Evaluating Contextual Embeddings and their Extraction Layers for Depression Assessment

Matthew Matero **Albert Hung** **H. Andrew Schwartz**
Department of Computer Science
Stony Brook University
{mmatero, has}@cs.stonybrook.edu

Abstract

Recent works have demonstrated ability to assess aspects of mental health from personal discourse. At the same time, pre-trained contextual word embedding models have grown to dominate much of NLP but little is known empirically on how to best apply them for mental health assessment. Using degree of depression as a case study, we do an empirical analysis on which off-the-shelf language model, individual layers, and combinations of layers seem most promising when applied to human-level NLP tasks. Notably, we find RoBERTa most effective and, despite the standard in past work suggesting the second-to-last or concatenation of the last 4 layers, we find layer 19 (sixth-to last) is at least as good as layer 23 when using 1 layer. Further, when using multiple layers, distributing them across the second half (i.e. Layers 12+), rather than last 4, of the 24 layers yielded the most accurate results.

1 Introduction

Over the past decade natural language processing (NLP) has increasingly set its sights on interdisciplinary tasks, notably those within the computational social sciences (Sap et al., 2014; Preoţiuc-Pietro et al., 2016; Zamani et al., 2018). As more and more language has been generated on social media sites such as Facebook, Twitter, and Reddit, researchers have had a wealth of personal discourse available to them that spans across thousands of users.

Many researchers focus on applying these social media datasets to predict user demographics, personality, or mental health (Matero et al., 2019; Iyyer et al., 2014; Lynn et al., 2020). Those predicting facets of mental health, such as depression and suicide risk, can help an over-burdened mental health industry by using automated screening (Coppersmith et al., 2018). Often these automated tools can be applied to forums where a user is an active member and their account could be flagged to be brought to the attention of a moderator. Thus, a personalized and potentially early intervention could be provided to the user in question.

Here, we investigate one prominent aspect of mental health: *degree of depression* (DDep) as measured by answers to an online questionnaire administered to Facebook users. Depression assessment of social media users is of interest for the following reasons: (1) Depression is often highly correlated with suicidal tendencies (Leonard, 1974) with deaths by suicide on the rise (Curtin et al., 2016) and (2) Automated assessment of depression is of high importance as it is often an under-diagnosed ailment, where such predictions could be useful to screen individuals who are at risk (Eichstaedt et al., 2018).

While many recent NLP pipelines have moved onto leveraging large pre-trained language models based on the transformer architecture (Vaswani et al., 2017), applying these models to human-level analysis, such as predicting a person's states or traits, has received little attention. Even the use of extracted embeddings, often called contextual embeddings, has yet to be fully explored in this level of analysis (V Ganesan et al., 2021). We expand this area of research by investigating how best to leverage the individual layers of off-the-shelf transformer models for depression assessment. Notably, we are interested in going beyond just a single layer and propose a greedy algorithm for selecting layers to extract contextual embeddings and aggregate them for large user-level embeddings.

Our contributions include: (1) A predictive model for depression assessment that out-performs the current state-of-the-art, (2) Evaluation of standard extraction techniques on contextual embeddings and their ability to detect depression levels and (3) Analysis on the effectiveness of layer selection to generate large contextual embedding representations of users.

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 89 - 94

2 Related Works

One of the downsides when modeling mental health data is often that it is very small, with only a few hundred participants per study (Guntuku et al., 2017). However, it is sometimes possible to get around this by using data from Social Media websites where participants can choose to opt in to share past language data and take a small survey or questionnaire (Coppersmith et al., 2014). Schwartz et al. (2014) applies this technique to Facebook users and evaluates their DDep over a continuous scale (1-5) rather than bucketing users into classes such as mild/moderate/severe.

Even somewhat recent human-level models in NLP have used bag-of-words style approaches for prediction (Lynn et al., 2019; Andy et al., 2021), while other areas such as word or document-level tasks have adopted contextual embedding representations (Bao and Qiao, 2019; Babanejad et al., 2020; Matero et al., 2021). As these are often output from very large models, with hundreds of millions or more parameters, they are able to encode syntactic and semantic information that transfer to downstream tasks either through word or sentence embeddings (Guu et al., 2020).

While there has been some work applying contextual embeddings and transformer language models to human-level predictions, the most in depth has been V Ganesan et al. (2021) who investigated the use of contextual embeddings in low-data scenarios across various areas including mental health, demographics, and personality assessment. However, they only focus on using the base-size variants with an emphasis on dimensionality reduction techniques to apply contextual embeddings to small datasets (N <= 1000). Here, we work with a medium size dataset of 3 million Facebook posts across 25 thousand users and apply both base and large sized language models, as well as investigate layer selection beyond using just the second to last layer of the model.

3 Methods

Task: A person's degree of depression score is estimated by their response to a subset of neuroticism questions on a personality assessment through Facebook's MyPersonality app (Schwartz et al., 2013). The responses were on a scale of 1 to 5 and averaged together to represent a person's overall degree of depression. Here, we formulate the task of depression assessment as building a single user representation where each status is processed through a language model as a sentence and then all words from a user are avg-pooled. We evaluate our models using mean squared error(MSE) and disattenuated pearson r(r_{dis}) to account for questionnaire reliability (Lynn et al., 2018). We perform all experiments using the DLATK (Schwartz et al., 2017) library.

Model	r_{dis}	MSE
Baselines		
Open-Ridge	.507	.7696
Schwartz et al.	.526	N/A
AvgPool-XLNet	.499	.7728
AvgPool-BERT	.528	.7575
AvgPool-ALBERT	.508	.7675
AvgPool-RoBERTa	**.542***	**.7497***

Table 1: Performance of extracting embeddings from second to last layer (11) from *base* sized variants of each language model on the held-out test set. Each model is used to encode a 768 dimensional vector for all words that are then averaged to a user representation. **Bold** indicates best in column and * indicates statistical significance $p < .05$ w.r.t AvgPool-BERT via paired t-test.

Transformer Language Models: From the wide selection of general purpose language models, we select the following: XLNet, RoBERTa, ALBERT and BERT (Yang et al., 2019; Liu et al., 2019; Lan et al., 2019; Devlin et al., 2019). These models are chosen as they cover common language model types (e.g. autoregressor vs autoencoder), have been pre-trained on various corpus sizes, and in the case of ALBERT offer a more lightweight footprint in terms of total model parameters.

When comparing which language model to perform our layer analysis on, we first evaluate performance using only the second to last layer on our held-out test set. This allows us to deduce which model may lead to better application to aggregate human-level predictions.

Layer Selection: To decide on which layers to extract for our final model, we perform a 10-fold cross-fold validation, for each individual layer or combination of layers. First we select the best performing layer, once found, we then concatenate all other layers to find the best 2-layer combination. This process is iterated on until we reach a number of layers where we cease to see a performance increase via the cross-folds. Once the best

Model	Hid. Size	r_{dis}	MSE
RoBERTa-B L11	768	.542	.7497
RoBERTa-L L23	1024	**.543**	**.7476**
DistilRoBERTa L5	768	.533	.7545

Table 2: Performance of extracting embeddings from second to last layer of RoBERTa variants, which was found to be the best performing among base models, on the held-out test set. DistilRoBERTa is also considered as a small sized alternative. **Bold** indicates best in column.

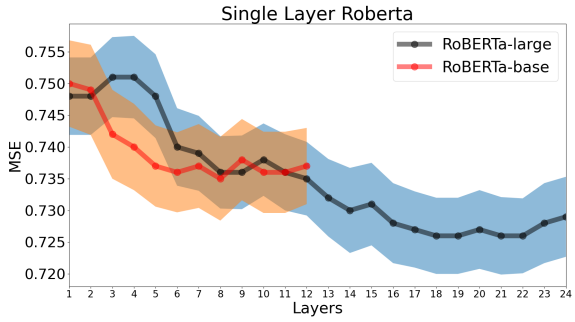

Figure 1: Layer-wise mean squared error performance across the 10-fold validation set with standard error shown by the shaded region for both RoBERTa-base and large. At lower layers (3-6), RoBERTa-base shows a much lower error rate. However at layer 13 and higher of RoBERTa-large there is lower error beyond any available base layer.

performing layers are found via cross-folds, we extract a final test set representation and run the final selection on our held-out test set. When comparing within cross-folds we only compare the MSE, rather than correlation, as that is the metric being optimized as well as being a less noisy evaluation of each model.

As well as our best performing layer combinations, for a final comparison on the test set, we also evaluate performance of standard layer extraction techniques. This includes the second-to-last layer and the concatenation of the top-4 layers enabling us to validate that our layer selection method and suggested layers are worthwhile.

Regression: Our model of choice is a regularized linear regression (ridge) with input being the mean aggregate of extracted contextual embeddings. To find the regularization parameter α, we use a 10-fold cross-validation technique searching between 10 and 1 million, increasing by powers of 10 each time, then selecting the α that gave the lowest mean squared error. A simple predictive model is chosen to highlight the improvements from the features themselves rather than any specific network architecture.

4 Dataset & Baseline

Dataset: The dataset is comprised of Facebook users who opted in to share their status updates between 2009 and 2011 and completed a personality questionnaire (Schwartz et al., 2014). There are 25,000 train users and 1,000 test users which are then filtered down to those who wrote at least 1,000 words across all of their status updates. The final result is a training set of 17,599 and test set of 986 users.

Baseline: We compare to the proposed model of Schwartz et al. (2014) which leverages both open-vocab and count based lexicons. Notably, the model is trained on 1 - 3 grams, a 2000 dimensional social media LDA topic vector, Lexical Inquiry and Word Count (LIWC) lexicon, and NRC sentiment lexicon (Pennebaker et al., 2001; Mohammad et al., 2013). We compare our models both to the reported scores in the original publication and to a version we recreated, referred to as Open-Ridge.

5 Evaluation

Our recreated Open-Ridge came within .019 r_{dis} of the original work, however, both the recreated and original model are outperformed by both BERT and RoBERTa base variants, as shown in table 1. Interestingly ALBERT, while being 10x smaller than the other language models, performs quite well; outperforming XLNET and baseline models. We also see that all models based on the autoencoder style architecture (BERT variant) perform better than autoregressors (XLNet). This suggests that for human-level analysis the autoencoder style models are better than autoregressors, agreeing with the findings of V Ganesan et al. (2021).

We also compare against possible variants of RoBERTa, which offer a computation versus performance trade-off, RoBERTa-large (24 layers) and DistilRoBERTa (6 layers) in table 2. Ultimately, RoBERTa-large performs only slightly better than the base model. While this small difference is found to not be statistically significant, due to the number of available layers of RoBERTa-large this gives more options for layer selection without a loss in performance and move forward with RoBERTa-large as our selected model.

Rank	*1 Layer*		*2 Layers*		*3 Layers*		*4 Layers*		*5 Layers*		*6 Layers*	
1	**19**	**0.7257**	**16**	**0.7234**	**24**	**0.7215**	**22**	**0.7208**	**18**	**0.7206**	**14**	**0.7207**
2	18	0.7264	15	0.7241↓	22	0.7216	21	0.7210	17	0.7206	15	0.7207
3	22	0.7263	17	0.7241↓	23	0.7218	18	0.7210	15	0.7207	12	0.7207
4	21	0.7265	22	0.7242↓	21	0.7220	14	0.7211	14	0.7207	17	0.7207
5	17	0.7272	14	0.7242↓	20	0.7225↓	17	0.7211	21	0.7208	21	0.7208
6	20	0.7275	23	0.7246↓	18	0.7225↓	15	0.7211	12	0.7208	23	0.7208
7	23	0.7282↓	18	0.7246↓	14	0.7226↓	23	0.7211↓	13	0.7209	9	0.7211
8	16	0.7284↓	21	0.7247↓	17	0.7226↓	12	0.7212	23	0.7210↓	7	0.7211
9	24	0.7286↓	13	0.7247↓	15	0.7226↓	13	0.7213↓	20	0.7210↓	6	0.7213
10	15	0.7305↓	24	0.7248↓	12	0.7227↓	20	0.7213↓	10	0.7211	4	0.7215
Layers Included	–		*19*		*19;16*		*19;16;24*		*19;16;24;22*		*19;16;24;22;18*	

Table 3: Comparison of performance between the top 10 best individual layers and additional layers on the 10-fold cross validation data, ordered by mean squared error. **Bold** indicates best in column and ↓ indicates significantly lower performing models $p < .05$ via paired t-test compared to best in column (rank 1). The best performing of the previous column is used to find the next best layer to add on (via concatenation indicated by ;). During cross-folds training N=16,694 and validation N=905.

Layer Combo	r_{dis}	**MSE**
Standard		
L23	.542	.7476
L21+22+23+24	.546	.7479
Optimized		
L19	.553	.7439
L16+19+22+24	.552*	.7208*
Other Sizes		
L16+18+19+22+24	**.554***	**.7206***
L14+16+18+19+22+24	.553*	.7433*

Table 4: Performance of extracting embeddings using standard techniques and from the optimized layers we find to be most promising via cross-fold selection. **Bold** indicates best in column and * indicates statistical significance $p < .05$ w.r.t standard top-4 (21-24) layer extraction via paired t-test.

As mentioned in section 3, for investigating layer selection we only evaluate on cross-fold validation results to avoid any overfitting to the test set. First, we look at all individual layers of RoBERTa, as shown in in figure 1, and the standard errors associated with each layer's performance across the 10 cross-folds. We find that performance slowly improves as you move up the model but begins to slow down around the middle layers and peaks at layer 19.

Next, we explore the question of how many layers should be used as well as which layers to extract in order to build a user representation. For this, we apply our layer selection technique based on empirical results of the cross-folds. We show results for the top 10 best combinations per layer amount in table 3. We find 3 interesting outcomes from our experiments: (1) When using only a single layer the second-to-last is not the best and is not even in the top 5, (2) We do not see a drop in performance from using more than 4 layers, in fact, we do not see a plateau until we try 6 total layers thus suggesting that for human-level predictions large representations are ideal and (3) The layers that boost performance all come from the top half of RoBERTa-large likely due to them including more semantic information than syntactic (Rogers et al., 2020), which could be more informative for modeling at the human-level.

Lastly, we compare our optimized extraction models to the standard approaches on the held-out test set; shown in table 4. We find that our layer 19 model performs quite well but is not a statistically significant finding (p=.08) when compared against layer 23. Our 4-layer model continues to give a boost in performance and is found to be statistically significant compared to standard top-4 extraction. The 5-layer version has a small improvement in both metrics and is found to be significant(p=.02) compared to our optimized 4-layer model. For the 6-layer model we see an expected drop in performance, based on cross-fold analysis, suggesting that the additional layer has hurt the model's ability to generalize.

6 Conclusion

With many tasks in NLP focused around human-level prediction, methods that can use state-of-the-art, off-the-shelf models in the best way are of interest to the community at large. In this work, we found that applying pre-trained transformer language models to depression assessment benefited from non-standard extraction techniques. Fur-

ther, applying a straight forward empirical analysis of layer performance could lead to noticeable boosts in downstream applications. Ultimately, we achieved sate-of-the-art performance of $r_{dis} = .554$ and $MSE = .7206$ using a 5-layer user representation from RoBERTa-large.

Ethics Statement: Our work is part of a growing body of interdisciplinary research that aims to improve the automatic assessment of a person's mental health. However, at this time we do not suggest our model(s) be used in practice to label mental health states. Instead, this should be viewed as a step toward a clinical tool that would be used with professional oversight. This research has been approved (deemed exempt status) by an academic institutional review board.

References

Anietie U Andy, Sharath C Guntuku, Srinath Adusumalli, David A Asch, Peter W Groeneveld, Lyle H Ungar, and Raina M Merchant. 2021. Predicting cardiovascular risk using social media data: performance evaluation of machine-learning models. *JMIR cardio*, 5(1):e24473.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243.

Xingce Bao and Qianqian Qiao. 2019. Transfer learning from pre-trained BERT for pronoun resolution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 82–88, Florence, Italy. Association for Computational Linguistics.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *Eighth international AAAI conference on weblogs and social media*.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Sally C Curtin, Margaret Warner, and Holly Hedegaard. 2016. Increase in suicide in the united states, 1999–2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019: The Annual Meeting of the North American Association for Computational Linguistics*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Sharath Chandra Guntuku, David B 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.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

CV Leonard. 1974. Depression and suicidality. *Journal of consulting and clinical psychology*, 42(1):98.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Veronica Lynn, Niranjan Balasubramanian, and H Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, New Orleans, LA. Association for Computational Linguistics.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammadzaman Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2021. MeLT: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2959–2966, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Mohammadzaman Zamani, H Andrew Schwartz, Veronica E Lynn, Salvatore Giorgi, and Niranjan Balasubramanian. 2018. Residualized factor adaptation for community social media prediction tasks. *arXiv preprint arXiv:1808.09479*.