Comparative Analysis of BERT-Based Models for Emotion Detection

Simran Gill **Vicky Liu** **Jo (Jiayue) Zhu**
University Of California, Berkeley
DATASCI 266: Natural Language Processing with Deep Learning
{simran.gill, vicklylliu, jzhu868}@berkeley.edu

Abstract

Emotion detection is critical in mental health support tools. This project explores multi-label emotion classification using transformer models on the GoEmotions dataset to support triage in university chatbots. We evaluate BERT-based transformer models across various training strategies. Our findings show that DistilBERT, fine-tuned with threshold optimization (F1 = 0.5778) and enhanced by layer freezing (F1 = 0.5799), outperforms other models and binary classifiers. Results highlight DistilBERT as an efficient and effective solution for detecting overlapping emotions in real-world mental health applications.

1 Introduction

Emotion recognition is a key component of emotionally aware artificial intelligence systems, especially in mental health applications where understanding user affect can inform meaningful interventions. In this project, we develop a multi-label emotion classification model fine-tuned on the GoEmotions dataset to detect several possible emotions from a single user message. Our focus is on a real-world use case involving a triage module in a mental health chatbot used by university counseling centers. Such systems help screen students for emotional distress and guide them toward appropriate resources, such as peer support, informational material, or clinical follow-up, based on their detected emotional state.

Unlike basic sentiment analysis tasks that classify text into broad categories like positive or negative, our task involves distinguishing among over twenty nuanced emotions such as grief, nervousness, embarrassment, and pride. The complexity lies in the fact that a single message may express more than one of these emotions simultaneously. This multi-label nature significantly increases the difficulty of the task compared to traditional single-label sentiment classification.

To evaluate our models, we use the F1 score as our primary metric. In a mental health triage setting, both types of errors carry risk: missing a distress-related emotion can result in someone not receiving help when needed, while incorrectly flagging non-distress emotions may cause unnecessary interventions. The F1 score balances precision and recall, helping us account for both false positives and false negatives in a way that supports responsible and efficient triage.

In our experiments, we explore a range of modeling approaches including BERT, RoBERTa, DistilBERT, and DeBERTa architectures, class balancing strategies, threshold tuning, and per-emotion binary classification. These experiments aim to identify the methods that best support emotion detection in real-world, high-stakes contexts.

2 Background

A number of recent studies have tackled multi-label emotion classification problems, aiming to detect multiple emotional states from text. Specifically, Demszky et al. (2020) made a significant contribution by introducing the GoEmotions dataset, a large-scale and fine-grained dataset annotated for 27 emotion categories plus neutrality. To establish baseline performance, they experimented with BiLSTM and BERT-base models, finding that transformer-based models consistently outperformed traditional architectures. While BERT model performance was promising, they observed that class imbalance posed challenges—particularly for low-frequency emotions, which were often misclassified as more common ones. Nevertheless, their work demonstrated the effectiveness of BERT for this task and provided a solid foundation for subsequent research in multi-label emotion detection.

Building on prior work, Kane et al. (2022) proposed a transformer-based ensemble for multi-label emotion detection. Their method leveraged data augmentation and sampling techniques to address class imbalance and involved training an ensemble of transformer models, resulting in improved overall F1 performance. However, the ensemble method incurs high computational cost due to repeated training across models.

In our work, we focus on a systematic comparison of several transformer backbones introduced in recent years. Specifically, we examine BERT (Devlin et al. (2019)), which laid the foundation for contextualized language representations; RoBERTa (Liu et al. (2019)), which enhances BERT through larger-scale pretraining and dynamic masking, and improved training efficiency and performance; DistilBERT (Sanh et al. (2019)), a lightweight alternative that preserves most of BERT's performance while being faster and smaller; and DeBERTa (He et al. (2021)), which introduces a disentangled attention mechanism for improved context modeling.

Rather than adopting computationally expensive ensemble models, we explore more lightweight and scalable strategies. These include experimenting with class weight implementation, threshold optimization for each label on individual pre-trained models, and selective layer freezing to reduce training cost and mitigate overfitting. Our approach offers an efficient solution to address both label imbalance and complex label interactions in multi-label emotion classification

3 Methods

3.1 Data

The dataset used is the publicly available GoEmotions dataset created by Google Research. It contains over 58k English Reddit comments, each manually annotated with one or more labels from 27 fine-grained emotion categories or Neutral. The dataset supports multi-label emotion classification, where a single sentence can express multiple, overlapping emotional states.

Our exploratory data analysis (EDA) revealed significant class imbalance, with emotions like *grief* underrepresented and *neutral* dominating (see Figure 3 in Appendix). This informed our need to test imbalance-handling techniques during modeling. Additionally, most samples were associated with multiple emotion labels, indicating the need for a multi-label classification approach (see Figure 4 in Appendix). A co-occurrence heatmap revealed frequent emotion pairs, such as *disapproval* and *annoyance* (see Figure 5 in Appendix). These co-occurrences indicate label dependencies, suggesting a fixed threshold may be suboptimal for all labels and motivating our test of label-specific thresholds.

3.2 Selecting Models

We evaluated four transformer models (BERT, RoBERTa, DistilBERT, and DeBERTa) to compare BERT to more recent models. Each model was chosen for its capability in handling complex multi-label emotion classification.

BERT was selected as the baseline due to its strong text classification performance and ability to capture nuanced context. Because of its bidirectional attention, it is able to understand the full context. BERT will be used as the reference for comparing improvements in performance with modern models.

RoBERTa was selected as one of the modern models because it builds on BERT by removing next-sentence prediction tasks using dynamic masking. Making it helpful when understanding complex emotional contexts better (Cortiz, 2021). Previous studies show RoBERTa often outperforms BERT in F1 score while maintaining efficiency.

DistilBERT, a lightweight alternative to BERT, was selected for its efficiency. It reduces model size by 40%, retains about 97% of BERT's language understanding capabilities, and runs approximately 60% faster, making it particularly well-suited for real-time applications such as sentiment and emotion analysis (Sanh et al., 2019).

DeBERTa was selected for its advanced architectural improvements. It improves language understanding by disentangling position and content information using an enhanced attention mechanism. DeBERTa also incorporates a modified mask decoder that strengthens pre-training. These innovations contribute to its strong performance across a range of natural language processing benchmarks and its ability to capture complex emotional context (He et al., 2021).

3.3 Experimentation

To identify factors driving performance in a multi-label emotion prediction, we conducted five experiments across all four models. The baseline used default settings. We then optimized thresholds to improve F1 scores and applied class weights. We also tested freezing early layers to reduce overfitting. These experiments help identify whether performance improvements stem from the model architecture or the training strategies.

Table 1: Summary of Experiments

#	Name	Description
1	Simple Baseline	Creating a simple model without adding weights, thresholds, or freezing layers.
2	Simple Baseline + Thresholds	Apply thresholds to the predictions of the baseline model to improve classification performance.
3	Weighted Loss	Add class weights to the loss function to address class imbalance.
4	Weighted + Threshold	Combine class weighting with threshold tuning (thresholds derived from weighted model).
5	Weighted + Threshold + Freezing Layers	Add freezing of lower layers in the model to reduce overfitting and improve generalization.

Our dataset exhibits class imbalance. To address this, we implemented custom models incorporating class weights during training. We computed a `pos_weight` vector based on the inverse frequency of each label. These weights are passed into `BCEWithLogitsLoss`, which penalizes misclassifications of underrepresented labels. This approach helps the model learn infrequent classes more effectively. Imbalanced data is common in multi-label emotion classification, necessitating methods to handle it. Custom models are used for BERT, RoBERTa, DistilBERT, and DeBERTa only when the underlying model needs to explicitly handle class imbalance, otherwise standard pretrained models were used.

Model outputs raw logits for each label; we applied sigmoid activation to obtain probabilities. To determine custom thresholds per label, we used validation set performance. For each label, we identified the threshold yielding the best individual F1 score. This method was chosen because different emotion classes exhibit varying distributions, and label-specific thresholds can optimize the precision-recall tradeoff.

3.4 Training N Binary Models

To assess if label-specific modeling performs better, we trained six separate binary classifiers using each pretrained transformer model. We selected the top three and bottom three performing labels based on the F1 score from BERT's best-performing model (Experiment 2). This isolates learning dynamics per emotion and tests if individual models perform better, and with which pretrained transformer. To account for class imbalance, each binary model was trained using a balanced subset of the training data, 50% positive and 50% negative samples per label.

4 Results and Discussion

4.1 BERT

Test set evaluations show Experiment 2 achieved the best overall multi-label classification performance (F1 = 0.5753). While Experiment 4 (F1 = 0.5633) showed strong performance, particularly on minority classes (*pride*, *relief*, and *grief*), our primary goal was accurate multi-label emotion detection across all labels. Therefore, we prioritized overall model performance. The general F1 improvement across all labels in Experiment 2 makes it more suitable. Experiment 5 showed freezing added minimal change; therefore, Experiment 4 remained the second-best model.

Table 2: BERT: Test Set Evaluation Metrics per Experiment

#	Subset Accuracy	Precision	Recall	F1	Runtime (min)
1	0.1508	0.7778	0.3564	0.4904	10:29
2	0.0918	0.5274	0.6329	0.5753	10:27
3	0.0511	0.4076	0.6403	0.4981	10:36
4	0.0730	0.5016	0.6423	0.5633	10:39
5	0.0730	0.5016	0.6423	0.5633	10:56

Table 3: F1 scores on the test set for majority and minority emotion labels across the best (Experiment 2) and second-best (Experiment 4) BERT-based models.

Label	F1 – Experiment 2	F1 – Experiment 4
	Majority Classes	
Neutral	0.7864	0.7618
Approval	0.5066	0.4730
Admiration	0.6900	0.6770
	Minority Classes	
Pride	0.1299	0.2093
Relief	0.2741	0.3459
Grief	0.1285	0.3317

Analyzing the best BERT model, we identified labels with high false negative rates. These labels were *grief*, *pride*, and *relief*. For all three labels, *neutral* was frequently predicted alongside the actual label. This may occur because *neutral* is the majority class, leading the model to predict it frequently.

For *grief*, the model sometimes predicted related emotions like *sadness*, *remorse*, and *disappointment*. These emotions share contextual similarities with *grief* in English, potentially causing the model to predict them instead. This highlights a challenge in multi-label emotion classification: significant semantic overlap between labels. Similar trends occurred for *pride* and *relief*. For *pride*, similar predicted labels included *approval* and *admiration*. For *relief*, they included *realization* and *approval*. This trend persisted for RoBERTa, DistilBERT, and DeBERTa.

Figure: Best BERT Model - Minority Class Predictions

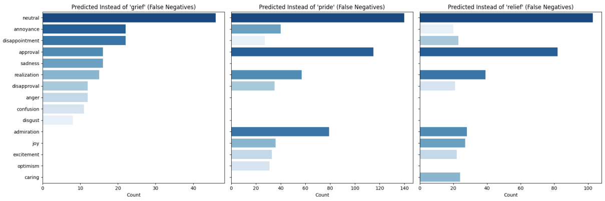


Figure 1: Other labels being chosen for minority class for best BERT model

4.2 Model Comparison: RoBERTa, DistilBERT, and DeBERTa

We evaluated RoBERTa, DistilBERT, and DeBERTa across five experimental setups to understand their relative performance on multi-label emotion classification. As shown in Table 4, each model was tested under a consistent experimental framework to examine the effects of class weighting, threshold tuning, and layer freezing. Across all three architectures, Experiment 2, which involved fine-tuning the base model with threshold optimization, consistently achieved the highest F1 scores (RoBERTa: 0.5761, DistilBERT: 0.5778, DeBERTa: 0.5604). This suggests that threshold tuning plays a critical role in balancing precision and recall across labels.

Experiment 1, a simple baseline without weighting or threshold tuning, produced relatively high precision but suffered from low recall, limiting overall effectiveness. Introducing class weighting alone (Experiment 3) generally improved recall at the cost of precision, leading to many false positives—especially notable with DeBERTa. This pattern is consistent with observations reported in Demszky et al. (2020), which caution that naive weighting may amplify noise in underrepresented labels. Combining weighting with threshold tuning (Experiment 4) yielded improvements over weighting alone, but typically did not surpass the results from Experiment 2. This reinforces that while class balancing is necessary, overly aggressive weighting can be counterproductive when not paired with proper calibration.

The addition of layer freezing to reduce overfitting (Experiment 5) had limited impact. Both RoBERTa and DistilBERT showed marginal declines in F1 compared to Experiment 4, suggesting that freezing offered little performance benefit even for smaller models. For DeBERTa, the drop was slightly more pronounced, reinforcing that layer freezing may not be effective for multi-label emotion classification on the GoEmotions dataset.

Comparing across models, DistilBERT, despite being smaller, matched or slightly exceeded RoBERTa's performance, making it a strong candidate when computational efficiency is considered. While RoBERTa had slightly stronger recall in some settings, and DeBERTa showed competitive recall results with class weighting, DistilBERT's ability to generalize well with fewer parameters makes it appealing for real-time or resource-constrained deployments, such as mental health chatbots.

Overall, these results reflect expected trade-offs between model size, precision, and recall sensitivity. The consistent performance patterns across experiments highlight the value of threshold tuning and class balancing strategies for multi-label emotion detection, regardless of the underlying transformer architecture. These findings are summarized in Table 4.

Table 4: Test Set Evaluation Metrics by Model and Experiment

Exp. #	Subset Accuracy	Precision	Recall	F1	Runtime (min)
RoBERTa (2 epochs, batch size = 16)					
1	0.1505	0.7808	0.3531	0.4862	11:32
2	**0.0860**	**0.5185**	**0.6479**	**0.5761**	**11:17**
3	0.0103	0.2814	0.8684	0.4251	13:01
4	0.0593	0.4710	0.6804	0.5567	12:41
5	0.0629	0.4532	0.6477	0.5552	8:13
DistilBERT (3 epochs, batch size = 16)					
1	0.1527	0.7564	0.3793	0.5053	8:28
2	**0.0956**	**0.5346**	**0.6286**	**0.5778***	**8:20**
3	0.0600	0.4190	0.6406	0.5067	8:51
4	0.0812	0.5111	0.6358	0.5667	8:23
5	0.0755	0.5179	0.6211	0.5648	4:58
DeBERTa (2 epochs, batch size = 16)					
1	0.1452	0.7806	0.3346	0.4684	16:40
2	**0.0777**	**0.4938**	**0.6478**	**0.5604**	**17:58**
3	0.0089	0.2664	0.8802	0.4090	19:24
4	0.0632	0.4847	0.6574	0.5580	19:29
5	0.0602	0.4837	0.6088	0.5391	12:05

4.3 N Binary Classification Models

To assess if binary classification models outperform multi-label models, we selected the three best and worst performing labels based on F1 scores from BERT's best model (Experiment 2). The goal was to determine if binary models could improve performance on minority and low-performing labels. Although *neutral* was a top-performing label by F1 score, we excluded it due to runtime and memory constraints with DeBERTa.

The best-performing labels selected were *love*, *gratitude*, and *amusement*. The worst-performing labels selected were *relief*, *pride*, and *grief*.

Table 5: Per-label performance for best and binary versions of each model (F1 Score).

Label	**BERT**		**RoBERTa**		**DistilBERT**		**DeBERTa**	
	Best	Binary	Best	Binary	Best	Binary	Best	Binary
Love	0.7850	0.5938	0.7853	0.6077	0.7794	0.6122	**0.7918**	0.6059
Gratitude	0.7592	0.6032	0.7467	0.5706	0.7546	0.6031	**0.7625**	0.5947
Amusement	0.7540	0.5523	**0.7638**	0.5664	0.7480	0.5243	0.7564	0.5435
Relief	0.2625	0.1014	**0.2957**	0.1536	0.2904	0.1084	0.2457	0.1291
Pride	0.1050	0.0920	0.1642	0.0401	**0.1884**	0.0970	0.1348	0.0823
Grief	0.1596	0.0548	0.2044	0.0847	**0.2397**	0.0672	0.1158	0.0612

Compared to the multi-label models, the binary models underperformed across all labels, as shown in Table 5 (see Table 7 in the Appendix for the full table). This may occur because multi-label models capture co-occurring emotions and shared contexts, while binary models treat each label independently. For the low-performing labels(*pride*, *grief*, and *relief*),the DistilBERT multi-label model performed best on *pride* and *grief*, while RoBERTa performed best on *relief*. These emotions are rare in our dataset and can be contextually subtle. These labels often co-occur with other emotions (Figure 2), making them difficult to detect with isolated binary models.

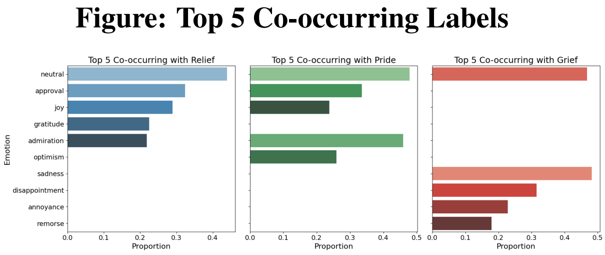


Figure 2: Using training set with actual values to see which labels often appear together

Overall, this suggests multi-label models are better suited for emotion classification tasks involving overlapping emotions and shared contextual cues.

4.4 Final Discussion

Our experiments show the pretrained DistilBERT model, fine-tuned for multi-label classification with optimized probability thresholds (Experiment 2), achieved the strongest overall performance (F1 = 0.5778). It outperformed all other BERT-based baselines, making it the best performer. Additionally, DistilBERT was the most efficient model, running approximately 60% faster than others, highlighting advantages in both accuracy and computational efficiency.

Comparing Experiment 2 (thresholds only) and Experiment 4 (weights + thresholds), we observed Experiment 4 improved performance on certain minority labels (e.g., *grief*) but at the expense of performance on several majority labels. While acknowledging the importance of detecting minority labels in mental health, our priority was selecting a model offering consistent, balanced performance across all emotional states. This provides a more reliable foundation for future enhancements or targeted fine-tuning. Therefore, we selected Experiment 2 as our preferred model.

Table 6: F1 scores on the test set for majority and minority emotion labels across the best (Experiment 2) and second-best (Experiment 4) DistilBERT models.

Label	**F1 – Experiment 2**	**F1 – Experiment 4**
	Majority Classes	
love	0.7794	0.7682
gratitude	0.7546	0.7567
amusement	0.7480	0.7426
	Minority Classes	
embarrassment	0.3372	0.3264
nervousness	0.2807	0.2838
grief	0.2397	0.3220

To further enhance this model, we applied selective layer freezing—freezing lower transformer layers (excluding embeddings) while fine-tuning only the top layers. This yielded a modest performance gain, increasing the F1 score to 0.5799. Additionally, it reduced training time and mitigated overfitting, suggesting better generalization and achieving the highest performance across all experiments.

A closer examination of the best-performing model (Experiment 2 with layer freezing) compared to Experiment 4 shows that its predictions more closely align with the ground truth, both in terms of label accuracy and the number of predicted labels per instance. Sample outputs provided in the Appendix Figure 31 & 32 further illustrate how this model's predicted label distributions correspond more accurately to the actual data.

A key contributor to success was threshold optimization for each label. Tuning individual thresholds, rather than using a uniform threshold, enabled better handling of class imbalance and optimization of the precision-recall tradeoff. This is particularly important in multi-label classification tasks where label frequencies vary widely and multiple labels often co-occur within a single instance.

Importantly, our final multi-label model also outperforms the N binary classifiers approach, where a separate classifier is trained for each label (see Table 5). While binary classifiers treat each label independently, the multi-label DistilBERT model captures inter-label dependencies, which are especially relevant in emotion classification tasks. Computationally, the multi-label model is far more efficient, requiring only one forward pass during inference versus 28 passes for the binary models. While binary classifiers were tested as an alternative approach to improve performance on minority and low-performing labels, the multi-label model ultimately demonstrated stronger consistency, scalability, and overall performance.

5 Conclusion

Our goal was to assess whether modern transformers (RoBERTa, DistilBERT, DeBERTa) outperform BERT in multi-label emotion classification, or if performance stems from training strategies. Across five experiments, DistilBERT performed best with threshold optimization alone (F1 = 0.5778). Further exploration showed applying layer freezing to this model increased the F1 score from 0.5778 to 0.5799. This demonstrates that a lighter model can achieve strong results with proper tuning. DistilBERT also had the fastest training runtime, beneficial for conducting experiments and reducing computational costs.

Performance on minority emotion labels remained relatively low even with per-label binary models. To improve identification of overlapping emotions, future work could explore contextual similarities between labels to better distinguish subtle emotional expressions.

Authors' Contributions

All authors collaboratively conducted initial EDA, examining label distributions and co-occurrence patterns. Simran handled dataset preprocessing and created the train/validation/test splits. She also developed and executed the five experiments using BERT and RoBERTa, trained the corresponding N binary classifiers, and performed additional EDA on the final model to assess performance and extract insights. Jo ran the five experiments using DistilBERT, implemented binary classifiers with that architecture, and conducted further analysis on the final model, which was selected as the best-performing model. Vicky fine-tuned DeBERTa models for all five experiments, trained N binary classifiers with DeBERTa, and carried out the cross-model performance analysis. All authors contributed to interpreting results and writing the final report.

References

Diogo Cortiz. 2021. Exploring transformers in emotion recognition: a comparison of bert, distilbert, roberta, xlnet and electra. *arXiv preprint arXiv:2104.02041*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Aditya Kane, Shantanu Patankar, Sahil Khose, and Neeraja Kirtane. 2022. Transformer based ensemble for emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, page 250—254.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Appendix

Table 7: Per-label performance for best and binary versions of each model (F1 Score).

Label	**BERT**		**RoBERTa**		**DistilBERT**		**DeBERTa**	
	Best	Binary	Best	Binary	Best	Binary	Best	Binary
Love	0.7850	0.5938	0.7853	0.6077	0.7794	0.6122	**0.7918**	0.6059
Gratitude	0.7592	0.6032	0.7467	0.5706	0.7546	0.6031	**0.7625**	0.5947
Amusement	0.7540	0.5523	**0.7638**	0.5664	0.7480	0.5243	0.7564	0.5435
Relief	0.2625	0.1014	**0.2957**	0.1536	0.2904	0.1084	0.2457	0.1291
Pride	0.1050	0.0920	0.1642	0.0401	**0.1884**	0.0970	0.1348	0.0823
Grief	0.1596	0.0548	0.2044	0.0847	**0.2397**	0.0672	0.1158	0.0612

Figure: Label Frequency

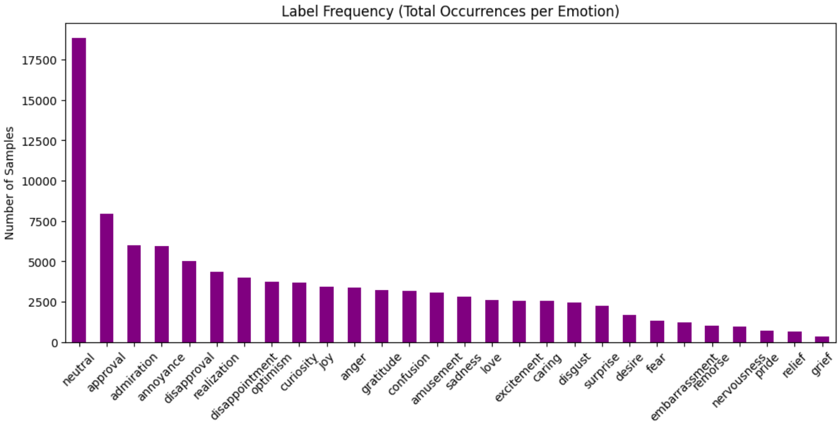


Figure 3

Figure: Number of Labels per Text

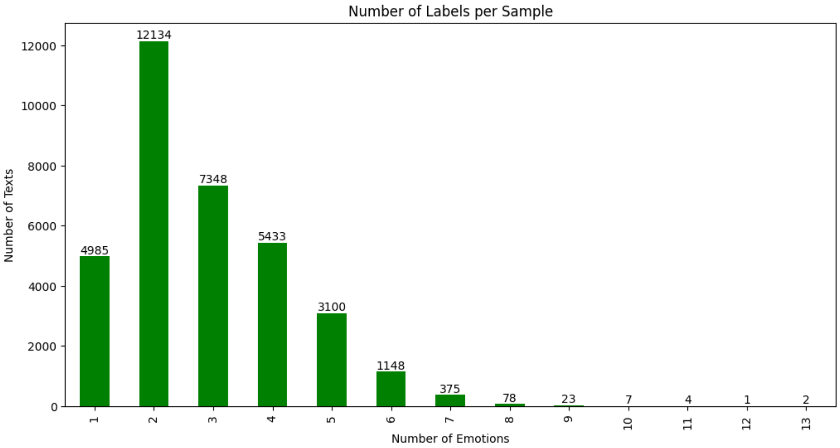


Figure 4

Figure: Label Correlation

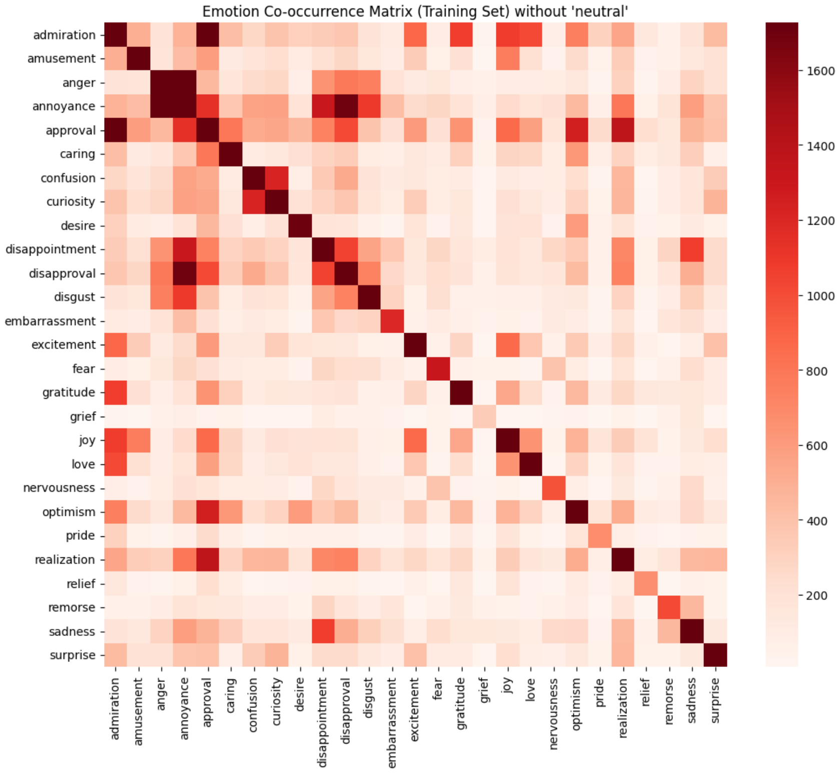


Figure 5

Figure: Token Length Distribution

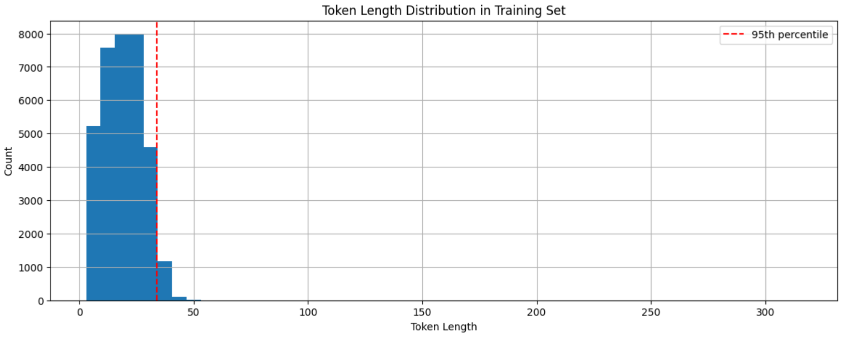


Figure 6

Figure: Top 5 Co-occurring labels

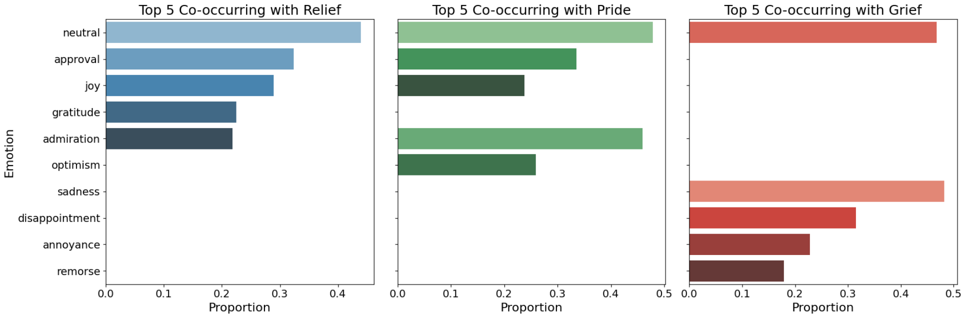
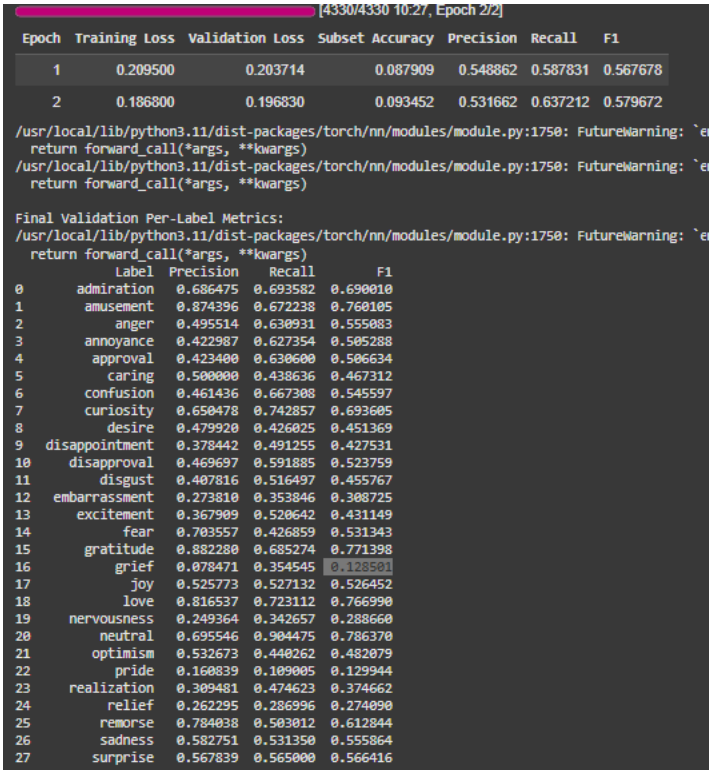


Figure 7: Using training set with actual values to see which labels often appear together

Figure: BERT Experiment 2 Training and Validation

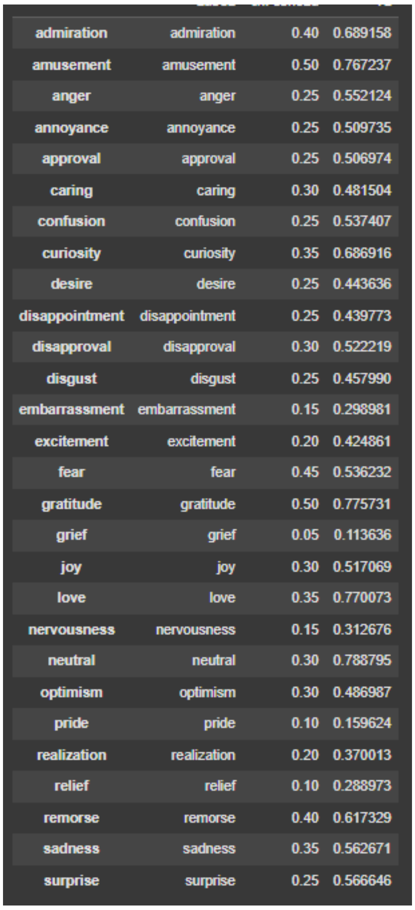

[4330/4330 10:27, Epoch 2/2]

Epoch	Training Loss	Validation Loss	Subset Accuracy	Precision	Recall	F1
1	0.209500	0.203714	0.087909	0.548862	0.587831	0.567678
2	0.186800	0.196830	0.093452	0.531662	0.637212	0.579672

```
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/module.py:1750: FutureWarning: `e
  return forward_call(*args, **kwargs)
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/module.py:1750: FutureWarning: `e
  return forward_call(*args, **kwargs)

Final Validation Per-Label Metrics:
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/module.py:1750: FutureWarning: `e
  return forward_call(*args, **kwargs)
             Label  Precision    Recall        F1
0       admiration   0.686475  0.693582  0.690010
1        amusement   0.874396  0.672238  0.760105
2            anger   0.495514  0.630931  0.555083
3        annoyance   0.422987  0.627354  0.505288
4         approval   0.423400  0.630600  0.506634
5           caring   0.500000  0.438636  0.467312
6        confusion   0.461436  0.667308  0.545597
7        curiosity   0.650478  0.742857  0.693605
8           desire   0.479920  0.426025  0.451369
9   disappointment   0.378442  0.491255  0.427531
10     disapproval   0.469697  0.591885  0.523759
11         disgust   0.407816  0.516497  0.455767
12   embarrassment   0.273810  0.353846  0.308725
13      excitement   0.367909  0.520642  0.431149
14            fear   0.703557  0.426859  0.531343
15       gratitude   0.882280  0.685274  0.771398
16           grief   0.078471  0.354545  0.128501
17             joy   0.525773  0.527132  0.526452
18            love   0.816537  0.723112  0.766990
19     nervousness   0.249364  0.342657  0.288660
20         neutral   0.695546  0.904475  0.786370
21        optimism   0.532673  0.440262  0.482079
22           pride   0.160839  0.109005  0.129944
23     realization   0.309481  0.474623  0.374662
24          relief   0.262295  0.286996  0.274090
25         remorse   0.784038  0.503012  0.612844
26         sadness   0.582751  0.531350  0.555864
27        surprise   0.567839  0.565000  0.566416
```

Figure 8

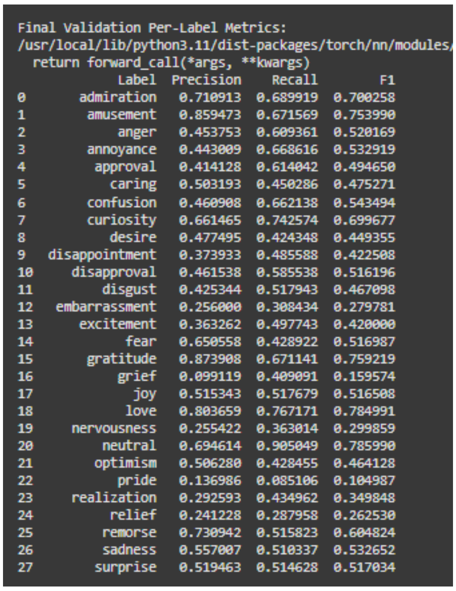
Figure: BERT Experiment 2 Thresholds



admiration	admiration	0.40	0.689158
amusement	amusement	0.50	0.767237
anger	anger	0.25	0.552124
annoyance	annoyance	0.25	0.509735
approval	approval	0.25	0.506974
caring	caring	0.30	0.481504
confusion	confusion	0.25	0.537407
curiosity	curiosity	0.35	0.686916
desire	desire	0.25	0.443636
disappointment	disappointment	0.25	0.439773
disapproval	disapproval	0.30	0.522219
disgust	disgust	0.25	0.457990
embarrassment	embarrassment	0.15	0.298981
excitement	excitement	0.20	0.424861
fear	fear	0.45	0.536232
gratitude	gratitude	0.50	0.775731
grief	grief	0.05	0.113636
joy	joy	0.30	0.517069
love	love	0.35	0.770073
nervousness	nervousness	0.15	0.312676
neutral	neutral	0.30	0.788795
optimism	optimism	0.30	0.486987
pride	pride	0.10	0.159624
realization	realization	0.20	0.370013
relief	relief	0.10	0.288973
remorse	remorse	0.40	0.617329
sadness	sadness	0.35	0.562671
surprise	surprise	0.25	0.566646

Figure 9

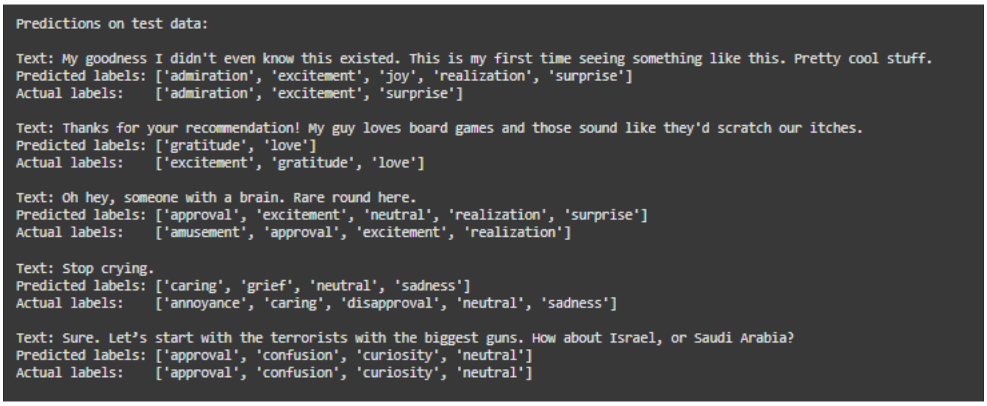
Figure: BERT Experiment 2 Test Set Label Evaluation



```
Final Validation Per-Label Metrics:
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/
  return forward_call(*args, **kwargs)
             Label  Precision    Recall        F1
0       admiration   0.710913  0.689919  0.700258
1        amusement   0.859473  0.671569  0.753990
2            anger   0.453753  0.609361  0.520169
3        annoyance   0.443009  0.668616  0.532919
4         approval   0.414128  0.614042  0.494650
5           caring   0.503193  0.450286  0.475271
6        confusion   0.460908  0.662138  0.543494
7        curiosity   0.661465  0.742574  0.699677
8           desire   0.477495  0.424348  0.449355
9   disappointment   0.373933  0.485588  0.422508
10     disapproval   0.461538  0.585538  0.516196
11         disgust   0.425344  0.517943  0.467098
12   embarrassment   0.256000  0.308434  0.279781
13      excitement   0.363262  0.497743  0.420000
14            fear   0.650558  0.428922  0.516987
15       gratitude   0.873908  0.671141  0.759219
16           grief   0.099119  0.409091  0.159574
17             joy   0.515343  0.517679  0.516508
18            love   0.803659  0.767171  0.784991
19     nervousness   0.255422  0.363014  0.299859
20         neutral   0.694614  0.905049  0.785990
21        optimism   0.506280  0.428455  0.464128
22           pride   0.136986  0.085106  0.104987
23     realization   0.292593  0.434962  0.349848
24          relief   0.241228  0.287958  0.262530
25         remorse   0.730942  0.515823  0.604824
26         sadness   0.557007  0.510337  0.532652
27        surprise   0.519463  0.514628  0.517034
```

Figure 10

Figure: BERT Experiment 2 Test Set Output



```
Predictions on test data:

Text: My goodness I didn't even know this existed. This is my first time seeing something like this. Pretty cool stuff.
Predicted labels: ['admiration', 'excitement', 'joy', 'realization', 'surprise']
Actual labels:    ['admiration', 'excitement', 'surprise']

Text: Thanks for your recommendation! My guy loves board games and those sound like they'd scratch our itches.
Predicted labels: ['gratitude', 'love']
Actual labels:    ['excitement', 'gratitude', 'love']

Text: Oh hey, someone with a brain. Rare round here.
Predicted labels: ['approval', 'excitement', 'neutral', 'realization', 'surprise']
Actual labels:    ['amusement', 'approval', 'excitement', 'realization']

Text: Stop crying.
Predicted labels: ['caring', 'grief', 'neutral', 'sadness']
Actual labels:    ['annoyance', 'caring', 'disapproval', 'neutral', 'sadness']

Text: Sure. Let’s start with the terrorists with the biggest guns. How about Israel, or Saudi Arabia?
Predicted labels: ['approval', 'confusion', 'curiosity', 'neutral']
Actual labels:    ['approval', 'confusion', 'curiosity', 'neutral']
```

Figure 11

Figure: Best BERT Model - Minority Class Predictions

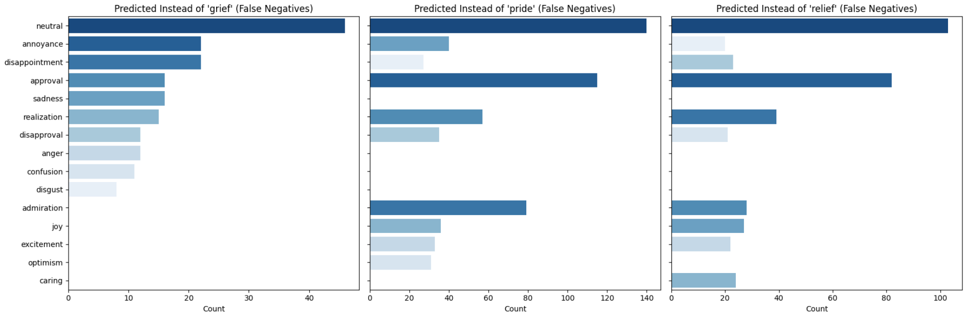


Figure 12: Other labels being chosen for minority class for best BERT model

Figure: BERT Best Model (Experiment 2) Evaluation

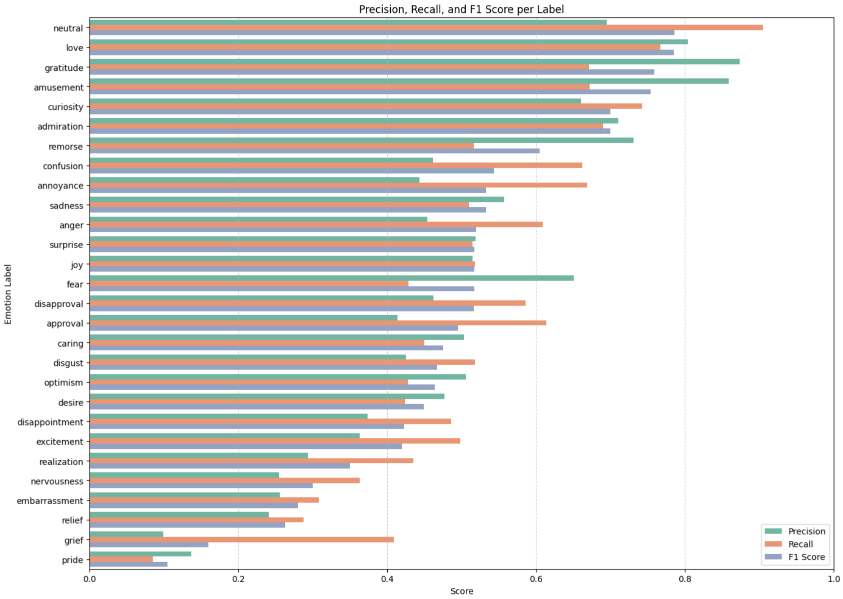


Figure 13

Figure: BERT Best Model (Experiment 2) F1

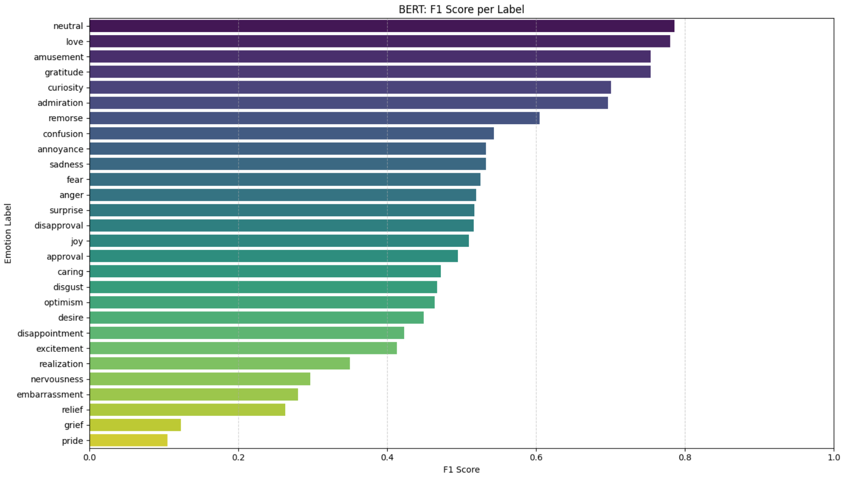


Figure 14

Figure: BERT Best Model (Experiment 2) Recall

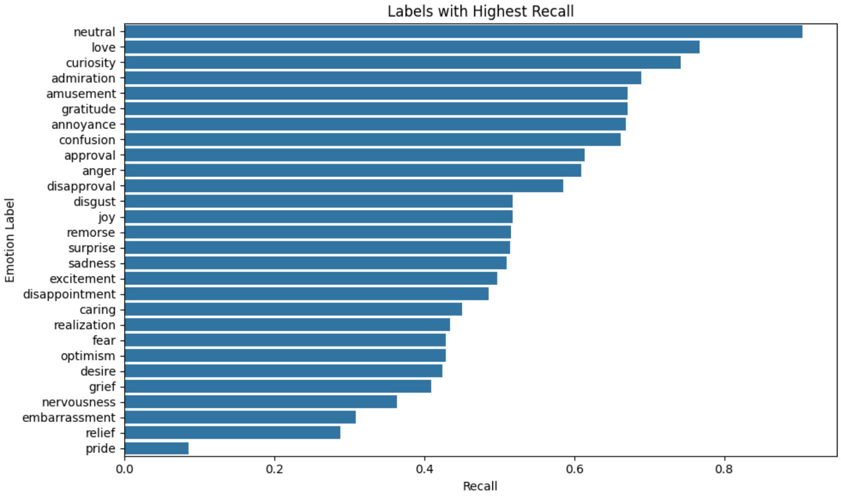
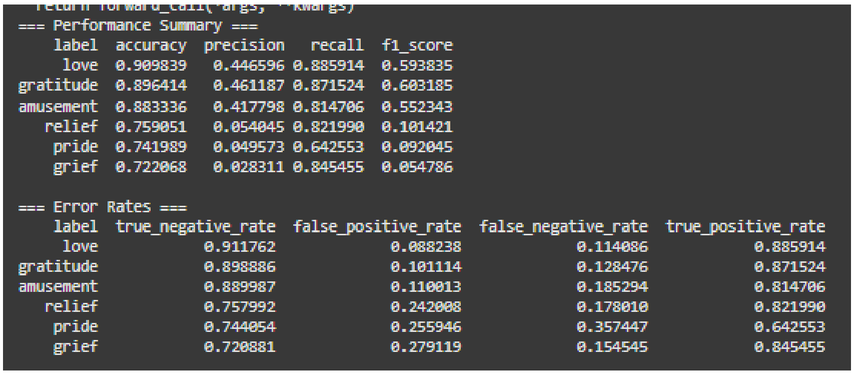


Figure 15

Figure: BERT - Training N Binary Model Results



```
  return forward_call(*args, **kwargs)
=== Performance Summary ===
    label  accuracy  precision   recall  f1_score
     love  0.909839   0.446596 0.885914  0.593835
gratitude  0.896414   0.461187 0.871524  0.603185
amusement  0.883336   0.417798 0.814706  0.552343
   relief  0.759051   0.054045 0.821990  0.101421
    pride  0.741989   0.049573 0.642553  0.092045
    grief  0.722068   0.028311 0.845455  0.054786

=== Error Rates ===
    label  true_negative_rate  false_positive_rate  false_negative_rate  true_positive_rate
     love            0.911762             0.088238             0.114086            0.885914
gratitude            0.898886             0.101114             0.128476            0.871524
amusement            0.889987             0.110013             0.185294            0.814706
   relief            0.757992             0.242008             0.178010            0.821990
    pride            0.744054             0.255946             0.357447            0.642553
    grief            0.720881             0.279119             0.154545            0.845455
```

Figure 16

Figure: BERT - Training N Binary Model F1 Comparison with Best Model

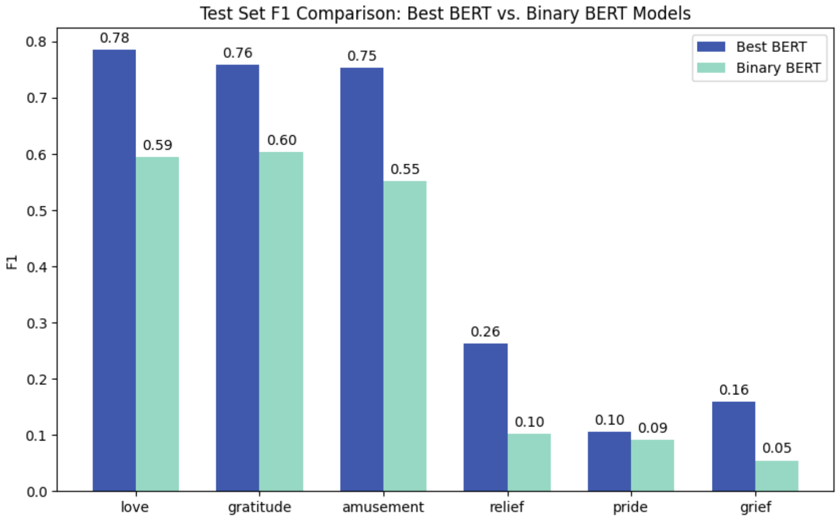
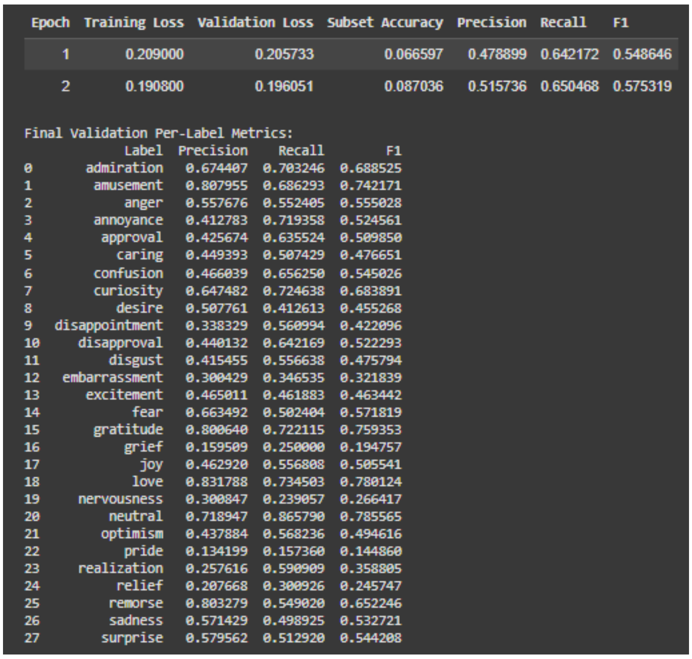


Figure 17

Figure: RoBERTa Experiment 2 Training and Validation

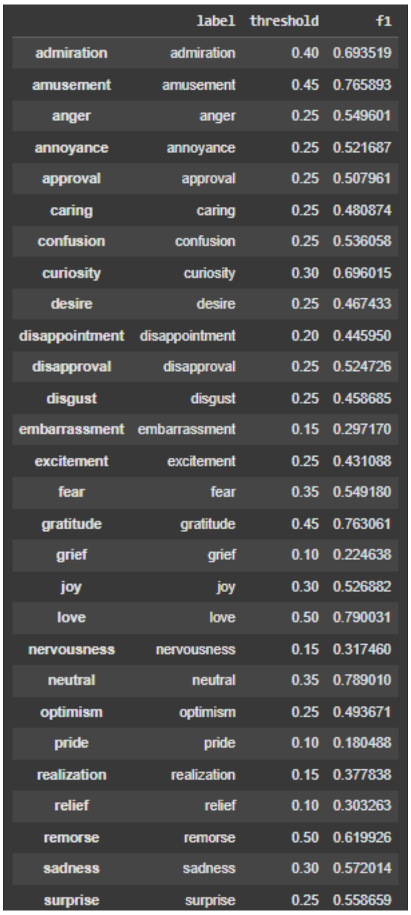


Epoch	Training Loss	Validation Loss	Subset Accuracy	Precision	Recall	F1
1	0.209000	0.205733	0.066597	0.478899	0.642172	0.548646
2	0.190800	0.196051	0.087036	0.515736	0.650468	0.575319

```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.674407  0.703246  0.688525
1        amusement   0.807955  0.686293  0.742171
2            anger   0.557676  0.552405  0.555028
3        annoyance   0.412783  0.719358  0.524561
4         approval   0.425674  0.635524  0.509850
5           caring   0.449393  0.507429  0.476651
6        confusion   0.466039  0.656250  0.545026
7        curiosity   0.647482  0.724638  0.683891
8           desire   0.507761  0.412613  0.455268
9   disappointment   0.338329  0.560994  0.422096
10     disapproval   0.440132  0.642169  0.522293
11         disgust   0.415455  0.556638  0.475794
12   embarrassment   0.300429  0.346535  0.321839
13      excitement   0.465011  0.461883  0.463442
14            fear   0.663492  0.502404  0.571819
15       gratitude   0.800640  0.722115  0.759353
16           grief   0.159509  0.250000  0.194757
17             joy   0.462920  0.556808  0.505541
18            love   0.831788  0.734503  0.780124
19     nervousness   0.300847  0.239057  0.266417
20         neutral   0.718947  0.865790  0.785565
21        optimism   0.437884  0.568236  0.494616
22           pride   0.134199  0.157360  0.144860
23     realization   0.257616  0.590909  0.358805
24          relief   0.207668  0.300926  0.245747
25         remorse   0.803279  0.549020  0.652246
26         sadness   0.571429  0.498925  0.532721
27        surprise   0.579562  0.512920  0.544208
```

Figure 18

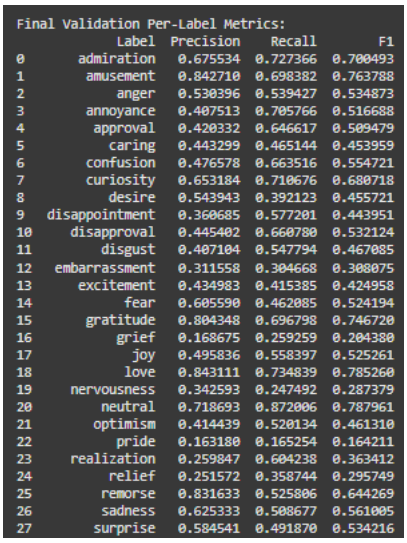
Figure: RoBERTa Experiment 2 Thresholds



	label	threshold	f1
admiration	admiration	0.40	0.693519
amusement	amusement	0.45	0.765893
anger	anger	0.25	0.549601
annoyance	annoyance	0.25	0.521687
approval	approval	0.25	0.507961
caring	caring	0.25	0.480874
confusion	confusion	0.25	0.536058
curiosity	curiosity	0.30	0.696015
desire	desire	0.25	0.467433
disappointment	disappointment	0.20	0.445950
disapproval	disapproval	0.25	0.524726
disgust	disgust	0.25	0.458685
embarrassment	embarrassment	0.15	0.297170
excitement	excitement	0.25	0.431088
fear	fear	0.35	0.549180
gratitude	gratitude	0.45	0.763061
grief	grief	0.10	0.224638
joy	joy	0.30	0.526882
love	love	0.50	0.790031
nervousness	nervousness	0.15	0.317460
neutral	neutral	0.35	0.789010
optimism	optimism	0.25	0.493671
pride	pride	0.10	0.180488
realization	realization	0.15	0.377838
relief	relief	0.10	0.303263
remorse	remorse	0.50	0.619926
sadness	sadness	0.30	0.572014
surprise	surprise	0.25	0.558659

Figure 19

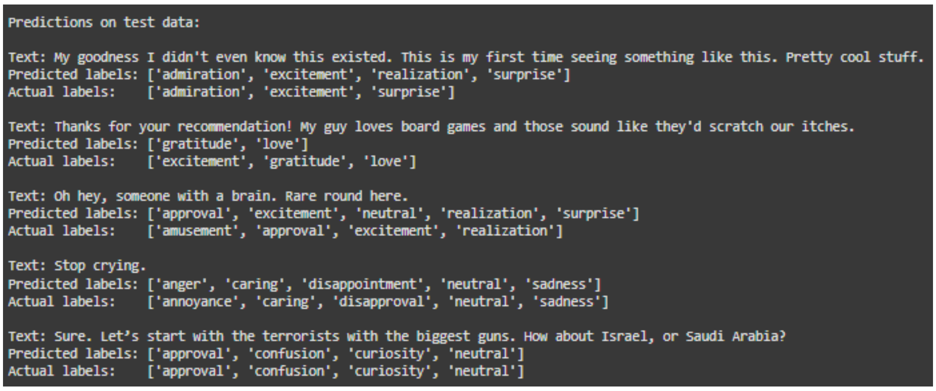
Figure: RoBERTa Experiment 2 Test Set Label Evaluation



```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.675534  0.727366  0.700493
1        amusement   0.842710  0.698382  0.763788
2            anger   0.530396  0.539427  0.534873
3        annoyance   0.407513  0.705766  0.516688
4         approval   0.420332  0.646617  0.509479
5           caring   0.443299  0.465144  0.453959
6        confusion   0.476578  0.663516  0.554721
7        curiosity   0.653184  0.710676  0.680718
8           desire   0.543943  0.392123  0.455721
9   disappointment   0.360685  0.577201  0.443951
10     disapproval   0.445402  0.660780  0.532124
11         disgust   0.407104  0.547794  0.467085
12   embarrassment   0.311558  0.304668  0.308075
13      excitement   0.434983  0.415385  0.424958
14            fear   0.605590  0.462085  0.524194
15       gratitude   0.804348  0.696798  0.746720
16           grief   0.168675  0.259259  0.204380
17             joy   0.495836  0.558397  0.525261
18            love   0.843111  0.734839  0.785260
19     nervousness   0.342593  0.247492  0.287379
20         neutral   0.718693  0.872006  0.787961
21        optimism   0.414439  0.520134  0.461310
22           pride   0.163180  0.165254  0.164211
23     realization   0.259847  0.604238  0.363412
24          relief   0.251572  0.358744  0.295749
25         remorse   0.831633  0.525806  0.644269
26         sadness   0.625333  0.508677  0.561005
27        surprise   0.584541  0.491870  0.534216
```

Figure 20

Figure: RoBERTa Experiment 2 Test Set Output



```
Predictions on test data:

Text: My goodness I didn't even know this existed. This is my first time seeing something like this. Pretty cool stuff.
Predicted labels: ['admiration', 'excitement', 'realization', 'surprise']
Actual labels:    ['admiration', 'excitement', 'surprise']

Text: Thanks for your recommendation! My guy loves board games and those sound like they'd scratch our itches.
Predicted labels: ['gratitude', 'love']
Actual labels:    ['excitement', 'gratitude', 'love']

Text: Oh hey, someone with a brain. Rare round here.
Predicted labels: ['approval', 'excitement', 'neutral', 'realization', 'surprise']
Actual labels:    ['amusement', 'approval', 'excitement', 'realization']

Text: Stop crying.
Predicted labels: ['anger', 'caring', 'disappointment', 'neutral', 'sadness']
Actual labels:    ['annoyance', 'caring', 'disapproval', 'neutral', 'sadness']

Text: Sure. Let’s start with the terrorists with the biggest guns. How about Israel, or Saudi Arabia?
Predicted labels: ['approval', 'confusion', 'curiosity', 'neutral']
Actual labels:    ['approval', 'confusion', 'curiosity', 'neutral']
```

Figure 21

Figure: RoBERTa Best Model (Experiment 2) Evaluation

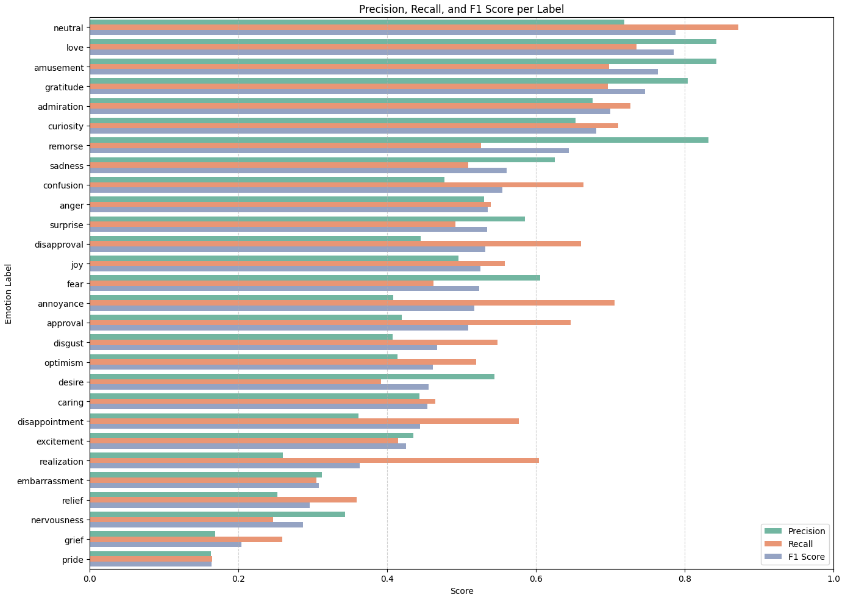


Figure 22

Figure: RoBERTa Best Model (Experiment 2) F1

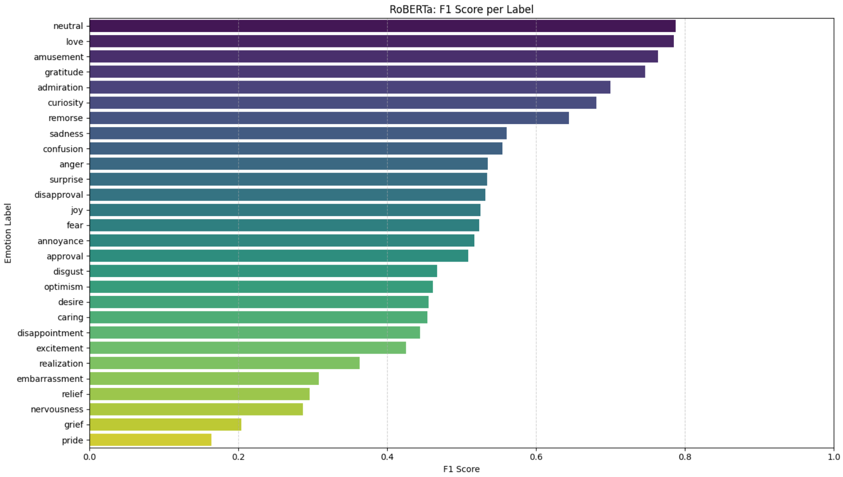


Figure 23

Figure: RoBERTa Best Model (Experiment 2) Recall

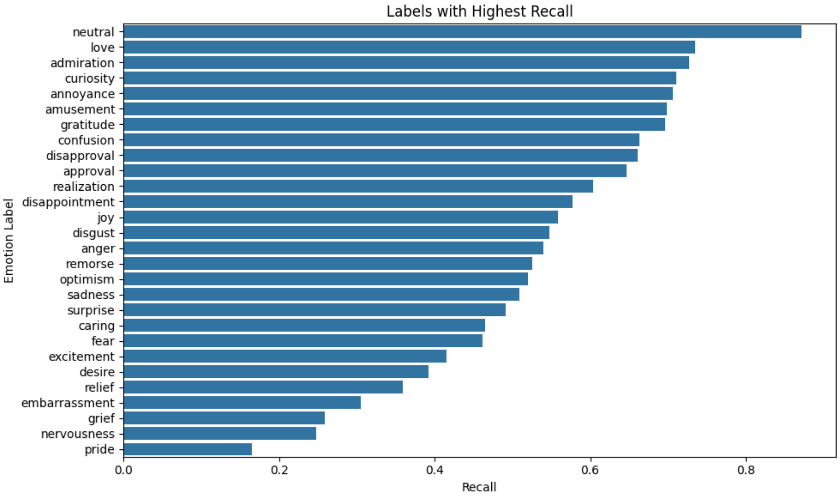
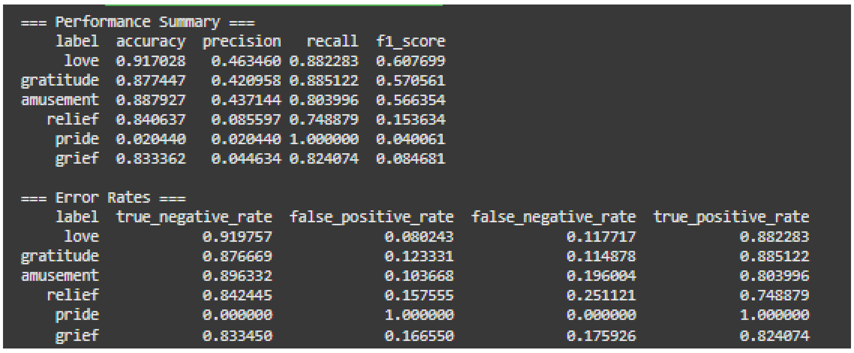


Figure 24

Figure: RoBERTa - Training N Binary Model Results



```
=== Performance Summary ===
     label  accuracy  precision   recall  f1_score
      love  0.917028   0.463460 0.882283  0.607699
 gratitude  0.877447   0.420958 0.885122  0.570561
 amusement  0.887927   0.437144 0.803996  0.566354
    relief  0.840637   0.085597 0.748879  0.153634
     pride  0.020440   0.020440 1.000000  0.040061
     grief  0.833362   0.044634 0.824074  0.084681

=== Error Rates ===
     label  true_negative_rate  false_positive_rate  false_negative_rate  true_positive_rate
      love            0.919757             0.080243             0.117717            0.882283
 gratitude            0.876669             0.123331             0.114878            0.885122
 amusement            0.896332             0.103668             0.196004            0.803996
    relief            0.842445             0.157555             0.251121            0.748879
     pride            0.000000             1.000000             0.000000            1.000000
     grief            0.833450             0.166550             0.175926            0.824074
```

Figure 25

Figure: Best RoBERTa Model - Minority Class Predictions

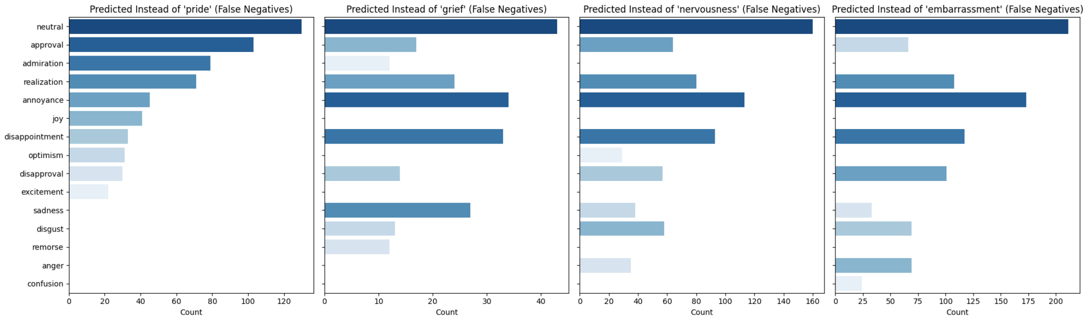


Figure 26: Other labels being chosen for minority class for best RoBERTa model

Figure: RoBERTa - Training N Binary Model F1 Comparison with Best Model

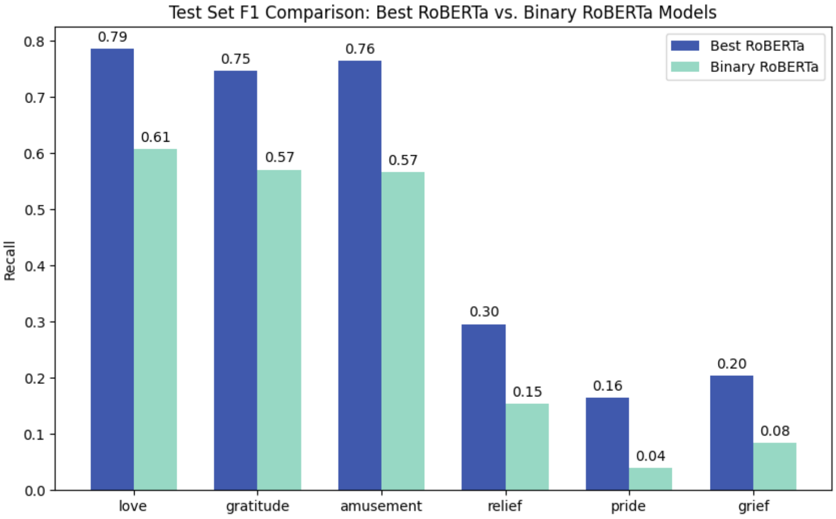
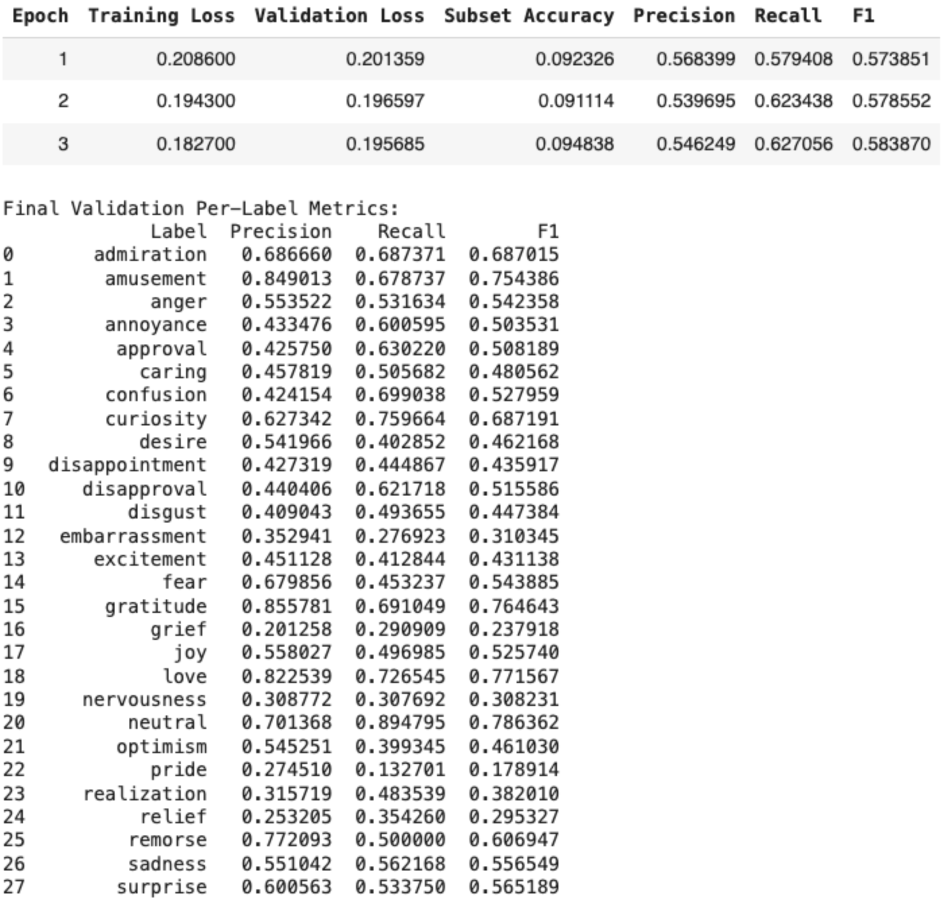


Figure 27

Figure: DistilBERT Experiment 2 + Freezing Training and Validation

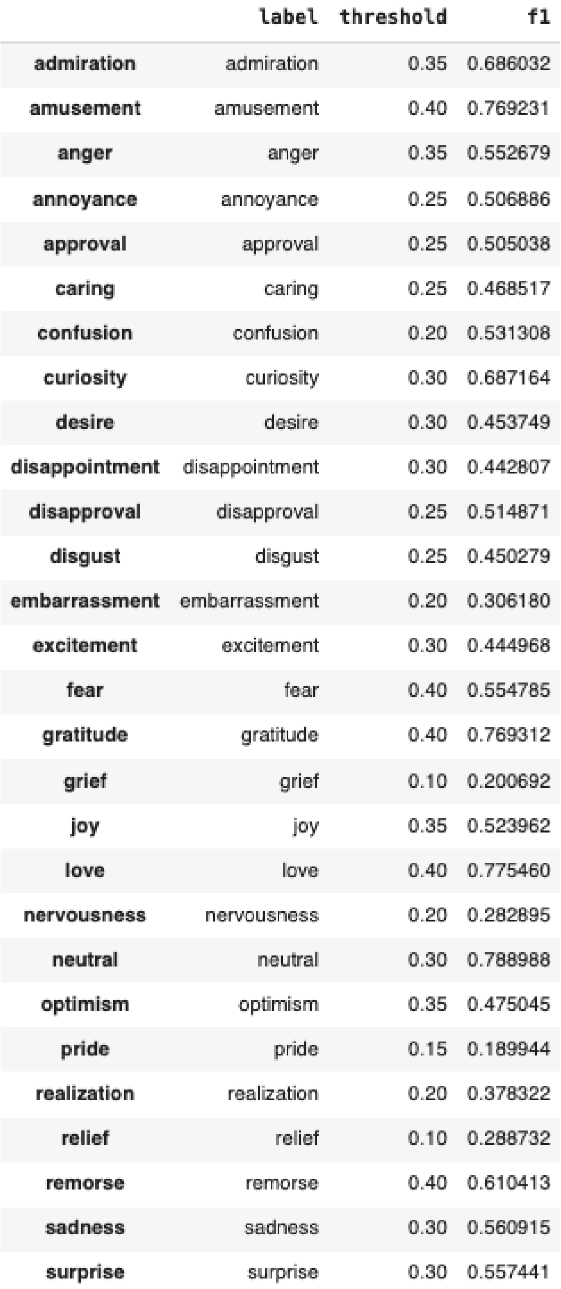


Epoch	Training Loss	Validation Loss	Subset Accuracy	Precision	Recall	F1
1	0.208600	0.201359	0.092326	0.568399	0.579408	0.573851
2	0.194300	0.196597	0.091114	0.539695	0.623438	0.578552
3	0.182700	0.195685	0.094838	0.546249	0.627056	0.583870

```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.686660  0.687371  0.687015
1        amusement   0.849013  0.678737  0.754386
2            anger   0.553522  0.531634  0.542358
3        annoyance   0.433476  0.600595  0.503531
4         approval   0.425750  0.630220  0.508189
5           caring   0.457819  0.505682  0.480562
6        confusion   0.424154  0.699038  0.527959
7        curiosity   0.627342  0.759664  0.687191
8           desire   0.541966  0.402852  0.462168
9   disappointment   0.427319  0.444867  0.435917
10     disapproval   0.440406  0.621718  0.515586
11         disgust   0.409043  0.493655  0.447384
12   embarrassment   0.352941  0.276923  0.310345
13      excitement   0.451128  0.412844  0.431138
14            fear   0.679856  0.453237  0.543885
15       gratitude   0.855781  0.691049  0.764643
16           grief   0.201258  0.290909  0.237918
17             joy   0.558027  0.496985  0.525740
18            love   0.822539  0.726545  0.771567
19     nervousness   0.308772  0.307692  0.308231
20         neutral   0.701368  0.894795  0.786362
21        optimism   0.545251  0.399345  0.461030
22           pride   0.274510  0.132701  0.178914
23     realization   0.315719  0.483539  0.382010
24          relief   0.253205  0.354260  0.295327
25         remorse   0.772093  0.500000  0.606947
26         sadness   0.551042  0.562168  0.556549
27        surprise   0.600563  0.533750  0.565189
```

Figure 28

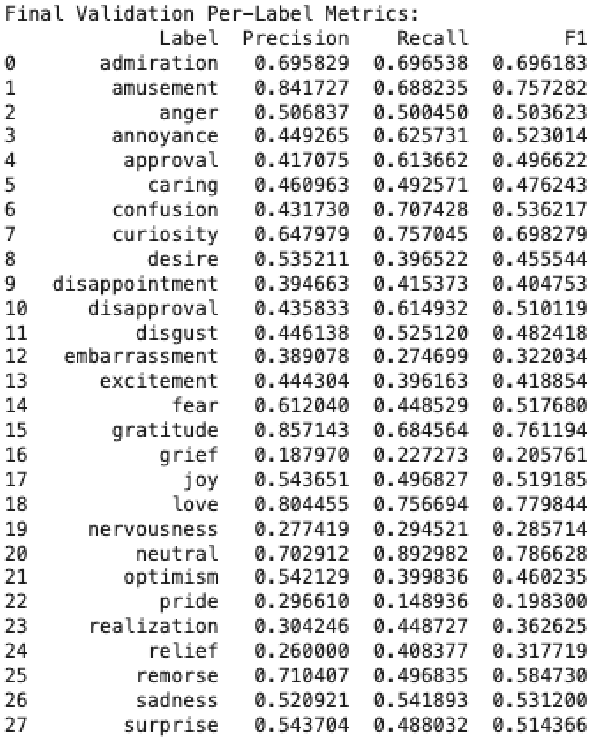
Figure: DistilBERT Experiment 2 + Freezing Thresholds



	label	threshold	f1
admiration	admiration	0.35	0.686032
amusement	amusement	0.40	0.769231
anger	anger	0.35	0.552679
annoyance	annoyance	0.25	0.506886
approval	approval	0.25	0.505038
caring	caring	0.25	0.468517
confusion	confusion	0.20	0.531308
curiosity	curiosity	0.30	0.687164
desire	desire	0.30	0.453749
disappointment	disappointment	0.30	0.442807
disapproval	disapproval	0.25	0.514871
disgust	disgust	0.25	0.450279
embarrassment	embarrassment	0.20	0.306180
excitement	excitement	0.30	0.444968
fear	fear	0.40	0.554785
gratitude	gratitude	0.40	0.769312
grief	grief	0.10	0.200692
joy	joy	0.35	0.523962
love	love	0.40	0.775460
nervousness	nervousness	0.20	0.282895
neutral	neutral	0.30	0.788988
optimism	optimism	0.35	0.475045
pride	pride	0.15	0.189944
realization	realization	0.20	0.378322
relief	relief	0.10	0.288732
remorse	remorse	0.40	0.610413
sadness	sadness	0.30	0.560915
surprise	surprise	0.30	0.557441

Figure 29

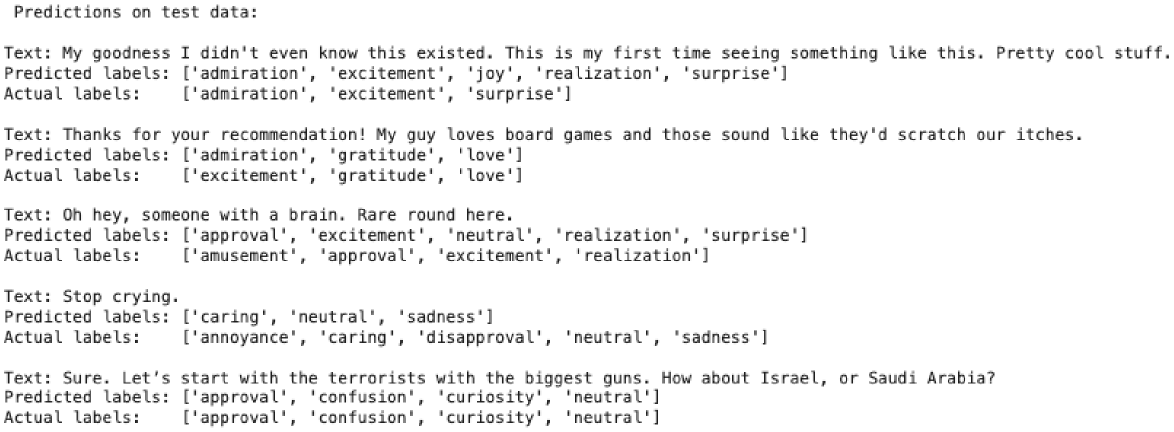
Figure: DistilBERT Experiment 2 + Freezing Test Set Label Evaluation



```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.695829  0.696538  0.696183
1        amusement   0.841727  0.688235  0.757282
2            anger   0.506837  0.500450  0.503623
3        annoyance   0.449265  0.625731  0.523014
4         approval   0.417075  0.613662  0.496622
5           caring   0.460963  0.492571  0.476243
6        confusion   0.431730  0.707428  0.536217
7        curiosity   0.647979  0.757045  0.698279
8           desire   0.535211  0.396522  0.455544
9   disappointment   0.394663  0.415373  0.404753
10     disapproval   0.435833  0.614932  0.510119
11         disgust   0.446138  0.525120  0.482418
12   embarrassment   0.389078  0.274699  0.322034
13      excitement   0.444304  0.396163  0.418854
14            fear   0.612040  0.448529  0.517680
15       gratitude   0.857143  0.684564  0.761194
16           grief   0.187970  0.227273  0.205761
17             joy   0.543651  0.496827  0.519185
18            love   0.804455  0.756694  0.779844
19     nervousness   0.277419  0.294521  0.285714
20         neutral   0.702912  0.892982  0.786628
21        optimism   0.542129  0.399836  0.460235
22           pride   0.296610  0.148936  0.198300
23     realization   0.304246  0.448727  0.362625
24          relief   0.260000  0.408377  0.317719
25         remorse   0.710407  0.496835  0.584730
26         sadness   0.520921  0.541893  0.531200
27        surprise   0.543704  0.488032  0.514366
```

Figure 30

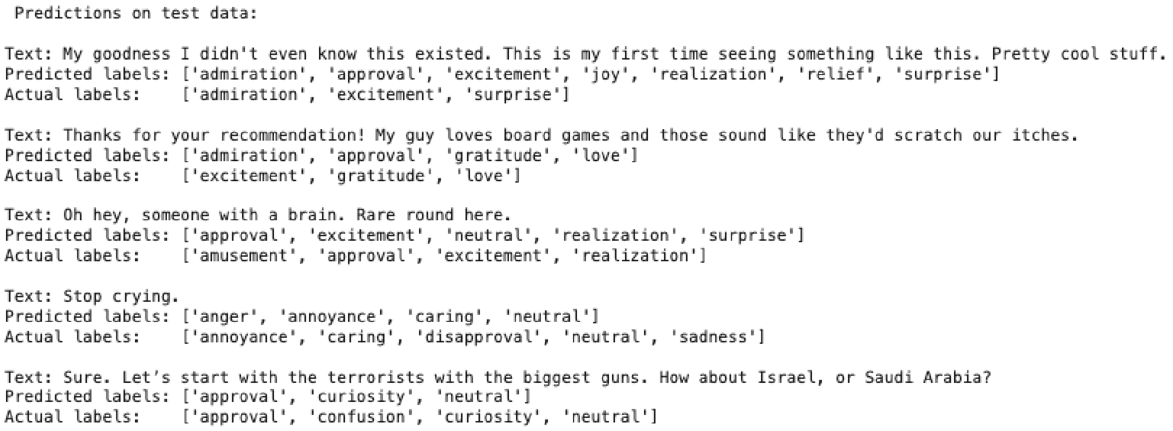
Figure: DistilBERT Experiment 2 + Freezing Test Set Output



```
 Predictions on test data:

Text: My goodness I didn't even know this existed. This is my first time seeing something like this. Pretty cool stuff.
Predicted labels: ['admiration', 'excitement', 'joy', 'realization', 'surprise']
Actual labels:    ['admiration', 'excitement', 'surprise']

Text: Thanks for your recommendation! My guy loves board games and those sound like they'd scratch our itches.
Predicted labels: ['admiration', 'gratitude', 'love']
Actual labels:    ['excitement', 'gratitude', 'love']

Text: Oh hey, someone with a brain. Rare round here.
Predicted labels: ['approval', 'excitement', 'neutral', 'realization', 'surprise']
Actual labels:    ['amusement', 'approval', 'excitement', 'realization']

Text: Stop crying.
Predicted labels: ['caring', 'neutral', 'sadness']
Actual labels:    ['annoyance', 'caring', 'disapproval', 'neutral', 'sadness']

Text: Sure. Let’s start with the terrorists with the biggest guns. How about Israel, or Saudi Arabia?
Predicted labels: ['approval', 'confusion', 'curiosity', 'neutral']
Actual labels:    ['approval', 'confusion', 'curiosity', 'neutral']
```

Figure 31

Figure: DistilBERT Experiment 4 Test Set Output



```
 Predictions on test data:

Text: My goodness I didn't even know this existed. This is my first time seeing something like this. Pretty cool stuff.
Predicted labels: ['admiration', 'approval', 'excitement', 'joy', 'realization', 'relief', 'surprise']
Actual labels:    ['admiration', 'excitement', 'surprise']

Text: Thanks for your recommendation! My guy loves board games and those sound like they'd scratch our itches.
Predicted labels: ['admiration', 'approval', 'gratitude', 'love']
Actual labels:    ['excitement', 'gratitude', 'love']

Text: Oh hey, someone with a brain. Rare round here.
Predicted labels: ['approval', 'excitement', 'neutral', 'realization', 'surprise']
Actual labels:    ['amusement', 'approval', 'excitement', 'realization']

Text: Stop crying.
Predicted labels: ['anger', 'annoyance', 'caring', 'neutral']
Actual labels:    ['annoyance', 'caring', 'disapproval', 'neutral', 'sadness']

Text: Sure. Let’s start with the terrorists with the biggest guns. How about Israel, or Saudi Arabia?
Predicted labels: ['approval', 'curiosity', 'neutral']
Actual labels:    ['approval', 'confusion', 'curiosity', 'neutral']
```

Figure 32

Figure: DistilBERT Best Model (Experiment 2 + Freezing) Evaluation

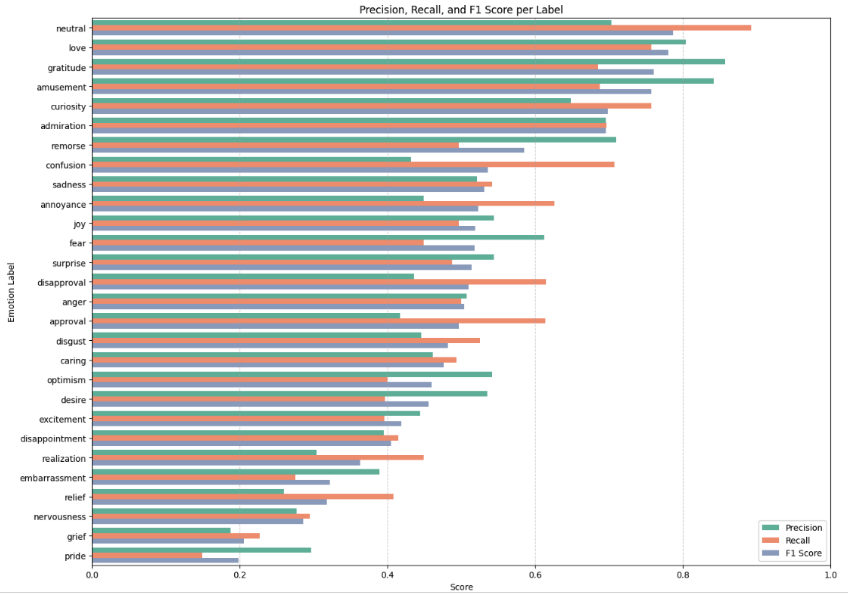


Figure 33

Figure: DistilBERT Best Model (Experiment 2 + Freezing) F1

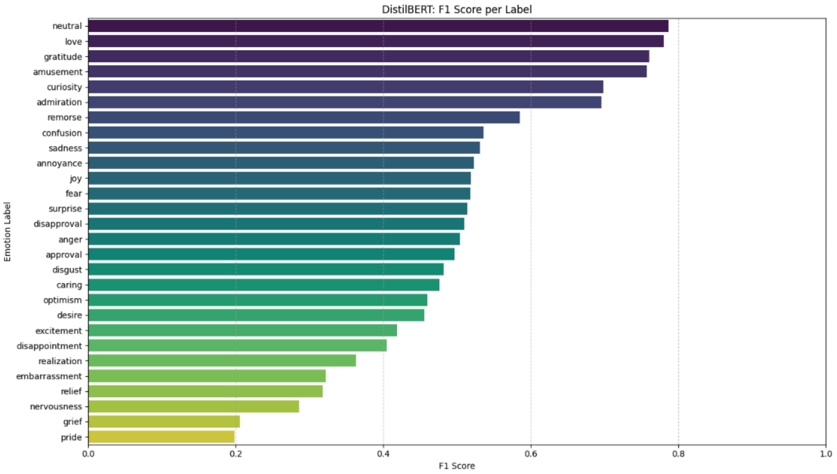


Figure 34

Figure: DistilBERT Best Model (Experiment 2 + Freezing) Recall

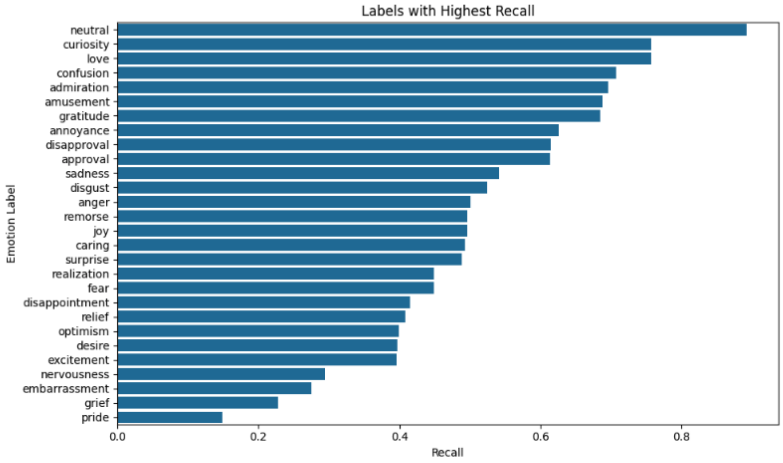
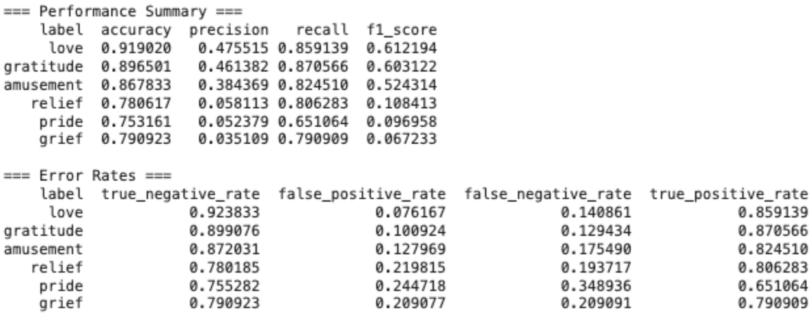


Figure 35

Figure: DistilBERT - Training N Binary Model Results



```
=== Performance Summary ===
     label  accuracy  precision   recall  f1_score
      love  0.919020   0.475515 0.859139  0.612194
 gratitude  0.896501   0.461382 0.870566  0.603122
 amusement  0.867833   0.384369 0.824510  0.524314
    relief  0.780617   0.058113 0.806283  0.108413
     pride  0.753161   0.052379 0.651064  0.096958
     grief  0.790923   0.035109 0.790909  0.067233

=== Error Rates ===
     label  true_negative_rate  false_positive_rate  false_negative_rate  true_positive_rate
      love            0.923833             0.076167             0.140861            0.859139
 gratitude            0.899076             0.100924             0.129434            0.870566
 amusement            0.872031             0.127969             0.175490            0.824510
    relief            0.780185             0.219815             0.193717            0.806283
     pride            0.755282             0.244718             0.348936            0.651064
     grief            0.790923             0.209077             0.209091            0.790909
```

Figure 36

Figure: Best DistilBERT Model - Minority Class Predictions

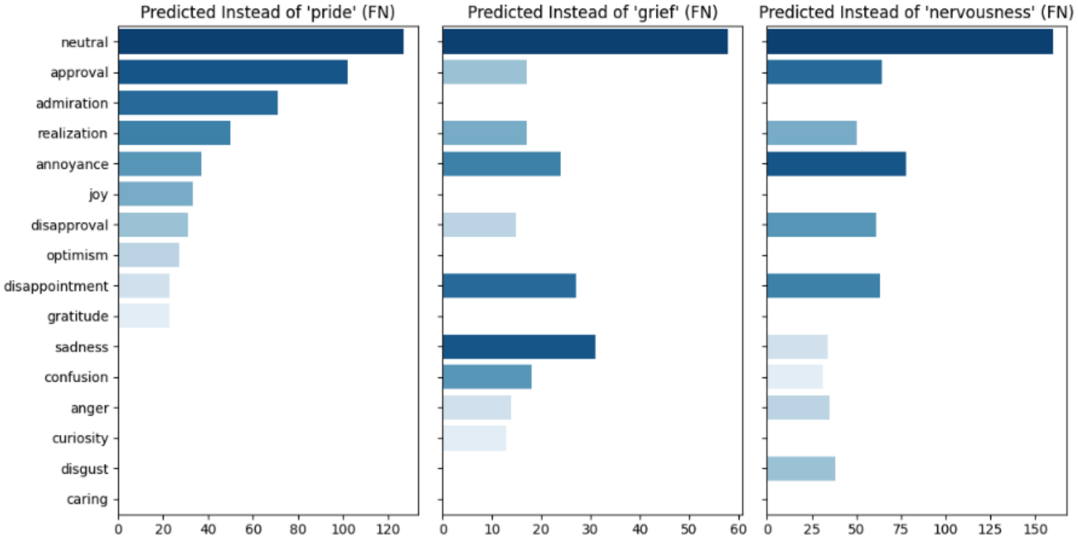


Figure 37: Other labels being chosen for minority class for best DistilBERT model

Figure: DistilBERT - Training N Binary Model F1 Comparison with Best Model

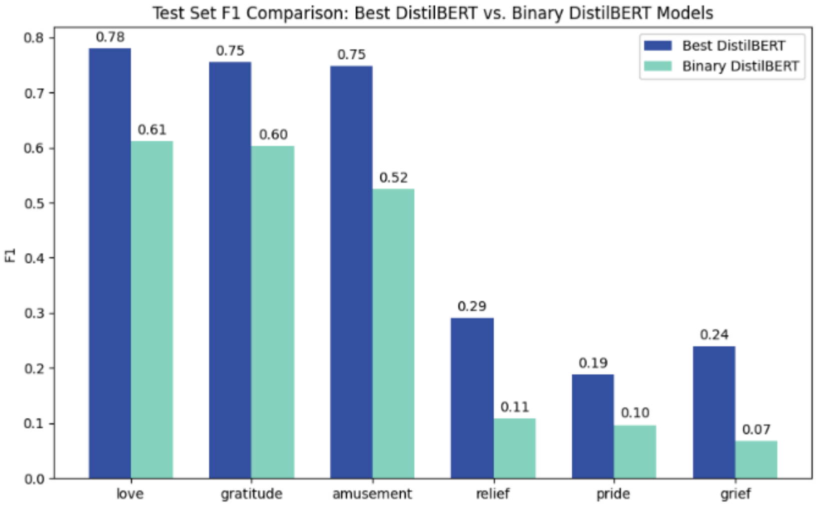
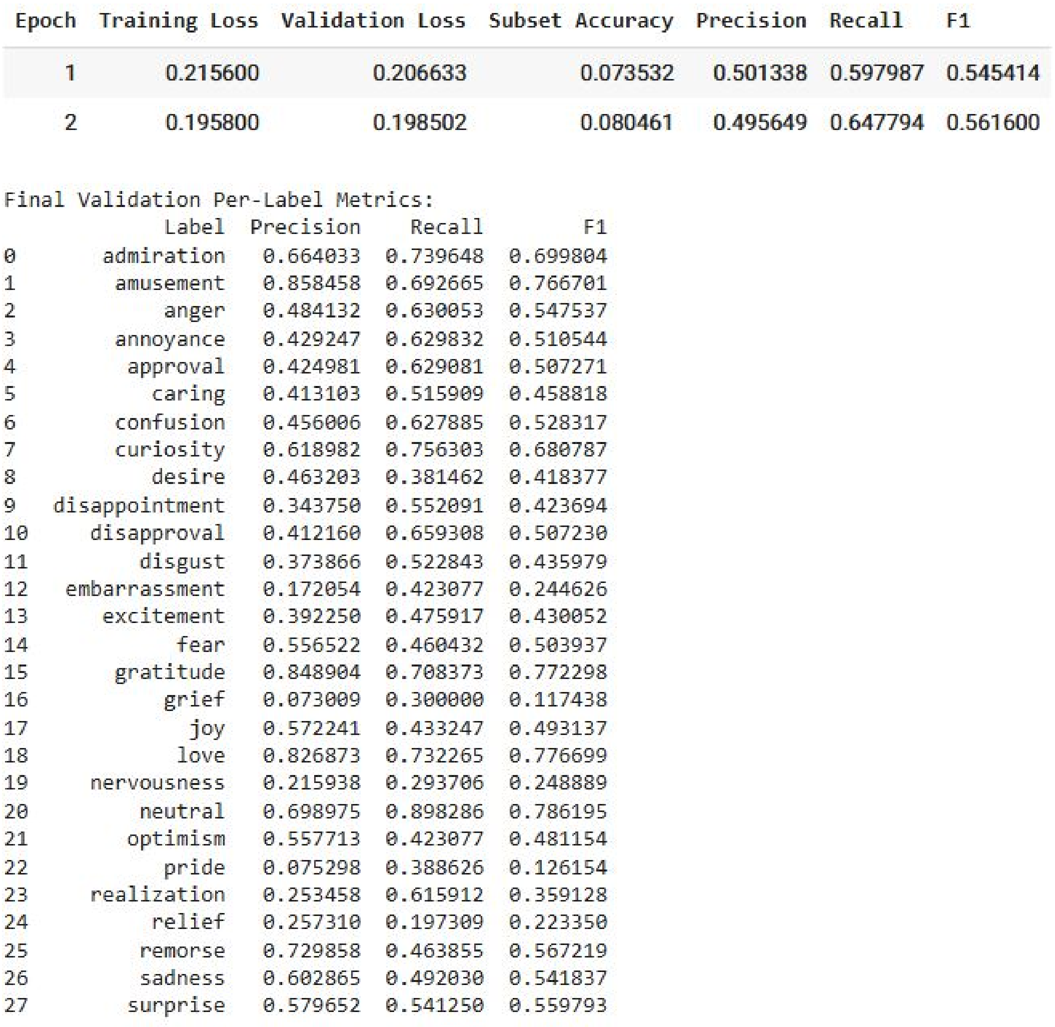


Figure 38

Figure: DeBERTa Experiment 2 Training and Validation

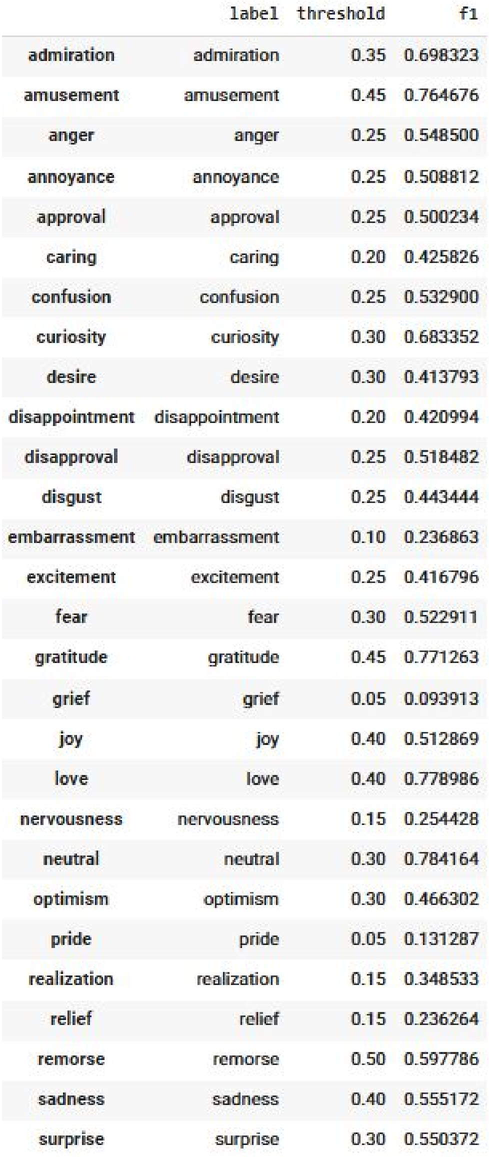


Epoch	Training Loss	Validation Loss	Subset Accuracy	Precision	Recall	F1
1	0.215600	0.206633	0.073532	0.501338	0.597987	0.545414
2	0.195800	0.198502	0.080461	0.495649	0.647794	0.561600

```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.664033  0.739648  0.699804
1        amusement   0.858458  0.692665  0.766701
2            anger   0.484132  0.630053  0.547537
3        annoyance   0.429247  0.629832  0.510544
4         approval   0.424981  0.629081  0.507271
5           caring   0.413103  0.515909  0.458818
6        confusion   0.456006  0.627885  0.528317
7        curiosity   0.618982  0.756303  0.680787
8           desire   0.463203  0.381462  0.418377
9   disappointment   0.343750  0.552091  0.423694
10     disapproval   0.412160  0.659308  0.507230
11         disgust   0.373866  0.522843  0.435979
12   embarrassment   0.172054  0.423077  0.244626
13      excitement   0.392250  0.475917  0.430052
14            fear   0.556522  0.460432  0.503937
15       gratitude   0.848904  0.708373  0.772298
16           grief   0.073009  0.300000  0.117438
17             joy   0.572241  0.433247  0.493137
18            love   0.826873  0.732265  0.776699
19     nervousness   0.215938  0.293706  0.248889
20         neutral   0.698975  0.898286  0.786195
21        optimism   0.557713  0.423077  0.481154
22           pride   0.075298  0.388626  0.126154
23     realization   0.253458  0.615912  0.359128
24          relief   0.257310  0.197309  0.223350
25         remorse   0.729858  0.463855  0.567219
26         sadness   0.602865  0.492030  0.541837
27        surprise   0.579652  0.541250  0.559793
```

Figure 39

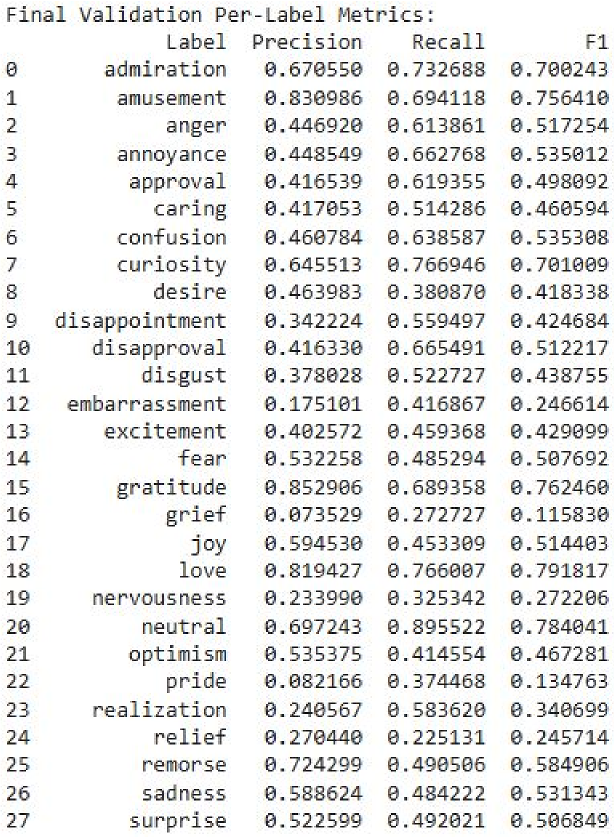
Figure: DeBERTa Experiment 2 Thresholds



	label	threshold	f1
admiration	admiration	0.35	0.698323
amusement	amusement	0.45	0.764676
anger	anger	0.25	0.548500
annoyance	annoyance	0.25	0.508812
approval	approval	0.25	0.500234
caring	caring	0.20	0.425826
confusion	confusion	0.25	0.532900
curiosity	curiosity	0.30	0.683352
desire	desire	0.30	0.413793
disappointment	disappointment	0.20	0.420994
disapproval	disapproval	0.25	0.518482
disgust	disgust	0.25	0.443444
embarrassment	embarrassment	0.10	0.236863
excitement	excitement	0.25	0.416796
fear	fear	0.30	0.522911
gratitude	gratitude	0.45	0.771263
grief	grief	0.05	0.093913
joy	joy	0.40	0.512869
love	love	0.40	0.778986
nervousness	nervousness	0.15	0.254428
neutral	neutral	0.30	0.784164
optimism	optimism	0.30	0.466302
pride	pride	0.05	0.131287
realization	realization	0.15	0.348533
relief	relief	0.15	0.236264
remorse	remorse	0.50	0.597786
sadness	sadness	0.40	0.555172
surprise	surprise	0.30	0.550372

Figure 40

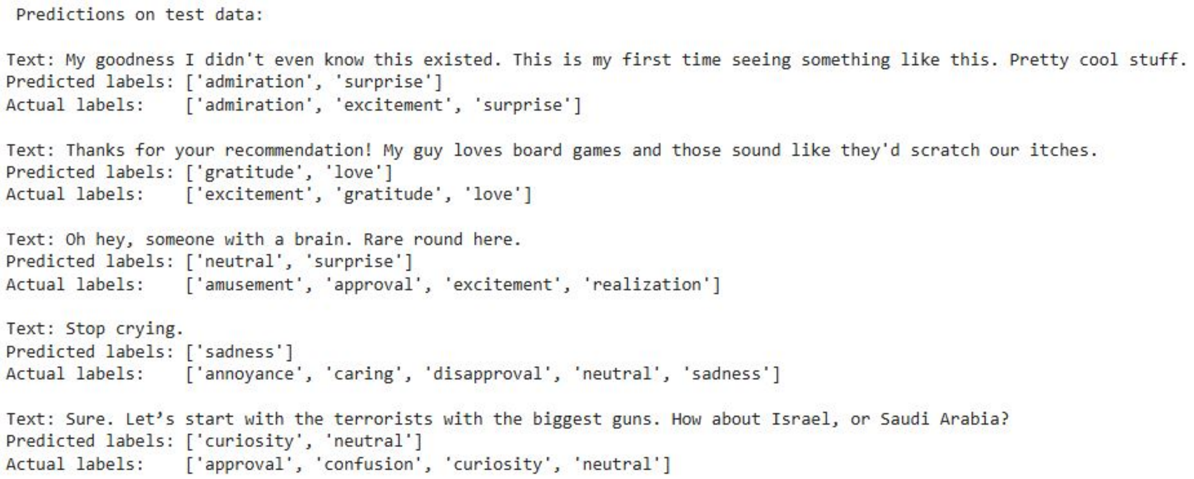
Figure: DeBERTa Experiment 2 Test Set Label Evaluation



```
Final Validation Per-Label Metrics:
             Label  Precision    Recall        F1
0       admiration   0.670550  0.732688  0.700243
1        amusement   0.830986  0.694118  0.756410
2            anger   0.446920  0.613861  0.517254
3        annoyance   0.448549  0.662768  0.535012
4         approval   0.416539  0.619355  0.498092
5           caring   0.417053  0.514286  0.460594
6        confusion   0.460784  0.638587  0.535308
7        curiosity   0.645513  0.766946  0.701009
8           desire   0.463983  0.380870  0.418338
9   disappointment   0.342224  0.559497  0.424684
10     disapproval   0.416330  0.665491  0.512217
11         disgust   0.378028  0.522727  0.438755
12   embarrassment   0.175101  0.416867  0.246614
13      excitement   0.402572  0.459368  0.429099
14            fear   0.532258  0.485294  0.507692
15       gratitude   0.852906  0.689358  0.762460
16           grief   0.073529  0.272727  0.115830
17             joy   0.594530  0.453309  0.514403
18            love   0.819427  0.766007  0.791817
19     nervousness   0.233990  0.325342  0.272206
20         neutral   0.697243  0.895522  0.784041
21        optimism   0.535375  0.414554  0.467281
22           pride   0.082166  0.374468  0.134763
23     realization   0.240567  0.583620  0.340699
24          relief   0.270440  0.225131  0.245714
25         remorse   0.724299  0.490506  0.584906
26         sadness   0.588624  0.484222  0.531343
27        surprise   0.522599  0.492021  0.506849
```

Figure 41

Figure: DeBERTa Experiment 2 Test Set Output



```
 Predictions on test data:

Text: My goodness I didn't even know this existed. This is my first time seeing something like this. Pretty cool stuff.
Predicted labels: ['admiration', 'surprise']
Actual labels:    ['admiration', 'excitement', 'surprise']

Text: Thanks for your recommendation! My guy loves board games and those sound like they'd scratch our itches.
Predicted labels: ['gratitude', 'love']
Actual labels:    ['excitement', 'gratitude', 'love']

Text: Oh hey, someone with a brain. Rare round here.
Predicted labels: ['neutral', 'surprise']
Actual labels:    ['amusement', 'approval', 'excitement', 'realization']

Text: Stop crying.
Predicted labels: ['sadness']
Actual labels:    ['annoyance', 'caring', 'disapproval', 'neutral', 'sadness']

Text: Sure. Let’s start with the terrorists with the biggest guns. How about Israel, or Saudi Arabia?
Predicted labels: ['curiosity', 'neutral']
Actual labels:    ['approval', 'confusion', 'curiosity', 'neutral']
```

Figure 42

Figure: DeBERTa Best Model (Experiment 2) Evaluation

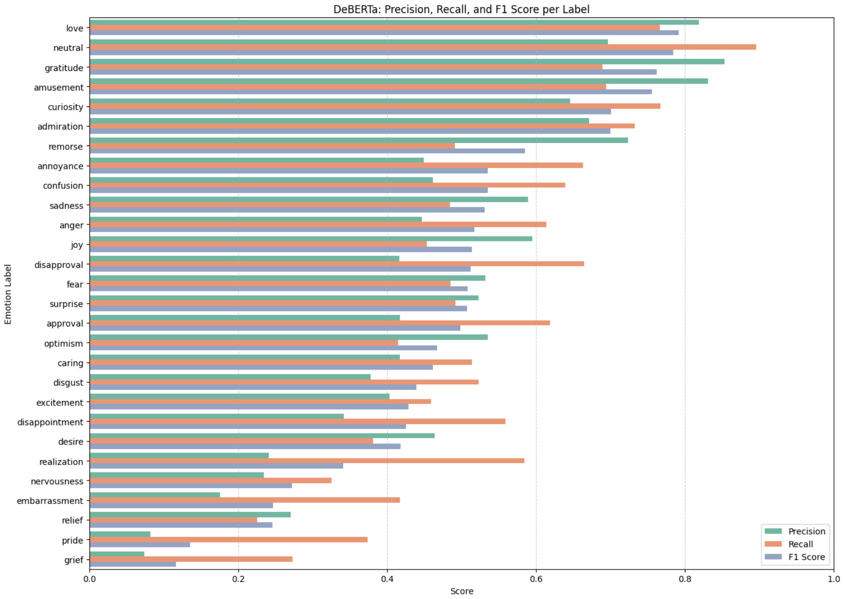


Figure 43

Figure: DeBERTa Best Model (Experiment 2) F1

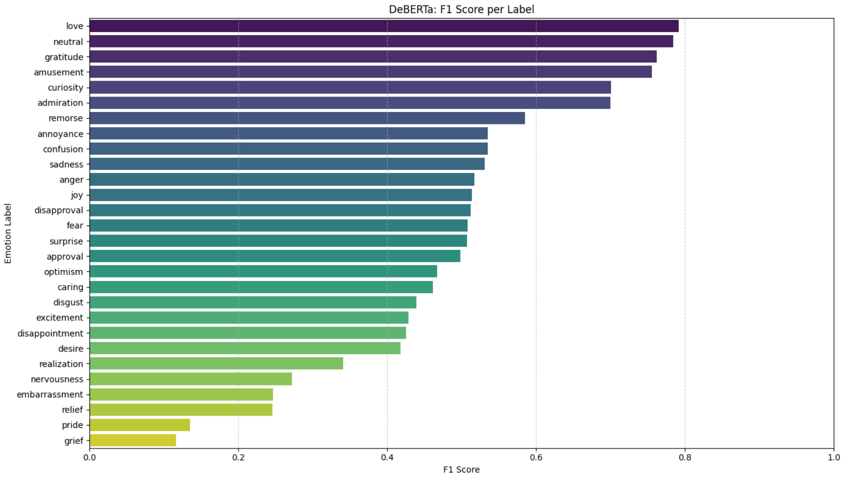


Figure 44

Figure: DeBERTa Best Model (Experiment 2) Recall

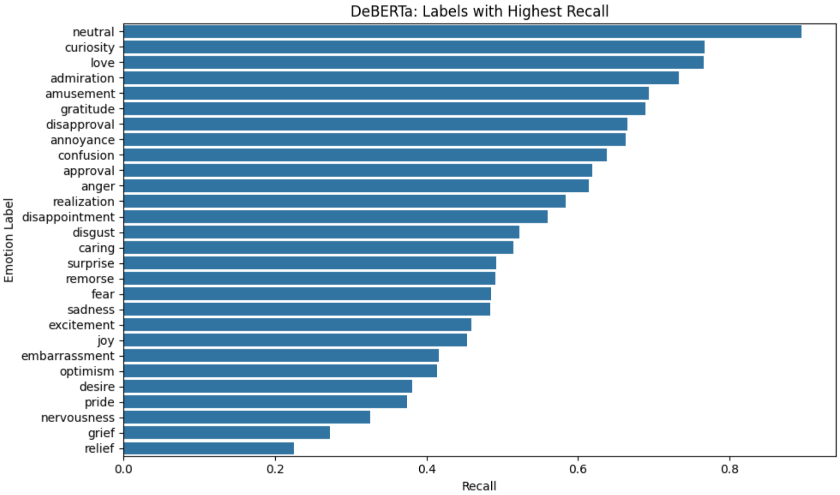


Figure 45

Figure: DeBERTa - Training N Binary Model Results

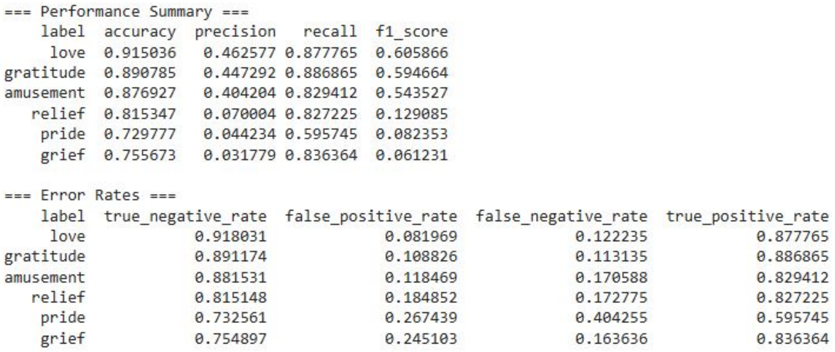

```
=== Performance Summary ===
    label  accuracy  precision   recall  f1_score
     love  0.915036   0.462577 0.877765  0.605866
gratitude  0.890785   0.447292 0.886865  0.594664
amusement  0.876927   0.404204 0.829412  0.543527
   relief  0.815347   0.070004 0.827225  0.129085
    pride  0.729777   0.044234 0.595745  0.082353
    grief  0.755673   0.031779 0.836364  0.061231

=== Error Rates ===
    label  true_negative_rate  false_positive_rate  false_negative_rate  true_positive_rate
     love            0.918031             0.081969             0.122235            0.877765
gratitude            0.891174             0.108826             0.113135            0.886865
amusement            0.881531             0.118469             0.170588            0.829412
   relief            0.815148             0.184852             0.172775            0.827225
    pride            0.732561             0.267439             0.404255            0.595745
    grief            0.754897             0.245103             0.163636            0.836364
```

Figure 46

Figure: Best DeBERTa Model - Minority Class Predictions

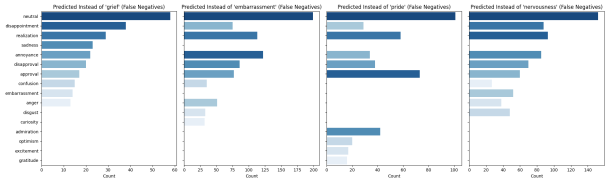


Figure 47: Other labels being chosen for minority class for best DeBERTa model

Figure: DeBERTa - Training N Binary Model F1 Comparison with Best Model

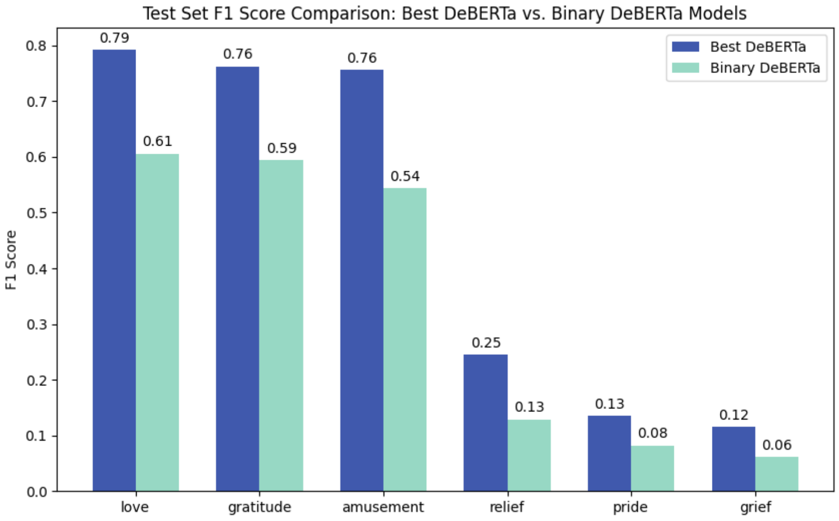


Figure 48