Transfer Learning for Eating Disorder Sentiment Analysis

By: Christine Manegan and Rhett Owen

Abstract

Our project aims to expand on the very small pool of research on the classification of eating disorder language on social media using natural language processing. In this paper, we use multiple 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 fine-tuned 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 on our task.

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 for the Reddit API. We cleaned the data to remove Nonetype/NaN posts, emojis, hyperlinks, and non-alphabetic chars and posts to ultimately have 14,194 cleaned posts that we partitioned into training, and validation sets, and hand-labeled 200 posts for our test set.

Learning functions for labeled training sets generated from SGD label model

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

We wrote the above learning functions using ED-identifying heuristics to create label models, which contain the resulting label from each of the corresponding learning functions for each post in our training set. Using the snorkel API's .fit() function, we ran an SGD optimizer with 500 epochs and a learning rate of 0.01 to train the label models to produce single sets of noise-aware training labels (snorkel 1, 2, … 5) using different combinations of learning functions. These labels were ultimately used to finetune our pretrained models.

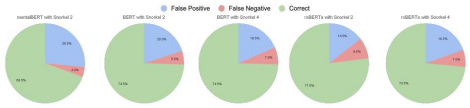
Results

	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

Our best performing combination of pre-trained model and training labels for fine-tuning was RoBERTa base with Snorkel data 2.

RoBERTa consistently outperformed BERT, likely due to its higher quantity of pre-training data, dynamic masking pattern, and training on longer sequences than BERT. The label models that worked best with RoBERTa did not need to use emotion or pronoun analysis in their labeling functions, while label models that did use those additional labeling functions performed better than their counterparts on traditional BERT.

This table also excludes our original round of testing, where instead of a hand-labeled test set we randomly split our data into 10000 posts training, 1000 for validation, and 1000 for testing. As expected, performance was worse across the board on the hand-labeled data, (with a maximum F1 score of 0.755 as opposed to 0.915), but we focused on improving performance on the hand-labeled test set instead to get more useful results for real-world applications.



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

Snorkel data 2 was generated using an SGD optimizer and learning functions that looked for keywords and sentiment polarity. Other label sets that used additional learning functions such as pre-trained emotion analysis, first vs third person pronoun analysis, subreddit prelabel, and post length were not as effective as simple keyword lookups for fine-tuning on roBERTa.

Analysis

Common traits for false positives: recovery posts with ED keywords, non-ED posts about mental health, posts about eating disorder resources

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

Common traits for false negatives: posts without ED keywords that were negative, posts talking about recovery as well as current struggles

> "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."

Conclusions

With our hand-labeled test set of 200 posts we achieved a maximum F1 score of 0.755, which we got by fine-tuning a RoBERTa model on our second pass of Snorkel data. We were able to get a maximum recall of 0.909 (with an accuracy of 0.697) with MentalBERT, albeit with an increased rate of false positives. We can conclude that adding more labeling functions and heuristics to our label model to train does not always lead to improvements in performance. However, labeling models trained using eating disorder and non-eating disorder keywords were consistently effective for generating training labels, as opposed to naively bucketing based on subreddit and other learning func. 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.

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. *International Journal of Eating Disorders*, 2019

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

[3] Huggingface docs: https://huggingface.co/docs

[4] Snorkel classification and learning documentation: https://snorkel.readthedocs.io/en/v0.9.7/

Stanford ENGINEERING