [4] and word_types.ipynb [5]. The results of the SentProp algorithm on the word embeddings and word types are located in the polarities [6] directory (inside the data_collection directory). There are several files, submission_polarity.csv, submission_sentiment_from_token_polarities.ipynb, compare_subreddit_word_polarities.ipynb, and non_2020_polarity_comparisons.ipynb which contain code using the results of the SentProp algorithm on our word embeddings and word types to analyze differences in sentiment of words that hold significance in eating disorder related Reddit submission text across subreddits and time periods.

5.3 Instruction Prompting

We used ChatGPT (OpenAI, 2023) and instruction prompting for trigger and factor extractions. **Trigger** is defined as a feeling or event that makes the person who suffers from EDs feel worse or relapse. A trigger can be classified as either internal (stress or lack of motivation) or external (shaming from others such as family or peer pressure from friends at a party). **Factor** is more generic, defined by anything that either helps or harms the person who suffers from EDs. Harmful factors include not only triggers but also treatments (such as therapy or medical treatments) that turn out to have a negative effect on the writer. We observed that ChatGPT is better at extracting more generic factors (both in terms of extraction rates and accuracy) so we used the factor extraction for the social science analysis. See our Github repository for further details about our prompts [7].

For the trigger extraction task, we devised an experiment to compare the performance of different prompting techniques as follows:

Step 1. Collect a diverse set of posts by sampling 20 posts from each of the 20 topics returned

[3] https://github.com/williamleif/socialsent

[4] https://github.com/nguyentr17/cs685-narrative-analysis/blob/main/data_collection/embeddings/word2vec_word_embeddings.ipynb

[5] https://github.com/nguyentr17/cs685-narrative-analysis/blob/main/data_collection/word_types.ipynb

[6] https://github.com/nguyentr17/cs685-narrative-analysis/tree/main/data_collection/polarities

[7] https://github.com/nguyentr17/cs685-narrative-analysis/blob/main/clinical-extractions/utils.py

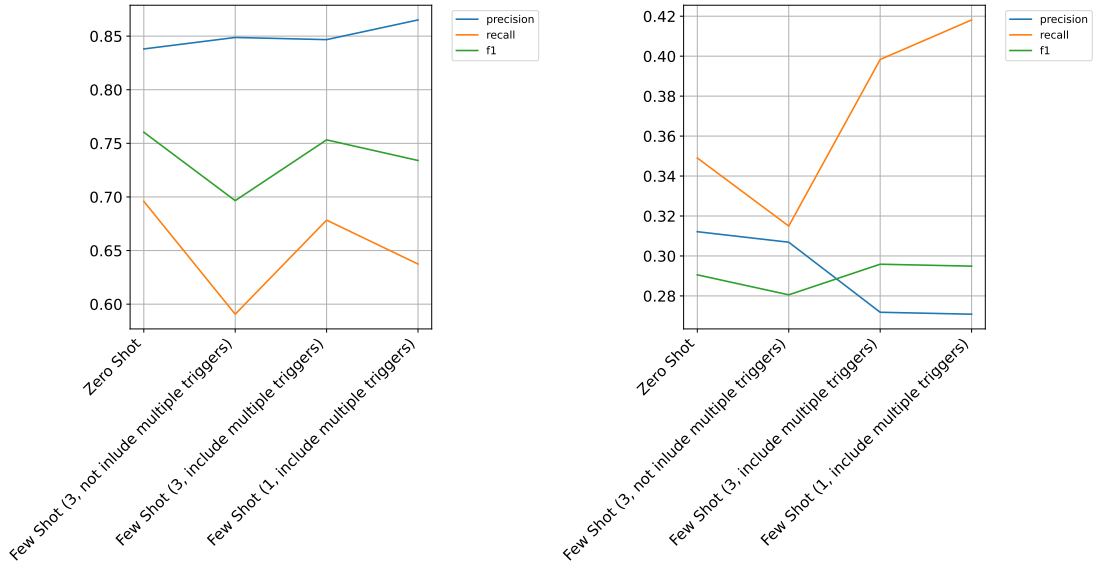

Figure 1: Comparison of Different Prompting Methods on Trigger Extraction Performance (left: trigger extraction rate; right: trigger extraction accuracy

by the **Topic Modeling** section and hand-picking 16 additional posts. We have 216 posts in total for evaluation.

Step 2. Collect ground truth by having members in the group read and extract the triggers and trigger types by the post. [8]

Step 3. Evaluate the following metrics:

1. Extraction rate (number of posts with at least extracted trigger confirmed by a human).

2. Unigram-based precision/recall/F1 score between the LLM-generated trigger and human-annotated trigger [9]

We experimented with the following prompting techniques:

1. Zero-shot: instructions and no example.

2. Few-shot with 3 single-trigger examples: instructions and 3 posts with one trigger each.

3. Few-shot with 3 mixed examples: instructions and 3 posts with both one and multiple triggers.

4. Few-shot with 1 example of a post with multiple triggers (that has both trigger types). We want to see whether 1 information-dense example can perform on par with 3 examples in method (2) and (3) above.

Results. Figure 1 shows the binary trigger detection performance on the trigger extraction task (whether or not the LLM extracts a trigger in a post that does have a trigger) as well as the accuracy of the extracted triggers compared to the ground truth labels. For the latter, we compute the scores based on a set of 87 posts (out of 216 in total) in which all methods and the annotators indicate that it includes at least one trigger. We can see that the few-shot methods not including an example about multiple triggers (method 2) performs the worst in terms of F1 score and Recall for both the binary task and the accuracy of extracted terms.

With respect to the accuracy performance, both few-shot methods that include posts with multiple triggers and trigger types (method 3 and 4) perform better than the baseline (zero-shot) and few-shot method that do not include posts with multiple triggers (method 1 and 2). Including only the post with multiple triggers not only decreases number of processed tokens but also shows a slight improvement in recall. Therefore, we proceed to choose the few-shot with one example posts with multiple triggers as our optimal methods.

Error Analysis. In this section, we take a deeper look at some examples that ChatGPT scores low in terms of accuracy for both the baseline and our optimal method. Table 3 shows a few

[8] Guidelines and annotated data can be found https://docs.google.com/spreadsheets/d/1bDKzmZQT0oDDkXsh-s7jEbL6n1qWAcTpqNkQ1K0vSO4/edit?usp=sharing

[9] Evaluation script from SQUAD https://rajpurkar.github.io/SQuAD-explorer/

Narrative	Ground truth Label	Zero-shot	Few-shot (1, including multiple triggers)
"I have BED and when I start restricting and my face is the first place people will notice. I have have naturally high and prominent cheekbones, which are 'trendy' right now, and **my mom and friends won't stop telling how 'pretty' I look these days.**"	Getting compliments when they start restricting, Feeling hypocritical because they study these problems	['restricting food intake', 'pressure to have trendy features']	['people complimenting her on her appearance', 'articles on "heroin chic" and body ideals']
"I recently started **playing league of legends and boy oh boy, the thinspo is real**, mamma. The girl champions are just perfect thinspo to me so i just keep going over the skins and the K/DA music videos fueling my fasting and my b/p :)"	video game thin character design	['playing league of legends', 'viewing girl champions and K/DA music videos']	['playing league of legends', 'league of legends girl champions and skins', 'K/DA music videos', 'friend comment on the non-realistic aspect of the bodies']
"Gods, it drives me crazy. **Fitness influencers, fine. I don't fucking care**, do your thing. But the ones that promote it as some warped form of recovery? Admit it or not, I know they know what they're doing."	social media posts about body tracking	seeing fitness influencers promote their restrictive lifestyles as recovery	['fitness influencers promoting fitness as a form of recovery']

Table 3: Examples of Posts with low F1 score for all methods

examples in which the F1 score is 0. We see that though the score is low, the response returned by ChatGPT is reasonably right. The score is low for these examples because the choice of words by ChatGPT and human annotators are different. While human annotators generalize into high-level triggers (i.e. league of legends $\rightarrow$ video games, fitness influencers $\rightarrow$ social media), ChatGPT cannot do that high-level generalization yet and instead only extract terms from the post. Since generalized terms are more valuable for social science analysis, we do want the extracted triggers to be more on the general side. However, this insight is valuable because it inspires future directions in this trigger extraction task. We can break down the trigger extraction problem into two separate steps: (1) extract triggers from the post (in which ChatGPT is fairly good at), and (2) generalize the extracted triggers into a curated list of trigger groups.

For this project, we decided to add one more qualitative evaluation by taking a look at the top-10 word analysis between the ground-truth trigger extraction and the few-shot extraction results. Our hypothesis is that if the extraction method works well, we should observe similar conclusion about the triggers provided by ground truth label and by ChatGPT (taking into account similar concepts but different words like "league of legends" and "video games"). Table 4 shows that in the list of top 10 keywords returned by ground truth labels, "people", "bad", "causing", and "writer" do not show up in the list returned by ChatGPT. However, there are some words that might describe a similar concept to those words like "mom" for "people", "change" for "causing", and "fear" for "bad". This type of conclusion-based evaluation (evaluation based on social science conclusions) might add another perspectives to evaluating computation methods for social science purposes.

Ground Truth Labels							
Keywords		Adjectives		Verbs		Nouns	
feeling	10	weight	5	feel	10	food	7
weight	8	bad	5	eat	8	body	6
food	7	anxious	3	cause	8	people	5
feel	7	diet	2	make	5	feeling	4
body	6	unhealthy	2	get	3	writer	4
eating	6	upset	2	compare	3	selfimage	4
causing	5	large	2	go	2	feel	3
people	5	hypocritical	1	stress	2	boyfriend	3
bad	5	unhappy	1	walk	2	weight	3
writer	4	short	1	read	1	post	2
COMBINED FEW SHOT							
Keywords		Adjectives		Verbs		Nouns	
weight	33	weight	8	eat	19	body	25
body	25	recent	5	feel	13	weight	23
feeling	21	mental	5	see	12	fear	18
fear	17	low	5	comment	9	comment	15
mom	15	chic	3	purge	8	mom	15
eating	15	loose	3	look	7	food	14
food	14	slim	3	go	7	change	11
seeing	12	due	3	want	6	gain	11
change	11	old	3	get	5	feeling	9
gain	11	new	3	work	5	post	8

Table 4: Top Terms in Combined Few Shot versus the Ground Truth Labels

5.4 Topic Modeling

For the task of Topic Modeling, we experimented with two different combinations of stop words: nltk package's English stop words and our custom stop words. To construct the latter group, we computed the TF-IDF score for each word and designated those in at least 50% of the corpus to be stop words. However, after running topic modeling on these different combinations, we observed that the topics are most coherent when no stop words were removed. We decided that removing stop words was not helpful as a pre-processing stage and skipped this step altogether. As a final step, we lowercase each word, replace numbers with the string "NUM", as well as remove punctuation and short words (fewer than 2 characters).

To train our topic models, we use the Python wrapper for MALLET package (McCallum, 2002), which implements LDA (Blei et al., 2003b) with Gibbs sampling.

We trained our model with different number of topics: 10, 15, 20, 30, 50. We then examined the resulting topics manually and observed that the model trained with 20 topics returned the most diverse and coherent topics. Table 7 lists the most coherent topics and their corresponding top 5 keywords. Overall, we found these topics to be aligned with our own intuition on the content of ED posts.

5.5 Power and Agency Analysis

Baseline We first ran Antoniak et al.'s pipeline on the 7676 posts classified as narratives. The only major change we made was to define our own personas specific to our dataset. We match personas using regular expressions, which can be found in Table 5. These same personas are used for all the following methods. A notable divergence we made from Antoniak et al.'s personas was to define two that are not people: ED and food. We made this decision after noticing that posters often personified their ED. We also wanted to attempt to measure posters' relationship with their ED and with food.

Antoniak et al.'s method uses spaCy to parse each story, matching noun chunks with the defined personas. It saves a set of (persona, verb) tuples, where the persona was either the subject or direct

Persona	Regex
narrator	I|me
reader	you
SO-male	BF|bf|boyfriend|husband
SO-female	GF|gf|girlfriend|wife
SO-neutral	partner|spouse
SO-any	partner|spouse|bf|boyfriend|husband|GF|gf|girlfriend|wife
ex-SO	ex|ex bf|ex gf|ex boyfriend |ex girlfriend|ex husband|ex wife|ex spouse
friend	friend|friends|bff|best friend|bestfriend
family	family|mom|mum|dad|mother|father|sibling|brother|sister|son|daughter| grandmother|grandfather|grandma|grandpa|grandson|gran
nurse	nurse|nurses|np
doctor	doctor|doctors|dr
therapist	therapist|therapists
medical-prof	nurse|nurses|np|doctor|doctors|dr|therapist|therapists|psychologist|psychiatrist
ED	ED|eating disorder|anorexia|bulimia|binge-eating|binge eating|binge|disorder
food	food|drink|snack|meal|breakfast|lunch|dinner|dessert|coffee|tea|alcohol|wine| beer|cocktail|soda|juice|water|seltzer

Table 5: Regexes used to match our defined personas

object of the verb. Then, these tuples are used to compute a power and agency score for each persona using Sap et al.'s predefined lexicon. For agency, a verb can be labeled either agency_pos (the verb gives the subject agency), agency_neg (the verb takes agency from the subject), or agency_equal (no or neutral effect on agency). For power, a verb can be labeled power_agent (gives power to subject), power_theme (gives power to direct object), power_equal (no effect or equal power). A verb will not have a label for "power" if the verb is intransitive.

Coreference Resolution We hypothesized that the baseline method could miss cases where pronouns are used to refer to a particular persona. To address this, we performed coreference resolution on our data with spaCy package[10].

GPT-Augmented Lexicon Another potential area of improvement is verb coverage. Sap et al.'s original lexicon has 2142 verbs, but when parsing our data with spaCy, we found an additional 2625 verbs in our dataset which are not included in the lexicon. Some of these missing verbs were parsed incorrectly, so to filter out any non-verbs we prompted ChatGPT to label whether each one could be used as a verb. We set p=0 to perform greedy decoding. Our prompt was: "Can the word "[word]" be used as a verb? Give a one word, yes or no answer." After this, we were left with 1358 verbs that were missing from the original lexicon.

We then tested ChatGPT's ability to label the power and agency of verbs. We split Sap et al.'s lexicon into train and test set (0.8/0.2). For each verb in the test set, we constructed 10 demonstrations by finding semantically similar verbs from the train set. Semantic similarity was measured with cosine similarity between the 300-dimensional GloVe embeddings (Pennington et al., 2014). If the verb had no GloVe embedding, we chose 10 random verbs from the train set.

We compared the accuracy of two prompts, which can be found in our source code.[11] Both prompts provide an explanation of the task, definitions of the different labels, and 10 demonstrations. Where they differ is that Prompt 1 gives examples in JSON format, and asks for the completion to also be in JSON format, whereas Prompt 2 puts each example in the context of a sentence. ChatGPT is asked to generate a sentence with the verb before labeling it. For both prompts, we set p=0 and took the majority vote of 3 responses (separately for power and agency). Even though p=0, there were 4 verbs that had different responses for Prompt 1, and 1 verb that had different responses for Prompt 2.

[10] https://spacy.io/universe/project/neuralcoref

[11] https://github.com/nguyentr17/cs685-narrative-analysis/blob/main/power_frames/gpt_augment_lexicon.ipynb

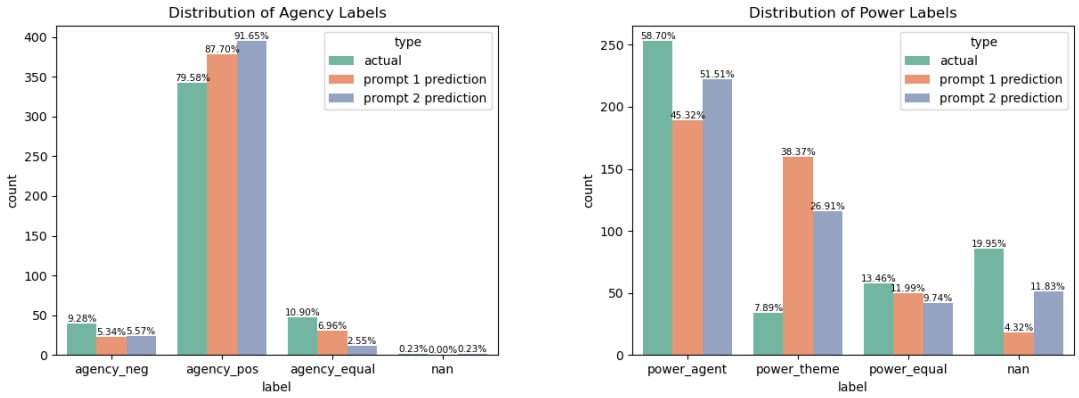

Figure 2: Distribution of actual and predicted power/agency labels in the test set (n=431)

The distributions of actual and predicted labels are shown in Figure 2. Prompt 1 had accuracy 0.79 (340/431) for labeling agency, and 0.45 (193/431) for labeling power. Prompt 2 performed better, with 0.81 (349/431) accuracy for agency and 0.57 (244/431) for power. These accuracies seem low, especially for power, but it is important to consider human performance on this task. Sap et al. reported the pairwise agreement between annotators as 0.51 for agency and 0.56 for power. When a neutral label is counted as agreeing with a positive or negative label, the agreement between annotators increases to 0.94 for agency and 0.96 for power. With these loosened restraints, our Prompt 1 has accuracy 0.94 (405/431) for agency and 0.63 (271/431) for power. Prompt 2 has accuracy 0.94 (403/431) for agency and 0.74 (318/431) for power. Although Prompt 2 has higher accuracy than Prompt 1, it still has a tendency to label verbs as giving power to the object (power_theme) more often than it should. The ground-truth labels of the test set are power_theme only 7.89% of the time, but Prompt 2 makes ChatGPT assign this label 26.91% of the time. This is an improvement over Prompt 1 (38.37%), but still not ideal.

Based on these results, we used Prompt 2 to have ChatGPT label the verbs in our dataset which were missing from Sap et al.'s lexicon. Of the 1358 missing verbs, ChatGPT labeled all but 17, either because they were explicit, had been censored by the original poster and thus were deemed inappropriate by ChatGPT (i.e. "v*mit"), were misspelled (i.e. "sais" which should be "said"), or were not valid verbs. We hand-labeled the 11 which were valid verbs. This brought the size of the labeled missing verbs to 1352. We added these to Sap et al.'s lexicon to form our final, GPT-augmented lexicon, which consists of 3327 verbs.

Comparison of Methods Table 6 shows the number of verbs we were able to match for each persona and method. Over all personas, the addition of coreference resolution resulted in a 33% increase in verb matches over the baseline. Adding the GPT-augmented lexicon resulted in a 6% increase. When both methods were combined, there was a 42% increase in matches over the baseline.

6 Discussion

Polarity Scores for the set of words with high significance in ED-related Reddit posts differ across subreddits. The result of utilizing the SentProp algorithm on word types and embeddings specific to each individual subreddit we collected submissions from was a set of word-score pairings belonging to each subreddit. The initially returned polarity scores were normalized such that there was a mean of zero and unit variance – following the methodology described in the SocialSent paper (Hamilton et al., 2016).

From here, polarity scores of the same set of words across subreddits could be observed. We found that there were a small set of words that had high significance in these ED-related subreddits and these words tended to vary in score value across different subreddits, often matching our expectations for how certain subreddits would be using particular words in their submissions. For instance, as we expected, the word "gain" had a more positive connotation in recovery or support focused subreddits such as r/AnorexiaRecovery and r/edsupport, while it may have a more negative connotation in subreddits dedicated to eating

Persona	Baseline	Coref	GPT	Coref+GPT	% Increase (Coref)	% Increase (GPT)	% Increase (Coref+GPT)
narrator	17863	19775	19107	21083	10.70	6.96	18.03
reader	5149	5392	5527	5778	4.72	7.34	12.22
SO_male	435	1309	450	1374	200.92	3.45	215.86
SO_female	108	638	111	678	490.74	2.78	527.78
SO_neutral	163	490	174	522	200.61	6.75	220.25
SO_any	706	2437	735	2574	245.18	4.11	264.59
ex_SO	1429	1801	1507	1918	26.03	5.46	34.22
friend	1022	3202	1069	3394	213.31	4.60	232.09
family	2353	5301	2502	5657	125.29	6.33	140.42
nurse	104	134	124	157	28.85	19.23	50.96
doctor	866	1169	898	1229	34.99	3.70	41.92
therapist	441	678	453	700	53.74	2.72	58.73
medical_prof	1499	2126	1568	2237	41.83	4.60	49.23
ED	2000	2298	2094	2426	14.90	4.70	21.30
food	5370	5925	5645	6264	10.34	5.12	16.65
TOTAL	39508	52675	41964	55991	33.33	6.22	41.72

Table 6: Number of **matched verbs** per persona and **percent increase** from baseline when applying each method

Top 5 Keywords	Topic Labels	Popular end topic	Polarity score
purging binge-purge binging stop	Eating Behavior	Eating Behavior	-0.83
xNUMb hair would teeth dentist	Body Parts	Feeling	-0.62
mom family fat dad sister	Family	Feeling	-1.25
like feel even want know	Feeling	Feeling	-1.28
food mad buy money cookies	Grocery Shopping	Time	-1.24
eat eating food feel like	Eating	Eating	-0.83
people post group recovery made	Community	Time	-0.57
weight gain lose gained eating	Weight	Weight	-1.24
body look weight see clothes	Appearance	Feeling	-1.28
foods ate NUM eat food	Food	Food	-0.94
treatment inpatient hospital therapist NUM	Formal Treatment	Work	-1.37
NUM years since time year	Time	Feeling	-1.31
work get want time going	Work	Work	-1.61

Table 7: Most coherent topics for our sample of narrative posts by long-standing users. Labels in 'Topic Labels' are manually annotated by one of the authors. Values in 'Polarity Score' are rounded to 2 decimal places.

disorders categorized by different behavior, such as r/BingeEatingDisorder and r/bulimia.

Figure 3 shows the polarity score differences on the same set of words between the r/AnorexiaRecovery and r/BingeEatingDisorder subreddits. These differences are calculated by subtracting the polarity score associated with r/BingeEatingDisorder from that of r/AnorexiaRecovery to demonstrate how words have different connotations across subreddits. Negative differences indicate that the polarity score was more negative in r/AnorexiaRecovery than in r/BingeEatingDisorder. Positive differences indicate that the polarity score was more positive in r/AnorexiaRecovery than in r/BingeEatingDisorder.

Figure 4 demonstrates the differences in polarity scores on the same set of words between the r/AnorexiaRecovery and r/EdAnonymousAdults subreddits. Similar to Figure 3, these differences are calculated by subtracting the polarity score associated with r/EdAnonymouAdults from that of r/AnorexiaRecovery to demonstrate how words have different connotations across subreddits. Negative differences indicate that the polarity score was more negative in r/AnorexiaRecovery than in r/EdAnonymousAdults. Positive differences indicate that the polarity score was more positive in r/AnorexiaRecovery than in r/EdAnonymousAdults.

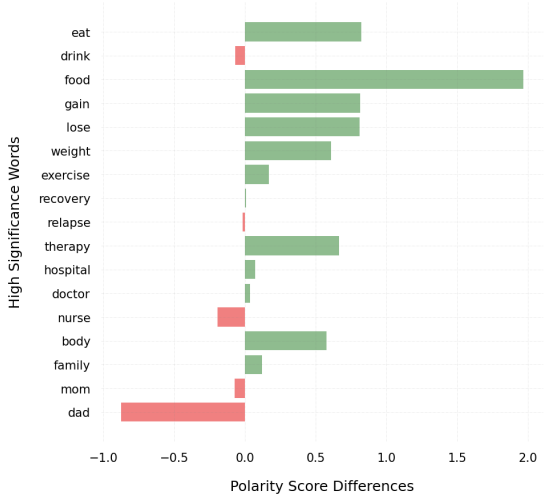

Figure 3: r/AnorexiaRecovery polarity score - r/BingeEatingDisorder polarity score

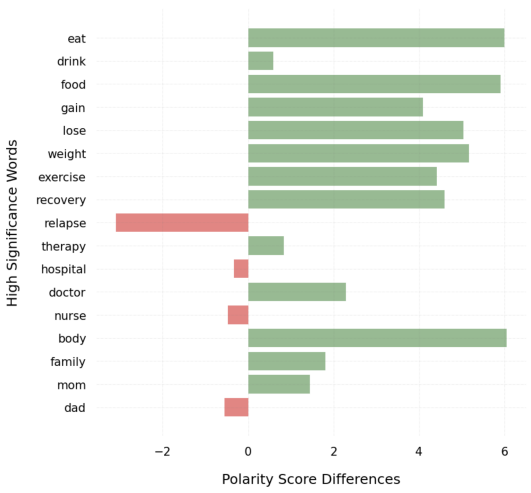

Figure 4: r/AnorexiaRecovery polarity score - r/EdAnonymousAdults polarity score

Users tend to follow up on their previous posts with a post about "Feeling". Using our inferred topics, we examined possible topical changes in posts written by the same user across time. We looked to Antoniak et al. 2019 and Akoury et al. 2020 for inspiration. We found that Antoniak et al. 2019's approach of dividing each post into 10 chunks and treating each chunk as an input document for topic modeling not applicable for two reasons. First, the average length of our posts (≈220 words) were much shorter compared to those of Antoniak et al. 2019 (≈1,311 words), making the input documents too short to contain interesting topics. Second, if we instead only look at users with a specific number of posts and treat each post as a input document, our sample size will be too small. Specifically, there are 541 users with exactly 2 posts, 164 with 3 posts, and fewer than 100 with 4 or more posts. We decided that such small sample sizes might not return interesting topics for our analyses.

For each user, we treated each of their posts as a document and examined the local topic transition between these entries written by the same user across time following Akoury et al. 2020. We then looked at the most popular topic transitions across all users. We observed that most of these transitions ended in the topic of Feeling. For example, users tend to discuss Eating, Appearance and Body Parts before transitioning to Feeling in their subsequent posts. ED recovery and discussion can be emotionally charged experiences, so it makes sense that people dedicate posts to talk about their feelings. Some other interesting transitions include Grocery Shopping to Time and Community to Time. Additional topic transitions can be found in Table 7.

All individual topics are associated with negative emotions. Using our sentiment lexicon (6.1.1) and inferred topics, we examined the dominant sentiments for each topic. Overall, the sentiments were negative across all topics (Table 7). Posts with Weight and Feeling receive the lowest polarity scores, which match our expectations that these topics aroused negative sentiments among people recovering from EDs.

Local topic transitions that involve the topic of "Formal Treatment" are more likely to be positive. We also examined the sentiments associated with each local topic transition. For post A and subsequent post B by the same user, we obtained the local topic transition between A and B as well as their difference in polarity score which were the average polarity score over words in the posts. We then tried to find the most popular transition-polarity association across all users. Among the most positive transitions, we found the transitions between Discomfort and Formal Treatment, Recovery and Formal Treatment as well as Weight and Community. These transitions indicated that formal treatments (keywords: treatment impatient hospital therapist NUM) might have a positive impact on users' sentiments. On the other hand, Grocery Shopping - Recovery and Food - Feeling transitions are among the most negative transitions, suggesting that food may have a bad effect on users' feelings.

Users tended to predominantly post if they

HELPFUL TABLE							
Keywords		Adjectives		Verbs		Nouns	
recovery	82	conscious	26	struggle	36	recovery	82
support	48	negative	26	work	31	support	47
purging	42	positive	25	stop	30	meal	30
struggling	36	supportive	20	purge	27	therapist	28
binge	36	healthy	16	get	25	calorie	28
fear	35	new	16	go	22	fear	28
therapist	33	mental	10	allow	19	effort	27
weight	31	emotional	10	binge	19	binge	26
calorie	30	good	9	find	19	weight	24
meal	30	normal	9	recover	18	purging	20
HARMFUL TABLE							
Keywords		Adjectives		Verbs		Nouns	
weight	1181	negative	698	struggle	632	weight	773
negative	698	weight	356	restrict	346	fear	442
struggling	615	low	117	purge	304	calorie	402
fear	501	physical	113	lose	225	comment	397
calorie	421	restrictive	99	trigger	213	image	328
purging	412	obsessive	91	experience	191	family	317
comment	397	guilty	83	gain	165	loss	295
restricting	389	social	77	use	142	lack	293
image	328	excessive	77	binge	109	gain	250
family	317	high	72	work	97	recovery	237

Table 8: Top Terms in Helpful and Harmful Context.

were going through some negative rather than if they were to post about something positive. We can see this when we look at the factors found by the instruction prompting and by grouping by users (removing any anonymized posts first and then anonymizing the grouped users to respect their privacy). Out of 2840 extracted factors, 1769 were marked as harmful, 225 were marked as helpful, and the rest were either unknown, neutral, or had to differences in their factor extraction which. Figure 5 shows the number of times a post has a harmful factor versus a helpful factor and the majority of the time users are posting, they're posting about things that harmed them rather than things that helped them.

The top words in the extracted helpful factors underscore the importance of creating safe spaces where people are encouraged to make positive changes as the top words include "recovery" and "support". However, there are many other words that contrast with this such as "purging", and "struggling" which reflect the difficulties of an eating disorder and how those in recovery still face an uphill battle. The words that appear in the helpful column are often associated with positive actions, supportive relationships, and the process of recovery from an eating disorder. In the helpful context, words like "conscious," "positive," "supportive," "healthy," "allow," and "find" stand out as they suggest a more proactive and constructive approach to dealing with eating disorders. Some factors include *"Supportive partner who stayed with the writer through their eating disorder"*, and *"Making a conscious effort to eat breakfast and improve nutrition"* which both make use of the "support" and "conscious" keywords. Words such as "therapist" emphasize that the writer is actively trying to work through their recovery, and we can see that with the factor: *"Working with a dietician, talking to different therapists"* which shows that the user was being proactive in addressing their issues instead of hiding it. "Fear" was an interesting top keyword as it had a different connotation than we thought, some users have a fear of certain foods because of negative thoughts about nutritional content, thus we see factors such as *"Eating a fear food and overcoming negative thoughts"* where users are

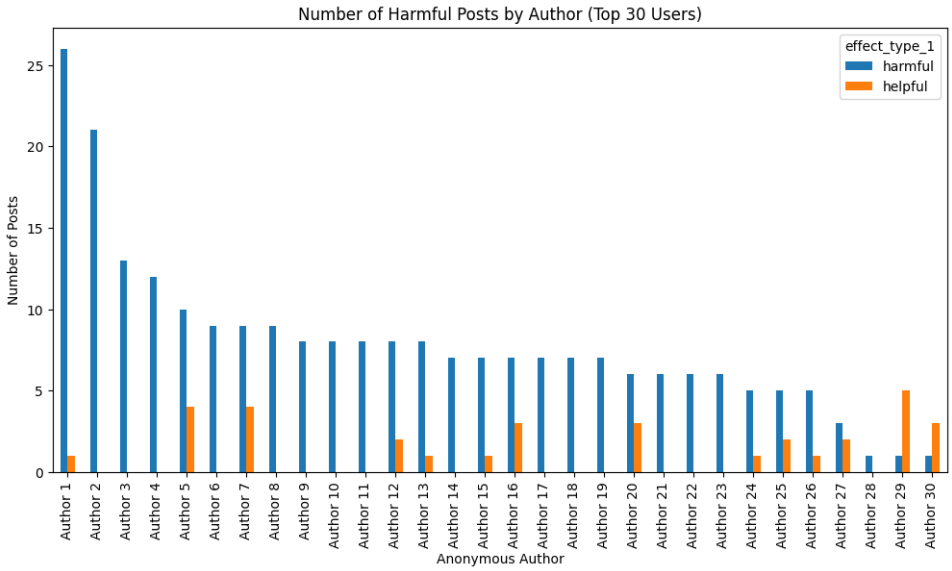

Figure 5: Overall trend of posts with Harmful vs Helpful factors

again actively trying to overcome their negative mindsets.

The top keywords from the harmful factor include "weight", "negative", "struggling", and "fear", signifying the common issues faced by these individuals. We can see that these are also very common words that we also see in the helpful table, however with less frequency and magnitude. There is overlap in the verbs used in both contexts, including "struggle," "work," "purge," "binge," and "recover." These verbs and top words reflect that these are actions that are part of the journey of dealing with an eating disorder, whether in a harmful or helpful context. Nouns such as "weight", "fear", and "calorie" make sense, as these are all food focused and highlight the common struggles of an eating disorder. Nouns such as include "family" and "comment" are more specific and when we look at some of the factors that include these words: *"Negative comments and social pressure exacerbating fears"* and *"Lack of support from family and friends during the recovery process"*, we see that this highlights the lack of support and the effect other people can have on people. The word "weight" appears 1181 times in the harmful keywords category, whereas in the helpful keywords, the most frequent word "recovery" appears 82 times. This reflects a focus on body image and weight, which is a known trigger for people with eating disorders. This also might suggest that harmful contexts are often more fixated on their weight while those that are posting something that has a helpful factor are focused more on their recovery. This makes sense given that the top word is "recovery" in factors. Stronger words such as "restrict," "obsessive," "guilty," and "trigger" in the harmful context stand out for their strong negative connotations, reflecting the potential for harm in discussions around these topics. Overall, these harmful factors give an insight into the significance of the negative influence that societal norms, weight obsession, and lack of adequate support can have on individuals suffering from eating disorders. Further analysis is needed to find more of these trends.

The narrator consistently has low power compared to the other personas. We report results after applying the combined **Coreference + GPT-Augmented Lexicon** method. Figure 6 shows the average agency and power scores for each persona. Notably, the narrator has the least agency among all the personas, even less than the inanimate personas food and ED. ED has the highest agency. The narrator also has the second-lowest power of the people personas, with ex_SO (ex significant other) having the lowest. Medical professionals have the second-highest agency, but less power than the other people personas (excluding the narrator). Significant others have the most power of all the personas except for the reader. However, the reader's power is likely high only because of how posters address questions to the reader or use the pronoun "you" in an abstract sense.

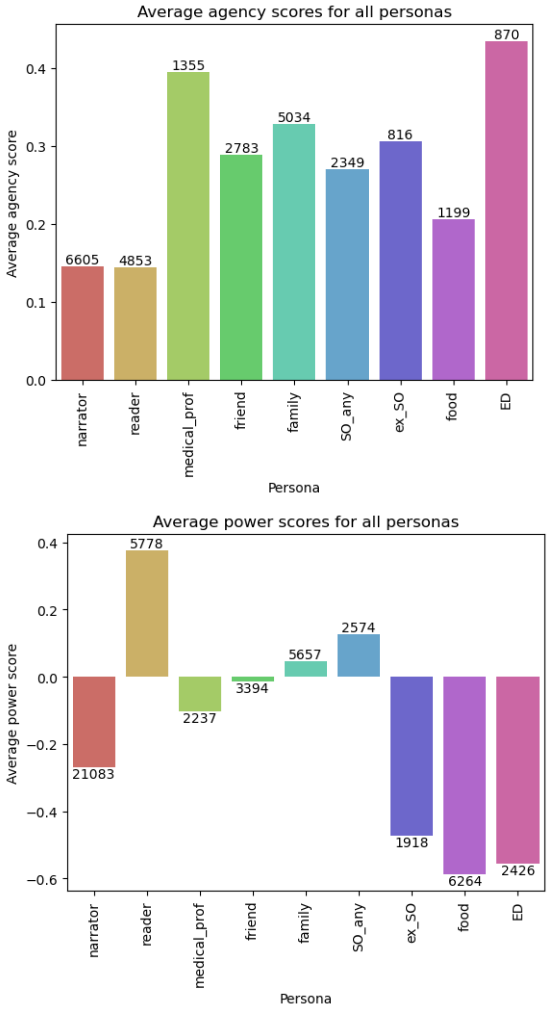

Figure 6: Average **agency and power scores** for the main personas. Note that less verbs are used in the computation of agency, since the persona must be the subject of the verb. For power, the persona can be the subject or object.

We also computed power and agency scores for each subreddit. Figure 7 shows these results for the narrator, medical professional, significant other, and ED personas. The narrator has the highest power in r/anorexiaflareuphelp and r/BulimiaRecovery. Medical professionals have high power in r/EDRecovery_public. Significant others have high agency but low power in r/EatingDisorderHope, and high power in r/edsupport. Finally, ED has low agency in r/EDRecovery_public, and the highest power in r/BingeEatingDisorder and r/edsupport.

7 Error analysis

We did an error analysis for the trigger extraction task. Details and examples can be found in Section 5.3. We include it in the Approach section because we have multiple tasks and the error analysis is specific to the trigger extraction task.

8 Contributions of group members

- Anthony Campbell: factor extraction using ChatGPT, prompting engineering
- Jeng-Yu Chou: data collection / processing, domain-specific analysis
- Marisa Hudspeth: power and agency of personas
- Trang Nguyen: narrative detection, trigger extraction
- Chau Pham: topic modeling and topic-related analyses

Each member of the group contributed equally to the project and the write-up of the final report. Each member of the group did annotations for the error analysis of the factor and trigger extraction test.

9 Conclusion

The project shows that there is an added value in using pre-trained LLMs and NLP methods in support of social science analysis. The insights we can get from the sentiment analysis, topic modeling, and power analysis seem to be aligned with the general views about eating disorders.

Below we will outline limitations and future directions for each of the subtasks.

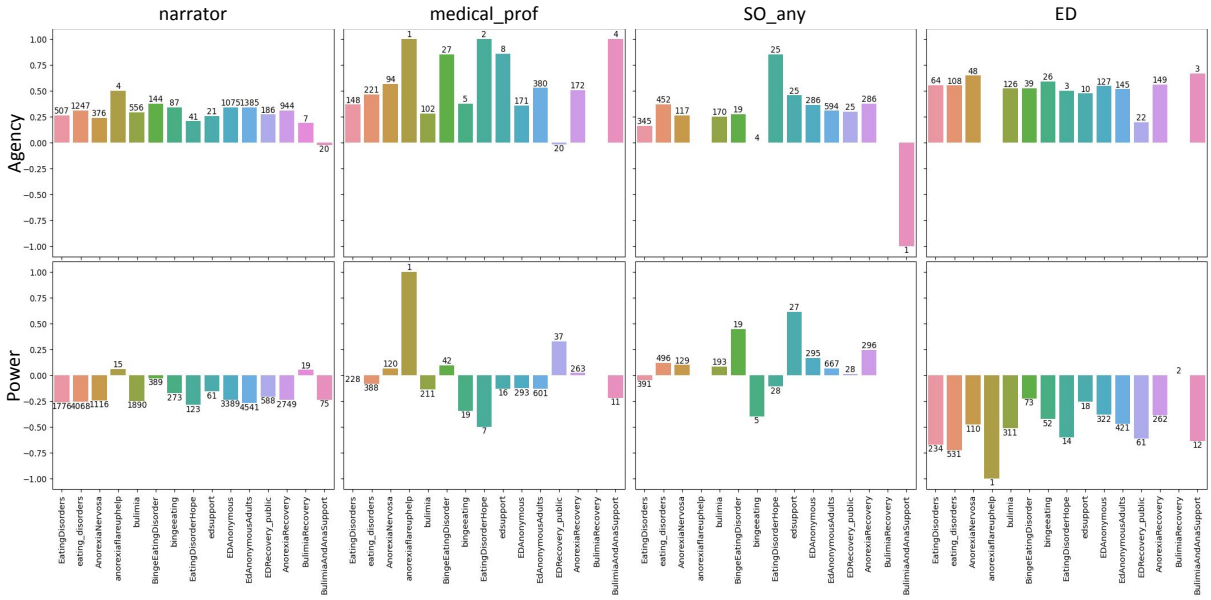

Figure 7: Power and Agency scores per subreddit for selected personas. Subreddits on the left are more general, and on the right are recovery/support-oriented.

Information Extraction One main limitation we ran into while using ChatGPT for clinical trigger and factor extraction is that the extracted factors are more specific than we hope they would be. This makes it difficult to compare our analysis against clinical trial analysis in which factors are usually higher-level (i.e. therapy, medication, family support). We did not review each post by more than one person but would have liked to as well or the human evaluation part of reviewing. We also would have liked to experiment with the temperature and the number of outputs as well, because we believe this may have resulted in better extraction outcomes. We also would have liked to train a BERT model to learn from our human evaluations to determine if a post has triggers or not and to identify the triggers as a "ground truth".

Data Collection Another added difficulty was the limitations of the resources used to gather the initial textual data. With the loss of a large portion of Reddit data using the PushshiftAPI, we were unable to make the comparisons across subreddit users' eating disorder experiences using pre-2020, post-2020, and peak COVID-19 periods. Also, not all subreddits contained the same high significance keywords related to eating disorders which restricted our ability to make comparisons between sentiments corresponding to a full list of keywords that we expected there to be differences and similarities with across all of the subreddits we collected posts from depending on those subreddit's purpose (e.g. venting versus support) and specific eating disorder.

Power Analysis Task The main limitations of the power and agency analysis are its subjectivity and inability to account for context. Although we were able to expand the verb coverage, this method still relies on a static lexicon of words. Similarly to how we found that sentiment could be domain-specific, it is likely that power and agency are, too. For example, consider the verb "gain." In Sap et al.'s lexicon, it is labeled as agency_pos and power_agent. This makes sense in a general context, but within the eating disorder community, "gain" is often associated with "gaining weight," which may be out of the subject's control. It may be possible to obtain more accurate power and agency scores by having an LM such as fine-tuned BERT or ChatGPT label verbs in context. We decided not to pursue these options; it would be difficult to fine-tune BERT because Sap et al.'s verbs are standalone, not in the context of sentences. Having ChatGPT label verbs in context is possible, but would require more time and money - this could be a future direction for research.

10 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - ChatGPT

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Prompts are included in the table and github code.
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - It was overall very helpful. The tables for data pre-processing and topics were generated by ChatGPT.

References

Akoury, N., Wang, S., Whiting, J., Hood, S., Peng, N., and Iyyer, M. (2020). Storium: A dataset and evaluation platform for machine-in-the-loop story generation.

Antoniak, M., Mimno, D., and Levy, K. (2019). Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Blei, D., Ng, A., and Jordan, M. (2003a). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003b). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Donati, M. and Strapparava, C. (2022). Coreds: a corpus on eating disorders. In *Proceedings of the RaPID Workshop-Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments-within the 13th Language Resources and Evaluation Conference*, pages 80–85.

Ganti, A., Wilson, S., Ma, Z., Zhao, X., and Ma, R. (2022). Narrative detection and feature analysis in online health communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65. Association for Computational Linguistics.

Gao, J., Zhao, H., Yu, C., and Xu, R. (2023). Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Giorgi, S., Himelein-Wachowiak, M., Habib, D., Ungar, L., and Curtis, B. (2022). Nonsuicidal self-injury and substance use disorders: A shared language of addiction. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 177–183.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Murray, C., Mitchell, L., Tuke, J., and Mackay, M. (2020). Symptom extraction from the narratives of personal experiences with covid-19 on reddit.

OpenAI (2023). ChatGPT 3.5. https://openai.com. Version 3.5.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., and Choi, Y. (2017). Connotation frames of power and agency in modern films. In *Conference on Empirical Methods in Natural Language Processing*.

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al. (2023). Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Yang, K., Ji, S., Zhang, T., Xie, Q., and Ananiadou, S. (2023). On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). Can large language models transform computational social science?