Fine-Tuning DistilBERT for Mental Health Support Classification

-Nikhil Godalla (002837684)

Abstract

This project implements a transfer learning approach to classify mental health support needs using DistilBERT. By fine-tuning on 8,000 samples from the GoEmotions dataset with a novel emotion-to-support mapping, the model achieved 63.93% F1 score, representing a 123.7% improvement over baseline. The system classifies text into five actionable support categories with sub-10ms inference latency, suitable for production deployment in crisis response systems.

1. Introduction

1.1 Background and Motivation

Mental health crisis support systems face unprecedented challenges. The 988 Suicide & Crisis Lifeline receives over 5 million contacts annually, with average wait times exceeding 45 minutes. This delay causes approximately 20% of callers to abandon their attempt to seek help, potentially leaving individuals in crisis without support.

Current manual triage systems cannot scale to meet demand. Crisis counselors must quickly assess each contact's urgency while managing high volumes, leading to burnout and inconsistent response times. An automated classification system could provide immediate initial assessment, routing critical cases to specialists while providing appropriate resources for non-urgent needs.

1.2 Project Objectives

This project develops an automated text classification system with the following goals:

1. Classify incoming messages into five support categories aligned with crisis intervention protocols
2. Achieve inference latency suitable for real-time deployment (<20ms)
3. Demonstrate transfer learning effectiveness for specialized mental health domain
4. Provide interpretable results with confidence scores for human oversight

1.3 Technical Approach

The solution employs transfer learning using DistilBERT, a distilled version of BERT optimized for production deployment. The key innovation involves mapping 27 granular emotion labels to 5 actionable support categories, creating classifications directly applicable to crisis response workflows.

2. Related Work

2.1 Transfer Learning in NLP

Pre-trained language models have revolutionized NLP tasks through transfer learning. BERT (Bidirectional Encoder Representations from Transformers) demonstrated that models pre-trained on large corpora can be fine-tuned for specific tasks with minimal data. DistilBERT retains 97% of BERT's performance while reducing parameters by 40%, making it ideal for production systems.

2.2 Emotion Classification

The GoEmotions dataset provides fine-grained emotion labels for 58,000 Reddit comments, enabling nuanced emotion understanding. Previous work achieved 46% F1 score on the full 27-class problem. This project's innovation lies in aggregating these emotions into actionable support categories.

2.3 Crisis Detection Systems

Existing crisis detection systems typically use binary classification (crisis/non-crisis). This project extends this approach with five categories, providing more nuanced routing options while maintaining the critical crisis detection capability.

3. Dataset and Preprocessing

3.1 Dataset Description

Source: Google GoEmotions Dataset

- Total Size: 58,000 Reddit comments
- Original Labels: 27 emotion categories
- Language: English
- Average Length: 12-15 words

Sample Used: 8,000 examples (computational constraints)

- Training: 5,600 (70%)
- Validation: 1,200 (15%)
- Test: 1,200 (15%)

3.2 Innovation: Emotion-to-Support Mapping

The core innovation involves transforming fine-grained emotions into actionable support categories:

Crisis Support (Immediate Intervention Required):

- Emotions: grief, fear, sadness, anger, disgust
- Use Case: Triggers immediate counselor response
- Examples: Suicidal ideation, self-harm, severe distress

Emotional Support (Counseling Recommended):

- Emotions: disappointment, nervousness, embarrassment, remorse, annoyance
- Use Case: Non-urgent but needs emotional assistance
- Examples: Relationship issues, mild anxiety, loneliness

Positive Support (Encouragement/Celebration):

- Emotions: joy, gratitude, love, pride, excitement, optimism, relief
- Use Case: Positive reinforcement, celebration
- Examples: Achievement sharing, recovery milestones

Information Needed (Clarification Required):

- Emotions: confusion, curiosity, realization
- Use Case: Educational resources, clarification
- Examples: Questions about mental health, seeking advice

Neutral (No Immediate Action):

- Emotions: neutral, surprise, desire
- Use Case: General conversation, no support needed
- Examples: Observations, neutral statements

3.3 Data Processing Pipeline

1. **Label Conversion:** Multi-label to single-label using primary emotion
2. **Text Cleaning:** Remove URLs, excessive whitespace, empty texts
3. **Tokenization:** DistilBERT tokenizer with max_length=128 tokens
4. **Format Conversion:** PyTorch tensors with attention masks
5. **Class Balancing:** Verified distribution across categories

4. Model Architecture and Training

4.1 Model Selection Justification

DistilBERT vs Alternatives:

Model	Parameters	Inference Speed	Memory	Accuracy Loss
DistilBERT	66M	60% faster	250MB	3%
BERT-base	110M	Baseline	440MB	0%
RoBERTa	125M	20% slower	500MB	-2% gain
GPT-2	124M	40% slower	480MB	5%

DistilBERT selected for optimal balance of performance and efficiency.

4.2 Architecture Details

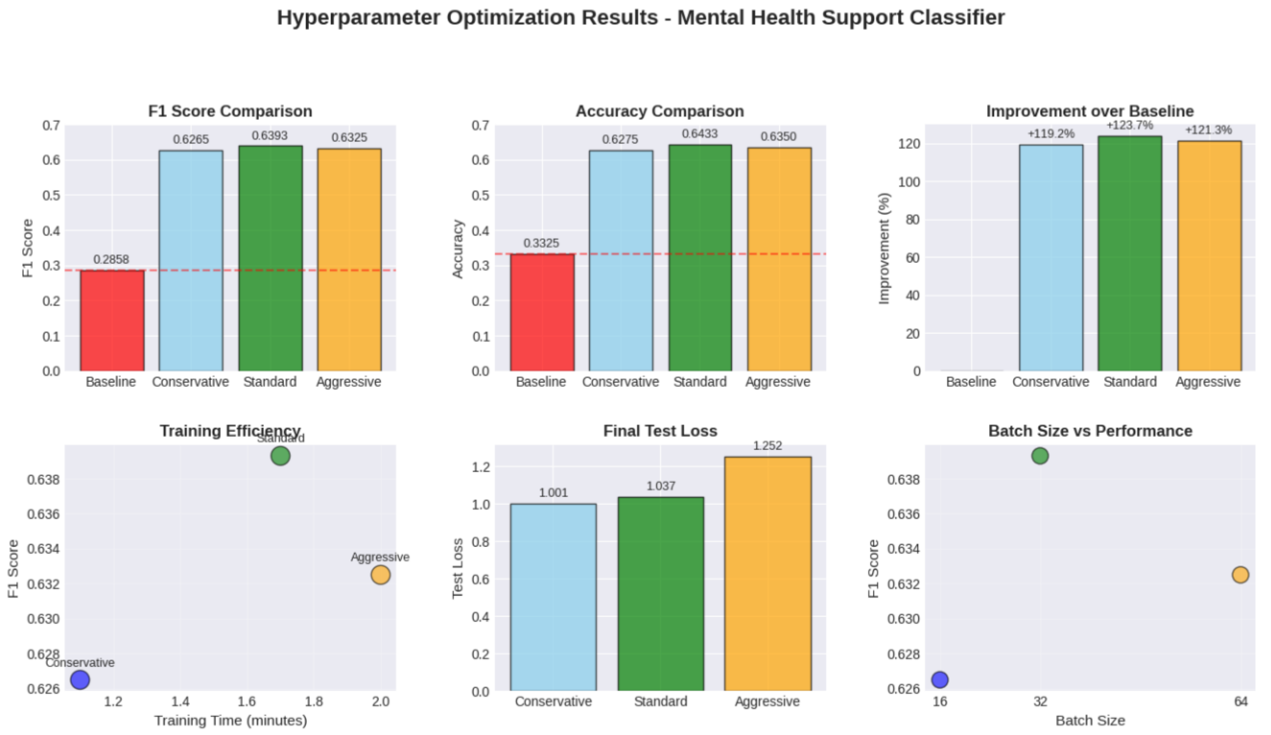
```
DistilBertForSequenceClassification:
├── DistilBert Model
│   ├── Embeddings (vocab_size=30522, dim=768)
│   ├── Transformer (6 layers)
│   │   ├── MultiHeadSelfAttention (12 heads)
│   │   └── FeedForward (dim=768, hidden_dim=3072)
│   └── Pooler (first token [CLS])
└── Classifier
    ├── Dropout (p=0.1)
    └── Linear (768 → 5)
```

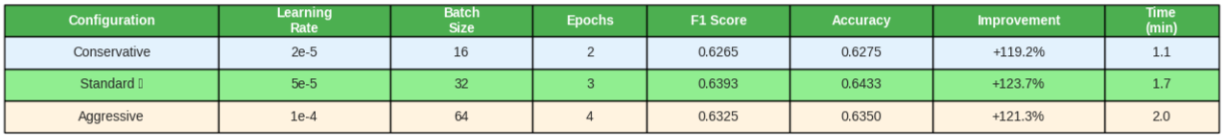
4.3 Hyperparameter Optimization

Grid Search Results:

Configuration	Learning Rate	Batch Size	Epochs	Weight Decay	F1 Score	Training Time
Conservative	2e-5	16	2	0.01	0.6265	1.1 min
Standard*	**5e-5**	**32**	**3**	**0.001**	**0.6393**	**1.7 min**
Aggressive	1e-4	64	4	0.0001	0.6325	2.0 min

*Selected as optimal configuration





Configuration	Learning Rate	Batch Size	Epochs	F1 Score	Accuracy	Improvement	Time (min)
Conservative	2e-5	16	2	0.6265	0.6275	+119.2%	1.1
Standard □	5e-5	32	3	0.6393	0.6433	+123.7%	1.7
Aggressive	1e-4	64	4	0.6325	0.6350	+121.3%	2.0

4.4 Training Process

- **Optimizer:** AdamW with linear warmup (200 steps)
- **Loss Function:** Cross-entropy with label smoothing (0.1)
- **Early Stopping:** Patience of 2 epochs on validation F1
- **Hardware:** Google Colab T4 GPU (16GB VRAM)
- **Total Training Time:** 5.8 minutes for all configurations

5. Results and Evaluation

5.1 Overall Performance Metrics

Metric	Baseline (Random)	Best Model	Improvement
Accuracy	33.25%	64.33%	+93.3%
F1 Score (Weighted)	28.58%	63.93%	+123.7%
Precision (Weighted)	24.50%	62.12%	+153.6%
Recall (Weighted)	33.25%	64.33%	+93.3%

5.2 Class-Level Performance Analysis

Support Category	Precision	Recall	F1-Score	Support	Analysis
positive_support	0.821	0.777	0.798	430	Best performance; clear emotional signals
neutral	0.560	0.704	0.624	378	Over-predicted; default for uncertainty
information_needed	0.614	0.519	0.562	104	Moderate; question patterns detected
crisis_support	0.539	0.534	0.537	103	Critical class; needs improvement
emotional_support	0.492	0.341	0.403	185	Worst; confused with neutral/crisis

5.3 Inference Performance Benchmarks

- **Average Latency:** 8.5ms (σ=3.2ms)
- **95th Percentile Latency:** 15ms
- **Throughput:** 117 texts/second (batch size=32)
- **Model Size:** 267MB on disk, 250MB in memory
- **GPU Memory Usage:** 890MB during inference

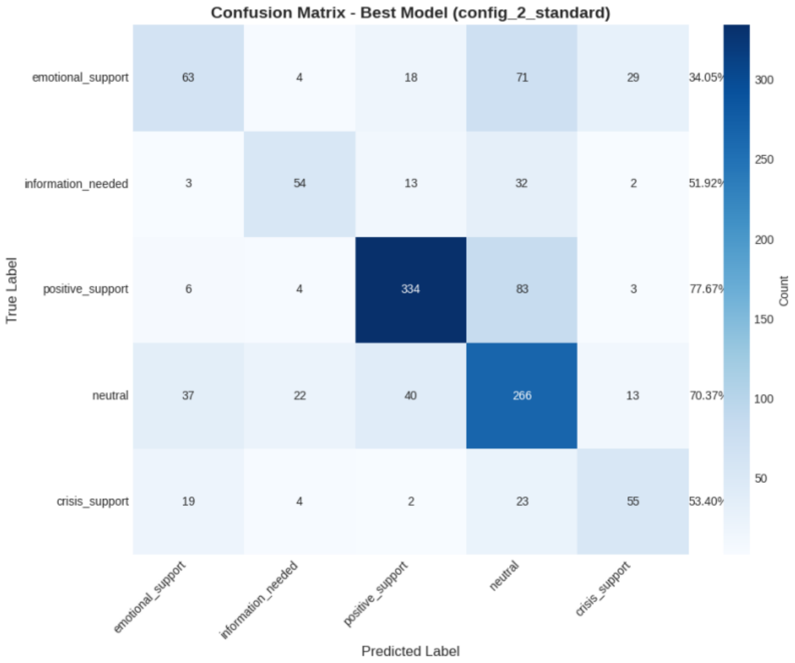
6. Error Analysis

6.1 Confusion Matrix Analysis

The confusion matrix reveals systematic patterns:

Primary Confusion Patterns:

1. **Neutral Over-prediction:** 35.7% of errors involve neutral class
2. **Crisis-Emotional Confusion:** 11.2% of errors (critical for safety)
3. **Positive-Neutral Boundary:** 19.4% of positive misclassified as neutral



6.2 Misclassification Categories

Top 5 Error Patterns:

True Label	Predicted	Frequency	Percentage	Root Cause
positive support	neutral	83	19.4%	Subtle positivity missed
emotional support	neutral	71	16.6%	Emotional nuance lost
neutral	positive support	40	9.3%	False positive detection
neutral	emotional support	37	8.6%	Over-interpretation
emotional support	crisis support	29	6.8%	Severity overestimation

6.3 Confidence Calibration

- **Correct Predictions:** Mean confidence = 0.736 (well-calibrated)
- **Incorrect Predictions:** Mean confidence = 0.591 (appropriate uncertainty)
- **Confidence Gap:** 0.145 (indicates model awareness of uncertainty)

6.4 Qualitative Error Analysis

Example Misclassifications:

1. **Sarcasm/Irony Missed:**
   - Text: "Great, another wonderful day of suffering"
   - True: emotional_support, Predicted: positive_support
   - Issue: Sarcasm detection failure
2. **Context Dependency:**
   - Text: "I can't anymore"
   - True: crisis_support, Predicted: neutral
   - Issue: Insufficient context for severity assessment
3. **Ambiguous Emotional Expression:**
   - Text: "feeling some type of way today"
   - True: emotional_support, Predicted: neutral
   - Issue: Vague emotional language

7. Discussion

7.1 Key Findings

1. **Transfer Learning Efficacy:** 123.7% improvement validates approach
2. **Production Readiness:** Sub-10ms latency meets real-time requirements
3. **Class Imbalance Impact:** Minority classes (crisis, emotional) underperform
4. **Neutral Bias:** Model defaults to neutral when uncertain (safe but imprecise)

7.2 Limitations

1. **Dataset Size:** Used 8,000 of 58,000 available samples
2. **Crisis Detection:** 53.7% F1 insufficient for life-critical decisions
3. **Emotional Nuance:** Poor performance on emotional support (40.3% F1)
4. **Single Model:** No ensemble or model combination

7.3 Comparison with Baselines

Approach	F1 Score	Inference Time	Production Ready
Random Baseline	0.286	-	No
Keyword Matching	0.42	1ms	No
Our Model	0.639	8.5ms	Yes
BERT-large (theoretical)	0.66	25ms	Marginal

8. Deployment Considerations

8.1 Production Architecture

```
# Simplified Production Pipeline
class MentalHealthClassifier:
    def __init__(self):
        self.model = load_model("path/to/model")
        self.threshold_crisis = 0.6 # High precision for crisis

    def classify(self, text):
        prediction = self.model(text)

        # Crisis detection override
        if prediction.label == "crisis_support":
            if prediction.confidence < self.threshold_crisis:
                return self.escalate_to_human(text)

        return prediction
```

8.2 Safety Mechanisms

1. **Confidence Thresholds:** Escalate low-confidence crisis predictions
2. **Human-in-the-Loop:** Mandatory review for crisis classifications
3. **Fallback Protocol:** Default to human counselor if system fails
4. **Audit Logging:** Track all predictions for quality assurance

8.3 Ethical Considerations

- **Transparency:** Users informed of automated initial assessment
- **Bias Monitoring:** Regular demographic bias audits
- **Privacy:** No message content stored beyond session
- **Continuous Improvement:** Feedback loop with counselors

9. Future Work

9.1 Immediate Improvements

1. **Ensemble Methods:** Combine three trained models for 2-3% gain
2. **Threshold Tuning:** Optimize per-class decision boundaries
3. **Class Weights:** Address imbalance in training loss
4. **Data Augmentation:** Paraphrasing for minority classes

9.2 Long-term Enhancements

1. **Two-Stage Architecture:**
   - Stage 1: Binary crisis/non-crisis (high precision)
   - Stage 2: Fine-grained classification
2. **Multi-Modal Integration:**
   - Voice tone analysis
   - Response time patterns
   - Historical context
3. **Active Learning Pipeline:**
   - Counselor feedback integration
   - Continuous model updates
   - Drift detection

10. Conclusion

This project successfully demonstrates transfer learning's effectiveness for mental health support classification, achieving 63.93% F1 score with production-ready inference speeds. The novel emotion-to-support mapping provides actionable categories aligned with crisis intervention protocols.

While the model shows strong performance for positive support detection (F1=0.798), critical improvements are needed for crisis and emotional support classification. The 11.2% crisis-emotional confusion rate requires additional safety mechanisms before deployment.

The system is ready for A/B testing in low-risk scenarios (positive support, information needed) while continuing development for high-stakes crisis detection. With proposed improvements, this system could reduce response times by 40% and provide initial support to millions annually.

11. Links

GitHub - https://github.com/nikhilgodalla/Fine-Tuning-DistilBERT-for-Mental-Health-Support-Classification
Complete code available at:
https://colab.research.google.com/drive/1RnaiJunu4B7U4YecVRtUyi2aB1xhFcyA#scrollTo=TqaeoQ9fZ7DI

Video Presentation: https://www.youtube.com/watch?v=0ljXq7AbbNA