A Pretrained Language Model for Mental Health Risk Detection

Diego Maupomé[1], Fanny Rancourt[1], Raouf Belbahar[1] and Marie - Jean Meurs[1]

[1]Université du Québec à Montréal, Montréal, QC, Canada

Abstract

Early detection of mental health issues is a key contributor to efficient treatment. Natural language processing-based approaches can provide automated means to facilitate access to appropriate services and support for at-risk individuals. Using pretrained language models provides state-of-the-art results in various downstream tasks as these models leverage significant amounts of textual content. They can be critical in data-scarce research areas, such as early detection of mental health issues. Nonetheless, exposing models to domain-specific language can be beneficial to their performance in downstream task. To this end, we release pretrained language models, MentalHealthBERT, leveraging content from Reddit fora discussing anorexia, depression and self-harm. These models are evaluated on risk detection tasks for the respective conditions.

1. Introduction

Early intervention in mental health and well-being has become a critical principle of mental health care, ushering in an international wave of service reform [1, 2]. Given the ever-growing use and diversity of online social media, there has been a vast increase in research interest for the use of Natural Language Processing (NLP) for the development of automated means of analyzing online textual content in the service of mental health care support and early intervention in particular [3, 4].

The inference of such predictive models requires the gathering of annotated data. These data map online textual content to an assessment of certain aspects of the mental health of the authors of this content. Such assessments are difficult to produce. Whereas for other common tasks in NLP, annotation can operate on the observation itself (*e.g.* the text), annotation relating to mental health generally requires further information about the author of the textual content. That is, the true aspects of interest pertain to the author rather than the text. In particular, clinically grounded assessments require access to the individual. As such, gathering annotated data is expensive and time-consuming.

In the absence of large quantities of annotated data, it is a well-established principle of machine learning that pretraining on an unsupervised task can help performance on a downstream supervised task. As such, there has been increased interest in the production of pretrained models leveraging large amounts of textual content [5, 6, 7, 8]. Such models are made available for use on a variety of specialized downstream tasks [9, 10]. The core tenet is that large models trained on sufficiently large data sets will learn to produce useful representations of text regardless of what specialized task these representations will serve. Such a framework leverages large quantities of data for models to learn aspects of language that are thought to precede the specifics of the specialized task.

While this assumption may hold for *tasks*, pretraining data can also issue from different *sources* than the specialized data. As such, representations produced by general-purpose models might be inadequate. Recent work has pointed to the benefits of domain specificity in large pretrained models. Broadly, the term *domain* refers to the topics, mode or register of documents. Domain specificity concerns can take the form of models pretrained entirely on domain-specific data or domain adaptation.In either case, gains in downstream task performance have been reported for several tasks and domains from the use of such domain-specific pretraining [11].

The textual data analyzed for mental health care purposes issues from Internet fora and social media. These data can differ both in register and topics from news or encyclopedia articles comprising significant parts of large corpora. Nonetheless, there is no established linguistic consensus on what constitutes a domain [see 12, Sec. 3.4.1]. Given this difficulty in defining the notion of domain, it is difficult to delineate given domains or to establish quantitative differences between them.

Pragmatically, one might ask whether a more narrow concept of a given domain may provide more benefit to downstream task performance than a broader one. The present work seeks to study this issue in the context of mental health risk assessment. Models are pretrained on data from Internet fora revolving around three different mental health concerns: anorexia, depression and self-harm. The models are evaluated on detection tasks surrounding these concerns and compared to models trained on broader data [8]. Our results corroborate the benefits of domain adaptation for general-purpose language

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

maupome.diego@uqam.ca (D. Maupomé); meurs.marie-jean@uqam.ca (M. -. J. Meurs)

0000-0003-2527-2515 (D. Maupomé); 0000-0003-0189-810X (F. Rancourt); 0000-0001-8196-2153 (M. -. J. Meurs)

models but only show advantages to pretraining-data specificity in one case: anorexia.

2. Data

2.1. Retrieval

The data were extracted from three Reddit[1] fora (known as *subreddits*): `depression`, `selfharm` and `AnorexiaNervosa`. This extraction was performed using Pushshift[2] [13]. For all three subreddits, it was limited to posts published from the 1st of January 2019 to the 25th of November 2020.[3] Further, posts struck off as "removed" were discarded. The fields associated with each post include the title and body of the post, as well as the timestamp, the score (aggregation of up- and downvotes), the number of replies and the identifier of the parent post. No additional filtering was applied.

2.2. Description

All subreddits considered are described as communities that offer a safe place and peer support for people affected by the aforementioned issues[4]. Summary statistics for the corpus are presented in Table 1.

The `depression` forum is by far the biggest community of the three with more than 736,000 members as of March 2nd 2021. Of those, about 45% authored at least one publication (*i.e.* a post or a comment) in the selected time frame. A similar proportion can be observed from `AnorexiaNervosa`. In turn, it jumps to almost two-thirds for `selfharm`. Across all three subreddits, approximately 40% of the authors published exactly once. Despite having the fewest overall publications, threads on `AnorexiaNervosa` seem to generate the most engagement, having a higher ratio of comments per thread and remaining active for longer periods. The smaller size of this community is a likely explanation for these observations.

3. Ethical Considerations

All posts collected in the aforementioned subreddits are public but our collection will not be publicly available. Further, resources discussed in this work will be released upon the signature of a User Agreement. The released model should only be used in combination with other screening tools for prevention purposes under the supervision of trained mental health professionals. Hence, this system does not aim to diagnose mental health disorders and should not be used to do so.

However, the misuse of this kind of work can have negative societal impacts. For example, an organization could use our pretrained language models to detect at-risk job applicants of mental health disorders before hiring. This practice, violating the terms of the release agreement, would further spur discrimination in hiring processes in addition to well-documented gender and racial unfairness [14, 15].

While this line of research could potentially advance early intervention and treatment processes, it does not directly address the stigma surrounding mental health issues and underlying the high rate of treatment avoidance and discontinuation [16, 17]. Further, widespread study and deployment of models in this direction could potentially lead to self-censorship, defeating its purpose.

It is also important to note that demographic data on the authors is missing. As noted by Shatz [18], most subreddits do not have data regarding their community demographics. Hence, it is impossible to ensure that the textual productions used to train the released model adequately represent content from diverse individuals. To the best of our knowledge, there is no readily available dataset containing information regarding the author's age, gender, ethnicity, or location. From inferred demographics, Amir et al. [19] presented that those sensitive attributes affect depression prevalence across social media users. Aguirre et al. [20] observed performance gaps related to gender and racial attributes. To address this gap, a data collection combining strict privacy policies and clinical supervision must be achieved. As noted by Aguirre et al. [20], storing such sensitive data comes with serious potential harms. Therefore, it is critical to enforce protective measures such as data anonymization.

4. Pretraining

4.1. Preprocessing

One key issue in modeling corpora from Internet fora rather than an edited outlet, such as a newspaper or encyclopedia, is the longer vocabulary tail caused by misspellings, neologisms and even usernames. Common practice would be to remove words having fewer than three occurrences [21]. Keeping such words would increase the computational burden of the model while having little chance of learning because of the limited number of occurrences. However, this is not suitable for our purposes: Important words might be misspelled or obfuscated, but their exclusion will hinder performance [22]. Similarly, usernames and neologisms might be composed from familiar, significant words. As such, we preserve the entire vocabulary of each dataset, relying on subword-

[1] https://www.reddit.com
[2] https://pushshift.io/
[3] The latest post from `AnorexiaNervosa` was published on the 3rd of December 2020.
[4] As per their respective "About Community" section of each subreddit

	AnorexiaNervosa	depression	selfharm
Tokens	3.7G	160.9G	18.8G
Vocabulary	38.4k	303.2k	87.1k
Posts	10.3k	412.4k	78.0k
average number of tokens	141	204	116
Comments	45.8k	1404.3k	236.4k
average number of tokens	49	54	41
Unique author	10.1k	338.1k	43.3k
Community size*	23.8k	736k	66.4k

Table 1
Subreddits statistics. Unique authors exclude deleted accounts. *As of March 2nd 2021.

level tokenization to capture these variations.

Before learning this tokenization, the data was split into training and validation sets by stratifying across length (word count) percentiles. This preserves the key length statistics, such as the median and interquartile range. In terms of vocabulary, words in the validation set not present in the training set make up 0.50%, 0.20% and 0.10% of occurrences in the anorexia, self-harm and depression sets, respectively.

The data was tokenized by Byte-Pair Encodings (BPEs) [23] at the byte level [6], with the merges extracted from all three datasets. This consolidation was done to provide a more robust tokenization scheme, less skewed towards any particular forum, while still learning the words and spellings of online parlance. For comparison, each dataset was also tokenized using merges learned exclusively from itself.

4.2. Training

Once tokenized, these datasets were used to train Transformers [24] using the RoBERTa approach [6]. Models are trained by the Adam optimizer [25] with a learning rate of 5E-4 on batches of 256 sequences of a maximum length of 256 tokens. Training takes place over a maximum of 300 epochs, applying early stopping based on validation set perplexity.

5. Mental Health Risk Detection

We evaluate the MentalHealthBERT models on the eRisk datasets [26, 27, 28]. These datasets comprise Reddit users (subjects) labeled as being at risk (positive) or not (negative) for depression, self-harm or anorexia, respectively. For each subject, a history of their writings is included, spanning a variety of subreddits. The proportion of positive subjects is fairly small and varies somewhat, as does the size of the datasets, as shown in Table 2.

The key issue is utilizing the document-level encoding afforded by MentalHealthBERT in predictions at the history level, which spans a variety of presumably independent writings. In order to make this subject-level prediction, information gathered across a set of writings needs to be aggregated. To achieve this, token embeddings are averaged together within posts and subsequently fed to a feed-forward network with a single hidden layer and hyperbolic tangent activation. The resulting document vectors are then aggregated by averaging into a single vector encoding a history of writings. This vector is then mapped to a binary prediction for the sequence of writings by a feed-forward network with a single hidden layer with hyperbolic tangent activation.

5.1. Experiments

The experiments compare the performance of MentalHealthBERT to the generic RoBERTa Transformer as well as the latter further pretrained on our data (domain adaptation). For MentalHealthBERT, experiments were carried out using BPEs learned from the combined dataset as well as from the individual collections. Additionally, we run experiments using MentalRoBERTa [8]. This model was pretrained on Reddit data from several different fora touching on mental health topics[5]. It should be noted that the results reported by the authors on the eRisk depression detection task are not comparable to those reported here, as they make use of a custom data split with some resampling [29]. Models are evaluated per the area under the precision-recall curve.

One difficulty of detecting potential threats to mental health is the small proportion of positive subjects that can be found in datasets and, indeed, in a real-world setting. Additionally, for the selected datasets, these proportions vary widely between the training and testing sets, as shown in Table 2. Models were evaluated using the latest set of data for each task: 2022 for Depression, 2021 for Self-Harm and 2019 for Anorexia. Training and validation

[5] An exhaustive list of the fora from which pretraining data were extracted is not available, but they include `depression`, `SuicideWatch`, `Anxiety`, `offmychest`, `bipolar`, `mentalillness`, and `mentalhealth`.

	Train		Test	
dataset	positive	negative	positive	negative
Depression	214	1493	98	1302
Self-Harm	145	618	152	1296
Anorexia	61	411	73	742

Table 2
Positive and negative subject counts from the eRisk training and testing sets

	Tokenization	Depression	Self-Harm	Anorexia
RoBERTa [6]	RoBERTa	0.487	0.434	0.401
RoBERTa with domain adaptation	RoBERTa	0.496	**0.494**	0.555
MentalRoBERTa [8]	RoBERTa	**0.536**	0.476	0.416
MentalHealthBERT	Separate	0.520	0.475	0.560
MentalHealthBERT	Combined	0.457	0.485	**0.569**

Table 3
Area under the precision-recall curve on the eRisk test sets

sets for each task were obtained by combining the data from all previous sets and randomly selecting 80% of subjects for training and 20% for validation, preserving equal proportions of positive and negative subjects.

To address this class imbalance in training, a number of strategies were deployed, including inverse class weighting, class weighting based on effective samples [30] and Focal Loss [31]. These proved to be ineffective in validation. The most effective mechanism proved to be sampling batches of even proportions of positive and negative subjects in training.

The number of writings used to arrive at a prediction for a subject was set to $m = 50$. In order to reduce overfitting, a contiguous sample of m writings was taken per subject at training time. In validation and testing, only the last m writings were taken. The classifiers are trained by the Adam optimizer [25] over 10 epochs. Given the relatively modest size of the datasets in terms of positive subjects, only the top two layers of the Transformer encoder were trained, with a learning rate of 1E-5. The remainder of the model had a learning rate set to 1E-4.

Results on the eRisk datasets are presented in Table 3. Results for the base RoBERTa model indicate improvements with domain adaptation, in agreement with the literature [11]. Perhaps counterintuitively, these improvements appear to decrease with the amount of domain adaptation data available. MentalRoBERTa and MentalHealthBERT achieve comparable results in all but the anorexia task, for which MentalRoBERTa and domain-adapted RoBERTa outperform MentalRoBERTa. This may be due to a deficiency in eating disorder content in pretraining MentalRoBERTa, though we cannot confirm this. Tokenization seems to be inconsequential, with a more marked decrease in performance for the combined tokenizer in depression. Given the difficulty that specialized tokenization puts in transferring learning, it is difficult to support in light of these results. Finally, there appears to be little difference in performance between domain-adapted RoBERTa and the best MentalHealthBERT model, suggesting no real benefit to training blank models over adapting pretrained models. In light of these results, it is difficult to establish whether the specific domain pretraining of MentalHealthBERT helps downstream performance more so than more the general domain adaptation found in MentalRoBERTa. As mentioned, benefits are only observable for the anorexia task. Given what is known of the pretraining of MentalRoBERTa, it is difficult to establish whether this may be due to any material characteristics of discourse around anorexia or its relatively smaller weight in pretraining.

6. Conclusion

There is increased research interest in the development of NLP approaches to assist in early risk assessment in mental health care. Gathering annotated data is a costly process, making pretraining a crucial step in the modeling process. Thus, pretrained language models can be a valuable resource. However, general-purpose language

models, while trained on large amounts of data, may not be suited to specific domains, such as mental health discussions. As such, there is interest in adapting language models to particular domains.

In the case of mental health risk assessment from text, domain-specific pretraining resources would contain discourse concerning mental health concerns. However, it is worth considering whether discourse issuing from outlets specific to a particular mental health concern are more adequate than discourse around mental health issues at large. Our experiments have thus made use of data extracted from fora dedicated to specific mental health concerns to pretrain models. These models are compared to general-purpose language models as well as language models pretrained on broader mental health content in a mental health risk assessment task. Our results indicate that domain adaptation does improve classification performance. However, a difference in performance between more narrowly pretrained models is only manifest in anorexia risk detection.

Further work is needed to understand how textual data from separate mental health topics interact in terms of benefits from pretraining: more experimentation is needed to find whether the detection of certain mental health concerns is improved by pooling pretraining data and whether these gains in detection performance align with the comorbidity of the underlying disorders. Were this the case, those benefits might be explained by the mention of related concerns in discussions about a specific mental health concern. While pretraining data for our experiments was extracted from dedicated fora, our experiments do not control for the mention of related disorders or threats to mental health.

Release of Resources

Given the sensitive nature of the resources introduced, the models and associated open-source code will be released upon signing of a User Agreement providing details on permitted uses.

References

[1] M. Schotanus-Dijkstra, C. H. C. Drossaert, M. E. Pieterse, B. Boon, J. A. Walburg, E. T. Bohlmeijer, An early intervention to promote well-being and flourishing and reduce anxiety and depression: A randomized controlled trial, Internet Interventions 9 (2017) 15–24. URL: https://www.sciencedirect.com/science/article/pii/S2214782916300288. doi:10.1016/j.invent.2017.04.002.

[2] P. D. McGorry, C. Mei, Early intervention in youth mental health: Progress and future directions, Evidence-Based Mental health 21 (2018) 182–184.

[3] H.-C. Shing, P. Resnik, D. W. Oard, A prioritization model for suicidality risk assessment, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8124–8137.

[4] D. Maupomé, M. D. Armstrong, F. Rancourt, M.-J. Meurs, Leveraging textual similarity to predict Beck Depression Inventory answers, Proceedings of the Canadian Conference on Artificial Intelligence (2021). doi:10.21428/594757db.5c753c3d.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, arXiv:1802.05365 [cs] (2018).

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, arXiv:1907.11692 (2019). arXiv:1907.11692.

[7] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, arXiv:2003.10555 [cs] (2020). URL: http://arxiv.org/abs/2003.10555, arXiv: 2003.10555.

[8] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, arXiv:1804.07461 [cs] (2019).

[10] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, SuperGLUE: A stickier Benchmark for general-purpose language understanding systems, arXiv:1905.00537 [cs] (2020).

[11] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, arXiv:2004.10964 [cs] (2020). URL: http://arxiv.org/abs/2004.10964.

[12] B. Plank, Domain Adaptation for Parsing, Ph.D. thesis, University of Groningen, 2011.

[13] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, J. Blackburn, The Pushshift Reddit dataset, in: Proceedings of the international AAAI conference on web and social media, volume 14, 2020, pp. 830–839.

[14] J. Sánchez-Monedero, L. Dencik, L. Edwards, What

does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 458–468.

[15] L. Quillian, D. Pager, O. Hexel, A. H. Midtbøen, Meta-analysis of field experiments shows no change in racial discrimination in hiring over time, Proceedings of the National Academy of Sciences 114 (2017) 10870–10875.

[16] O. F. Wahl, Stigma as a barrier to recovery from mental illness, Trends in Cognitive Sciences 16 (2012) 9–10.

[17] C. Henderson, S. Evans-Lacko, G. Thornicroft, Mental illness stigma, help seeking, and public health programs, American Journal of Public Health 103 (2013) 777–780.

[18] I. Shatz, Fast, Free, and Targeted: Reddit as a Source for Recruiting Participants Online, Social Science Computer Review 35 (2017) 537–549.

[19] S. Amir, M. Dredze, J. W. Ayers, Mental health surveillance over social media with digital cohorts, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 114–120.

[20] C. Aguirre, K. Harrigian, M. Dredze, Gender and racial fairness in depression research using social media, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021, pp. 2932–2949.

[21] S. Merity, C. Xiong, J. Bradbury, R. Socher, Pointer sentinel mixture models, arXiv:1609.07843 [cs] (2016). arXiv:1609.07843.

[22] B. Plank, What to do about non-standard (or non-canonical) language in nlp (2016). arXiv:1608.07836.

[23] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword unit, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1715–1725. URL: https://www.aclweb.org/anthology/P16-1162. doi:10.18653/v1/P16-1162.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[25] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014). arXiv:1412.6980.

[26] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early risk prediction on the Internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2018, pp. 343–361.

[27] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019: Early risk prediction on the Internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, 2019, pp. 340–357.

[28] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2020: Early risk prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), 2020.

[29] S. Ji, Private Correspondence, 2022.

[30] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, arXiv:1708.02002 [cs] (2018). arXiv:1708.02002.