Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction Supplemental Material

Ayaan Haque[1], Viraaj Reddi[1], and Tyler Giallanza[2]

[1] Saratoga High School, Saratoga, CA, USA
[2] Department of Psychology & Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

Abstract. This supplemental material document include additional ablation studies on specific tests.We make our code and dataset available at https://github.com/ayaanzhaque/SDCNL

A Baseline Model Evaluations

Appendix A displays the full experimentation of baseline results for 6 transformers and 7 classification algorithms (Table A). The results are evaluated with 5 metrics: Accuracy (Acc), Precision (Prec), Recall (Rec), F1 score (F1), and Area Under Curve (AUC). With these results, we selected the four strongest performing combinations of models to perform the remainder of the experimentation.

B Comparison to Conventional Task

Appendix B presents an additional table showing the performance of our 4 best models on both the conventional task of suicidal vs clinically healthy classification against our task of suicidal vs depression classification. The results demonstrate the difficulty of our task and justify our clinical motivations.

C Vectorizers and Clustering

Appendix C shows how the number of extracted features from the vectorizers was chosen (Figure 1). The AUC scores from the MNB classifier after using the vectorizers with different number of embeddings are shown. After around 400 features for each model, the performance converges, so we finalized on extracting 768 features for consistency. Figure 2 visualizes the clustering of a GMM using BERT embeddings with PCA reduction.

Embedding Model	Metrics	Classifiers						
		CNN	Dense	BiLSTM	GRU	MNB	SVM	LogReg
BERT	Acc	72.14	70.50	71.50	71.50	57.78	68.07	68.60
	Rec	73.99	71.92	67.78	68.91	53.37	73.58	70.98
	Prec	72.18	70.77	74.28	73.86	59.54	66.98	68.50
	F1	72.92	71.25	70.70	71.05	56.28	70.12	69.72
	AUC	76.35	75.43	77.11	76.66	54.21	55.43	54.72
SentenceBERT	Acc	68.65	68.87	69.55	70.77	59.37	68.34	63.85
	Rec	73.37	74.61	67.98	67.36	46.63	73.06	69.95
	Prec	67.88	67.94	71.22	73.35	63.83	67.46	63.08
	F1	70.40	70.82	69.41	70.01	53.89	70.15	66.34
	AUC	73.52	73.70	74.00	74.99	56.13	53.12	51.33
GUSE	Acc	72.66	72.24	72.82	73.19	69.39	71.50	71.50
	Rec	78.96	76.37	78.03	77.10	67.36	75.65	74.61
	Prec	70.79	71.38	71.36	72.31	71.04	70.53	70.94
	F1	74.62	73.61	74.49	74.52	69.15	73.00	72.73
	AUC	77.82	77.76	77.41	76.33	47.68	49.47	50.75
TFIDF	Acc	67.12	69.28	67.97	67.92	69.39	70.45	69.13
	Rec	81.35	67.78	72.33	74.61	76.68	74.09	73.58
	Prec	65.98	70.73	67.41	66.61	67.58	69.76	68.27
	F1	71.89	69.19	69.64	70.21	71.85	71.86	70.83
	AUC	70.04	75.23	71.70	72.23	51.65	51.38	51.35
CountVec	Acc	69.18	68.92	68.13	67.86	66.75	65.43	63.32
	Rec	82.07	73.47	75.23	74.82	72.54	69.43	66.84
	Prec	65.91	68.32	66.62	66.41	65.73	65.05	63.24
	F1	73.06	70.61	70.59	70.31	68.97	67.17	64.99
	AUC	74.44	73.34	72.66	72.30	51.47	51.08	47.35
HashVec	Acc	67.86	67.44	65.49	65.17	65.96	66.23	66.75
	Rec	71.50	69.02	66.74	68.19	69.43	69.95	72.54
	Prec	67.58	67.79	66.34	65.30	65.69	65.85	65.73
	F1	69.27	68.31	66.17	66.47	67.51	67.84	68.96
	AUC	71.47	71.63	68.98	69.58	52.94	54.06	52.85

Table 1. Performance of all 42 combinations. 7 classifiers and 6 embedding models are used, with 4 deep classifiers and 3 deep embedding models.

Metrics (%)	Proposed				**Standard**			
	guse-dense	bert-dense	bert-bilstm	bert-cnn	guse-dense	bert-dense	bert-bilstm	bert-cnn
Acc	72.24	70.50	71.50	72.14	92.28	57.46	92.78	93.11
Prec	76.37	71.92	67.77	73.99	93.56	80.70	92.16	92.26
Rec	71.38	70.77	74.28	72.18	92.46	58.28	94.54	95.07
F1	73.61	71.25	70.70	72.92	93.00	67.39	93.35	93.63
AUC	77.76	75.43	77.11	76.35	96.65	54.91	97.24	96.49

Table 2. Comparison of classification metrics between conventionally researched task of suicide vs clinically healthy against our proposed task of suicide vs depression. The better performing category is bolded. The standard task performs far better on the same models, highlighting how our proposed task is more difficult to categorize.

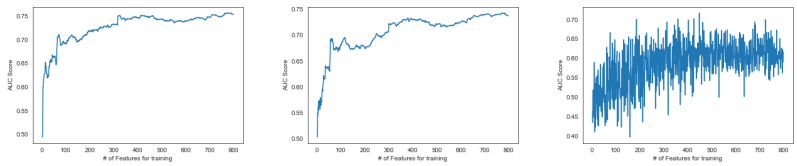

Fig. 1. Area Under Curve (AUC) values at different number of extracted word embeddings/features from the three vectorizers (TFIDF, CVec, HVec) inputted into Bayesian classifier. AUC plateaus at around 400 features, so we used 768 features to be consistent with other word embedding models.

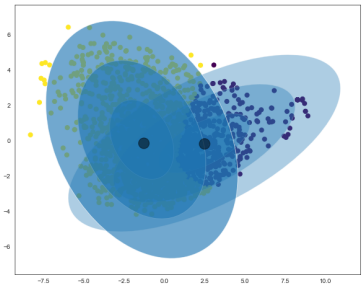

Fig. 2. GMM clustering using BERT embeddings and PCA reduction to 2 dimensions show the difficulty of the clustering task, as there is little variety in the clusters and they heavily overlap. We use co-variance type "full".