Final Project report – Team 13 – EDS 6397 : Natural Language Processing

MENTAL HEALTH TEXT CLASSIFICATION USING NLP

Lekha Kancharla, Ruchith Reddy Parnem, Saketh Desini

I. INTRODUCTION

A lot of people now talk about their mental health on social media. They write short posts saying they feel depressed, anxious, stressed, or sometimes suicidal. There are thousands of such posts, so it is not possible for humans to manually read everything and identify who might need urgent help. Automatic text classification can be one step toward supporting early detection. In this project, we try to build models that can read such texts and predict what kind of mental state they show.

A. Problem Statement

The objective of the project is to train three different Neural Network models to perform mental health text classification on short user posts. Each model should be able to take a given text and assign it to one of the 7 labels, namely Normal, Depression, Suicidal, Anxiety, Stress, Bipolar and Personality Disorder. The output for each input post will be exactly one of these seven classes, showing the most likely mental health category expressed in the text.

B. Literature Review

The user-generated content serves as a valuable resource for understanding public sentiment regarding mental health[11]. Traditional text classification systems usually used bag-of-words or TF-IDF features with machine learning models like logistic regression or SVM [10]. These methods ignore word order and have trouble capturing context. Later, word embeddings like Word2Vec, GloVe and FastText [2] became popular[10]. They give each word a dense vector. RNN models such as LSTMs and GRUs can use these vectors and read sentences as sequences.

Several works have tried to detect mental health issues like depression or suicide risk from social media posts. Many of these systems are based on RNNs or CNNs with pretrained embeddings. They get reasonable results but struggle with long, noisy, or highly informal text. In recent years, transformer models such as BERT [6], RoBERTa [3] and DeBERTa [4] have become the main tools in NLP. These models are pretrained on very large corpora and then fine-tuned for specific tasks. RoBERTa improves BERT by training longer on more data and removing the next sentence prediction objective [8]. DeBERTa further improves the attention mechanism by separating content and position information [9]. Because these models work well on many NLP tasks, we decided to use them for mental health text classification and compare them with a simpler BiLSTM baseline.

II. DATASET

For this project, we have used *Sentiment Analysis for Mental Health* dataset from Kaggle [1]. Each row has an id, a Statement which was the text, and a Status which was the label. The seven labels are Normal, Depression, Suicidal, Anxiety, Stress, Bipolar and Personality Disorder. The posts mainly come from social media platforms like Reddit and Twitter. So the language is informal and sometimes noisy with slang and spelling mistakes.

Some posts, like “I am done with everything”, are ambiguous and could mean depression or suicidal. This label ambiguity is one reason why the models later confuse some classes, especially Depression and Suicidal.

The data is not evenly distributed across the seven labels. Normal and Depression have more examples compared to the other classes. While labels like Stress and especially Personality Disorder have much fewer samples. The distribution of the labels in the dataset is shown in *Figure 1*, which is a bar chart of the class counts.

We shuffled the full dataset and then split it into three parts. We used 64% of the data for training, 16% for validation and 20% for testing. We used a stratified split so that the class ratios stayed similar in all three sets.

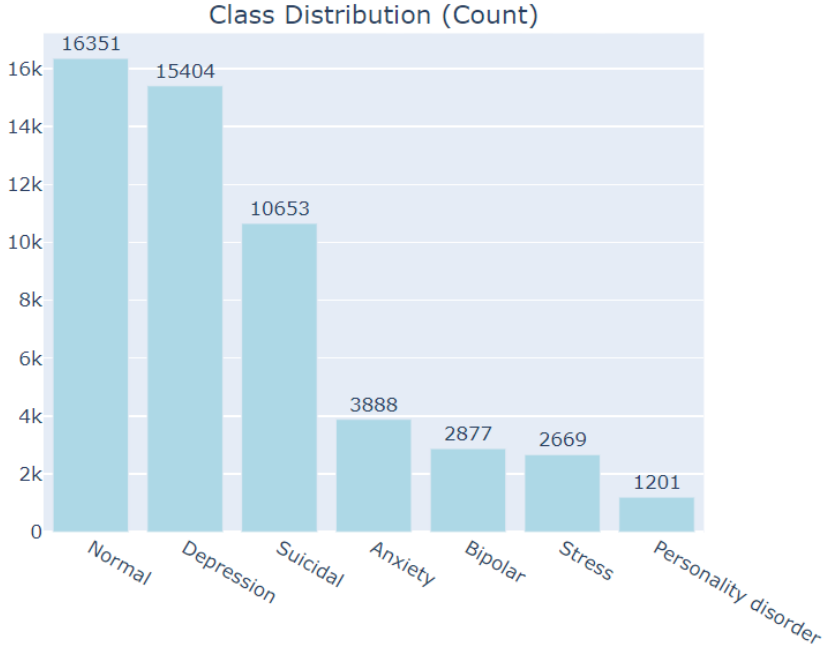


Figure 1: Bar chart showing the number of examples for each of the 7 mental health labels in the dataset.

III. METHODOLOGY

A. Pre-processing

We applied light pre-processing on the text. We removed extra spaces, line breaks and simple URL patterns like “http://…” or “https://…”. We did not remove punctuation or informal words, because they could carry important emotional information. For the BiLSTM model we converted the text to lowercase. For the transformer models (RoBERTa and DeBERTa) we have used their own tokenizers from Hugging Face [5].

B. Input Representation and Context Length

For the BiLSTM model, we used word-level tokenization. We built a vocabulary on the training set and mapped each word to an integer id. These integer sequences were then used as input to the embedding layer. For embedding we used Fasttext embeddings [2] of 300*1 vector.

For RoBERTa and DeBERTa we used the official tokenizers from Hugging Face. They split the text into subword tokens, added special tokens and returned token ids and attention masks. The context length (maximum sequence length) was a hyperparameter. We tried several values like 128, 256, 384 and 512 tokens. Smaller values like 128 and 256 often cut off important parts of longer posts and gave slightly worse validation scores. A larger value like 512 made training slower and used more memory without clear improvement. Based on these trials, we fixed the maximum sequence length at 384 tokens and after looking the token statistics our 95% of the token are around 309. For all models, we mapped the seven labels to integers from 0 to 6.

C. Training Setup and Loss Functions

All three models were implemented in PyTorch. For RoBERTa and DeBERTa we also used the Hugging Face Transformers library [5]. We trained on the training set and used the validation set to tune hyperparameters. For each model we tried three loss settings to handle class imbalance:

- Categorical cross entropy (CCE)
- Weighted CCE
- Oversampling with CCE

In the CCE runs we used the same loss for all classes. In the weighted CCE runs we gave higher weights to rare classes and lower weights to common classes. In the oversampling runs we sampled more examples from rare classes in each training batch while using CCE.

C. Neural Network Architecture

For classifying the mental health labels, the model is divided into four layers, excluding the input and output layer. First layer and second handles the tokenization, vectorization in Bilstm and embedding of the input sequence. This includes FastText for BiLSTM, built-in embeddings for transformer-based models. Third Layer is the Transformer/Bisltm Model, where we will use three NN models which include a RoBERTa_base, Bi-LSTM and DeBERTa_v3_Large. At the output, there is a classification head for the transformer and dense layer for the bilstm producing the output with softmax activation and classifying them into one of the 7. The overall diagram of the Model used for training the three different models is shown below in Figure 2 and Figure 3. The hyper-parameters used for each model are discussed in the following sections.

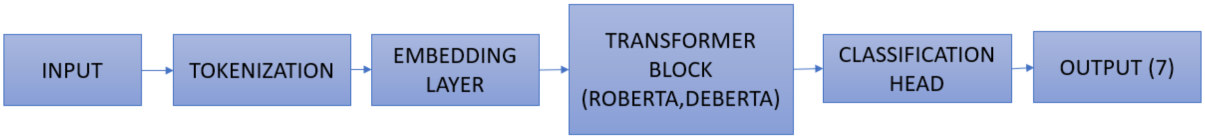


Figure 2: Model diagram of RoBERTa and DeBERTa

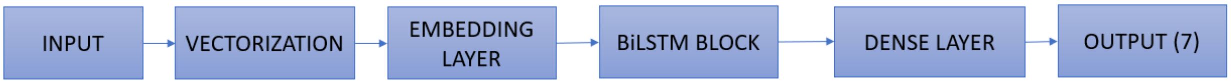


Figure 3: Model diagram of BiLSTM

i. MODEL 1: RoBERTa-base

- 12 transformer layers with hidden size 768
- 125M parameters
- Learning rate of 2×10^{-5}, batch size 16, epochs 5 with early stopping,weight decay 0.01 and context length 384

ii. MODEL 2: DeBERTa-v3-large

- 24 transformer layers with hidden size 1024
- 300M parameters
- Learning rate of 1×10^{-5}, batch size 16, epochs 5 with early stopping and weight decay 0.01 and context length 384

iii. MODEL 3: Bi-directional LSTM (BiLSTM)

- One BiLSTM layer with 64 units,
- followed by one LSTM layer with 32 units
- 5M parameters
- Adam optimizer, batch size 32, dropout 0.1, Epochs 20with early stopping based on validation macro F1

IV. RESULTS

Before talking about the results, we first explain how we counted the scores for each class.
For every post in the test set, we had a true label from the dataset and a predicted label from the model.
For a given class, for example Depression, we did the following:

- If the true label was Depression and the model also predicted Depression, we counted it as a True Positive (TP) for Depression.
- If the true label was Depression but the model predicted some other class, we counted it as a False Negative (FN) for Depression.

- If the true label was not Depression but the model predicted Depression, we counted it as a False Positive (FP) for Depression.

We repeated this for all seven labels- Normal, Depression, Suicidal, Anxiety, Stress, Bipolar and Personality Disorder. From TP, FP and FN, we then calculated precision, recall and F1-score for each label.
We also plotted confusion matrices and precision–recall (PR) curves for each model. The confusion matrix shows where the model is getting confused between labels. The PR curve shows how precision and recall change when we move the decision threshold and gives a better idea of performance for imbalanced data. The weighted CCE version was the best on the Test set for all the models. So we report that version in this results section.

i. MODEL 1: RoBERTa-base

Metric	Label (RoBERTa-base)							Micro	Macro	Accuracy
	Anxiety	Bipolar	Depression	Normal	Personality disorder	Stress	Suicidal			
Precision	0.87	0.84	0.84	0.96	0.79	0.75	0.71	0.84	0.82	84.33
Recall	0.89	0.86	0.74	0.95	0.76	0.84	0.81	0.84	0.83	
F1-Score	0.88	0.85	0.79	0.95	0.78	0.79	0.76	0.84	0.83	

Table 1: Precision, recall and F1-score for each label for the RoBERTa-base model

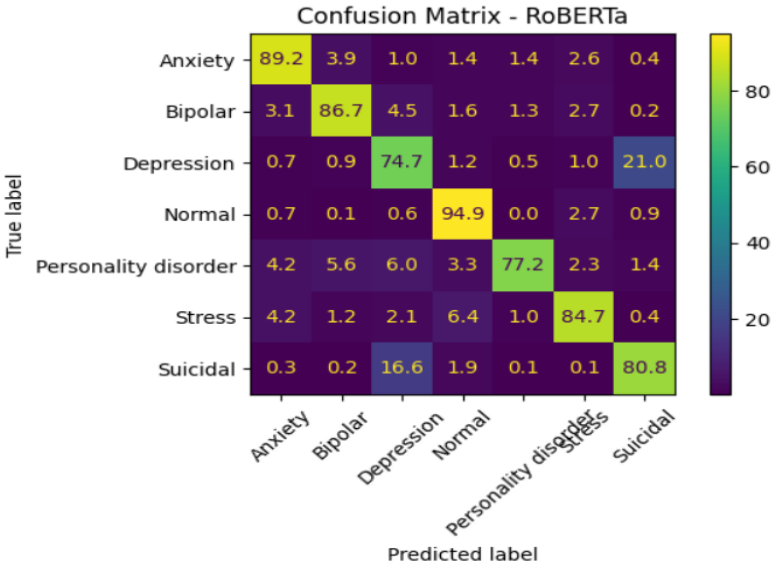


Figure 4: Confusion matrix for the RoBERTa-base model.

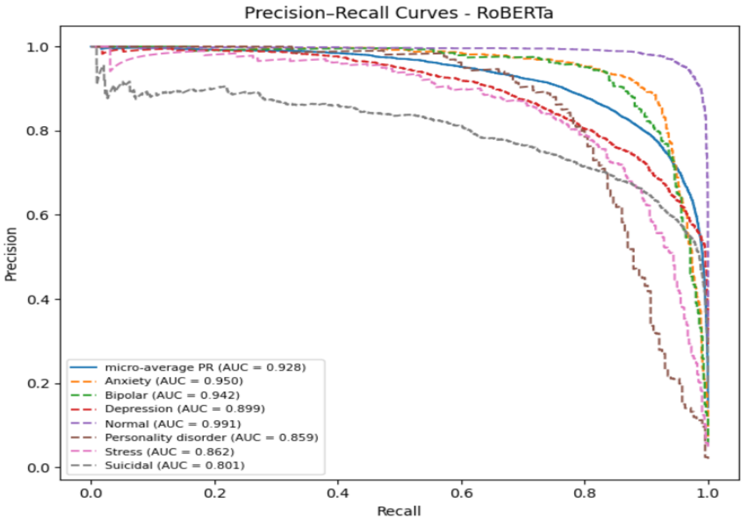
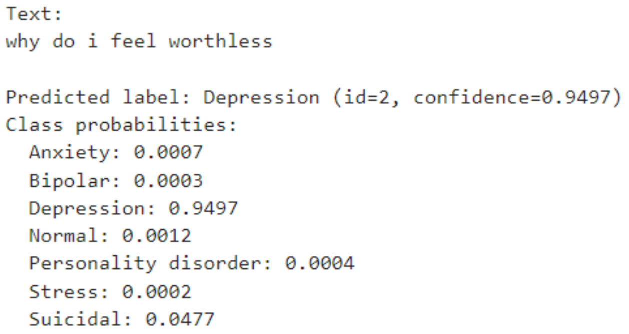


Figure 5: Precision–recall curves for the RoBERTa-base model.



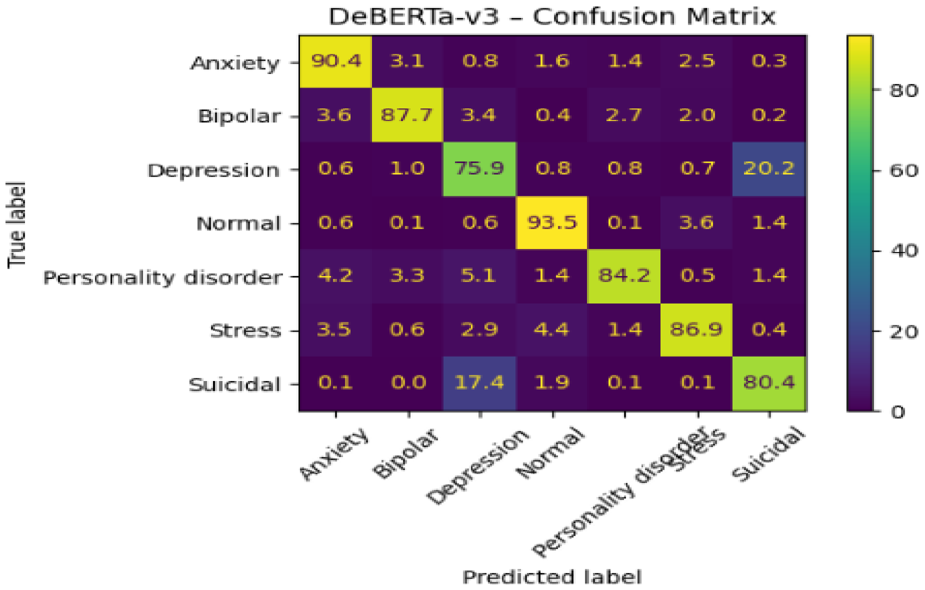
```
Text:
why do i feel worthless

Predicted label: Depression (id=2, confidence=0.9497)
Class probabilities:
  Anxiety: 0.0007
  Bipolar: 0.0003
  Depression: 0.9497
  Normal: 0.0012
  Personality disorder: 0.0004
  Stress: 0.0002
  Suicidal: 0.0477
```

Figure 6: Prediction for RoBERTa-base model.

ii. MODEL 2: DeBERTa-v3-large

Metric	Label (DeBERTa-v3-large)							Micro	Macro	Accuracy
	Anxiety	Bipolar	Depression	Normal	Personality disorder	Stress	Suicidal			
Precision	0.89	0.88	0.84	0.97	0.74	0.72	0.72	0.85	0.82	84.66
Recall	0.90	0.88	0.76	0.93	0.84	0.87	0.80	0.85	0.86	
F1-Score	0.90	0.88	0.80	0.95	0.79	0.79	0.75	0.83	0.84	

Table 2: Precision, recall and F1-score for each label for the DeBERTa-v3-large model



	Anxiety	Bipolar	Depression	Normal	Personality disorder	Stress	Suicidal
Anxiety	90.4	3.1	0.8	1.6	1.4	2.5	0.3
Bipolar	3.6	87.7	3.4	0.4	2.7	2.0	0.2
Depression	0.6	1.0	75.9	0.8	0.8	0.7	20.2
Normal	0.6	0.1	0.6	93.5	0.1	3.6	1.4
Personality disorder	4.2	3.3	5.1	1.4	84.2	0.5	1.4
Stress	3.5	0.6	2.9	4.4	1.4	86.9	0.4
Suicidal	0.1	0.0	17.4	1.9	0.1	0.1	80.4

(Figure 7: Confusion matrix for the DeBERTa-v3-large model.)

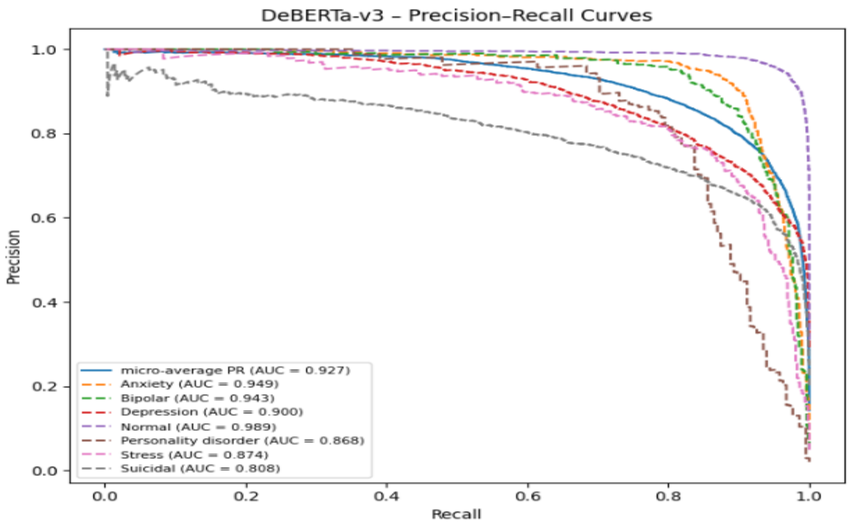
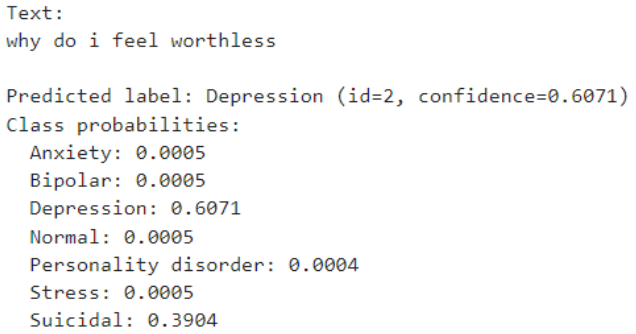


Figure 8: Precision–recall curves for the DeBERTa-v3-large model.



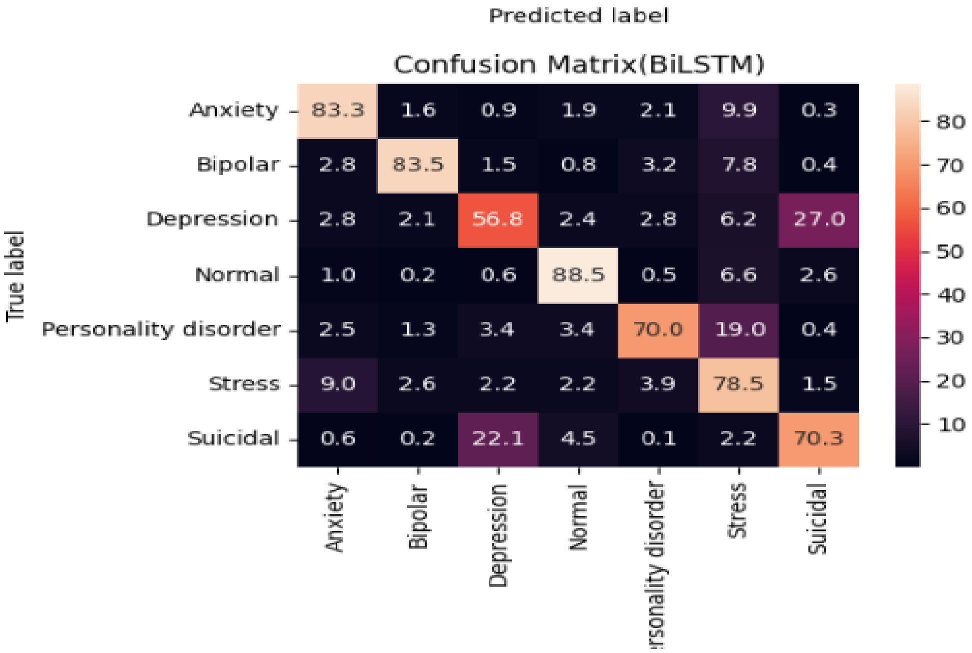
```
Text:
why do i feel worthless

Predicted label: Depression (id=2, confidence=0.6071)
Class probabilities:
  Anxiety: 0.0005
  Bipolar: 0.0005
  Depression: 0.6071
  Normal: 0.0005
  Personality disorder: 0.0004
  Stress: 0.0005
  Suicidal: 0.3904
```

Figure 9: Prediction for DeBERTa-v3-large model.

iii. MODEL 3: Bi-directional LSTM (BiLSTM)

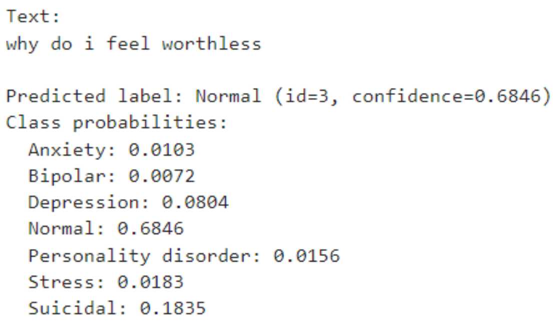
Metric	Label (BiLSTM)							Micro	Macro	accuracy
	Anxiety	Bipolar	Depression	Normal	Personality disorder	Stress	Suicidal			
Precision	0.76	0.81	0.76	0.93	0.52	0.41	0.62	0.74	0.68	0.74
Recall	0.83	0.83	0.57	0.88	0.70	0.78	0.70	0.74	0.75	
F1-Score	0.80	0.82	0.65	0.91	0.60	0.54	0.66	0.74	0.71	

Table 3: Precision, recall and F1-score for each label for the BiLSTM model



	Anxiety	Bipolar	Depression	Normal	rsonality disorder	Stress	Suicidal
Anxiety	83.3	1.6	0.9	1.9	2.1	9.9	0.3
Bipolar	2.8	83.5	1.5	0.8	3.2	7.8	0.4
Depression	2.8	2.1	56.8	2.4	2.8	6.2	27.0
Normal	1.0	0.2	0.6	88.5	0.5	6.6	2.6
Personality disorder	2.5	1.3	3.4	3.4	70.0	19.0	0.4
Stress	9.0	2.6	2.2	2.2	3.9	78.5	1.5
Suicidal	0.6	0.2	22.1	4.5	0.1	2.2	70.3

Figure 10: Confusion matrix for the BiLSTM model.



```
Text:
why do i feel worthless

Predicted label: Normal (id=3, confidence=0.6846)
Class probabilities:
  Anxiety: 0.0103
  Bipolar: 0.0072
  Depression: 0.0804
  Normal: 0.6846
  Personality disorder: 0.0156
  Stress: 0.0183
  Suicidal: 0.1835
```

Figure 11: Prediction for BiLSTM model.

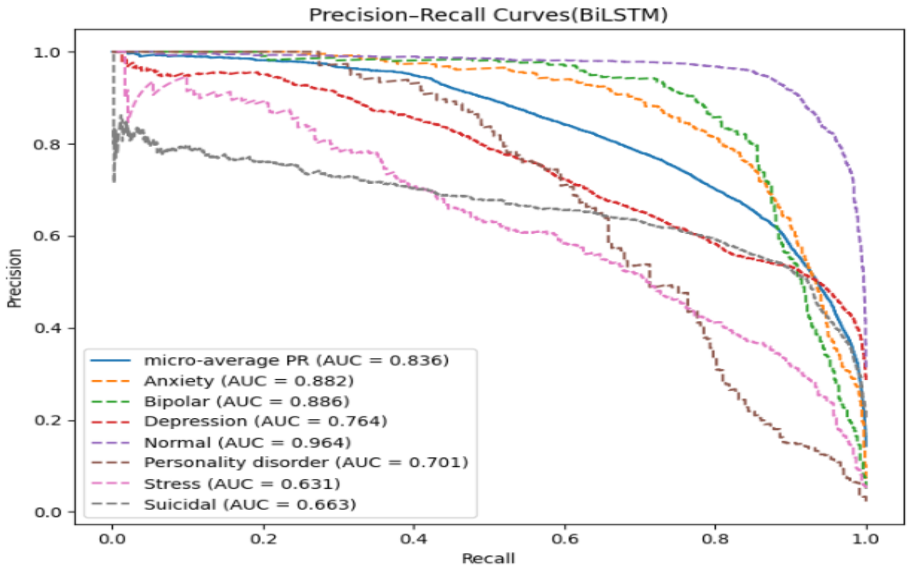


Figure 12: Precision–recall curves for the BiLSTM model.

V. DISCUSSION

From the weighted CCE results we can see that all three models behaved in a similar way on some labels and very differently on others. Deberta gives best overall performance among all the 3 models with a macro-F1 of 0.83 and accuracy of 0.8466, Roberta has a slightly less values with macro F1 of 0.83 and accuracy of 0.84 but for the BiLSTM it is almost 10 percentage behind with macro-f1 of 0.71 and accuracy of 0.74. In the precision recall curves we can see that the micro average pr-auc is 0.928 for the Roberta and the 0.927 for deberta these two transformers has both high precision and recall, but the bilstm has the lower micro average with 0.83.

The easiest label for all models was Normal. It always got the highest precision and recall in all models, highest with 0.97 and 0.93 in DeBERTa with f1-score of 0.95, this is followed by the Roberta with f1 0.95 and bilstm with f1 0.90. This is expected because Normal has the most examples in the dataset, so all three models saw many Normal posts during training. Next the classes anxiety and bipolar also performs well deberta achieves f1 score 0.89(Anxiety) and 0.877(Bipolar), Roberta has 0.88 and 0.85 and BiLSTM was a bit lower but still gave moderate results with f1 scores of 0.79 and 0.82. This shows that the language patterns for Anxiety and Bipolar are easier for the models to learn compared to some other labels.

The hardest labels were mostly the minority ones, especially Stress and Personality Disorder. Even with the weighted CCE the both the transformer f1 scores stay in between 0.76 – 0.79. In Roberta stress and personality disorder has values 0.78 and 0.79, In deberta stress and personality disorder has 0.79. whereas BiLSTM it drops to 0.5366 for stress and 0.5961 for personality disorder. In all three models, there is a strong two way confusion is in between depression and suicidal, For the F1 scores of these classes in deberta its 0.80 and 0.75, Roberta its 0.79 and 0.76 and in Bilstm its 0.65 and 0.66. while for the confusion in Roberta 21% of true depression posts are predicted as suicidal and 16.6% of suicidal posts are predicted as Depression, For DeBERTa the values are 20.2% and 17.4%. BiLSTM is afttected still more, misclassified 26.99% of depression posts as suicidal and 22.10% of suicidal as depression (around 814 and 477 instances). These two labels use very similar words and tone. This is why the models keep mixing these two classes even when the overall scores are good. There is also slight confusion between normal and stress around 3.6% normal to stress and 4.4% stress to normal in deberta and also between personality disorder and Bipolar/depression.

The AUC and precision–recall curves support these observations. For labels like Normal, Anxiety and Bipolar, the ROC-AUC values and PR curves were high and smooth for all the models. For more difficult labels like Suicidal, Stress and Personality Disorder, the curves were lower and noisier.

For prediction sample we gave the text of depression "why do I feel worthless", for the Roberta it predicted correctly with the confidence (softmax) of 0.94, where as deberta also predicted correctly but with less confidence of 0.60 and the 0.39 of suicidal, but with BiLSTM it predicted incorrectly with confidence of 0.68 as normal.

When we compare all the three loss versions, weighted CCE helped make the performance more balanced across classes. CCE tended to push the model more towards the majority labels like Normal and Depression. Oversampling changed some numbers but did not give a clear improvement and sometimes made training less stable. Weighted CCE gave a small boost to the rare labels without hurting the big ones.

Overall, all the three models can handle the majority mental health labels reasonably well but have trouble with rare and overlapping labels. The imbalance in the dataset and the similarity between some categories, especially Depression and Suicidal, make the task hard. More data for the small classes, better handling of label noise, or treating the task as multi-label instead of single-label could help in the future. We can also try to make the BiLSTM model more complex. For example, we can add CNN layers on top of the embeddings or between LSTM layers or add attention blocks. This might help the BiLSTM model learn better features and reduce the gap between the simple RNN models and the transformer models

VI. CONCLUSION

After comparing the performance of all the models, we find that the RoBERTa-base with weighted BCE is the best model to use for our mental health text classification task. It gave almost the same scores as DeBERTa-v3-large on most labels, but RoBERTa has around 125M parameters, while DeBERTa has about 300M. So RoBERTa is much smaller and easier to train and run but still gives very strong results on the dataset used. The BiLSTM model was clearly weaker than both transformer models, especially on the rare labels. DeBERTa-v3-large did give slightly better numbers on paper, but the gain was small compared to the extra size and compute. So overall, looking at both performance and model size together, we consider RoBERTa-base with weighted CCE as the best choice for this project.

REFERENCES

[1] "Sentiment Analysis for Mental Health," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health (accessed Nov. 2025).

[2] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, “Advances in Pre-Training Distributed Word Representations,” in *Proc. Int. Conf. Lang. Resour. Eval. (LREC)*, 2018.

[3] Y. Liu *et al*., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692, 2019.

[4] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-enhanced BERT with Disentangled Attention,” arXiv:2006.03654, 2020.

[5] T. Wolf *et al*., “Transformers: State-of-the-Art Natural Language Processing,” in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL)*, 2019, pp. 4171–4186.

[7] M. Schuster and K. K. Paliwal,"Bidirectional recurrent neural networks," IEEE Trans. Signal Process., vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[8] R. A. Pramunendar, “A comparative study of BERT and RoBERTa for sentiment analysis on mental health social media data,” International Journal of Electrical and Computer Engineering (IJECE), vol. 15, no. 2, pp. 155–164, 2025.

[9] S. Alshehri, “Sentiment analysis for mental health using boosting, bagging, and DeBERTa,” in Proc. 2024 Int. Conf. Artif. Intell. Data Sci. (ICAIDS), Riyadh, Saudi Arabia, 2024, pp. 1–6.

[10] A. Ince, A. G. Baydili, and D. Tuncer, “Advancing mental disorder detection: A comparative evaluation of LSTM and Transformer models,” *arXiv* preprint arXiv:2507.19511, 2025.

[11] B. A. Primack, A. Shensa, J. E. Sidani, N. Bowman, J. Knight, S. A. Karim, et al., “Reducing risk for mental health conditions associated with social media use: encouraging ‘REAL’ communication,” in Families and Technology, J. Van Hook, S. M. McHale, and V. King, Eds. Cham, Switzerland: Springer International Publishing, 2018, pp. 155–176.