Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings

Pegah Abed-Esfahani[1,2], Derek Howard[1,2], Marta Maslej[1], Sejal Patel[1,2], Vamika Mann[3], Sarah Goegan[3], and Leon French[1,2,4,5]

[1] Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada

[2] Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, ON, Canada

[3] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, ON, Canada

[4] Institute for Medical Science, University of Toronto, Toronto, Canada

[5] Division of Brain and Therapeutics, Department of Psychiatry, University of Toronto, Toronto, Canada

Abstract. Online social media platforms allow open sharing of thoughts and dialogue. These platforms generate large amounts of data about their users' online behaviour, which can be repurposed for the development of technologies which can detect mental health disorders. Towards this goal, we applied transfer and supervised learning techniques for predicting the severity and risk of depression for the eRisk 2019 Lab at the CLEF workshop. Both tasks were very difficult due to lack of training data, motivating our efforts to learn signals from other pre-trained models and datasets. For the early detection of signs of self-harm (Task 2), our classifiers that operated at the level of posts were too sensitive, resulting in low precision. For the task that evaluated ability to measure the severity of the signs of depression, we found that our submissions did not outperform chance or simple predictions for three of the four metrics. As pre-trained language models improve, we are optimistic that transfer learning will accelerate progress in early risk prediction on the internet.

Keywords: Transfer learning, depression, natural language processing

1. Introduction

Globally, the World Health Organization found that depression is the largest contributor to years lived with disability [1]. Early detection is an important goal as it can improve treatment and outcomes. Online social media platforms allow people to share their thoughts and dialogue openly with others. These platforms generate large amounts of data about their users' online behaviour, which can be repurposed for the

development of technologies which can detect mental health disorders. For example, language signals from Facebook posts have shown to be more predictive of depression diagnoses than existing screening surveys [2]. The tracking of users' language over time is of particular interest for understanding the development of their mental health states. Access to large chronological sequences of writings is fundamental to facilitate systems which can identify early signs of risk and propose interventions. Furthermore, understanding and quantifying the severity of different symptoms underlying a mental health disorder can be helpful for targeted approaches to determine the appropriate actions for an intervention.

In 2018, impressive gains in language modelling have been realized by training large unsupervised systems on large datasets [3, 4]. These approaches perform well on language understanding tasks after fine-tuning to a specific dataset. Such fine-tuning enables transfer learning from a large corpus to a smaller labelled dataset.

The CLEF eRisk shared tasks [5] were created to critically assess methods for early detection of depression on the internet. We participated in two tasks of the eRisk 2019 Lab [6]: Task 2 – Early detection of signs of self-harm and Task 3 – Measuring the severity of the signs of depression. Both tasks provide chronological collections of texts extracted from Reddit for each user. In Task 2, no training data is provided for the development of an early risk detection system for signs of self-harm. The goal of the task is to process a user's texts item by item, in the order they were created, to mimic a system which could be deployed to sequentially monitor a user's interactions in an online social media platform. Evaluation of Task 2 takes into account the correctness of the system's decision as well as the delay taken to emit a decision (measured by ERDE metric defined in [7]). Task 3 consists of estimating the severity of depression given a user's set of writings on Reddit. Various symptoms of depression are assessed with 21 questions on the Beck Depression Inventory (BDI), and they are to be filled based on evidence found in the user's writings. This task is also unsupervised as no training data is provided.

Broadly, our approach was to use similar datasets to provide training labels for supervised machine learning. To leverage transfer learning, we primarily used features extracted from a language model that was learned via unsupervised training (GPT-1 [2]).

2. Data and Methods

2.1. Task 2: Early Detection of Signs of Self-harm

Supervised

Dataset.

The University of Maryland (UMD) Reddit Suicidality Dataset (version 1) was used as training data for this task. This dataset was obtained from Reddit and is focused on posts from the r/SuicideWatch subreddit [8]. It was annotated for potential instances

of suicidality by crowdsourced workers and experts. The Centre for Addiction and Mental Health Research Ethics Board approved the use of this dataset for these analyses.

To simplify the classification problem, we built a classifier that considered only single posts instead of making classifications across a user's history. This better fit with the UMD dataset that was annotated at the post level. We collapsed the a, b, c, and d labels that range from no to severe risk into binary labels by selecting various combinations of crowdsourced and expert labels. We also propagated labels of risk backwards in time for a specific user to learn early risk (2-5 posts back). For example, we set any post classified as low or higher risk by an expert or moderate or higher risk by a crowdsourced annotator to be a post from a depressed individual (positives). In addition, we would set the preceding three posts to also be from a depressed individual. We varied the amount of 'control' users/posts by selecting different proportions of users from the UMD dataset that did not post to the r/SuicideWatch subreddit. Primarily we trained on datasets that contained 33% positively labelled posts.

Features. To extract features, we used the GPT-1 (Generative Pre-trained Transformer version 1) language model that was trained on a corpus of books [3, 9]. We further fine-tuned the language model on the full UMD dataset using the finetune python library (3 epochs). For each post, we extract the 768 GPT-1 encoded features. We did not experiment with other feature sets as the 768 are fixed in the pretrained GPT-1 transformer that OpenAI provided.

Training. After constructing several different versions of the UMD dataset, we used AutoSkLearn to learn to classify at the post level on each [10]. We used a repeated stratified k-fold cross validator with macro F1 as the target metric (using 5 repeats of 5-fold). AutoSkLearn was set to run for 6 to 24 hours per experiment. Across our many AutoSkLearn experiments, we selected five with high F1 and approximated ERDE5 values while taking into account the diversity of the dataset used for training. These five classifiers were then used to predict a given post at evaluation time (independent of preceding posts).

2.2. Task 3: Measuring the severity of the signs of depression

Features. Similar to Task 2, we extracted features with GPT-1. The pre-trained GPT model was fine-tuned on the provided text from the 20 subjects in the eRisk dataset (3 epochs). We additionally extracted features with the Linguistic Inquiry Word Count tool (LIWC) [11], which calculated the proportions of words from each post belonging to various word categories. We used all available LIWC categories, resulting in 70 features per post.

Unsupervised. To predict the BDI responses in an unsupervised manner, we used two approaches. In the first approach, we used the sentence-level feature vectors of

each user's writings which were provided by GPT-1. We then aggregated the sentence-level feature vectors into single user-level feature vectors by calculating the mean of each feature. As a result, each user's writing history was summarized into a single vector of size 768. Using GPT-1, we also generated a feature vector for the responses of the BDI questions. In total, we had 90 feature vectors for the BDI questionnaire (19 questions × 4 possible answers + 2 questions × 7 possible answers each).

To complete the questionnaire for a user, for each question, we computed the Cosine Similarity between the user's feature vector and the feature vector of each of the possible responses to that question. The response that gave the highest Cosine Similarity value was selected to be the user's response to the question.

One disadvantage of this approach is that a single 768-dimension vector may not adequately capture specific context and details. In particular, if the user has many sentences, representing the long history by only one feature vector seems inaccurate.

In our second approach, we did not aggregate the sentence-level feature vectors into a single user-level feature vector. Instead, to predict a user's response to each of questions, we computed the Cosine Similarity of each of the possible responses to all of this user's sentences. We picked the answer that had the highest Cosine Similarity to any of the sentences that this user had ever written. This is a nearest neighbour approach that asks which possible response to the question is closest to what the subject has previously written in the space of the GPT-1 features. This approach is more computationally intensive; however, it considers all sentences.

Supervised

Dataset. To train our Task 3 models, we used data from a study that one of the team members conducted during her PhD. In this study, undergraduate Psychology students filled out the BDI and several other questionnaires. The students also completed sessions of expressive writing, in which they wrote their thoughts and feelings about a negative event or personal difficulty [12]. Data were available from 236 students (197 females, 39 males). Their mean age was 19.00 years (*SD*=1.61), and their mean BDI score was 12.57 (*SD*=8.67). The McMaster University Research Ethics Board approved the studies, and all students provided their consent for their data to be used in research.

*Training***.** To generate predictions for each user's response to a given question on the BDI, we first extracted LIWC features from each user's post and averaged the features across all their posts. Next, we trained support vector machines (with linear kernels and L2 regularization) using the labelled (expressive writing) dataset (i.e., on LIWC features extracted from the texts and the corresponding responses or labels). We used the resulting model to generate a predicted response to that question for each user, and we repeated this process for each question on the BDI. For this model, we used the caret package in R.

We also tested a second supervised approach that used GPT-1 features and AutoSklearn. For these submissions, we used features extracted with GPT-1 that was fine-tuned on the writings of the 20 test subjects. Based on our unsupervised approach, we tested features representing a user's average GPT-1 vector and closest

relationships between each possible response and a user's writings. For these relationships, we used the minimum Euclidean distance and the maximum correlation between a user's input sentences and all possible responses in the BDI questionnaire. This resulted in 180 features. AutoSklearn was then applied to the average features (768 features), relationship features (180) and the combined set (948) to learn a classifier for each question. We again used a 5 repeats of stratified 5-fold cross validation with accuracy as the target metric.

3. Results

3.1. Task 2: Early Detection of Signs of Self-harm

While our five classifiers classified the majority of the user writings in under two days, it lacked precision. This is evidenced by our systems recalling 90%-100% of the true positive posts while achieving 12% precision. In terms of error, compared to the lowest ERDE5 of 0.08 from the UNSL team, our lowest ERDE5 value was higher at 0.15. Our best scoring classifier was trained on posts marked as moderate or severe risk by the experts and only severe risk by the crowd-sourced annotators. These positive training labels were extended to the preceding five posts.

3.2. Task 3: Measuring the severity of the signs of depression

Unsupervised. Our first approach, which generates a single user-level feature vector for each person, did not perform well on our labelled dataset and was not submitted. On the test dataset, the results of applying this approach are: average hit rate (AHR): 21.4%, average closeness rate (ACR): 55.9%, average difference between overall depression levels (ADODL): 79.0%, depression category hit rate (DCHR): 35.0%.

Our second approach, which uses the distance between the answers and all the sentences of a user's writing history, does slightly better. On the test dataset, the results of applying this approach are: AHR: 23.8%, ACR: 57.1%, ADODL: 81.0%, DCHR: 45%. At test time, this approach had higher ADODL and DCHR scores than all other submissions. In contrast, the per question AHR and ACR metrics were the lowest. Plotting of our scores on the background of 1,000 randomly generated submissions resolved this contradiction (Figure 1). Specifically, the ADODL and DCHR scores were overlapping with these random submissions. For our ADODL scores, 34 of 1,000 random runs achieve a higher score, for DCHR, 120 of the random runs are more accurate. While 34 or 3.4% is low, it doesn't take into account multiple submissions. If five random runs were submitted, there is a 17% chance that one would have a higher ADODL score. In summary, our unsupervised submission did not significantly perform better than chance.

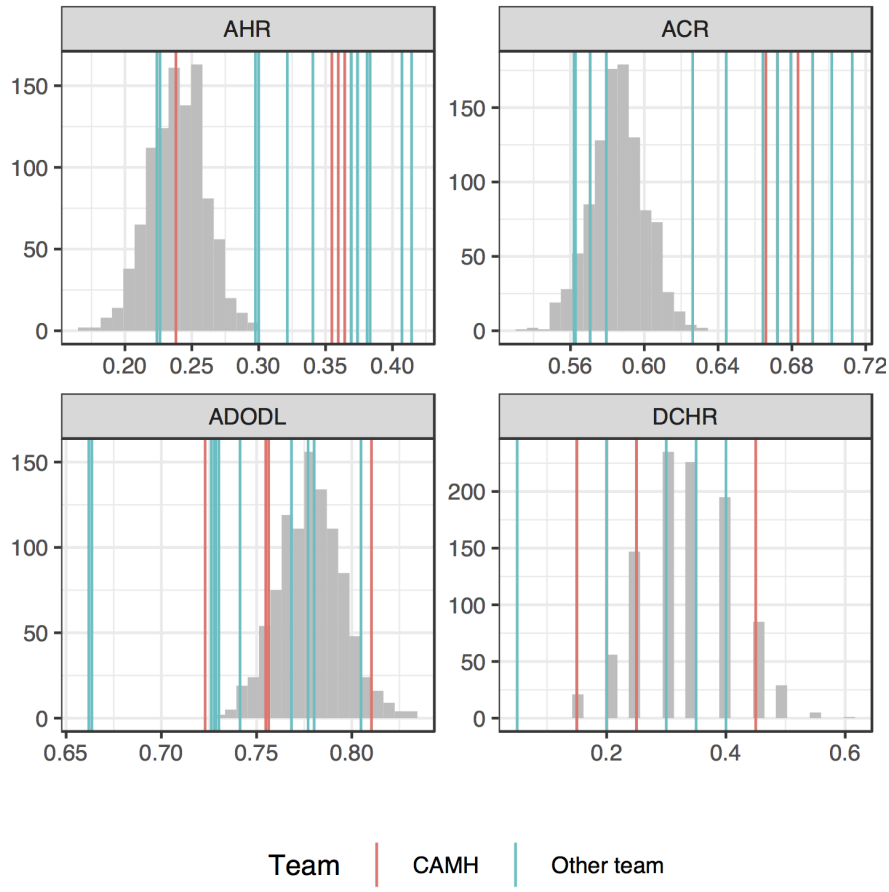

Fig. 1. Histograms of randomly generated submissions with team submissions marked by vertical lines. Average hit rate (AHR), average closeness rate (ACR), average difference between overall depression levels (ADODL), and depression category hit rate (DCHR) are plotted separately with different axes. One thousand randomly submissions are visualized with grey histograms. Our submissions are marked by vertical lines in red and other teams in blue.

Supervised. Our supervised submissions, like all other submissions, are not accurate at predicting depression category (DCHR) or overall score (ADODL) beyond chance levels. However, the per question metrics are high. Our submission that used AutoSklearn on the GPT-1 average and relationship features was ranked 7th of the 18 submissions for the AHR metric and 4th for the ACR metric. Analysis of this submission reveals little diversity in its predicted responses. For seven of the 21 questions, it predicted the same answer for all subjects. Building on this observation, we note that a submission of purely '1' gives an ACR score of 71.75%, surpassing all the submissions on this metric.

4. Discussion

In Task 3, our supervised approaches may not have performed well due to differences between our labelled dataset and the Task 3 posts. The labelled dataset was generated by Psychology students (mostly young women) attending a large Canadian university, and associations between their writing and depression scores may not generalize to an online sample. The students in the study were asked to engage in 20 minutes of focused journal-type writing in a lab setting, where they specifically described their thoughts and feelings about a negative personal issue. The content of the Task 3 texts was likely more diverse since it was compiled from a variety of user interactions on social media. The students also completed the BDI immediately prior to writing, so their responses and the writing content corresponded to the same underlying mental state. However, posts for Task 3 were compiled over time, and they may reflect a variety of mental states and interactions that do not always correspond to depression severity (or responses on the BDI). Observational studies of individuals with depression suggest that depressive symptoms tend to fluctuate over time [13, 14], making it very challenging to predict depression severity based on a history of posts. For these or other reasons, the associations between the features extracted from the expressive writing texts and responses on the BDI captured by our models may not have been applicable to the Task 3 posts or users. Nonetheless, we believe additional annotated data will improve depression prediction.

5. Conclusions and Future Work

For Task 2, which sought to detect early signs of self-harm, our transfer learning approach that used pre-trained language and classification models was too sensitive. This is likely because our system made predictions at the level of single posts and did not consider the preceding posts by a specific user. Also, we suspect our training data was not well matched with the task even though it was also from reddit. Taking into account temporal information about the input text should be investigated to extract recent signals.

We conclude that Task 3 is very difficult even when training data is available. We found that none of the submissions were able to predict overall depression better than chance. In addition, a high ACR can be reached by submitting a response of 1 for every question and subject. In our case, this was partially learned by our supervised approaches, which produced homogeneous predictions.

As pre-trained language models improve, we are optimistic that transfer learning combined with supervised learning will accelerate progress in early risk prediction on the internet.

Author Contributions

DH and LF designed and implemented our approach to Task 2. MM, VM and SG collected and transcribed the data used to train the Task 3 classifiers. DH and MM extracted features from the Task 3 data. PA, LF, MM and SP designed and

implemented our approaches to Task 3. Error analyses on our Task 3 results were performed by PA and LF. LF supervised the project.

Acknowledgments

The CAMH Specialized Computing Cluster, which is funded by The Canada Foundation for Innovation and the CAMH Research Hospital Fund, was used to run AutoSklearn. We thank the NVIDIA Corporation for the Titan Xp GPU that was used for this research. We acknowledge the assistance of the American Association of Suicidology in making the University of Maryland Reddit Suicidality Dataset available.

References

1. Friedrich MJ.: Depression is the leading cause of disability around the world. JAMA 2017;317(15):1517.
2. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proc. Natl. Acad. Sci. U. S. A. 115, 11203–11208 (2018).
3. Radford, A., Narasimhan, K., Sutskever, I., Salimans, T.: Improving Language Understanding by Generative Pre-Training. (2018).
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805, (2018).
5. Losada, D.E., Crestani, F., Parapar, J.: Early Detection of Risks on the Internet: An Exploratory Campaign. In: Advances in Information Retrieval. 41st European Conference on Information Retrieval, ECIR 2019. pp. 259-266. Springer International Publishing, Cologne, Germany (2019).
6. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019).
7. Losada, D.E., Crestani, F.: A Test Collection for Research on Depression and Language Use. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 28–39. Springer International Publishing (2016).
8. Shing, H.-C., Nair, S., Zirikly, A., Friedenberg, M., Daumé, H., III, Resnik, P.: Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 25–36. Association for Computational Linguistics, Stroudsburg, PA, USA (2018).
9. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 19–27 (2015).
10. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and Robust Automated Machine Learning. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 28. pp. 2962–2970. Curran Associates, Inc. (2015).

11. Pennebaker, J. W., Chung, C. K., Ireland, M. E., Gonzales, A. L., & Booth, R. J.: The Development and Psychometric Properties of LIWC2007. LIWC, Austin, Texas. (2007).
12. Pennebaker, J.W.: Writing About Emotional Experiences as a Therapeutic Process. Psychol. Sci. 8, 162–166 (1997).
13. Tang, T.Z., DeRubeis, R.J.: Sudden gains and critical sessions in cognitive-behavioral therapy for depression. J. Consult. Clin. Psychol. 67, 894–904 (1999).
14. Kelly, M.A.R., Roberts, J.E., Bottonari, K.A.: Non-treatment-related sudden gains in depression: the role of self-evaluation. Behav. Res. Ther. 45, 737–747 (2007).