

Cordyceps@LT-EDI : Depression Detection with Reddit and Self-training

Dean Ninalga

justin.ninalga@mail.utoronto.ca

Abstract

Depression is debilitating, and not uncommon. Indeed, studies of excessive social media users show correlations with depression, ADHD, and other mental health concerns. Given that there is a large number of people with excessive social media usage, then there is a significant population of potentially undiagnosed users and posts that they create. In this paper, we propose a depression severity detection system using a semi-supervised learning technique to predict if a post is from a user who is experiencing severe, moderate, or low (non-diagnostic) levels of depression. Namely, we use a trained model to classify a large number of unlabelled social media posts from Reddit¹, then use these generated labels to train a more powerful classifier. We demonstrate our framework on *Detecting Signs of Depression from Social Media Text - LT-EDI@RANLP 2023* (Sampath et al., 2023) shared task, where our framework ranks 3rd overall.

1 Introduction

1.1 Depression and Social Media

A unique feature of depression is its effect on cognitive and verbal patterns. For example, depression diagnosis is correlated to the frequency of personal pronoun usage and the usage of positive-negative words (Edwards and Holtzman, 2017; Tølbøll, 2019). Additionally,

persons suffering from depression often connect vicious yet potentially fictional narratives to benign experiences, generally increasing the number of overwhelming situations they may experience (Kanter et al., 2008). People may then go to social media and online forums like Reddit to discuss and post about traumatizing experiences and may publicly reflect on their thoughts and behavior. It is unsurprising, therefore, to find a wealth of attempts (as surveyed by Hasib et al. (2023)) to use social media posts to create a potential diagnostic screening tool through language modeling. Recently, language models can accurately predict symptoms before practitioners record them (Eichstaedt et al., 2018; Reece et al., 2016).

There are significant challenges to data collection in the depression detection setting despite a potential abundance of data that likely exists on social media. Indeed, excessive social media usage itself correlates with depression, ADHD, and other serious mental health diagnoses (Hussain and Griffiths, 2019). However, Guntuku et al. (2017) observe that most attempts at data collection rely on a self-declaration or a past diagnosis of depression, allowing for the possibility of non-actively depressed individuals creating depression-positive data. In this paper, we will attempt to apply an automatic data collection process from social media through a semi-supervised approach.

¹<https://www.reddit.com/>

1.2 Background on Self-training

Self-training techniques (Scudder, 1965) are a type of semi-supervised learning and are well known in various areas of research (e.g. Zoph et al. (2020); Xie et al. (2019); Sahito et al. (2021)). These techniques in broad terms, take a trained model, generate labels for a large set of unlabeled data, then train a new model incorporating the clean labels, generated labels, and unlabeled data. Where the new model is typically of the same size, or bigger, as the original trained model. Surprisingly, however, little work has been done exploring how to apply this process in the specific case of depression detection across many social media.

To summarize, our main contributions are the following:

- We describe our framework based on self-training.
- We demonstrate our framework on *Detecting Signs of Depression from Social Media Text - LT-EDI@RANLP 2023* (Sam-path et al., 2023) shared task, comparing to recent work.
- We describe areas where pseudo-labeling can advance depression detection modeling.

2 Related Work

Recent work has demonstrated semi-supervised learning techniques using unlabeled Twitter data for depression detection as surveyed by (Zhang et al., 2022). However, these studies tend to solely rely on Twitter² data as their source of unlabeled texts (Zhang et al., 2022; Yazdavar et al., 2017). Here, we will use Reddit for our semi-supervised approach.

Poswiata and Perelkiewicz (2022) also uses the *Reddit Mental Health Dataset* (Low et al.,

²<https://twitter.com/>

2020) in their depression detection system for last year’s iteration of the shared task. However, Poswiata and Perelkiewicz (2022) do not generate pseudo-labels but instead use the data for a pre-training task that is specifically designed for depression detection. Pirina and Çağrı Çöltekin (2018) suggested that the selection of Reddit forums (or *subreddits*) in the training data may influence the quality of classifiers. Here, our goal is to automate this selection process without having to rely on *subreddit* specific information and rely solely on the posts themselves.

3 Methodology

Here we will provide the major implementation details of our solution in this section. See Table 2 for further information on training hyper-parameter details used throughout.

3.1 Data Cleaning

We perform a few basic data-cleaning steps for any samples fed to the classifier. That is, we remove any newline and tab characters, strip leading and trailing white spaces, and replace all links with an identical string. Additionally, we remove duplicated texts and drop samples in the shared-task training set if it is also contained in the shared-task development set. In total, we dropped 128 duplicated samples.

3.2 Pre-Trained Models

Leveraging pre-trained language representations is a proven way to boost performance on essentially any given NLP task. Downstream task performance gains are even more prominent if the pre-training task is identical to the downstream tasks and uses large amounts of data. To that end, we use MentalRoBERTa (Ji et al., 2022) as our model of choice for training and inference. MentalRoBERTa (Ji et al., 2022) is a RoBERTa (Liu et al., 2019b) model

Name	Dev	Test
MentalRoBERTa (Ji et al., 2022) + <i>pl</i> + <i>ft</i> (ours)	0.7407	0.4309
MentalRoBERTa (Ji et al., 2022) + <i>pl</i>	0.5359	0.3975
MentalRoBERTa (Ji et al., 2022)	0.578	0.44
MentalXLNet (Ji et al., 2023)	0.5714	0.4443
MentalBERT (Ji et al., 2022)	0.5648	0.3901
RoBERTa (Liu et al., 2019a)	0.5627	0.3953
BERT (Devlin et al., 2018)	0.5512	0.3981

Table 1: *Macro*-averaged F1-Score results on the development and test set of the shared task. The best score on each set is highlighted. The top two rows highlight a single run of our approach: training on only pseudo-labels (*pl*) and then finetuning (*ft*). The next three rows detail the finetuning results of recently release pre-trained models for mental health. In the last two rows, we present a baseline using well-known models.

Hyper-Parameter	Value
Optimizer	Adam
Learning Rate	1e-5
Max Input Length	256
Batch Size	8

Table 2: Training Hyper-parameter Details

that is further pre-trained on Reddit mental-health-related data.

3.3 Self-Training

The details of our self-training and pseudo-labeling procedure are as follows. Firstly, we train a teacher model using the annotated training data. Next, we use the trained teacher model to generate predictions on the unlabeled data: *Reddit Mental Health Dataset* (Low et al., 2020). Here, we want to keep the highest-ranked 30,000 samples with the highest-valued predicted logit for any of the three label categories. For example, we only include a post in the severe depression category if the teacher model is very confident that a sample belongs in the depression category relative to all other posts. Subsequently, the resulting 90,000 posts are then assigned pseudo-labels based on the previously assigned groupings, where we as-

sume that each sample belongs to its respective category. Here, we do not consider the categorical probability distribution (as predicted by the teacher) since we are only keeping samples with high confidence. In practice, the predicted output probabilities of the 90,000 posts are very close to 1 for their respective category, hence, using the predicted probabilities adds very little information. Next, we use the 90,000 posts alongside the pseudo-labels to construct a new dataset which is used to train a new student model. Note, here we use the same model architecture for both the teacher and student. Finally, the student model is finetuned with the clean training data and then used for inference on the test set.

4 Experiments

4.1 Experimental Setup

We compare our setup to several other state-of-the-art pre-trained models we finetuned for the shared task. We report the macro-averaged F1-Score on the test and development sets. Where report the average score over five runs, unless otherwise stated. We perform all experiments on a single T4 GPU.

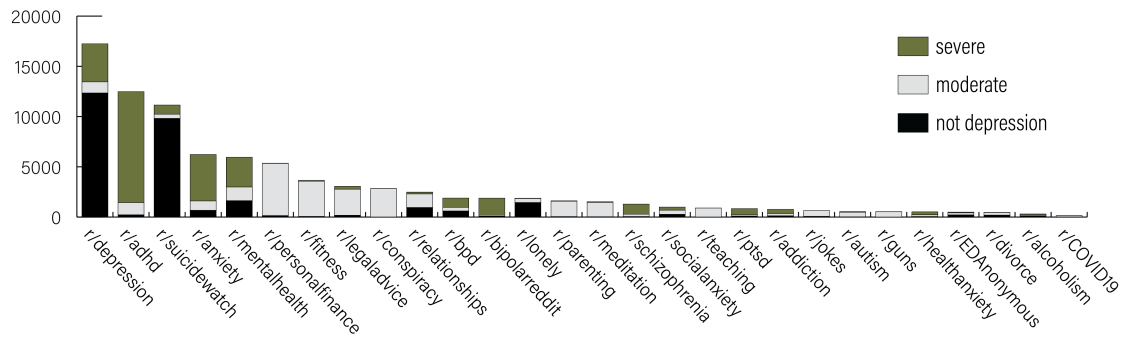


Figure 1: Breakdown of the pseudo-labels on each subreddit in the *Reddit Mental Health Dataset* (Low et al., 2020)

4.2 Results

We present our full results in Table 1. Indeed, our complete approach of self-training with MentalRoBERTa (Ji et al., 2022) performs the best on the development set by a wide margin. However, our approach performs narrowly worse than MentalXLNet (Ji et al., 2023) on the test set. Given this disparity in development and test set performance, future work should explore regulation techniques (e.g. augmentation and ensembling methods) to accompany the self-training approach. Nonetheless, our approach still places 3rd overall in the shared task.

5 Exploratory analysis

We present an analysis of our generated pseudo-labels on the *Reddit Mental Health Dataset* (Low et al., 2020). Recall, that we assign a pseudo-label to a post only if the post is ranked in the top 30,000 in any of the three depression severity labels. In Figure 1 we break down the distribution of the labels across the sources of these labels. Notably, we find about 60% of our generated labels are contained in five subreddits: ‘*r/depression*’, ‘*r/adhd*’, ‘*r/suicidewatch*’, ‘*r/anxiety*’, ‘*r/mentalhealth*’. In particular, the subreddit ‘*r/adhd*’ hosts the most pseudo-labels in the *severe* category out of any subreddit by a wide margin, account-

ing for about 37% of all pseudo-labels in the category.

There are multiple explanations for the above findings. Indeed, ADHD can co-occur with depression and can be seen as an early indication of a future depression diagnosis (Meinzer and Chronis-Tuscano, 2017). Additionally, ADHD and depression have overlapping symptoms (Riglin et al., 2020). Thus, it is possible that there is some level of overlapping language or similar verbal processes shared between the two disorders. We encourage future work to explore alternative explanations and leverage this connection between ADHD and depression in the depression-detection setting.

6 Conclusion

In this paper, we present our framework based on self-training and demonstrate its performance on the *Detecting Signs of Depression from Social Media Text - LT-EDI@RANLP 2023* (Sampath et al., 2023) shared task. Given the disparities observed in the development set and test set F1-score performance, future work should explore regulation techniques (e.g. augmentation and ensembling methods) to accompany the self-training approach. Nonetheless, our approach still places 3rd overall in the shared task.

With our use of pseudo-labeling on Reddit,

we highlighted ADHD-focused forums as a major source of (non-diagnostic) severe depression classifications and discussed some explanations. We hope our work serves as a starting point for further investigation of the linguistic patterns of depression overlapping with other mental disorders.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- T. Edwards and Nicholas S. Holtzman. 2017. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68:63–68.
- Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A. Asch, and H. A. Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, 115:11203 – 11208.
- Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.
- Khan Md Hasib, Md Rafiqul Islam, Shadman Sakib, Md. Ali Akbar, Imran Razzak, and Mohammad Shafiul Alam. 2023. Depression detection from social networks data based on machine learning and deep learning techniques: An interrogative survey. *IEEE Transactions on Computational Social Systems*.
- Zaheer Hussain and Mark D. Griffiths. 2019. The associations between problematic social networking site use and sleep quality, attention-deficit hyperactivity disorder, depression, anxiety and stress. *International Journal of Mental Health and Addiction*, 19:686 – 700.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.
- Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. [Domain-specific continued pre-training of language models for capturing long context in mental health](#). *arXiv preprint arXiv:2304.10447*.
- Jonathan W. Kanter, Andrew M. Busch, Cristal E. Weeks, and Sara J. Landes. 2008. The nature of clinical depression: Symptoms, syndromes, and behavior analysis. *The Behavior Analyst*, 31:1–21.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized pre-training approach. *ArXiv*, abs/1907.11692.
- Daniel M Low, Laurie Rumker, John Torous, Guillermo Cecchi, Satrajit S Ghosh, and Tanya Talkar. 2020. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. *Journal of medical Internet research*, 22(10):e22635.
- Michael C. Meinzer and Andrea Chronis-Tuscano. 2017. Adhd and the development of depression: Commentary on the prevalence, proposed mechanisms, and promising interventions. *Current Developmental Disorders Reports*, 4:1–4.
- Inna Loginovna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Conference on Empirical Methods in Natural Language Processing*.
- Rafal Poswiata and Michal Perelkiewicz. 2022. [Opi@lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models](#). In *LTEDI*.
- Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M.

- Danforth, and Ellen J. Langer. 2016. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports*, 7.
- Lucy Riglin, Beate Leppert, Christina Dardani, Ajay K Thapar, Frances Rice, Michael C. O'Donovan, George Davey Smith, Evie Stergiakouli, Kate Tilling, and Anita Thapar. 2020. Adhd and depression: investigating a causal explanation. *Psychological Medicine*, 51:1890 – 1897.
- Attaullah Sahito, Eibe Frank, and Bernhard Pfahringer. 2021. Better self-training for image classification through self-supervision. *ArXiv*, abs/2109.00778.
- Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the second shared task on detecting signs of depression from social media text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, Varna, Bulgaria. Recent Advances in Natural Language Processing.
- H. J. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Inf. Theory*, 11:363–371.
- Katrine Bønneland Tølbøll. 2019. Linguistic features in depression: a meta-analysis.
- Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and A. Sheth. 2017. Semi-supervised approach to monitoring clinical depressive symptoms in social media. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.
- Tianlin Zhang, Annika Marie Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ Digital Medicine*, 5.
- Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc V. Le. 2020. Rethinking pre-training and self-training. *ArXiv*, abs/2006.06882.