A Pretrained Language Model for Mental Health Risk Detection

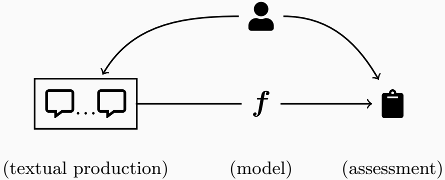
Diego Maupomé Fanny Rancourt Raouf Belbahar Marie-Jean Meurs

February 2024

UQÀM

Background

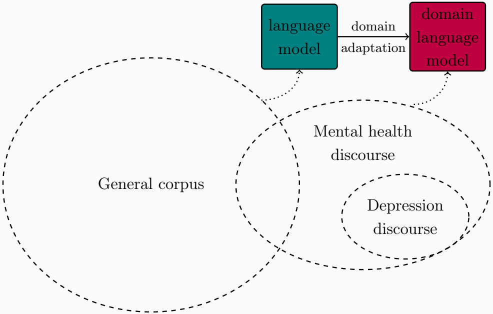
- Increased interest in early intervention in mental health care [Schotanus-Dijkstra et al., 2017, McGorry and Mei, 2018].
- Ever-growing use and diversity of online social media → research interest for automated analysis of online textual content for mental health care support [Shing et al., 2020, Maupomé et al., 2021].



- Annotation is expensive, semi-supervised learning is needed

Motivation

- Large models trained over large datasets perform better.
- Computation amortized by versatility.
- Some **domain adaptation** is needed: what does this mean in mental health?
- Idea: compare models pretrained on specific vs. general corpora.



Pretraining data

Pretraining data extracted from separate fora: `AnorexiaNervosa`, `depression`, `selfharm`

	`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 authors	10.1k	338.1k	43.3k
Community size*	23.8k	736k	66.4k

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

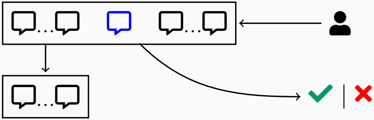
Pretraining process

- Based on RoBERTa [Liu et al., 2019]
- Adapt models with single set
- What role does tokenization play? Train "blank" models with tokens learned from joint data, separate data

Evaluation

Risk detection (binary) with eRisk datasets

[Losada et al., 2018, Losada et al., 2019, Losada et al., 2020]



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

Evaluation

- Building the classifier:
  - Document encoding: average tokens from post, feed-forward network
  - Average document vectors together
  - Training: contiguous sample of 50 posts, Test: last 50 posts
- Addressing class imbalance:
  - Inverse class weighting, effective sample weighting [Cui et al., 2019] and Focal Loss [Lin et al., 2018] proved ineffective
  - Even batches worked best
- Details:
  - Only two top layers of Transformer
  - Adam over ten epochs

Results

Baselines:

- General-purpose language model: RoBERTa [Liu et al., 2019]
- Mental health domain-adapted model: MentalRoBERTa [Ji et al., 2022]

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

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

Takeaways

- Domain-specific pretraining helps
- No real difference between blank and adapted models
- Mitigated results for more specific pretraining
- Future work:
  - Which disorders combine well in pretraining?
  - Does this relate to their theoretical relationship?
  - Would this be entirely attributable to topics?
  - Additional tasks

THANK YOU

Bibliography

[Cui et al., 2019] Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). **Class-balanced loss based on effective number of samples.** In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

[Ji et al., 2022] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). **MentalBERT: Publicly available pretrained language models for mental healthcare.** In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

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

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

[Losada et al., 2018] Losada, D. E., Crestani, F., and Parapar, J. (2018). **Overview of eRisk: Early risk prediction on the Internet.** In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 343–361.

[Losada et al., 2019] Losada, D. E., Crestani, F., and Parapar, J. (2019).
Overview of eRisk 2019: Early risk prediction on the Internet.
In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357.

[Losada et al., 2020] Losada, D. E., Crestani, F., and Parapar, J. (2020).
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).*

[Maupomé et al., 2021] Maupomé, D., Armstrong, M. D., Rancourt, F., and Meurs, M.-J. (2021).
Leveraging textual similarity to predict Beck Depression Inventory answers.
Proceedings of the Canadian Conference on Artificial Intelligence.

[McGorry and Mei, 2018] McGorry, P. D. and Mei, C. (2018).
Early intervention in youth mental health: Progress and future directions.
Evidence-Based Mental health, 21(4):182–184.

[Schotanus-Dijkstra et al., 2017] Schotanus-Dijkstra, M., Drossaert, C. H. C., Pieterse, M. E., Boon, B., Walburg, J. A., and Bohlmeijer, E. T. (2017).
An early intervention to promote well-being and flourishing and reduce anxiety and depression: A randomized controlled trial.
Internet Interventions, 9:15–24.

[Shing et al., 2020] Shing, H.-C., Resnik, P., and Oard, D. W. (2020).
A prioritization model for suicidality risk assessment.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137.