Transfer Learning for Eating Disorder Sentiment Analysis

Stanford CS224N Custom Project
TA mentor: Gaurab Banerjee

Christine Manegan
Department of Computer Science
Stanford University
manegan@stanford.edu

Rhett Owen
Department of Computer Science
Stanford University
rcowen@stanford.edu

Abstract

Our project aims to expand on the very small pool of existing research on the classification of eating disorder language use on social media using natural language processing. In this paper, we use several different methods of semi-supervised learning to create reasonably sized datasets with predicted labels for whether a post contains harmful eating disorder rhetoric. We scraped over 17,000 Reddit posts, 200 of which were hand-labeled to use as test data, and used the generated weakly-supervised training data to fine-tune three different transformer-based models: BERT, RoBERTa, and MentalBERT. Our results evaluate which weakly-supervised labeling methods, using learning functions and an SGD optimizer, generate the most effective training labels for our training data to use during fine-tuning. We aim to show how deep learning approaches compare to one another and improve upon previous papers' approaches that use logistic regression, word movers' distance, and simple semi-supervised learning [1] on eating disorder text classification tasks.

1 Introduction

Modern social media platforms enable users to post to audiences entirely separate from their in-person social graphs using any chosen username, creating the opportunity to share more deeply personal thoughts online than through traditional communication with peers, friends, or family members. This semi-anonymous nature of online social media posts provides a way to identify users who are in need of immediate mental health intervention for eating disorders such as anorexia, bulimia, and binge eating disorder, whose struggles may have gone unnoticed otherwise. Eating disorder forums and posts that involve "pro anorexia" and "pro bulimia" trends, the glorification of thinness, and inspiration for disordered eating behaviors are prevalent on many social media platforms, and this ego-syntonic nature of eating disorders makes social media posts particularly valuable for identifying those in need of professional help. Existing models that use deep learning for mental health classification, such as MentalBERT [2] are too broad to identify eating disorders specifically, and have high rates of false negatives on eating disorder tasks. However, there is extremely limited research dedicated to classification on eating disorders specifically, and existing research is based on extremely small amounts of training data and do not employ deep learning methods. Given the existing research in this field, our two goals were to 1) Fine-tune existing pre-trained language models to apply to our eating disorder classification task, and 2) Use semi-supervised learning to generate high quality, noise-aware training labels using psychology-based heuristics. Our research aims to expand on existing mental health classification tasks on social media posts by extending coverage to eating disorder-specific content and analyzing which combinations of label-generation learning functions and pre-trained models most accurately achieve this task.

2 Related Work

Despite the large amount of social media activity related to eating disorders, there is a surprising lack of published work relating to the classification of pro-anorexia content. The primary paper that we used as a reference was “Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention,” by Yan, Fitzsimmons-Craft, Goodman, Craft, Das, and Cavagos-Rehg [1].

Many of the papers that do explore this topic explore the characteristics of different types of posts relating to eating disorders, but do so via human review instead of training a classification model [3]. One possible reason for the lack of NLP research in the area of eating disorders is the difficulty in obtaining large amounts of training data. In this paper, they solved that problem by having two clinical psychologists hand-label a random selection of reddit posts from various subreddits related to eating disorders, and then used a large amount of unlabeled data to carry out positive and unlabeled (PU) learning. They labeled posts as positive if they indicated that the user is “in need of immediate intervention” and negative otherwise. In the span of three hours, they were able to label 53 posts, 38 that were positive and 15 that were negative. Hand-labeling potentially harmful posts related to eating disorders is a resource-intensive task – it requires qualified human laborers, guidelines for what to label as positive and negative, and hours of human labor. For these reasons, it is a major contribution that this paper was able to construct models from only a small set of labeled data.

The first major limitation of this paper was that none of the models they used took advantage of any form of deep learning. While they were able to construct models that achieved relatively low error rates with just linear regression, performance may be improved by looking at models that leverage recurrent neural networks to do classification of posts. Along these same lines, another limitation of this paper was the lack of useful evaluation metrics. It is mentioned in the discussion section that for social media users who may be in need of immediate help, false negatives (missing the opportunity to extend help to someone who needs it) can be more of a problem than false positives. Since this is the case, error rate alone may not be the best way to judge models – something like recall may be better.

There are three main ways in which we expanded upon the work in this paper. The first was that we used a much larger set of labeled training data by using a weakly-supervised learning method trained using the Snorkel API [4], which let us train on thousands of examples rather than just the few dozen used in the paper by Yan et al [1]. We also expanded upon the models used, and introduced neural networks in the form of a variety of transformer-based models such as BERT, MentalBERT, and RoBERTa. The final way that we expanded upon the paper was by changing the scope of classification – not only did we attempt to classify posts that indicated need for immediate medical attention, we attempted to classify any post that promotes unhealthy views of eating disorders.

3 Approach

Due to the high cost of obtaining sufficient hand-labeled data for classification training, we employed weakly-supervised methods of generating labels for fine-tuning existing language models to apply to our eating disorder classification task. To generate labels, we used learning functions based on heuristics from existing eating disorder psychological research as well as experimental learning functions written after evaluating random samples of training data.

Prior research [5] has shown that pro-eating disorder bloggers are generally less expressive, social, and emotional than their pro-recovery counterparts, indicating that machine learning techniques may be used to identify those in need of mental health intervention and can help differentiate those who are particularly unwilling to overcome their eating disorders from those in recovery, so we used prior emotion analysis models such as python text-to-emotion and TextBlob sentiment polarity as part of our learning functions.

In addition to emotion analysis, we wrote learning functions to analyze text length, eating disorder keywords, non-ED keywords, reddit subreddit name, and pronouns in the posts.

The learning functions that we created are as follows:

1.) **contains_EDkeywords:** return positive if a post contains any of 21 disordered eating keywords, return abstain otherwise
2.) **contains_nonEDkeywords:** return negative if a post vontains any of 3 dinstinctly non-disordered eating keywords, return abstain otherwise
3.) **lf_textblob_polarity:** using the TextBlob pre-trained sentiment classifier [**?**], return negative if the text sentiment polarity is greater than 0.3, and positive otherwise.
4.) **emotion:** using Python text-to-emotion sentiment classifier, return negative or positive given detected proportion "Sad", "Happy", "Surprise", "Sad", "Angry" and "Fear" emotions.
5.) **prelabel:** return negative if the post is in a non-eating-disorder subreddit, and positive otherwise.
6.) **length:** return negative if the post is shorter than 5 words, and abstain otherwise
7.) **containsPerson:** return negative if the post does not use pronouns, abstain otherwise
8.) **thirdPerson:** return negative if there are pronouns and they are mostly in third-person, and abstain otherwise.

We tested different combinations of these learning functions to generate multiple sets of labels for training using Snorkel AI's SGD optimizer to train our label model and generate labels for training. Then, we used these labels to fine tune pre-trained models such as BERT, mental BERT, and mental roBERTa to assess which combination of pre-trained model and training labels would provide the most accurate classification results according to our test set of 200 hand-labeled posts.

4 Experiments

4.1 Data

For our training and validation data, we scraped 17,000 posts from 26 Reddit subreddits, including eating disorder specific subreddits such as r/eating_disorders and r/ED_anonymous, as well as self-love and anti-ED subreddits such as r/self_love and r/body_positivity using the PRAW Python wrapper [6] for the Reddit API. We cleaned the data to remove Nonetype/NaN posts, emojis, hyperlinks, and non-alphabetic chars and posts to ultimately have 14,394 cleaned posts that we partitioned into training, validation, and test sets.

Our test set was comprised of 200 posts that we hand-labeled using recommendations from Stanford Eating Disorder psychologist Eric Stice, searching for words describing eating disordered behaviors (e.g., binge, binge eating, vomiting, laxative, diuretic, fasting), as well as language that captures social comparisons with body image and eating disturbances.

4.2 Evaluation method

For the quantitative analysis of the performance of our models, we reported three scores: accuracy, recall, and F1 score. Accuracy was the primary statistic reported by the paper by Yan et al [1], so by including accuracy scores we were able to measure our performance against other papers in the field. However, in the context of detecting harmful eating disorder language, it may be the case that false negatives have a larger impact than false positives. This was the reasoning behind the inclusion of recall, and F1 was included to compare model performance based on a combination of both accuracy and recall.

Originally, we gathered these statistics by splitting our datasets into training, validation, and test sets, then reported the accuracy, recall, and F1 scores of the models on the test set. After scraping and cleaning, we were left with a dataset of 12,000 Reddit posts that we split into 10,000 posts for training, 1000 for validation, and 1000 for test data. However, this approach relies on using non-gold-standard data for testing, which might not give a good picture of how our models would do in a real-world setting. To circumvent this problem, we hand-labeled 200 posts (with 88 positives and 112 negatives) to use as a test set. This guaranteed that our model was being tested on corrrectly-labeled data.

For qualitative analysis, we manually reviewed all of the examples from our hand-labeled test set that the models predicted incorrectly. From this, we were able to identify many different patterns that we were able to incorporate into new learning functions for labeling using Snorkel, which we then used to create training data for a new round of fine-tuning and analysis. This process was repeated for a total of 5 different rounds.

4.3 Experimental details

We used the Huggingface transformers framework to fine tune 3 different models: BERT Base (uncased), MentalBERT Base (uncased), and RoBERTa Base [7]. All models were loaded with Huggingface's auto class for sequence classification, and we decided that BERT base would give us a model that could be easily adapted to learn the sentiment behind posts related to eating disorder language. MentalBERT is a version of BERT base that was additionally fine-tuned on a corpus of over 13 million sentences from a variety of subreddits relating to mental health, which used a text encoder learning rate of 1e-05 and a classification layer learning rate of 3e-05 [2]. MentalBERT was included because we believed that the overlap between mental health language and eating disorder language would be substantial. Lastly, RoBERTa was included to give us a model that might generalize better to real-world data – we believed that the dynamic masking of RoBERTa might help to avoid overfitting on our training data and lead to better performance on our hand-labeled test data.

For our training data, we used the functions outlined in section 3 create a label model, which contains the resulting label from each of the above learning functions for each post in our training set. Using the snorkel API's .fit() function [8] we ran an SGD optimizer with 500 epochs and a learning rate of 0.01 to train the label model to ultimately produce a single set of noise-aware training labels. These labels were ultimately used for finetune our pre-trained models.

Learning Functions	original data	snorkel 1	snorkel 2	snorkel 3	snorkel 4	snorkel 5
prelabel	✓					✓
contains_EDkeywords		✓	✓	✓	✓	
contains_nonEDkeywords		✓	✓	✓	✓	
If_textblob_polarity		✓	✓	✓	✓	✓
containsPerson				✓	✓	✓
thirdPerson				✓	✓	✓
length					✓	✓
emotion					✓	✓

Figure 1: Training datasets using weakly-supervised label learning

Note: snorkel 1 and 2 differ in that snorkel 2 had hand-labeled test data removed from the training set, as follows for snorkel 3-5

Since we experimented with so many different sets of training data, we spent a large amount of time fine tuning – we used 3 different models with approximately 7 different rounds of training data. However, we did not fine-tune all 3 models on every dataset. Instead, we started by using MentalBERT and BERT base, but after witnessing poor performances from MentalBERT, we shifted to focusing on BERT base and RoBERTa base. For each session, we used the Huggingface trainer class with a learning rate of 2e-5, with a total of 2397 optimization steps, which took approximately 20 minutes to run on the Azure virtual machine.

4.4 Results

As described in section 4.2, we originally evaluated our data by splitting our set of reddit posts into training, validation, and test sets, then calculated metrics on the test data. The following table displays the results from that portion of fine-tuning. Original data refers to the naively bucketed data that was labeled by subreddit, and the numbered Snorkel data entries are the different iterations of the data labeled using Snorkel.

	Accuracy	Recall	F1
BERT base, original data	0.879	0.918	0.902
BERT base, Snorkel data 1	0.872	0.895	0.893
BERT base, Snorkel data 2	**0.893**	0.949	**0.915**
MentalBERT, original data	0.886	0.923	0.907
MentalBERT, Snorkel data 1	0.884	0.903	0.903
MentalBERT, Snorkel data 2	0.888	**0.972**	0.913

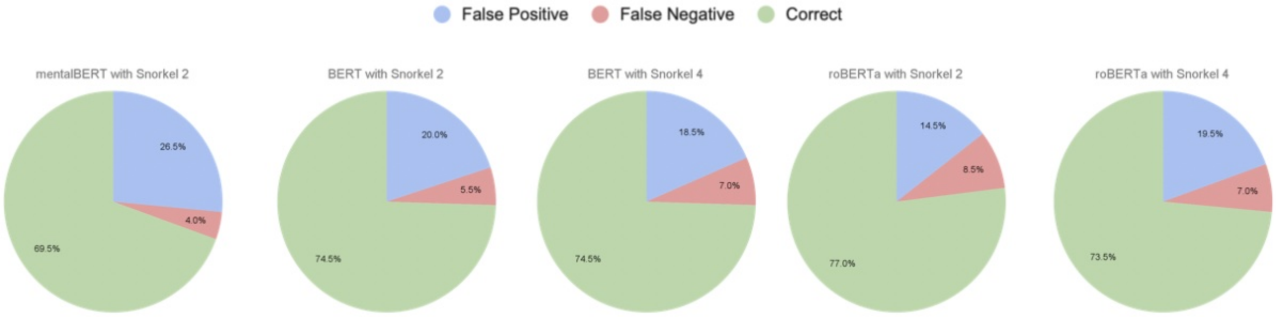
Figure 2: Test Results for Train/val/test Split

While we were able to adjust our labeled training data in order to improve performance on a randomly selected test set, we did not feel that this accurately captured the goals that we set out to achieve with this project. To shift our focus to measuring performance on human-labeled test data, as described in section 4.2, we removed 200 posts from our training data and labeled them by hand. The following table contains the resulting scores. Note: the first pass of data labeled using snorkel was excluded from this table since it contained the examples from the hand-labeled test set. All other rounds of training data for this table had those examples removed.

	Accuracy	Recall	F1
MentalBERT, original data	0.731	0.864	0.738
MentalBERT, Snorkel data 2	0.697	**0.909**	0.724
BERT base, original data	0.751	0.852	0.75
BERT base, Snorkel data 2	0.746	0.875	0.751
BERT base, Snorkel data 3	0.746	0.841	0.744
BERT base, Snorkel data 4	0.756	0.852	0.754
BERT base, Snorkel data 5	0.731	0.818	0.727
RoBERTa base, Snorkel data 2	**0.771**	0.807	**0.755**
RoBERTa base, Snorkel data 3	0.736	0.807	0.739
RoBERTa base, Snorkel data 4	0.736	0.841	0.736
RoBERTa base, Snorkel data 5	0.751	0.841	0.748

Figure 3: Test Results for Hand-labeled Data

As expected, the performance across the board is worse on the hand-labeled test set than on the test data from the train/val/test split. The best F1-score achieved in the first table was 0.915, compared to 0.755 in the second table. One interesting observation is that the different iterations of Snorkel labels didn't consistently lead to improvements in performance for all models. One possible reason for this could be that there is a tradeoff in how different labeling functions perform for different models. For example, consider the effects of using a labeling function based on keywords in Snorkel. Since RoBERTa model uses dynamic masking as opposed to the single fixed masking used for BERT and MentalBERT, we may expect this addition to have a larger effect on BERT and MentalBERT than for RoBERTa. This could be a potential reason why the recall of both the BERT and MentalBERT models is higher than that of RoBERTa for the second round of Snorkel data.



False, positive, false negative, and correct distributions for combinations of pre-trained models and generated training labels

Above is a visualization of the distributions of false positive, false negative, and correct predicted labels for the best performing snorkel labels on each pre-trained model. Across all pre-trained models and training labels, outcomes were consistently over-inclusive with high false positive rates and high recall.

5 Analysis

For every round of testing above, over 60% of incorrect classifications were false positives as opposed to false negatives, which is also evident in the recall scores in figures 2 and 3 that significantly outperform their corresponding accuracy scores. This over-inclusivity stems from multiple factors, but the top three types of false positives were posts that contained eating-disorder keywords but that were instead talking about recovery or healthy dieting, general mental health disorder posts about depression and/or anxiety that did not specifically have to do with eating disorders, and posts that spoke about eating disorder resources from an objective stance. Many recovery posts were also coupled with posts about their negative past eating disorder experiences, which further made them more difficult to classify into one category. Examples of these false positives are as follows:

- **General dieting posts:**

  “I understand you have to be in a caloric deficit to lose weight and track your foods. I personally have lost about 11 pounds so far but I’ve eating semi-clean. There are people that say they ate nothing but fast food and kept their deficits and lost tons of pounds."

  “I’ve decided to take it seriously, lose the weight and change the lifestyle before I am older and regret my choices.”

- **Recovery posts with ED keywords:**

  “I guess I am starting to realise that my own satiation and cravings are justification enough to eat”

  "I am NOT spending what life I have left being afraid of SPECIAL FUCKING K!”

  “I ate everything I craved. I ate past fullness a lot and have felt uncomfortably full many times... You will stop eating, you won’t binge forever and you *will* be satisfied, it might just catch you by surprise."

- **ED Resources:**

"I just started a podcast (today!) where I plan on sharing my experience with eating disorders, mental health"

"You can talk about whatever tf you want with regards to whatever illness you have and the only rule is you can't offer diet advice in response to that problem."

"ok I made /r/chronicallyantidiet! Pls join and I'll figure this all out in the morning :)"

False negatives were less common, but typically were more nuanced eating disorder posts that did not have obvious ED keywords but rather demonstrated one's negative emotions towards themselves and their body. Others had keywords but had a distinctly positive sentiment in favor of recovery despite their current struggles. Examples of false negatives are as follows:

- **Body issues:**

  "Also, whilst all the other girls in my class are sporty and slim, I am a little round and nerdy. It sucks."

  "Since then, i have eaten less and lost some weight due to the pressure, which i thought would make me happy, but i,am still so unhappy with myself."

- **Positive outlook:**

  "That I, have spent so much of my life trying NOT to eat that I have no idea how to eat enough to nourish myself.

  So I am wondering, if anybody might have recommendations for books that just teach me a bit about nut"

The outputs in Figure 2 significantly outperformed those in Figure 3 because these were results evaluated prior to creating a hand-labeled test set, and used subreddit name as the classifier for the test data. However, using the "prelabel" learning function, which labels training data based on subreddit, caused the models to perform significantly worse on actual labeled data. This means, the models trained in Figure 2 were better able to determine which type of subreddit a post might fall in (ED-related vs non ED-related), but is not as indicative of its ability to classify harmful eating disorder text itself.

Figure 3 more aptly demonstrates the models' ability to classify eating disorder text specifically since these models were evaluated on hand-labeled test data. Our best performing combination of pre-trained model training labels for finetuning was RoBERTa base with Snorkel data 2, with the highest accuracy and F1 score. This shows that emotion, pronoun analysis, and subreddit label were less influential on correct classification than simple keyword lookup and sentiment polarity, which were used to calculate Snorkel data 2. RoBERTa also consistently out-performed BERT, likely due to its higher quantity of pre-training data. RoBERTa also uses a dynamic masking pattern which increases its effectiveness in analyzing text emotion, as opposed to BERT which uses static masking. RoBERTa has also trained on longer sequences than had BERT, which was useful for our context since reddit posts are relatively long for social media posts (frequently > 20 words).

The highest performing training dataset using the BERT model, however, was Snorkel data 4. Snorkel 4 training labels were generated using additional heuristics to provide the context of emotion, first vs third person pronounds, and text length. As compared to the other BERT outputs, these additional learning functions helped reduce false positives that occurred due to incorrect sentiment analysis, were posts about third parties, or were extremely short posts that were irrelevant to eating disorders. However, these issues were not as prevalent in RoBERTa, so the training labels generated without these learning functions sufficed.

There was a 71% overlap between snorkel 2 and snorkel 3 incorrectly labeled posts when training on the BERT model, a 70% overlap between snorkel 2 and snorkel 4 incorrectly labeled posts, and a 95% overlap between snorkel 3 and snorkel 4 incorrectly labeled posts. This demonstrates that snorkel 2 not only significantly improved upon snorkel 3 + 4, but resolved specific false positives and false negatives without replacing them many new incorrect classifications.

MentalBERT was less accurate overall because of its extremely high rate of false positives, as seen in its high recall scores. This is because MentalBERT is prone to accidentally classifying any post exhibiting poor mental health as an eating disorder post, despite fine-tuning.

6 Conclusion

With this paper, we were able to use weakly-supervised learning techniques to generate sets of labeled data that improved the performance of transformer-based models like BERT and RoBERTa on the task of classification of harmful eating disorder rhetoric. We scraped over 15,000 posts from various subreddits on the topic of eating disorders, originally using the subreddit of a post as the only feature for labeling. After integrating many more labeling functions using Snorkel, we were able to improve the performance of our models on a randomly sampled test set, and achieved a maximum recall of 0.972 (with an accuracy of 0.888) by applying MentalBERT to our second round of data labeled using Snorkel. However, to get more informative results, we hand-labeled 200 posts, on which we were able to get a maximum recall of 0.909 (with an accuracy of 0.697) by applying MentalBERT. One limitation of our paper was that adding new labeling functions to our training data didn't always lead to improvements in performance across the board, and simple keyword lookups were ultimately the most important method for generating labels. Since there is such a lack of useful training data for eating disorder language, there is room for future research on how to generate more labeling functions that could yield large datasets useful for fine-tuning models on the classification of harmful eating disorder language. However, our fine-tuned models were able to achieve recall up to 90% and accuracy up to 77%, meaning that despite rates of false positives in the 14 - 16% range across the board, our best performing models were able to accurately predict almost all of the test set labels, which is a significant improvement from prior models' performance on this task.

References

[1] Goodman Craft Das Cavagos-Rehg RYan, Fitzsimmons-Craft. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. In *International Journal of Eating Disorders*, 2019.

[2] Luna Ansari Jie Fu-Prayag Tiwari Erik Cambria Shaoxiong Ji, Tianlin Zhang. Mentalbert: Publicly available pretrained language models for mental healthcare. In *https://arxiv.org/pdf/2110.15621.pdf*, 2021.

[3] Dawn B. Branley; Judith Covey. Pro-ana versus pro-recovery: A content analytic comparison of social media users' communication about eating disorders on twitter and tumblr. In *Frontiers in Psychology*, 2017.

[4] et al. Re, Chris. Snorkel ai. In *https://www.snorkel.org*, 2021.

[5] Theis F. Kordy H. Wolf, M. Language use in eating disorder blogs: Psychological implications of social online activity. In *Journal of Language and Social Psychology*, 2013.

[6] Bryce Boe. Praw api wrapper. In *https://praw.readthedocs.io/en/stable/*, 2021.

[7] https://huggingface.co/docs.

[8] https://snorkel.readthedocs.io/en/master/packages/_autosummary/labeling/snorkel.lab
eling.model.label_model.labelmodel.htmlsnorkel.labeling.model.label_model.labelmodel.