2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871120

Ensembles of BERT for Depression Classification

Saskia Senn[1], ML Tlachac[2], Ricardo Flores[2], and Elke Rundensteiner[2]

***Abstract*— Depression is among the most prevalent mental health disorders with increasing prevalence worldwide. While early detection is critical for the prognosis of depression treatment, detecting depression is challenging. Previous deep learning research has thus begun to detect depression with the transcripts of clinical interview questions. Since approaches using Bidirectional Encoder Representations from Transformers (BERT) have demonstrated particular promise, we hypothesize that ensembles of BERT variants will improve depression detection. Thus, in this research, we compare the depression classification abilities of three BERT variants and four ensembles of BERT variants on the transcripts of responses to 12 clinical interview questions. Specifically, we implement the ensembles with different ensemble strategies, number of model components, and architectural layer combinations. Our results demonstrate that ensembles increase mean F1 scores and robustness across clinical interview data.**

***Clinical relevance*— This research highlights the potential of ensembles to detect depression with text which is important to guide future development of healthcare application ecosystems.**

I. Introduction

Depression is one of the most prevalent mental illnesses, according to World Health Organisation (WHO) [1]. The number of people living with this mental illness increased by more than 18% between 2005 and 2015. Approximately 280 million people in the world are living with a depressive disorder [1]. As a result of an ongoing depression, the abilities of a person in performing daily activities can be critically decreased and negatively impact the patients life severely. In the worst case, depression can lead to suicide. Every year, more than 700'000 people globally die by suicide which is the fourth leading cause of death among people aged 15-29 [1]. Early detection is crucial for the prognosis of depression treatment [2], nevertheless diagnosis is difficult so depression often remains undiagnosed for many years [3]. Thus, research [4], [5], [6], [7], [8], [9], [10] has begun exploring the modeling of voice recordings and transcripts as a strategy to detect depression earlier.

For instance, Audio-Assisted BERT (AudiBERT) [8] was recently leveraged to optimize depression classification from clinical interview recordings. While the Bidirectional Encoder Representations from Transformers (BERT) models [11] for text classification were combined with audio classification architectures, the ablation study indicates that the BERT component was most influential to AudiBERT's success. However, RoBERTa, a more robustly trained BERT model, has demonstrated better performance than BERT for mental health applications [12]. Also, previous studies have revealed that ensembles of models can produce more robust classifications than individual models [13], [14].

Thus, within the scope of this research, we investigate the ability of BERT variants and BERT ensembles to classify depression from the transcripts of responses to 12 clinical interview questions. In particular, we hypothesize that ensembling several BERT variants will result in better performance. We compare the depression classification ability of:

1) three different individual BERT variants,
2) two different ensemble method strategies,
3) ensembles containing different BERT models, and
4) three different combinations of architectural layers.

II. Clinical Interview Transcript Data

For this research, we use the transcripts from the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [15], [16]. The DAIC-WOZ corpus consists of 189 clinical interviews conducted by a virtual agent. Participants were virtually asked a subset of *core* questions with varying amounts of follow-up questions to elicit more details [16].

The interviews are labeled with PHQ-8 depression screening scores. The PHQ-8 contains the first eight Likert scales in the PHQ-9 [17]. The PHQ-8 score ranges from 0 to 24 with a score of 10 being indicative of depression.

We treat each core question in DAIC-WOZ as an individual thematic dataset, as defined in the related literature [8]. Each dataset contains the responses to a single core question and related follow-up questions. In this research, we use the 12 thematic datasets with the most responses, as detailed in Table I. These datasets contain 94 to 105 responses with 21% to 31% of respondents labeled as depressed (PHQ-8$\geq$ 10).

III. Deep Learning Methodology

Our goal is to predict depression with transcripts of responses to clinical interview questions asked by a virtual agent. To accomplish this, we explore the depression screening capabilities of 3 BERT variants and 4 ensembles of BERT variants. Further, we compare two different ensemble method architectures. We focus on BERT classifiers in this research given their demonstrated success with smaller datasets and mental health applications [12], [8].

This work was supported by Fulbright Foreign Student Program, National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2015-56150007, and US Department of Ed. P200A180088: GAANN Fellowship. Results were obtained using a computing cluster acquired with NSF MRI grant DMS-1337943 to WPI.

[1]Saskia Senn is with Applied Computational Life Sciences, Zürcher Hochschule Angewandte Wissenschaften (ZHAW), Wädenswil, ZH, Switzerland sennsaskia@gmail.com

[2]ML Tlachac, Ricardo Flores, and Elke Rundensteiner are with the Departments of Data Science and Computer Science, Worcester Polytechnic Institute (WPI), Worcester, MA 01604, USA {mltlachac,rflores,rundenst}@wpi.edu

TABLE I

THEMATIC DATASET DESCRIPTIONS.

Core Question Description	Count	Depressed
How are you doing today?	105	28.6%
The last time you argued with someone?	103	29.1%
What advice would you give yourself?	102	28.4%
What are you most proud of?	100	28.0%
How are you at controlling your temper?	100	30.0%
When was the last time you felt really happy?	99	28.3%
How easy is it for you to get good sleep?	98	27.6%
How would your best friend describe you?	96	26.0%
What's your dream job?	95	30.5%
What'd you study at school?	95	30.5%
Do you travel a lot?	94	27.7%
Have you been diagnosed with depression?	94	21.3%

A. Individual BERT Variants

BERT. The Bidirectional Encoder Representations of Transformers is a pretrained model for language representation [11]. BERT was revealed to have a superior performance due to its architecture leveraging multi-layer bidirectional Transformer encoder combined with multiple attention heads. The model is pretrained on BooksCorpus and English Wikipedia (16GB) using Masked Language Modeling (*MLM*) and Next Sentence Prediction (*NSP*). This pretraining allows for previously unprecedented success when classifying smaller datasets.

RoBERTa. A previous study [12] suggests that the Robustly Optimized BERT Pretraining Approach (RoBERTa) model [18] is better than BERT for mental health applications. The training approach for RoBERTa is different from BERT base model as there is no *NSP* and instead an extented *MLM* procedure is integrated. Pretrained on a bigger corpus than BERT, RoBERTa includes more informal text data in the pretraining such as a Reddit corpus.

DistilBERT. DistilBERT [19] is a less computationally costly alternative to the BERT base model. Retaining 97% of BERT's language understanding and general-purpose pretraining, DistilBERT reduces the size of the BERT model by 40%. Thus, to reduce computational costs, DistilBERT is more attractive than the BERT base model in ensembles.

B. Ensemble Strategies

Simple Averaging. For the simple averaging ensemble strategy, the final logit result was created by averaging the numeric results in the classification layer of the models, as depicted in Fig.1. As with the individual models, the resulting average was then classified with a threshold of 0.5.

One Final Classifier. For the one final classifier ensemble strategy [8], we concatenated the CLS tokens from all of the individual classifiers in the ensemble, as depicted in Fig.1. The final classification layer was then applied to this concatenated vector to output the final logit for the ensemble.

C. Ensembles of BERT Models

We experimented with four different ensembles, noted in Table 2. In particular, to determine the impact of the ensemble strategy, we ensembled BERT and RoBERTa with both strategies, resulting in Ens 1 and Ens 2. Further, we determined whether BERT can be replaced with the less computationally costly DistilBERT by comparing the results of Ens 2 and Ens 3. Lastly, we explored the impact of adding a third model to the ensemble by comparing Ens 2 and Ens 3 with Ens 4.

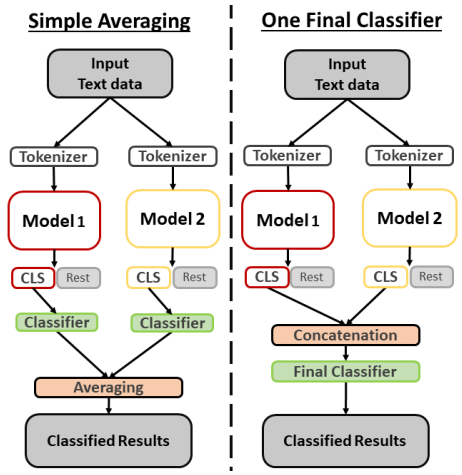

Fig. 1. Comparison of the two ensemble method strategies.

TABLE 2

THE COMPONENTS COMPRISING EACH OF THE DIFFERENT ENSEMBLES.

Ensemble	Method	BERT	RoBERTa	DistilBERT
Ens 1	Simple Averaging	✓	✓	
Ens 2	One Final Classifier	✓	✓	
Ens 3	One Final Classifier		✓	✓
Ens 4	One Final Classifier	✓	✓	✓

D. Combinations of Architectural Layers

We fine-tuned our models to increase classification ability by implementing each of the aforementioned classifiers with three different combinations of architectural layers. Notably, these included the base model (B), base model with an extra self-attention layer (BA), and the base model with an LSTM layer and an extra self-attention layer (BLA). This fine-tuning was performed on top of the concatenated vector for the *One Final Classifier* ensembles and on top of each CLS token for the *Simple Averaging* ensembles and BERT variants.

E. Model Implementation and Evaluation

Due to our hyperparameter tuning, we ran models with batchsize=8, learning rate=$2x10^{-5}$, and step size=$2x10^{8}$. Our models used 128 tokens, a cross entropy loss function, and an Adam optimizer with weight decay. Each model is trained for 10 epochs. For each thematic dataset, the test sets contain 20% of the responses [8]. To mitigate bias from unequal classes, training sets were upsampled. Experiments are repeated 10 times with randomly initialized weights.

Model performance is evaluated based on metrics calculated with the number of true positive TP, false positive FP, false negative FN, and true negative TN predictions. Our goal is to maximize $F1$ (Eq. 1), a popular metric for diagnostic tasks with unbalanced data. $F1$ is the harmonic mean between the positive predictive value and sensitivity.

TABLE 3

MEAN ± STANDARD DEVIATION ACROSS ALL 12 THEMATIC DATASETS. B = BASE MODEL, A = ATTENTION LAYER, L = LSTM LAYER.

Model Architecture	F1			Sensitivity			Specificity		
	B	BA	BLA	B	BA	BLA	B	BA	BLA
BERT	0.35±0.29	0.47±0.24	0.60±0.18	0.31±0.27	0.45±0.25	0.61±0.19	0.52±0.39	0.62±0.30	0.60±0.25
RoBERTa	0.58±0.22	0.43±0.33	0.47±0.31	**0.73**±0.29	0.56±0.44	0.54±0.36	0.30±0.32	0.17±0.29	0.34±0.32
DistilBERT	0.54±0.21	0.50±0.18	0.59±0.16	0.49±0.22	0.44±0.20	0.59±0.18	0.73±0.22	0.74±0.19	0.62±0.23
Ens 1	0.52±0.18	0.49±0.20	0.60±0.13	0.47±0.19	0.43±0.22	0.62±0.15	0.75±0.18	**0.76**±0.16	0.61±0.21
Ens 2	0.53±0.17	0.49±0.20	**0.62**±0.12	0.49±0.17	0.44±0.21	0.64±0.14	0.74±0.18	**0.76**±0.18	0.61±0.20
Ens 3	0.51±0.18	0.51±0.18	**0.62**±0.13	0.47±0.19	0.48±0.19	0.64±0.14	0.73±0.19	0.72±0.18	0.61±0.26
Ens 4	0.49±0.21	0.48±0.20	0.60±0.14	0.46±0.21	0.44±0.22	0.63±0.15	0.75±0.19	0.75±0.18	0.57±0.20

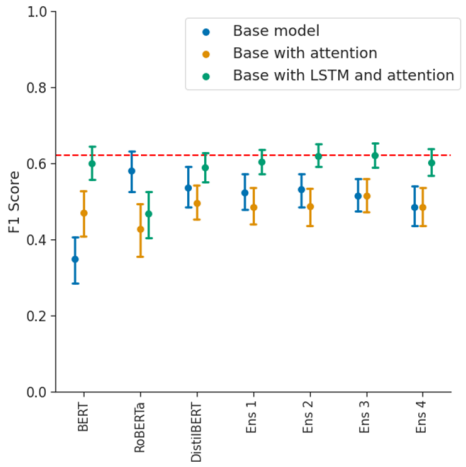

Fig. 2. Mean $F1$ ± standard deviation for all thematic datasets. Red line indicates highest mean value of 0.62.

Also popular for diagnostics, $sensitivity$ is the true positive rate and $specificity$ is the true negative rate. We report the metrics for models with the five highest F1 scores.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TN + FN};\ Specificity = \frac{TN}{TN + FP} \tag{2}$$

F. Availability

Upon publication, we will release the ensemble code at *github.com/sennsaskia/EnsemblesBERT.git* Further research updates will be available at *emutivo.wpi.edu*.

IV. RESULTS

The models' performances aggregated across the 12 thematic datasets are shown in Table 3 and Fig. 2. We observe the highest mean $F1$ of 0.62 for our ensemble models Ens 2 (BLA) and Ens 3 (BLA). These ensembles also have the lowest standard deviations, indicating robustness. Both ensembles have mean sensitivities of 0.64 and specificities of 0.61 which are well balanced for inversely related metrics. Given that DistilBERT in Ens 3 is less computationally costly than BERT in Ens 2, we consider Ens 3 to be the most preferable of our ensembles.

BERT Variants. The individual BERT (BLA), RoBERTa (B), and DistilBERT (BLA) models performs almost as well as best performing ensembles with mean $F1$ scores of 0.60, 0.58, and 0.59, respectively. However, the standard deviation of these individual models are larger so they are less robust than the ensembles. Unlike for the other individual models, RoBERTa performed better with the basic architecture than with additional LSTM and attention layers. The extra fine tuning layers had the most impact for the individual BERT model, increasing the mean F1 from 0.35 to 0.60.

Ensemble Strategies. To determine the effect of the ensemble strategy, we compared the $F1$ scores for Ens 1 with the Simple Averaging strategy and Ens 2 with the One Final Classifier strategy. Ens 2 (BLA) has a slightly higher mean $F1$ than Ens 1 (BLA), thus we proceed with the Simple Averaging ensemble strategy for further ensembles.

Ensembles of BERT Variants. The effect of different model combinations is further explored by comparing the $F1$ scores of Ens 2 and Ens 3. Both ensembles (BLA) achieved mean F1 of 0.62 and similar standard deviations. Hence, our results suggest that BERT can be successfully substituted in ensembles by the less computationally expensive DistilBERT. Interestingly, adding a third model to the ensemble decreased the mean F1 slightly to 0.60, so ensembling more models is not necessarily better.

Architectural Layers. The performance of all the ensembles are greatly improved by adding addition LSTM and attention layers (BLA); the mean F1 scores are increased and the standard deviations decreased. While also true for BERT and DistilBERT, adding both extra layers notably decreases the mean F1 of RoBERTa. Apart from BERT, adding an extra attention layer without an LSTM layer is not advantageous.

Individual Thematic Datasets. Fig.3 compares the $F1$ scores for the 12 individual thematic datasets. We observe that the ensembles (BLA) have very small standard deviation and never perform comparatively poorly for any dataset, though they may not always achieve the highest mean F1 scores. This indicates that the ensembles produce robust results, even for small datasets. However, there is a huge variation in depression detection capabilities across different datasets. For example, all models for the blunt diagnosed_depression dataset have very high mean $F1$ scores, up to 0.93 for BERT (BLA); the ensembles achieved a maximum mean $F1$ of 0.92 with smaller standard deviation. The ensembles (BLA) slightly outperformed BERT (BLA) for the controlling_temper dataset with mean F1 scores up to 0.65. This dataset highlights the importance of adding LSTM and attention layers to ensembles.

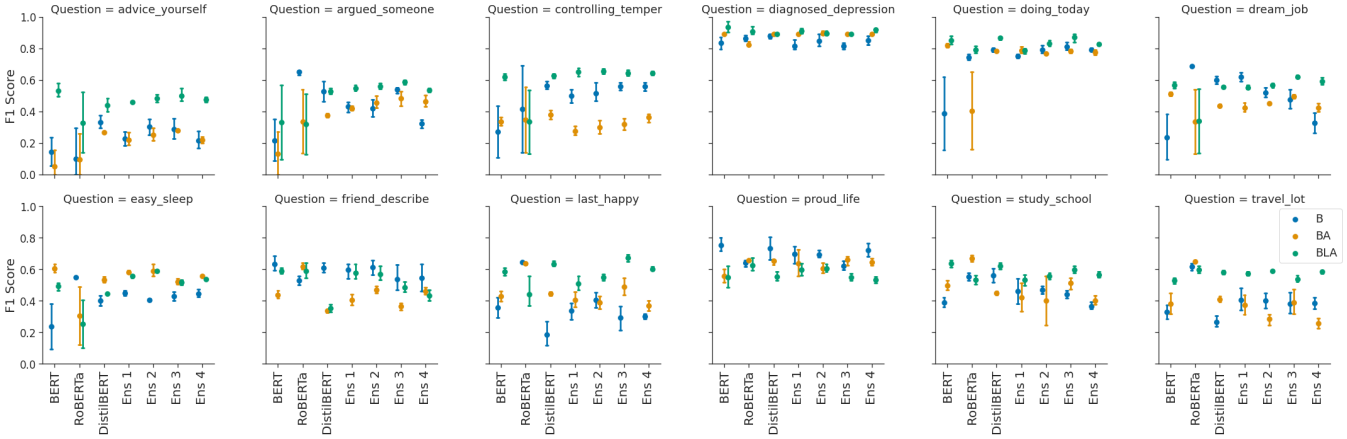

Fig. 3. Mean $F1 \pm$ standard deviation for each dataset, ordered alphabetically. B = base model, A = attention layer, L = LSTM layer.

V. Discussion, Limitations, & Future Work

The results support our hypothesis and previous research [13], [14] that ensembles of BERT variants perform better than individual models for depression classification. While the mean $F1$ scores of our ensembles (BLA) were only slightly higher than those of the best performing individual models, our ensembles were notably more robust, as evidenced by their small standard deviations. As suggested by related literature [12], the base RoBERTa model performed decently for the mental health datasets. RoBERTa did not benefit from fine tuning so BERT (BLA) surpassed it's performance. Yet, RoBERTa proved to be a critical component of the successful ensembles (BLA).

AudiBERT [8] improved depression detection ability by combining BERT with an audio model. We discovered that ensembling BERT models can also improve classification. Thus, integrating multiple text models with multiple audio models could further enhance the performance of AudiBERT.

While DAIC-WOZ [15], [16] is the best available dataset for this research, the small number of participants limits model performance. Our ensemble models were quite stable on the small datasets but collecting larger datasets will help further establish the robustness of our ensembles in this domain. While we used transcripts, our ensembles are applicable to any text data such as social media posts.

VI. Conclusion

Our research demonstrates that ensembling text classification models indeed improves performance for depression screening. In particular, we recommend ensembling AudiBERT and DistilBERT with One Final Classifier strategy for this task. While not always advantageous for individual models, fine-tuning with an additional LSTM and attention layers before classification benefited the ensembles. This research could guide future development of successful depression classification models for healthcare application ecosystems.

Acknowledgment

We thank Ermal Toto, Souyma Joshi, and the DAISY lab at WPI. We thank Claus Horn at ZHAW.

References

[1] World Health Organization, "Depression and other common mental disorders: global health estimates," Tech. Rep., 2017. [Online]. Available: https://apps.who.int/iris/handle/10665/254610

[2] A. Halfin, "Depression: the benefits of early and appropriate treatment," *American Journal of Managed Care*, vol. 13, no. 4, 2007.

[3] R. M. Epstein *et al.*, ""i didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress," *Journal of general internal medicine*, vol. 25(9), 2010.

[4] N. Cummins *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, 2015.

[5] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[6] M. Rodrigues Makiuchi *et al.*, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[7] E. Toto *et al.*, "Audio-based depression screening using sliding window sub-clip pooling," in *IEEE ICMLA*, 2020.

[8] E. Toto, M. Tlachac, and E. A. Rundensteiner, "AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4145–4154.

[9] R. Flores *et al.*, "Depression screening using deep learning on follow-up questions in clinical interviews," in *IEEE ICMLA*, 2021.

[10] M. L. Tlachac *et al.*, "Emu: Early mental health uncovering framework and dataset," in *IEEE ICMLA Special Session Machine Learning in Health*, 2021.

[11] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv*, 2019.

[12] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and Classification of mental illnesses on social media using RoBERTa," *arXiv*, 2020.

[13] H. Dang *et al.*, "Ensemble BERT for Classifying Medication-mentioning Tweets," in *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 2020.

[14] S. Ghosh and A. Chopra, "Using Transformer based Ensemble Learning to classify Scientific Articles," *arXiv*, 2021.

[15] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews." in *Proceedings of the 9th International Conference on Language Resources and Evaluation*. CiteSeer, 2014.

[16] D. DeVault *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 2014.

[17] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, 2001.

[18] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*, 2019.

[19] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv*, 2020.