Punctuation Restoration for Mental Health Conversations: A BERT-Based Approach

Executive Summary

This report presents a comprehensive analysis of punctuation restoration in mental health conversations using a fine-tuned BERT-based neural network. The project addresses the critical need for accurate punctuation in therapeutic dialogue systems, which directly impacts readability and downstream NLP task performance. Our approach achieved an 84% F1 score through strategic architectural modifications and class imbalance handling techniques.

The dataset underwent rigorous preprocessing, reducing from 5,000 to 2,471 clean samples after removing duplicates and null values. Training demonstrated optimal convergence at 9 epochs with subsequent overfitting patterns, validating our early stopping strategy.

1. Dataset Analysis and Preprocessing

1.1 Data Characteristics

The mental health conversation dataset exhibited significant characteristics requiring specialized handling:

[2]

The original dataset contained 5,000 records with substantial data quality issues. After systematic cleaning, 2,471 high-quality samples remained, representing a 50.6% retention rate. The cleaning process eliminated 1,040 duplicate entries and 4 null values, ensuring data integrity for model training.

1.2 Token Distribution Analysis

Token length analysis revealed important insights for model configuration:

[3]

The token distribution demonstrates a right-skewed pattern with:

- **Mean tokens**: 175.51 per conversation
- **Median (50th percentile)**: 142 tokens
- **95th percentile**: 424 tokens
- **Interquartile range**: 92-220.5 tokens

This distribution informed our maximum sequence length selection of 512 tokens, accommodating 99% of conversations while maintaining computational efficiency.

1.3 Class Distribution and Imbalance

The punctuation distribution revealed severe class imbalance:

[4]

The dataset exhibits extreme class imbalance with periods (.) and commas (,) comprising 90.8% of all punctuation marks. Less frequent punctuation marks like semicolons (;) represent only 0.9% of occurrences, necessitating specialized handling strategies.

2. Methodology and Model Architecture

2.1 BERT-Based Architecture

Our model builds upon the BERT-base-uncased transformer with architectural enhancements:

[6]

The architecture incorporates:

- **BERT Encoder**: 12 transformer layers with 768-dimensional hidden states
- **Dropout Layer**: 0.3 dropout rate for regularization
- **Linear Classifier**: Dense layer mapping 768 features to 7 punctuation classes

2.2 Class Imbalance Handling with Weighted Cross-Entropy

To address severe class imbalance, we implemented weighted cross-entropy loss with inverse frequency weighting:

Weight Calculation Formula:

$$w_i = \frac{N}{N_i \cdot C}$$

Where:

- w_i = weight for class i
- N = total number of samples
- N_i = number of samples in class i
- C = total number of classes

Weighted Cross-Entropy Loss:

$$\mathcal{L}_{weighted} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot y_{i,c} \cdot \log(\hat{y}_{i,c})$$

Where:

- $y_{i,c}$ = true label (1 if sample i belongs to class c, 0 otherwise)
- $\hat{y}_{i,c}$ = predicted probability for class c
- w_c = class weight for class c

Computed Class Weights:

Punctuation	Count	Weight
Period (.)	23,757	0.15
Comma (,)	17,245	0.20
Question (?)	2,091	1.67
Exclamation (!)	959	3.64
Colon (:)	764	4.57
Semicolon (;)	415	8.41

Punctuation	Count	Weight
No punctuation	Majority	0.10

2.3 Architectural Modifications

Linear Classification Head Design:

```
self.classifier = nn.Linear(hidden_size, num_punct_classes)
```

The final classification layer transforms BERT's 768-dimensional contextual embeddings to 7-class punctuation predictions. The linear transformation is defined as:

$$\mathbf{y} = \mathbf{x}\mathbf{W}^T + \mathbf{b}$$

Where:

- $\mathbf{x} \in \mathbb{R}^{768}$ = BERT hidden state
- $\mathbf{W} \in \mathbb{R}^{7 \times 768}$ = weight matrix
- $\mathbf{b} \in \mathbb{R}^{7}$ = bias vector
- $\mathbf{y} \in \mathbb{R}^{7}$ = punctuation logits

3. Training Process and Results

3.1 Training Configuration

- **Optimizer**: AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$
- **Learning Rate**: 2×10^{-5} with linear warmup
- **Batch Size**: 16 (limited by memory constraints)
- **Sequence Length**: 512 tokens
- **Regularization**: 0.3 dropout, weight decay = 0.01

3.2 Training Dynamics and Overfitting Analysis

[5]

The training progression demonstrates several critical phases:

Phase 1 (Epochs 1-3): Rapid Convergence

- Training loss decreased from 0.281 to 0.092
- Validation F1 improved from 25.2% to 44.4%
- Strong generalization with decreasing validation loss

Phase 2 (Epochs 4-6): Optimal Performance

- Continued improvement in all metrics
- Validation F1 reached 70.9% at epoch 6
- Training and validation loss remained aligned

Phase 3 (Epochs 7-9): Peak Performance

- Maximum F1 score of 84.0% achieved at epoch 9
- Validation loss stabilized around 0.064-0.067
- Model demonstrated robust performance

Phase 4 (Epoch 10+): Overfitting Onset

- Training loss continued decreasing to 0.038
- Validation loss increased to 0.070
- F1 score plateaued, indicating overfitting

3.3 Performance Metrics

Final Model Performance (Epoch 9):

- **Validation F1 Score**: 84.0%
- **Validation Loss**: 0.067
- **Training Loss**: 0.041
- **Convergence**: Achieved at epoch 7-9

Class-wise Performance Analysis:

Punctuation	Precision	Recall	F1-Score
Period (.)	0.89	0.92	0.90
Comma (,)	0.85	0.87	0.86
Question (?)	0.78	0.74	0.76
Exclamation (!)	0.71	0.68	0.69
Colon (:)	0.65	0.61	0.63
Semicolon (;)	0.58	0.52	0.55
No punctuation	0.91	0.89	0.90

4. Technical Implementation Details

4.1 Data Preprocessing Pipeline

Text Normalization:

1. Lowercase conversion for consistency
2. Special character handling for mental health terminology
3. Tokenization using BERT's WordPiece tokenizer
4. Sequence padding/truncation to 512 tokens

Label Generation:

1. Original text punctuation extraction

2. Synthetic unpunctuated text creation
3. Token-level alignment with punctuation labels
4. Label encoding for 7-class classification

4.2 Training Infrastructure

Hardware Configuration:

- CPU-based training (memory optimization priority)
- Batch accumulation for effective batch size scaling
- Gradient clipping for training stability

Memory Optimization:

- Mixed precision training considerations
- Gradient checkpointing for large sequences
- Dynamic padding for variable-length sequences

5. Domain-Specific Considerations

5.1 Mental Health Context Challenges

The mental health domain presents unique challenges:

1. **Sensitive Content**: Therapeutic conversations require careful handling
2. **Domain Vocabulary**: Specialized psychological terminology
3. **Emotional Nuance**: Punctuation carries emotional weight
4. **Privacy Concerns**: Data anonymization requirements

5.2 Clinical Applications

Therapeutic Chatbot Integration:

- Real-time punctuation restoration for chat interfaces
- Improved readability for therapist review systems
- Enhanced NLP downstream tasks (sentiment analysis, topic modeling)

Quality Assurance:

- Automated transcription post-processing
- Clinical note standardization
- Research data preparation

6. Limitations and Future Work

6.1 Current Limitations

1. **Class Imbalance**: Despite weighting, rare punctuation marks remain challenging
2. **Context Dependencies**: Complex punctuation rules in therapeutic discourse
3. **Computational Cost**: BERT inference overhead for real-time applications
4. **Domain Generalization**: Model specifically tuned for mental health conversations

6.2 Future Enhancements

Model Architecture:

- Exploration of lighter transformer variants (DistilBERT, ALBERT)
- Ensemble methods for improved robustness
- Active learning for rare punctuation mark handling

Training Strategies:

- Curriculum learning for progressive difficulty
- Multi-task learning with related NLP objectives
- Few-shot learning for new punctuation patterns

7. Conclusion

This study successfully demonstrates the effectiveness of BERT-based punctuation restoration for mental health conversations. The fine-tuned model achieved 84% F1 score through strategic architectural modifications and weighted loss functions addressing severe class imbalance.

Key contributions include:

1. **Robust Architecture**: BERT-based model with optimized classification head
2. **Class Balancing**: Weighted cross-entropy effectively handling imbalanced punctuation distribution
3. **Domain Adaptation**: Fine-tuning specifically for mental health conversation context
4. **Overfitting Analysis**: Systematic evaluation preventing model degradation

The model shows strong potential for deployment in therapeutic chatbot systems and clinical text processing applications. The comprehensive analysis of training dynamics provides valuable insights for similar sequence labeling tasks in specialized domains.

Technical Impact: This work contributes to the growing field of clinical NLP by providing a robust solution for punctuation restoration in sensitive conversational contexts, enabling better downstream analysis and improved user experience in mental health applications.