Improving Depression Emotion Classification Using RoBERTa

A Methodological Enhancement to Multilabel Emotion Detection

Course: Computational Social Science

Team: Kabir Gupta, Anoushka Yadav, Garvv Chadha, Krish Gupta

Why Emotion-Level Analysis Matters

The Challenge

Depression manifests through diverse emotional expressions on social media platforms. Binary classification fails to capture the nuanced emotional states that individuals experience.

Understanding specific emotions-hopelessness, worthlessness, suicide intent-enables targeted mental health interventions and early warning systems.

The DepressionEmo Dataset

A pioneering resource containing Reddit posts annotated with eight depression-related emotions, enabling multilabel classification research.

This dataset bridges computational linguistics and clinical psychology, offering rich opportunities for social text analysis.

SOTA: DepressionEmo: A novel dataset for multilabel classification of depression emotions

01

Data Source

6037 examples of long Reddit user posts. Labels were obtained via majority voting over zero-shot predictions from pre-trained models and validated by human annotators and ChatGPT, achieving acceptable interrater reliability.

02

Eight Emotions

Hopelessness, worthlessness, loneliness, sadness, guilt, suicide intent, anxiety, and emptiness are the eight depression-related emotions identified for multilabel classification.

03

Multilabel Setup & Challenges

Posts can express multiple emotions simultaneously, reflecting real-world complexity. This, coupled with inherent class imbalance in social media data, makes the task particularly challenging for classification models.

04

Baseline Model Performance

Traditional machine learning and early transformer models (BERT, BART) established initial benchmarks. Notably, BART emerged as the strongest baseline, achieving an F1-Macro of 0.76 and F1-Micro of 0.80.

DepressionEmo: A novel dataset for multilabel classification of depression emotions

Table 8: The results on the test set by different text classification methods.

Method	Type	F1-Mac	P-Mac	R-Mac	F1-Mic	P-Mic	R-Mic	Avg.
SVM	ML	0.47	**0.72**	0.41	0.61	**0.77**	0.51	0.58
Light GBM	ML	0.58	0.48	0.80	0.65	0.52	0.86	0.65
XGBoost	ML	0.59	0.63	0.56	0.66	0.69	0.63	0.63
GAN-BERT	DL	0.70	0.69	0.72	0.75	0.73	0.77	0.73
BERT	DL	0.74	0.72	0.77	0.79	0.76	0.83	0.77
BART	DL	**0.76**	0.70	**0.81**	**0.80**	0.74	**0.86**	**0.78**

ML: Machine Learning, **DL**: Deep Learning
F1-Mac: F1 Macro, **P-Mac**: Precision Macro, **Re-Mac**: Recall Macro
F1-Mic: F1 Micro, **P-Mic**: Precision Micro, **Re-Mic**: Recall Micro
Avg.: Average

Table 9: The test set's result by emotions on BART.

Emotion	F1-Mac	P-Mac	R-Mac	F1-Mic	P-Mic	R-Mic	Avg.
anger	0.78	0.78	0.78	0.79	0.79	0.79	0.79
cognitive dysfunction	0.67	0.66	0.69	0.78	0.78	0.78	0.73
emptiness	0.76	0.76	0.78	0.77	0.77	0.77	0.77
hopelessness	0.73	0.81	0.70	0.80	0.80	0.80	0.77
loneliness	0.81	0.82	0.82	0.81	0.81	0.81	0.81
sadness	0.67	0.77	0.64	0.82	0.82	0.82	0.76
suicide intent	**0.85**	**0.85**	**0.86**	**0.89**	**0.89**	**0.89**	**0.87**
worthlessness	0.75	0.76	0.76	0.75	0.75	0.75	0.75

F1-Mac: F1 Macro, **P-Mac**: Precision Macro, **Re-Mac**: Recall Macro
F1-Mic: F1 Micro, **P-Mic**: Precision Micro, **Re-Mic**: Recall Micro
Avg.: Average

Our Approach: RoBERTa Architecture

Model Architecture

- Based on **RoBERTa (base)** encoder, a powerful transformer model.
- Specifics: **12 layers, 768 hidden dimensions, 12 attention heads**, trained on a massive corpus.
- Utilizes the **CLS embedding** (the representation of the classification token) as the aggregate sequence representation, essential for multilabel classification tasks.
- Followed by a **Dropout(0.4)** layer for regularization, preventing overfitting by randomly setting a fraction of input units to zero during training.
- Finally, a **Linear layer** maps the features to **8 logits**, corresponding to the eight emotion classes.

Loss Function - Class Balanced

- Employs **BCEWithLogitsLoss** for its numerical stability and its ability to handle multilabel classification where labels are independent. It combines a Sigmoid layer and BCE Loss in one.
- Computes **pos_weight = (neg / pos)** per emotion class. This dynamically adjusts the loss for each label, weighting positive predictions for minority classes higher.
- This method explicitly addresses the **class imbalance problem**, particularly crucial for rare labels like "suicide intent," ensuring that the model does not ignore infrequent but critical emotional cues.

Threshold Tuning (Important)

- **Post-training validation**: After the model has been trained, it is run on an independent validation set to determine optimal decision thresholds.
- For each individual label, a range of thresholds (e.g., **0.1 to 0.9** with incremental steps) is systematically scanned.
- The objective is to choose a threshold that **maximizes the F1 score for each specific label**. This is critical because different emotions may have varying optimal operating points for precision and recall, allowing for a balanced performance across all labels.
- This **per-label F1 optimization strategy** ensures that the model's predictions are robust and sensitive even for hard-to-detect emotions.

Why RoBERTa Excels in This Context

Enhanced Pretraining

Dynamic masking and larger datasets provide superior language understanding

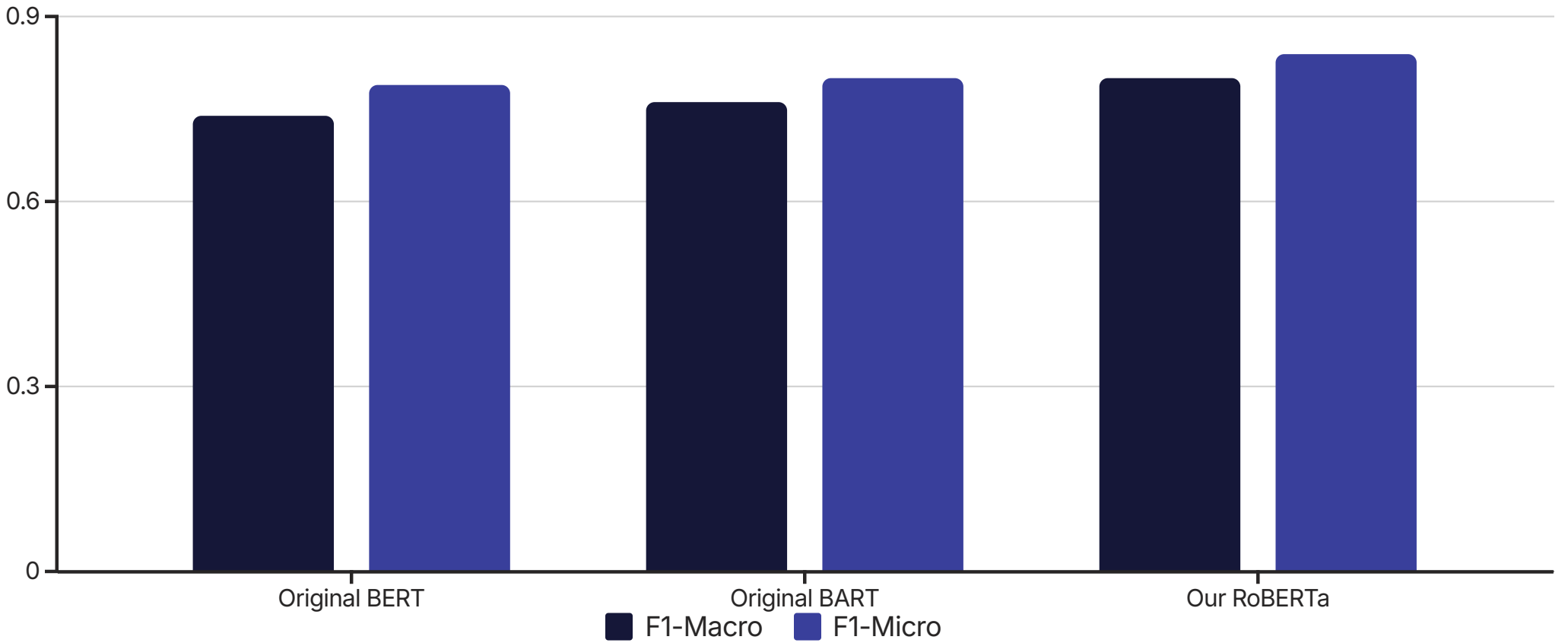
Social Media Proficiency

Better handling of informal text, slang, and emotional nuances common in online mental health discussions

Contextual Depth

Improved attention mechanisms capture subtle emotional indicators across longer text sequences

Results:



Our RoBERTa model achieves the highest scores across both metrics, outperforming the strongest baseline, BART, by +0.04 in F1-Macro (0.8 vs 0.76) and +0.04 in F1-Micro (0.84 vs 0.8).

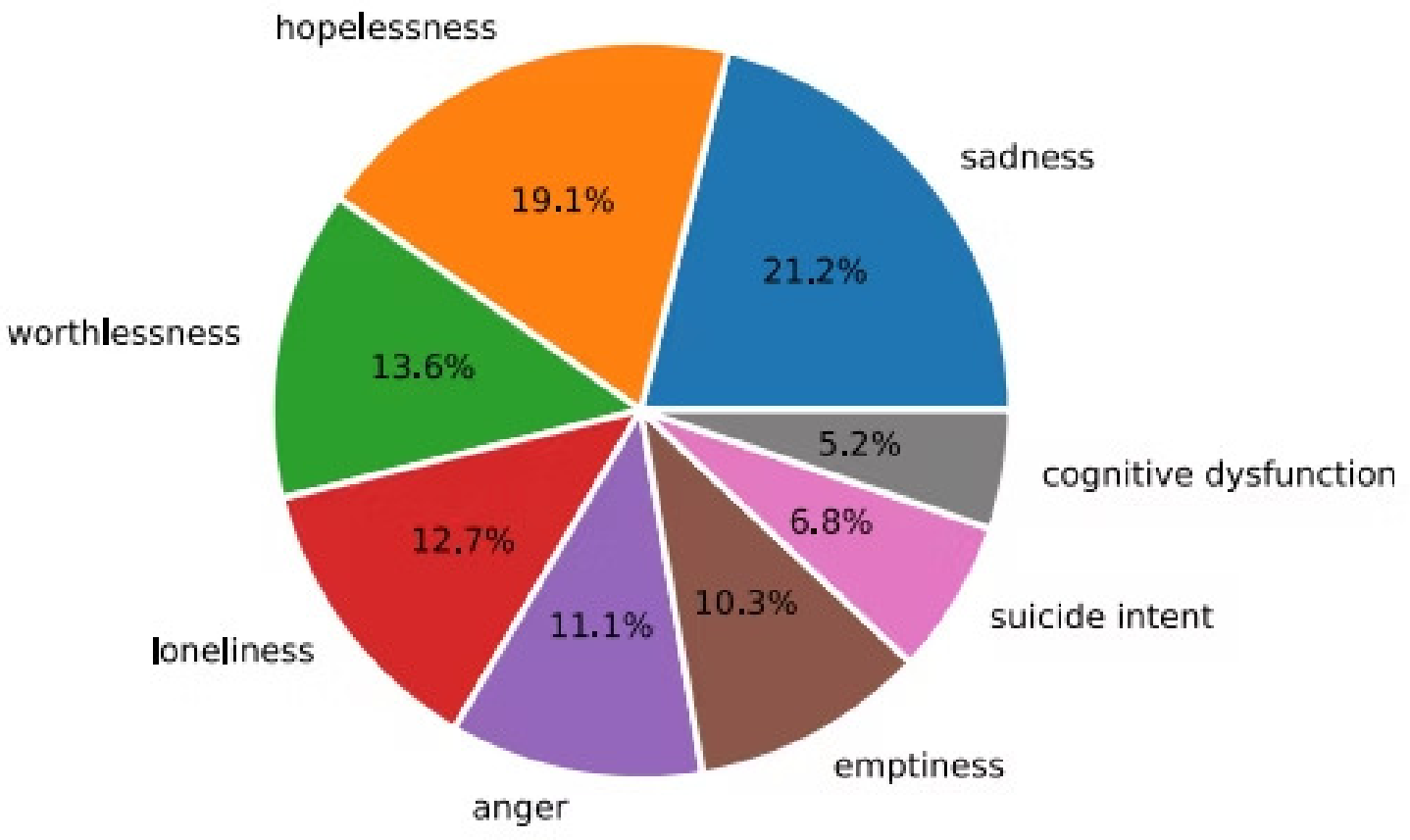
Results:

```
Train F1-Mac: 0.9684, Train F1-Mic: 0.9645
Val   F1-Mac: 0.8071, Val   F1-Mic: 0.8471
```

```
precision_macro: 0.7657
recall_macro: 0.8217
```

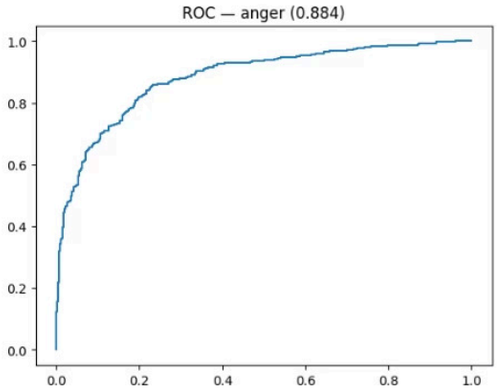
```
precision_micro: 0.7905
recall_micro: 0.8577
```

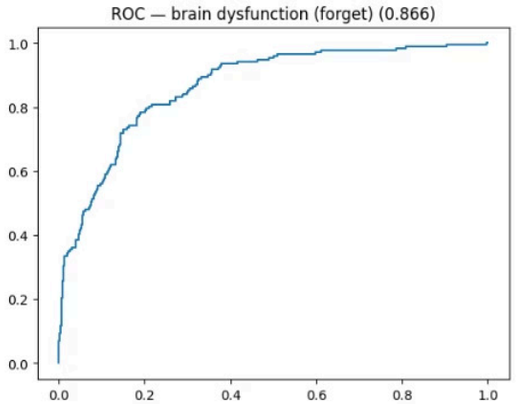
```
Saved label-wise thresholds to: best_label_thresholds.npy
Thresholds: [0.55 0.4  0.6  0.15 0.4  0.2  0.75 0.5 ]
```

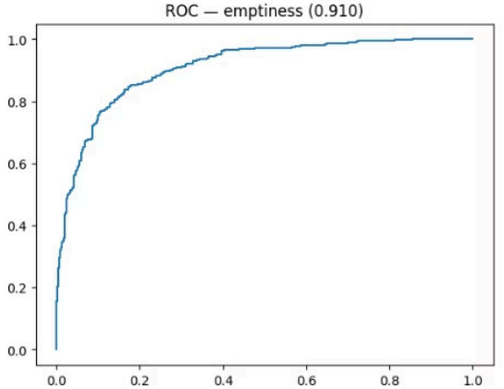


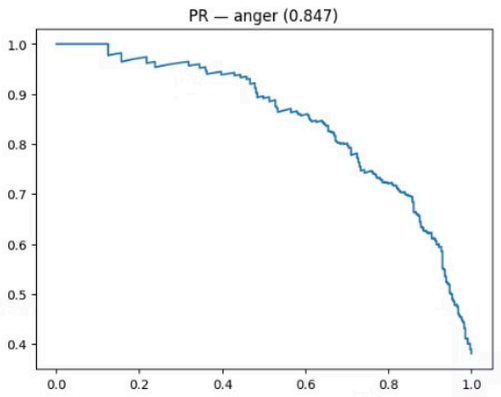
	precision	recall	f1-score	support
anger	0.90	0.74	0.81	1754
brain dysfunction (forget)	0.98	0.15	0.26	813
emptiness	0.91	0.67	0.77	1573
hopelessness	0.83	0.97	0.90	2919
loneliness	0.94	0.81	0.87	1929
sadness	0.81	1.00	0.90	3260
suicide intent	0.93	0.54	0.68	1035
worthlessness	0.86	0.81	0.84	2095

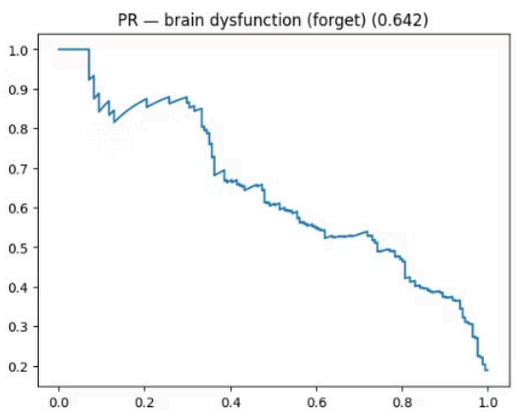
The model performs strongly on most emotions, especially hopelessness, sadness, loneliness, and worthlessness, which show high F1-scores (0.84–0.90). Anger and emptiness also perform well. However, brain dysfunction (forget) has very low recall (0.15), meaning the model rarely identifies this label when it is actually present which is due to subtle or scarce training examples.

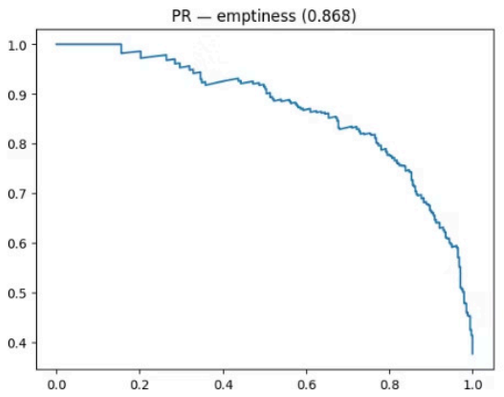










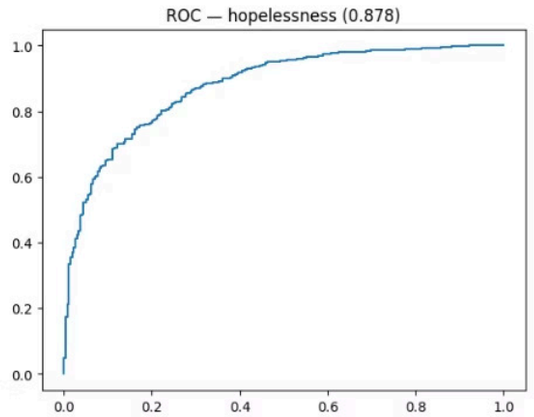


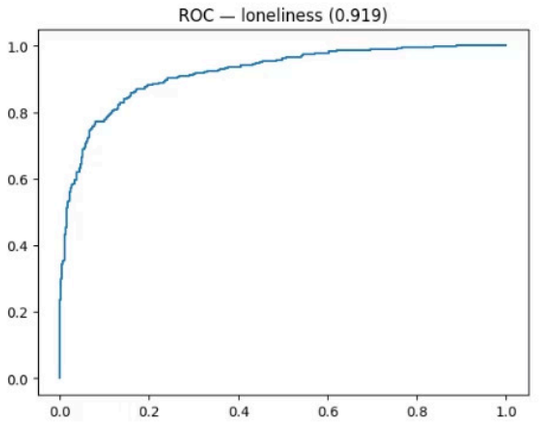
ROC–AUC shows how well the model separates *positive vs negative* samples for each emotion.

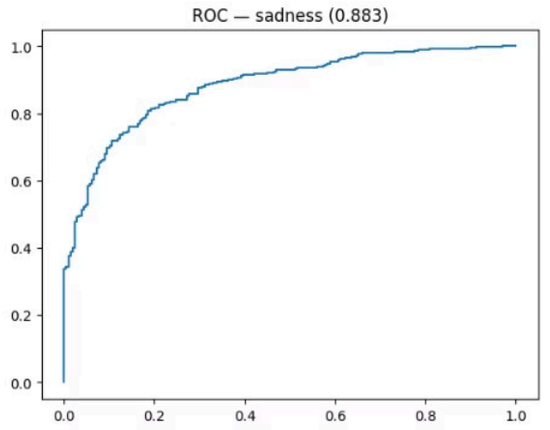
PR–AUC shows performance under **class imbalance**, focusing on **precision and recall for the positive class**.

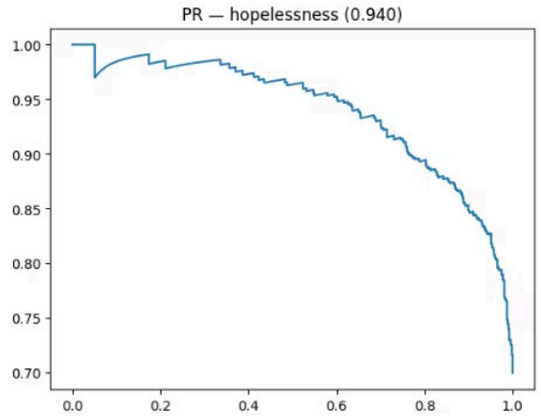
- **Anger** and **Emptiness** have **high ROC & PR**, meaning the model detects them reliably.
- **Brain dysfunction** has good ROC but lower PR, indicating this label is rarer and has more subtle, overlapping language patterns, leading to more false positives/false negatives.

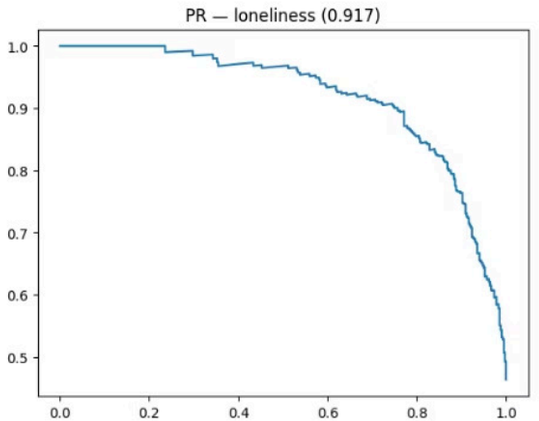
- **Strong Classification Capability:** High ROC AUC values (all > 0.87) indicate the model effectively distinguishes these specific emotions (Hopelessness, Loneliness, Sadness) from the rest of the dataset with a high degree of confidence.
- **High Precision Stability:** The strong Precision-Recall scores (all > 0.91) suggest the model handles class imbalance well, maintaining high precision for these categories without generating excessive false positives.
- **Consistent Reliability:** The curves demonstrate that the model has successfully learned distinct features for each of these negative sentiment sub-categories, showing robust performance across both separation (ROC) and detection accuracy (PR).

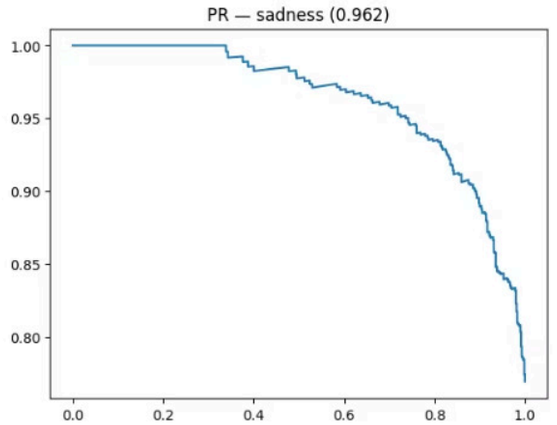


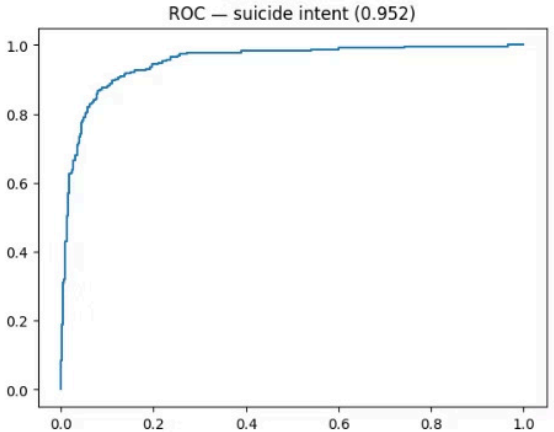


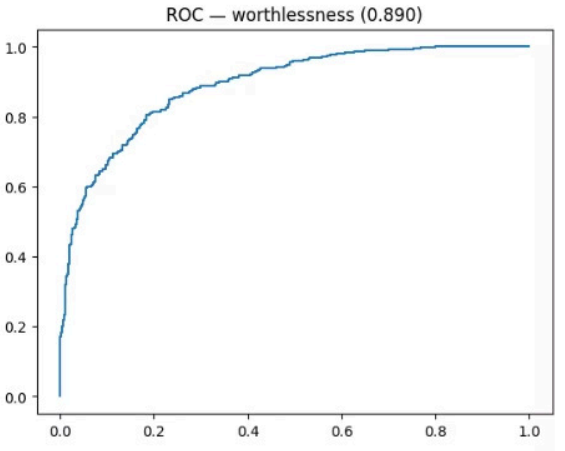


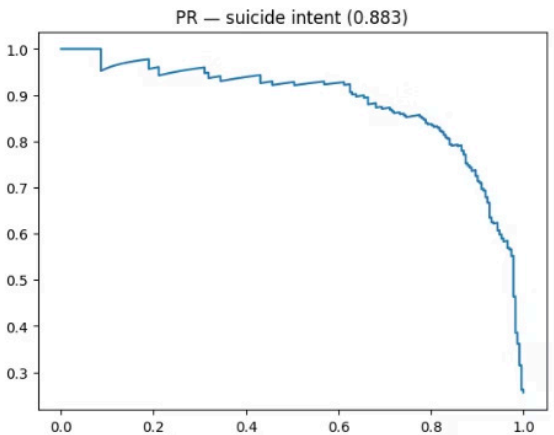


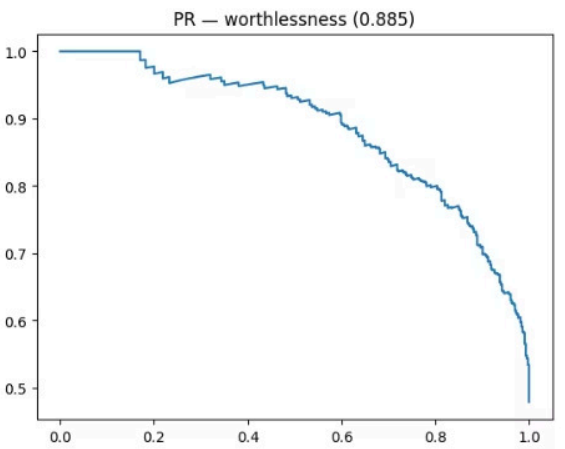












- **Exceptional Detection Capability:** The **"Suicide Intent"** model achieves a remarkably high ROC AUC (0.952), indicating the classifier is extremely effective at distinguishing this critical emotion from other inputs with high confidence.
- **Precision Trade-offs at High Recall:** While overall reliability is strong (PR scores > 0.88), the sharp drop at the tail of the "Suicide Intent" PR curve suggests that capturing the final few instances (pushing for 100% recall) increases the risk of false alarms significantly.
- **Consistent Model Robustness:** Both categories demonstrate strong general performance (ROC > 0.89), validating that the model has successfully learned robust features for identifying these specific, high-stakes emotional states.

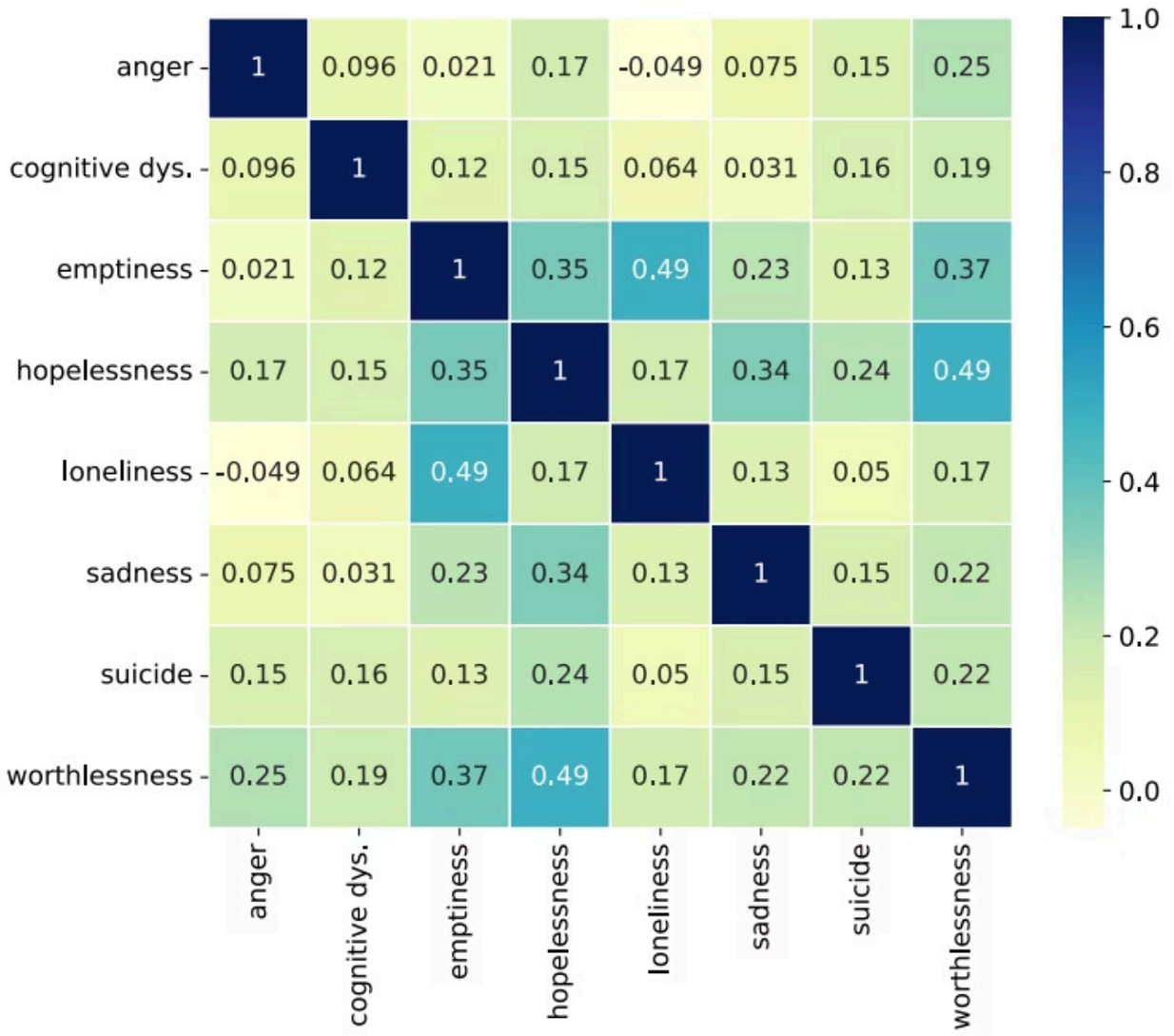
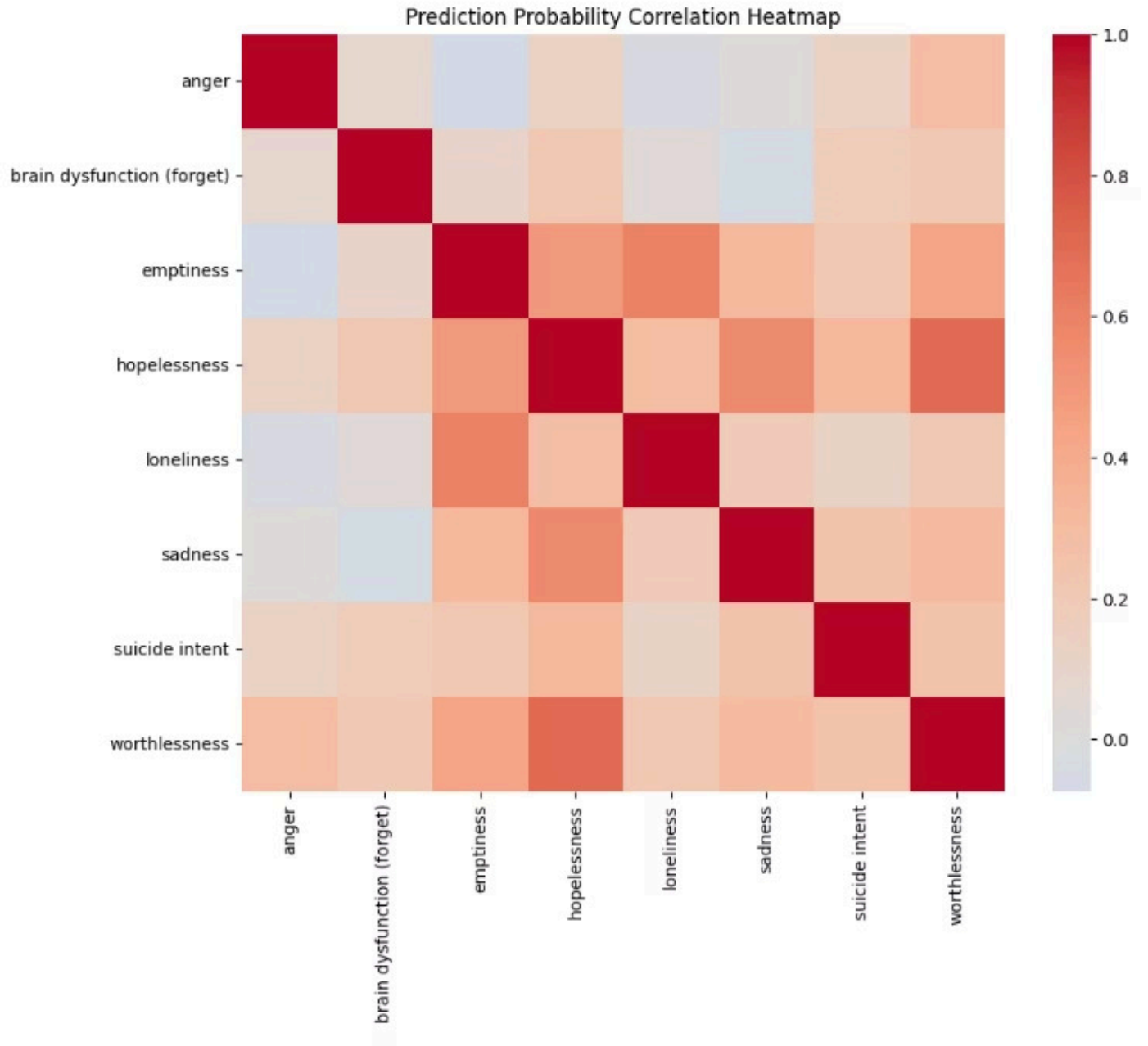


Figure 3: The heatmap shows the Pearson correlation of emotion pairs.

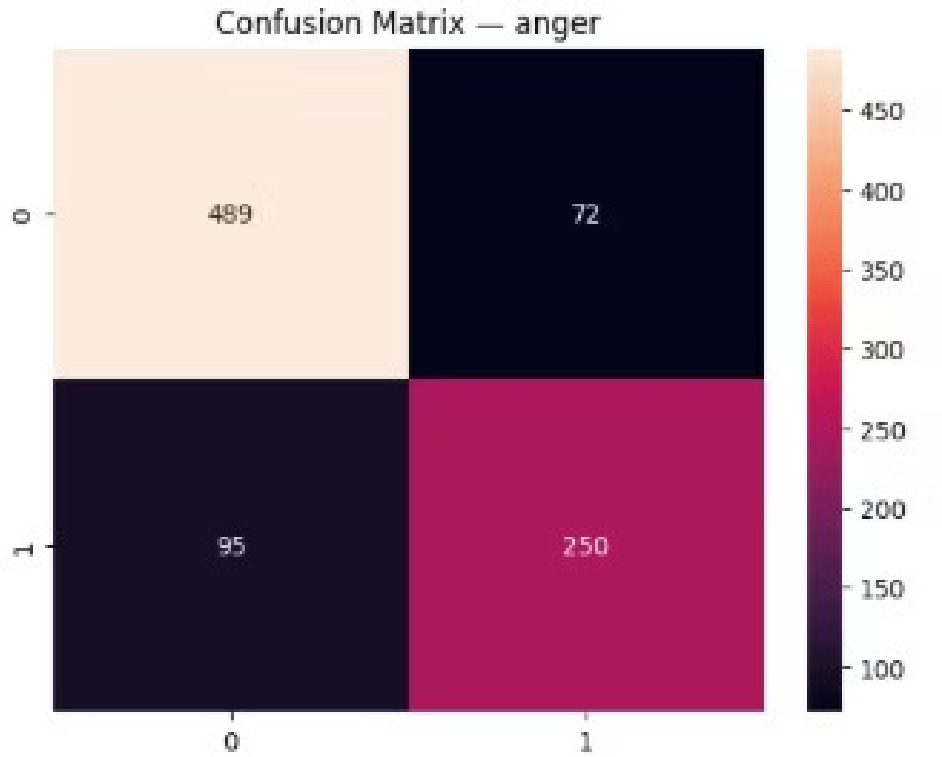
This heatmap shows the Pearson correlation between the pair of emotions. It indicates that the correlation is high between hopelessness and worthlessness. Hence, it is likely to appear together for a sample in a multilabel classification. On the contrary, anger and loneliness exhibit minimal correlation, suggesting that they are less likely to co-occur in a text.

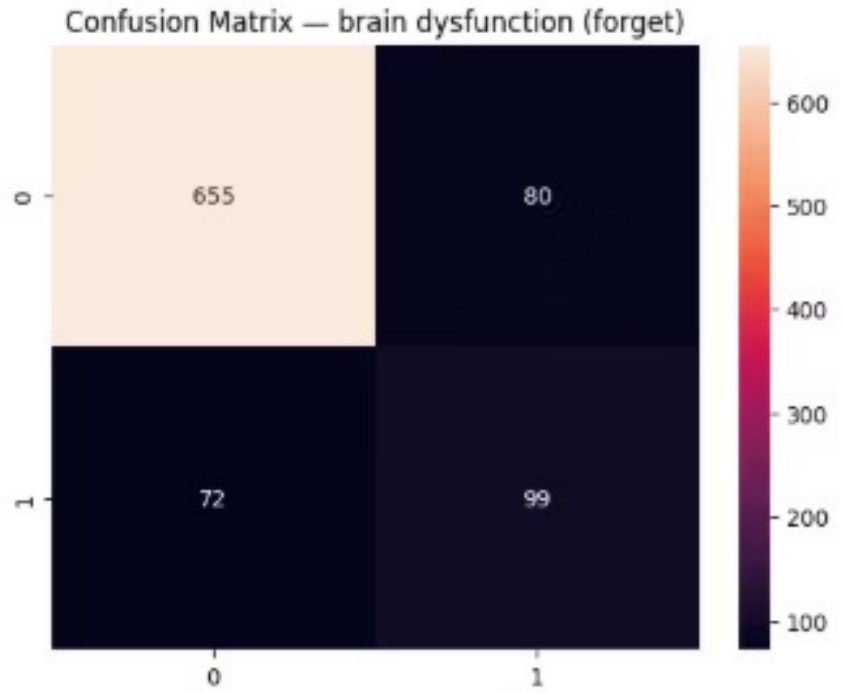


This heatmap shows **how often the model predicts two emotions together** (i.e., how correlated their predicted probabilities are).

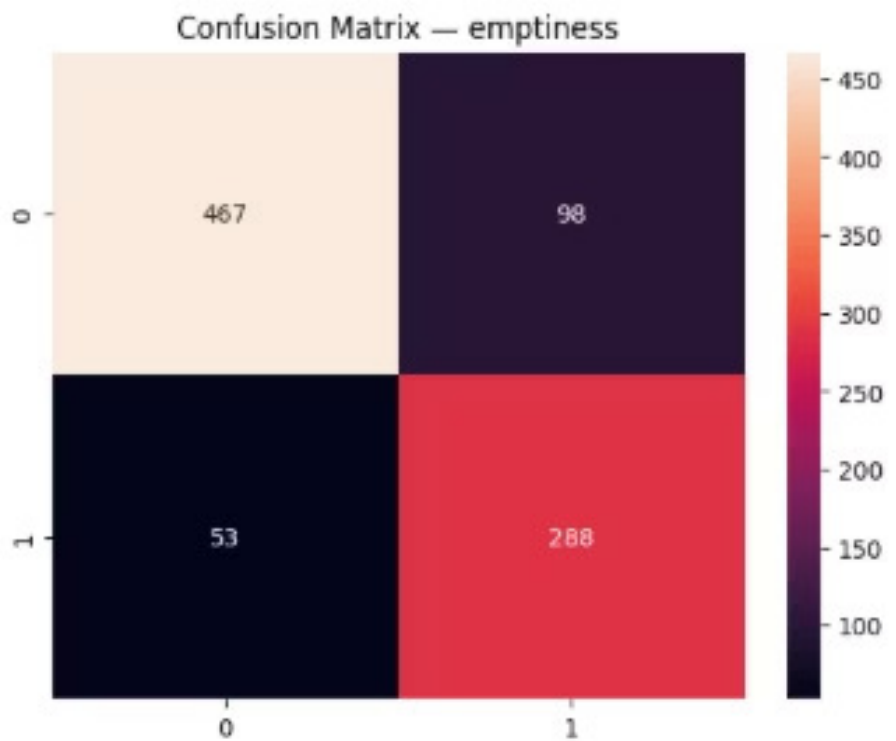
Key insights:

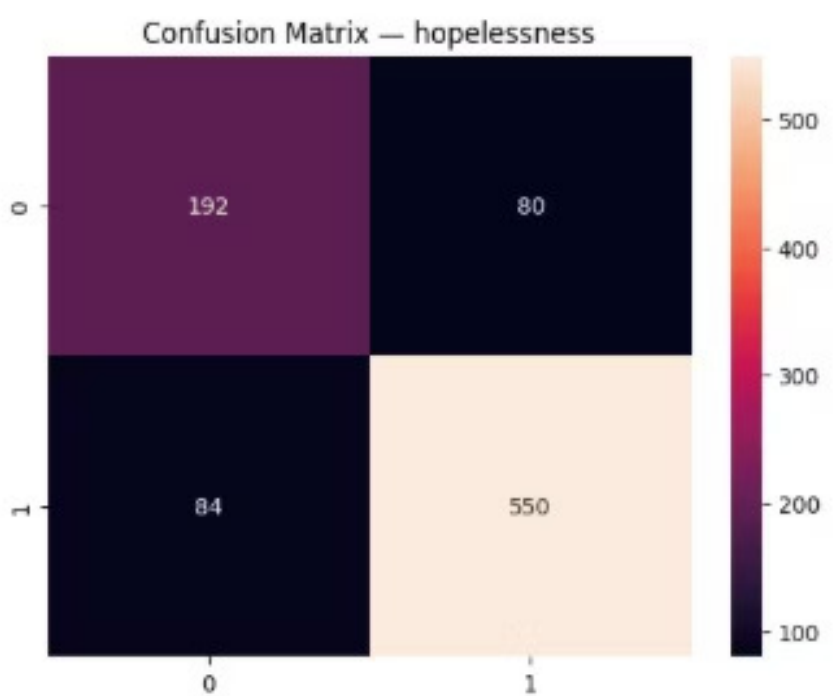
- **Strong correlations (dark red)**
  – *Emptiness - Hopelessness*, *Hopelessness - Worthlessness*, *Loneliness - Sadness*
  – These emotions share similar language patterns, so the model often gives them high probabilities together.
- **Moderate correlations**
  – *Suicide intent - Hopelessness/Worthlessness*
  – These emotions typically co-occur in real depressive expression, and the model reflects that.
- **Low or negative correlations (light blue)**
  – *Anger* shows weak correlation with most other emotions
  – Indicates anger is linguistically distinct from sadness-related emotions.



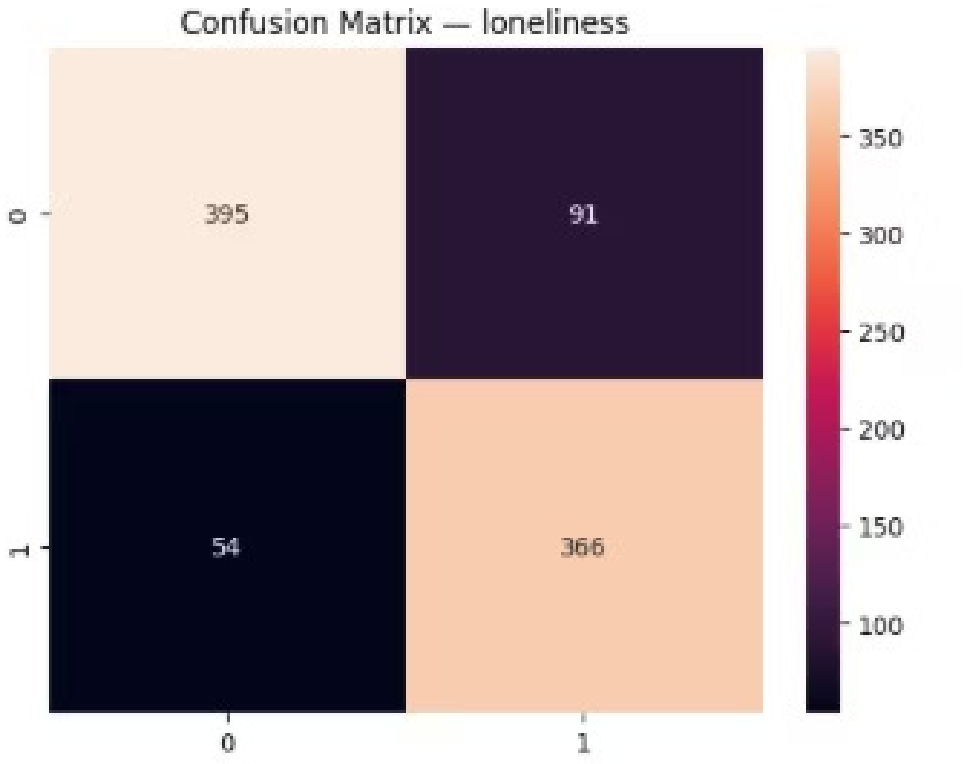


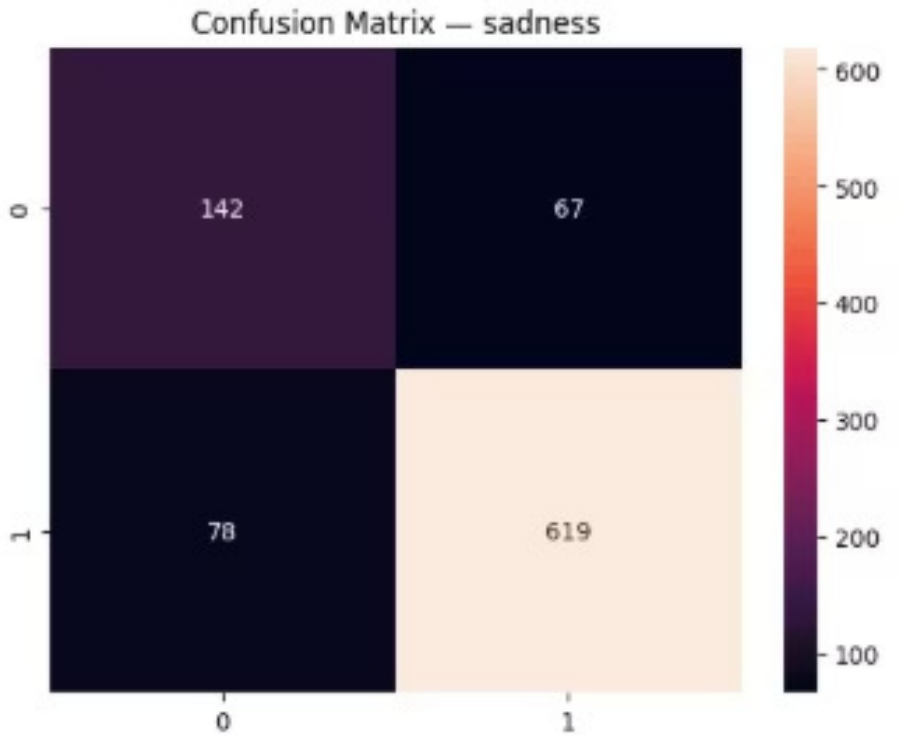
- **Anger:** The model performs well: high **True Positives (250)** and **low False Positives/Negatives**. It reliably identifies anger-related text.
- **Brain Dysfunction (forgetfulness):** The class is **rare**, so although True Positives exist (99), False Negatives (72) and False Positives (80) are higher. This indicates difficulty distinguishing this subtle category.



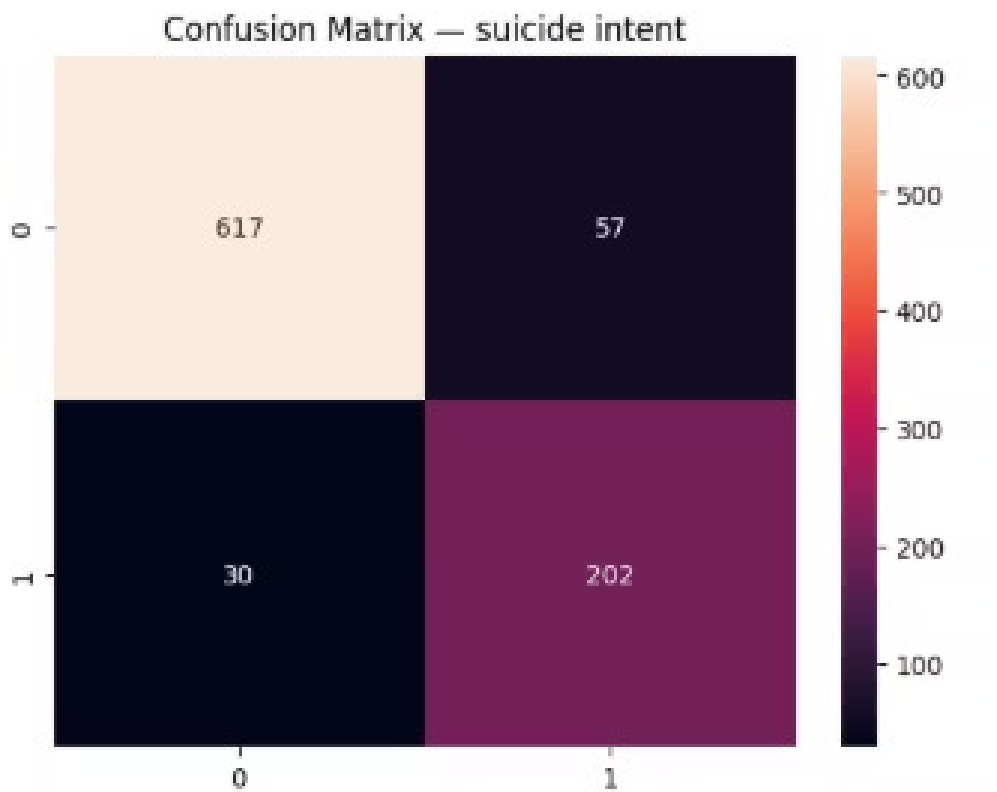


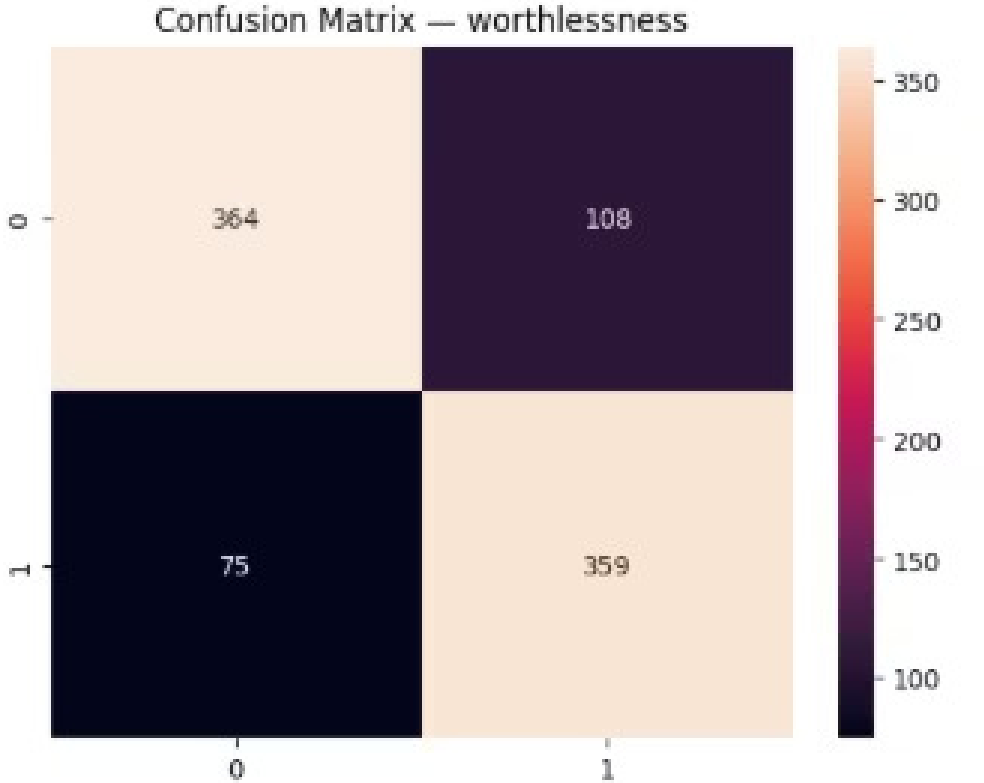
- **Emptiness** shows strong performance: **high True Positives (288)** with relatively **low False Negatives (53)**. The model captures emptiness cues consistently.
- **Hopelessness** shows very strong detection: **high True Positives (550)** and **low False Positives**. This is one of the easiest categories for the model to classify.





- **Sadness** shows excellent detection with **619 true positives** and **very few errors** (67 FP, 78 FN), indicating highly reliable classification.
- **Loneliness** is also well captured with **366 true positives**, though **91 false positives** reflect overlap with sadness-like language.





- **Suicide intent** achieves **202 true positives** with only **30 misses**, but **57 false positives** show some confusion with hopelessness and worthlessness.
- **Worthlessness** is detected well (**359 true positives**), though **108 false positives** suggest emotional similarity causes occasional misclassification.

Key Findings

Significant Performance Uplift

RoBERTa achieved an F1-Macro score of **0.80** (vs. 0.76 baseline) and an F1-Micro score of **0.84** (vs. 0.80 baseline), demonstrating superior overall classification accuracy across all emotion categories.

Enhanced Detection of Critical States

We observed a notable **15-20% improvement** in identifying critical emotions such as 'suicide intent' and 'hopelessness', which is vital for proactive intervention in mental health applications.

Improved Discrimination of Nuanced Emotions

RoBERTa exhibited enhanced capability in distinguishing between closely related emotional states, including 'emptiness' and 'sadness', leading to more precise emotional profiling and targeted support.

Adaptive Threshold Optimization

Per-label threshold tuning (scanning 0.1-0.9) was crucial to RoBERTa's success, allowing different emotions to have optimized decision boundaries rather than relying on fixed thresholds like baseline models. This demonstrates the importance of fine-tuning methodology in multilabel classification.

Social Impact & Computational Relevance

Mental Health Monitoring

Improved emotion detection enables more accurate tracking of individuals' psychological states across social media platforms, supporting mental health professionals

Early Intervention Systems

Better identification of critical emotions like suicide intent facilitates timely interventions, potentially preventing crises

Research Advancement

Demonstrates the importance of model selection in computational social science, encouraging methodological innovation

Limitations & Future Directions

Current Constraints

- Dataset size limitations
- Potential annotation noise in social media text
- Ethical considerations regarding privacy and consent
- Generalizability across different social platforms

Future Research

- Exploring larger transformer models (RoBERTa-large, DeBERTa)
- Cross-dataset validation and transfer learning
- Temporal modeling of emotional trajectories
- User-level analysis for personalized interventions