A New Training Method for Text Classification in Mental Health: LatentGLoss

Korhan Sevinç(✉)

TOBB University of Economics and Technology, Ankara, Turkey
ksevinc@etu.edu.tr

Abstract. Mental health text classification remains difficult when decisions must rely on short, naturally phrased user language. We study a seven-way task on a benchmark built by merging nine public mental-health text collections into one label schema. The paper reports a staged comparison: linear and ensemble models on TF–IDF features, convolutional networks, recurrent networks (including LSTM and GRU), and fine-tuned BERT-family encoders, so that performance can be read as a progression across representation choices. The main contribution is DualLatentGNet, a teacher–student framework in which a dual-objective CNN teacher (classification plus reconstruction) supervises a student. Unlike typical distillation pipelines that emphasize softened class targets, we route supervision through both the teacher's predictions and its latent geometry: a Gaussian mixture model is fit to concatenated teacher latent vectors and logits, and the student objective adds a latentG term tied to the assigned mixture component and to Euclidean distance in that feature space. On the merged benchmark, DualLatentGNet reaches 95.09% accuracy compared with 94.83% for the strongest fine-tuned BERT baseline under the same protocol, which supports the view that explicit latent alignment complements the baseline hierarchy we evaluate.

Keywords: mental health · text classification · teacher-student · knowledge distillation · latent representations

1 Introduction

Mental well-being promotion is central to the action plans of both the World Health Organization (WHO) for the years 2013–2020 [22] and the European Union's European Pact on Mental Health and Well-being [5]. The key breakthrough in addressing mental health issues lies in developing tools for early detection and preventive measures [8]. The WHO action plan highlights the importance of health policies and programs that not only address the needs of those affected by mental health disorders but also work to preserve mental well-being. These policies emphasize evidence-based, non-pharmacological interventions, focusing on early intervention and preventing unnecessary institutionalization and medicalization. Successful interventions are particularly dependent

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 261–273, 2027.
https://doi.org/10.1007/978-3-032-30494-0_17

on how frequently the therapy can be accessed [15]. In this context, automated systems have significant advantages over traditional therapies due to their ability to provide continuous support at minimal additional cost. Consequently, health assistants capable of delivering therapeutic interventions have gained considerable attention in recent years [1,7]. However, these systems are predominantly based on manually designed rules. In contrast, research in statistical methods for conversational systems has largely been limited to narrow-domain information-seeking dialogues [6,16,20,23].

We focus on inferring mental-health category cues from language people already produce—conversations, chat logs, forum posts, and similar text—because such material can surface signals between infrequent clinical contacts. For this work we did not collect new interviews; instead we merged nine publicly released datasets into one benchmark with a shared text field and seven labels, applied a single normalization and back-translation pipeline, and then trained a broad set of models so that conclusions are tied to one harmonized resource (Sect. 3). Classical machine learning on TF–IDF features, Word2Vec-based convolutional and recurrent models (including LSTM and GRU), and fine-tuned transformers (BERT and related encoders) are all included, which lets us relate our proposal to both lightweight and contextual baselines (Sect. 5).

We then describe DualLatentGNet: a dual-objective teacher network supplies class logits and internal representations, and a student is trained with a loss that combines the usual classification term with mixture-based and distance-based signals derived from the teacher's latent space (Sect. 4). The aim is to transfer information through predictions *and* through where those predictions live in representation space, rather than through logits-only distillation alone. Code, data links, training configurations, and model checkpoints are maintained at https://github.com/korhansevinc/LatentG-Loss.

2 Related Work

Interest in computational support for mental health has motivated many text-classification studies in which online writing is treated as a weak but abundant signal [2]. Cohan et al. analyzed mental-health forum posts with traditional machine-learning models, including support vector machines, and showed that lexical choices, sentiment, and simple frequency cues remain informative even though unstructured social language is noisy. Their discussion underscores how feature engineering and classifier choice interact when the target condition is expressed indirectly.

Recurrent architectures address limits of bag-of-words views by modeling token order. Rea et al. used LSTM networks for mental-health-related text and reported gains over classical baselines when sequential context is preserved [21]. Dey and Desai combined LSTMs with GloVe embeddings to obtain stronger semantic vectors when large transformer fine-tunes are expensive to deploy [9]. The CNN, RNN, LSTM, and GRU baselines we train in Sect. 5 sit in the same methodological line: they test how far convolutional and recurrent encoders can go without contextual pre-training.

Pre-trained transformers fine-tuned in domain have become the default strong comparators in NLP. Saleela et al. adapted BERT to mental-health text and reported strong discrimination across conditions such as depression, anxiety, and stress [17]. We mirror that practice by fine-tuning BERT, DistilBERT, ALBERT, and RoBERTa on our merged corpus (Table 1).

Multimodal systems widen the input beyond text: Kumar et al. combined written language with speech features to exploit acoustic correlates of distress alongside wording [4]. Our scope is deliberately text-only so that any gain from DualLatentGNet is attributable to the training objective and architecture rather than to additional sensors.

Across these lines of work, two constraints recur: labels and domains differ across corpora, and high accuracy from raw sentences remains challenging when classes are rare or language is indirect. Some pipelines therefore augment text with engineered linguistic or clinical features. Here we keep the input as plain text and ask whether a student model can absorb richer supervision from a teacher's internal representations as well as from its class predictions.

Teacher–student and distillation methods are familiar in vision and general NLP, but mental-health text classification has seen less work in which the student is explicitly regularized toward the *geometry* of the teacher's latent space via a mixture density over teacher features. DualLatentGNet (Sect. 4) targets that gap: a Gaussian mixture is fit to concatenated teacher latents and logits, and density- and distance-based terms enter the student loss alongside standard classification, which also offers an additional handle when training data are finite relative to model capacity.

3 Dataset

The dataset for this study combines nine publicly available mental-health-oriented text resources: *Depression Reddit Cleaned*, *Human Stress Prediction*, *Predicting Anxiety in Mental Health*, *Mental Health Bipolar*, *Reddit Mental Health Data*, *Students Anxiety and Depression*, *Suicidal Mental Health*, *Suicidal Tweet Detection*, and *3k Conversations Dataset for Chatbot*. Sources span forum-style threads, microblog posts, and short dialogues; we mapped them into one table with a single text column per example and a shared seven-way label set. All splits were passed through the same preprocessing and augmentation pipeline—lowercasing; stripping brackets, hyperlinks, HTML or special tags, punctuation, and redundant whitespace—followed by back-translation (English–French–English with TextBlob) and a second pass through the same cleaner so augmented lines match the rest of the corpus. Classical models consume TF–IDF vectors, CNN and RNN baselines use pre-trained Word2Vec embeddings, and transformer runs use the respective subword tokenizers, as summarized in the Preprocessing list below and implemented in Sect. 5. Where class frequencies skewed the metrics, we applied resampling for classical baselines and imbalance-aware losses for deep models, as noted later in the paper.

The final version of the dataset consists of unique person IDs, their statements obtained from conversations, chat logs, and other text-based platforms,

and the corresponding mental health labels. The classification includes the following seven mental health categories:

- **Normal**
- **Depression**
- **Suicidal**
- **Anxiety**
- **Stress**
- **Bipolar**
- **Personal Disorder**

Each sample consists of one or a few sentences describing an individual's mental state. Representative excerpts are included in our public repository.

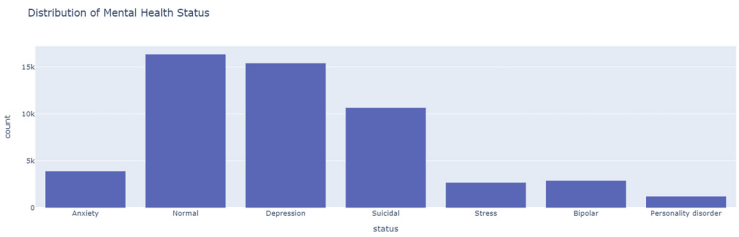

Fig. 1. Distribution of classes in the final dataset.

As shown in Fig. 1, the dataset is imbalanced, with Normal and Depression overrepresented. We mitigate this using imbalance-aware losses: Focal Loss [11], Dice Loss [14], and Tversky Loss [18]; for classical ML baselines we additionally test under-/over-sampling. We preserve proportional class distributions when splitting train/test and use k-fold cross-validation for robust evaluation.

Figure 2 shows that most samples are relatively short, consistent with conversational mental-health text.

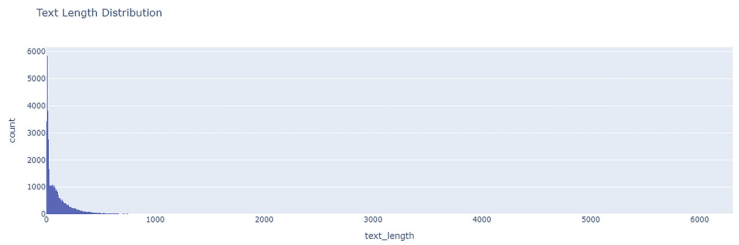

Fig. 2. Text Length distribution of the final dataset.

The combined and refined dataset, along with all the corresponding data, embeddings and TF-IDF vectors, are publicly available and the links for these sources can be found on our GitHub repository.

Preprocessing

- Text normalization: lowercase; remove brackets, hyperlinks, HTML/special tags, punctuation, and redundant whitespace.
- Data augmentation: back-translation EN→FR→EN using TextBlob.
- Representations: TF-IDF for classical ML; pre-trained Word2Vec embeddings [13] for CNN, RNN, LSTM, and GRU baselines; and the corresponding subword tokenizers for the transformer encoders (BERT, ALBERT, DistilBERT, RoBERTa).
- Class imbalance: optional under-/over-sampling for classical baselines and imbalance-aware losses [11,14,18] for deep models.

4 Method

In this section, we describe the methodology used to transfer knowledge from the teacher model to the student model, as well as the architecture and loss functions employed in the DualLatentGNet framework. Our approach leverages a teacher-student knowledge transfer mechanism, where the teacher model, trained with a CNN backbone and two decoders, provides guidance to the student model. This guidance is based on the teacher's feature vectors, which are passed through a Gaussian Mixture Model (GMM) to create a probability distribution. The student model is trained to match the teacher's feature distribution by minimizing the Euclidean distance and the difference in predicted Gaussian components. Additionally, we introduce a loss term $\mathcal{L}_{\text{latentG}}$ that combines the probability density function and Euclidean distance to shape student updates. We also use Focal, Tversky, and Dice losses as alternatives to cross-entropy for class imbalance during training.

Algorithm Explanation

The knowledge transfer is performed in three steps. First, we build teacher feature vectors by concatenating the teacher latent embeddings and prediction logits and fit a GMM over this feature space. Second, for each input, the student computes its feature vector and obtains (i) the most-likely GMM component PDF value and (ii) the Euclidean distance to the teacher feature representation. Finally, these stored quantities guide the student optimization through the latentG loss.

Units of the Architecture

The CNN backbone produces latent vectors that feed two decoders: a classification head and a reconstruction decoder. Training with both objectives shapes a discriminative latent space (see Fig. 3).

Algorithm 1: Teacher-Student Knowledge Transfer Algorithm

Input: Teacher feature vectors: `teacher_latent_v`, `teacher_pred_logits`; dataset (X, y)

Output: Student model predictions and stored PDFs and Euclidean distances.

```
teacher_feature_vectors ← concat[teacher_latent_v, teacher_pred_logits];
gmm ← FitGaussianMixtureModel(teacher_feature_vectors);
Initialize the Student Model;
for each (X, y) in dataset do
    student_feature_vector ← StudentModel(X);
    dist ← EuclideanDistance(student_feature_vector, teacher_feature_vectors);
    pred_pdf ← gmm.predict(student_feature_vector);
    pdf ← MostLikelyGaussian(pred_pdf);
    Store pdf, dist;
end
```

Intuition Behind the Method

The teacher is trained with a dual objective (autoencoding + classification) so that latent representations capture class-discriminative structure (see Fig. 3). We then concatenate teacher latent vectors and logits and fit a GMM to model the feature-space distribution across classes. During student training, the GMM provides the most-likely Gaussian component for each input; we compute both the component PDF value and the Euclidean distance between student and teacher representations, and use them to form the latentG-guided loss alongside the task objectives.

After calculating the pdf and euclid distance, we can denote our latentG (latentGaussian) term as follows:

$$\mathcal{L}_{\text{latentG}} = \alpha\,(1 - p) + \beta \cdot \text{Dist}_{\text{euclid}}$$

where α and β scale the two terms derived from the component PDF complement $(1 - p)$ and the Euclidean distance between teacher/student logits.

So the Dual architecture's total loss will be:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} \times \left(1 + \frac{e}{E} \times \mathcal{L}_{\text{latentG}}\right) + \text{MSE} \times \gamma$$

where e is the current epoch, and E is the total number of epochs. $\mathcal{L}_{\text{cls}}$ is the classification loss, and MSE is used for reconstruction; γ balances the two losses.

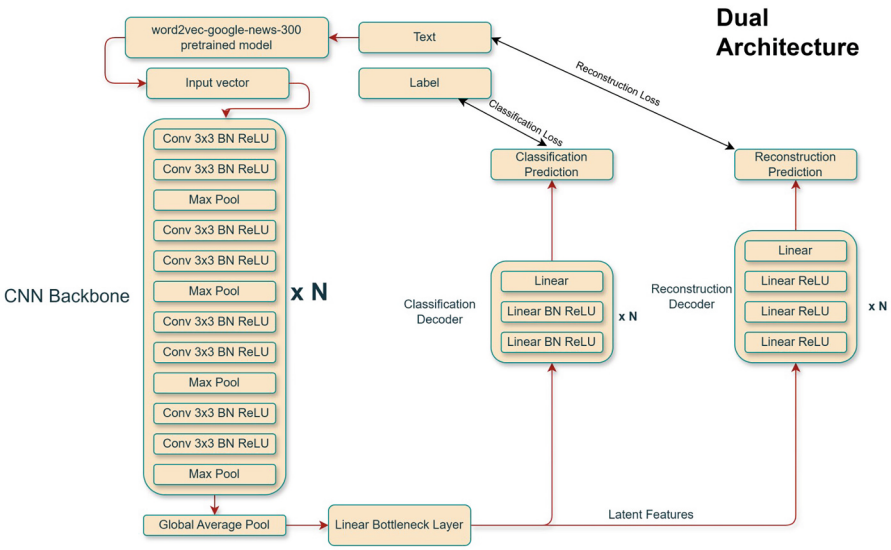

Fig. 3. Dual Architecture Diagram.

In imbalanced settings, we replace $\mathcal{L}_{\text{cls}}$ with Focal, Dice, or Tversky losses [11,14,18].

5 Experiments

This section summarizes how we instantiated and compared models on the merged benchmark. We trained classical TF–IDF pipelines, Word2Vec-based convolutional and recurrent networks, fine-tuned transformer encoders, and our proposed DualLatentGNet under identical splits and the preprocessing choices in Sect. 3. All scripts, notebooks, and hyperparameter configurations are archived in the GitHub repository linked in the introduction.

5.1 Machine Learning Models with TF-IDF Vectors

Using TF-IDF vectors, we train classical baselines (Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest, KNN, XGBoost, and Gradient Boosting). For each model, hyperparameters are selected via grid search with 3-fold cross-validation, and the resulting feature/class correlations motivate the latent-alignment idea used by DualLatentGNet.

5.2 Initial Deep Learning Experiments with Word2Vec Embeddings

We evaluate CNN and sequential baselines with Word2Vec embeddings: the cleaned text is mapped to 300-dim Word2Vec vectors and trained with TextCNN,

along with a lightweight autoencoder for latent extraction. We also train RNN, LSTM, and GRU models using a learnable embedding layer. These baselines capture temporal patterns, but their performance remains lower than the proposed DualLatentGNet and transformer variants.

5.3 Dual Architecture and Latent Representation Alignment

We train a teacher with a dual (autoencoding + classification) objective to shape discriminative latent vectors. We then fit a GMM on the concatenated teacher latent vectors and logits, and use the component PDF complement together with the Euclidean distance between teacher and student representations to form the latentG-guided student objective.

5.4 Transformer-Based Models

We fine-tune transformer baselines (BERT [3], ALBERT [10], DistilBERT [19], and RoBERTa [12]) using their respective pre-trained tokenizers.

5.5 Comparative Framework

DualLatentGNet is compared against classical ML baselines, CNN/RNN deep models, and transformer methods; metrics are reported in the next section.

6 Results

We report accuracy, precision, recall, and F1 on one held-out test split so that numbers remain comparable across model families. For each architecture we ran hyperparameter search and kept the best test result for that family; Table 1 collects these maxima. The training notebooks and configuration files used to produce the table are available on GitHub.

Table 1 summarizes the main numbers across all model families. Within the TF–IDF block, decision trees and random forests reach the low 93% accuracy range, SVM and logistic regression sit in the high 80s to low 90s, while naive Bayes, KNN, and several gradient-boosted configurations fall lower. Among recurrent baselines with trainable embeddings, LSTM reaches 94.14% accuracy and GRU 93.63%, with the vanilla RNN trailing at 88.36%. The fine-tuned transformer block is consistently strong: BERT peaks at 94.83% accuracy, with DistilBERT at 94.43%, RoBERTa at 93.38%, and ALBERT at 92.40%. The CNN-based block shows TextCNN at 91.41% and the dual-objective CNN teacher (DualTextCNN) at 93.01%. **DualLatentGNet** attains the highest scores in the table at **95.09%** accuracy and **95.10%** F1, with precision 95.28% and recall 95.09%, i.e., about 0.26% points above the strongest BERT baseline in absolute terms.

Table 1. Classification performance comparison of different models.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86.80	87.00	87.00	87.00
Naive Bayes	71.88	74.00	72.00	72.00
SVM	88.52	88.53	89.34	88.70
Decision Tree	92.25	92.00	92.00	92.00
Random Forest	93.41	94.00	93.00	93.00
KNN	39.35	59.00	39.00	30.00
XGBoost	74.76	75.00	75.00	74.00
Gradient Boosting	77.61	78.00	78.00	77.00
RNN	88.36	86.18	83.72	84.90
LSTM	94.14	93.61	92.06	92.81
GRU	93.63	93.43	91.29	92.30
BERT (Finetuned)	94.83	94.81	94.83	94.80
DistilBERT (Finetuned)	94.43	94.42	94.43	94.42
ALBERT (Finetuned)	92.40	92.39	92.40	92.37
RoBERTa (Finetuned)	93.38	93.37	93.38	93.37
TextCNN	91.41	91.43	91.41	91.38
DualTextCNN (Teacher)	93.01	92.99	93.01	92.97
DualLatentGNet (Proposed)	**95.09**	**95.28**	**95.09**	**95.10**

Table 2. Classification performance comparison of different loss functions.

Model	Loss	Accuracy	Precision	Recall	F1-Score
DualLatentGNet	Dice	94.30	94.43	94.41	94.42
DualLatentGNet	Tversky	94.56	94.75	94.22	94.49
DualLatentGNet	CE	**95.09**	**95.28**	**95.09**	**95.10**

1 CE: Cross Entropy

Table 2 reports an ablation in which the cross-entropy classification term inside DualLatentGNet is replaced by Dice or Tversky loss. Cross-entropy retains the highest accuracy at 95.09%, while Tversky reaches 94.56% and Dice 94.30%; the same ordering holds for precision, recall, and F1.

7 Discussion

Where the Gain Comes from. The 0.26-point margin of DualLatentGNet over the best BERT baseline is small in absolute terms but consistent across precision, recall, and F1 (Table 1). Because the only difference between the proposed model and its CNN teacher (DualTextCNN, 93.01%) is the addition of the latentG-guided student objective, the roughly two-point lift over the teacher and the

0.26-point lift over BERT are most readily attributed to that mechanism rather than to the underlying CNN backbone. We read this as evidence that explicit alignment to the teacher's latent geometry contributes signal that contextual token embeddings alone do not fully capture on this aggregate benchmark, but not as a claim that contextual modeling should be abandoned. The two views are complementary, and combining a transformer encoder with the latentG term is a natural next step.

Loss-Function Ablation. The ranking in Table 2 is informative even though the differences are modest. On highly imbalanced segmentation problems Dice and Tversky usually outperform cross-entropy precisely because they downweight easy majority pixels; here the opposite holds. A plausible reading is that latentG-guided alignment already reshapes the student's feature space so that minority-class samples are pulled towards the right Gaussian component, and the standard classification term then suffices. Adding Dice or Tversky on top appears to interact unfavorably with that geometric pressure on this split, rather than complementing it.

Recurrent vs. Contextual Baselines. The vanilla RNN lags the gated variants noticeably (Table 1), in line with the well-known difficulty of retaining information over variable-length posts with simple recurrence. LSTM slightly edges GRU here, plausibly because explicit memory cells help on mixed-length snippets. Even without pre-trained contextual embeddings, trainable embeddings combined with gated recurrence reach the low 94% range, which justifies keeping these models in the comparison alongside transformers and indicates how much of the headroom on this benchmark is reachable without large-scale pre-training.

Cost of the Proposed Method. A reviewer noted that the gain over BERT comes with a complexity increase, and we want to be explicit about it. Training Dual-LatentGNet requires three steps that fine-tuned BERT does not: (i) training the dual-objective CNN teacher with both classification and reconstruction heads, (ii) fitting a Gaussian mixture model on the concatenated teacher latent vectors and logits, and (iii) training the student under the additional latentG term, which involves a per-batch lookup of the most-likely Gaussian component and a Euclidean-distance computation against teacher representations. At inference time, however, only the student CNN runs, so deployment cost is on the same order as a single CNN forward pass and is markedly lower than running a fine-tuned BERT-base. The trade-off is therefore extra training cost and an additional fitted GMM in exchange for cheaper inference and the observed accuracy gain.

Threats to Validity. High offline accuracy on social-media text does not establish clinical utility. Mental-health categories such as *Suicidal* and *Personal Disorder* carry consequences that no aggregate accuracy number captures: false negatives matter more than false positives, calibration on the deployment population matters more than test-set ranking, and fairness across demographics matters more

than overall metrics. The seven categories in our label schema are also defined at a granularity that mixes diagnostic-style labels (e.g., *Bipolar*) with state-style labels (e.g., *Stress*); an inferred class is therefore best read as a routing cue toward human review rather than as a diagnosis. We also report point metrics on a single held-out split; reporting variance across seeds and folds, and uncertainty estimates per prediction, would strengthen the evidence base.

8 Conclusion

This paper assembled a seven-way benchmark by harmonizing nine public mental-health text sources under a single preprocessing and augmentation pipeline, compared classical TF–IDF models, CNN/RNN/LSTM/GRU baselines, and fine-tuned transformer encoders, and introduced DualLatentGNet, a teacher–student framework in which a Gaussian mixture is fit to the teacher's concatenated latent vectors and logits and a latentG term ties the student to that geometry alongside the standard classification loss. Under shared splits, DualLatentGNet reached 95.09% accuracy, narrowly above the strongest fine-tuned BERT baseline at 94.83% and about two points above its own CNN teacher at 93.01%, with the gain consistent across precision, recall, and F1.

We do not read this margin as a claim that contextual models should be replaced. The point of interest is that an explicit alignment to a teacher's latent geometry, in addition to its predictions, can carry signal that contextual token embeddings alone do not fully capture on this benchmark, while keeping inference cost on the order of a single CNN forward pass. The most informative next experiments are, in our view, an integration of pre-trained encoders into the student or teacher with the latentG term retained; reporting variance across seeds and folds together with calibration and per-prediction uncertainty estimates; and, for resource-constrained deployment, compression techniques such as pruning or quantization applied to DualLatentGNet (or to its teacher), provided they preserve the behavior observed here. Text classification for mental-health monitoring should remain a triage aid rather than a diagnostic substitute, and the work above is offered in that spirit.

References

1. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ. Couns. (2005). https://doi.org/10.1016/j.pec.2004.09.008
2. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N.: SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING), pp. 1485–1497 (2018). https://aclanthology.org/C18-1126/
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2019). https://doi.org/10.18653/v1/N19-1423
4. Drougkas, G., Bakker, E.M., Spruit, M.: Multimodal machine learning for language and speech markers identification in mental health. BMC Med. Inform. Decis. Mak. (2024). https://doi.org/10.1186/s12911-024-02772-0
5. European Commission: European pact for mental health and well-being (2008). https://health.ec.europa.eu/publications/european-pact-mental-health-and-well-being_en
6. Fatemi, M., El Asri, L., Schulz, H., He, J., Suleman, K.: Policy networks with two-stage training for dialogue systems. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2016). https://doi.org/10.18653/v1/W16-3613, https://aclanthology.org/W16-3613/
7. Fitzpatrick, K.K., Darcy, A., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Mental Health (2017). https://doi.org/10.2196/mental.7785
8. Insel, T.R., Scolnick, E.M.: Cure therapeutics and strategic prevention: raising the bar for mental health research. Mol. Psychiatry (2006). https://doi.org/10.1038/sj.mp.4001777
9. Kour, H., Gupta, M.K.: An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. Multimedia Tools Appl. (2022). https://doi.org/10.1007/s11042-022-12648-y
10. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019). https://arxiv.org/abs/1909.11942
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). https://arxiv.org/abs/1708.02002
12. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019). https://arxiv.org/abs/1907.11692
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013). https://arxiv.org/abs/1301.3781
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision (3DV) (2016). https://arxiv.org/abs/1606.04797
15. Olfson, M., Marcus, S.C., Druss, B., Pincus, H.A.: National trends in the use of outpatient psychotherapy. Am. J. Psychiatry (2002). https://doi.org/10.1176/appi.ajp.159.11.1914
16. Pietquin, O., Geist, M., Chandramohan, S.: Sample efficient on-line learning of optimal dialogue policies with Kalman temporal differences. In: IJCAI (2011). https://ijcai.org/Abstract/11/314
17. Saleela, D., et al.: BERT-based text classification pipeline for categorising mental health statements. In: Information Systems for Intelligent Systems (2026). https://doi.org/10.1007/978-3-032-12993-2_3, https://link.springer.com/chapter/10.1007/978-3-032-12993-2_3
18. Salehi, S.S.M., Erdogmus, D., Gholipour, A.: Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging (2017).https://doi.org/10.48550/arXiv.1706.05721, https://arxiv.org/abs/1706.05721

19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019). https://arxiv.org/abs/1910.01108
20. Schatzmann, J., Thomson, B., Young, S.: Statistical user simulation with a hidden agenda. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue (2007). https://doi.org/10.18653/v1/2007.sigdial-1.48, https://aclanthology.org/2007.sigdial-1.48/
21. Schoene, A.M., Lacey, G., Turner, A.P., Dethlefs, N.: Dilated LSTM with attention for classification of suicide notes. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pp. 136–145 (2019). https://doi.org/10.18653/v1/D19-6217, https://aclanthology.org/D19-6217/
22. World Health Organization: Mental health action plan 2013-2020 (2013). https://www.who.int/publications/i/item/mental-health-action-plan-2013-2020
23. Young, S., Gasic, M., Thomson, B., Williams, J.D.: POMDP-based statistical spoken dialogue systems: a review. Proc. IEEE (2013). https://doi.org/10.1109/JPROC.2012.2225812