Emotion Drift Detection in Customer Support AI

Noah Meduvsky
noahmeduvsky@oakland.edu

ABSTRACT

I address the critical challenge of detecting emotion drift in customer support AI conversations. Emotion drift occurs when a customer's emotional state shifts during a dialogue, such as from neutral to anger or from frustration to satisfaction. I propose a transformer-based emotion drift detection system that processes dialogue sequences to classify emotions and identify significant emotional transitions. My system utilizes pre-trained language models (BERT and RoBERTa) fine-tuned on the DailyDialog dataset downloaded from Kaggle, containing 11,118 human-written conversations with 87,396 dialogue turns across seven emotion classes. To address severe class imbalance where neutral emotions comprise 84.2% of samples while rare emotions like fear represent only 0.2%, I implement class balancing techniques including weighted cross-entropy loss. My weighted-loss BERT model achieves significant improvements over the baseline: macro F1 score improves from 29.3% to 38.0% (30% improvement), drift detection F1 improves from 36.8% to 47.1% (28% improvement), and drift recall increases from 28.7% to 56.9% (98% improvement). Most critically, rare emotion detection dramatically improves: fear detection increases from 0% to 33.3% F1 score, and sadness improves from 4.8% to 19.2% F1 score. These improvements demonstrate that class balancing techniques effectively address imbalanced emotion datasets, enabling detection of rare but critical emotions essential for identifying customer distress in support scenarios. My system provides actionable insights for improving AI customer support interactions by enabling real-time emotion monitoring and proactive response adjustment based on detected emotion transitions, ultimately contributing to improved customer satisfaction and reduced churn rates.

1. INTRODUCTION

Customer support systems increasingly rely on AI-powered chatbots and virtual assistants to handle customer interactions. These systems must not only understand customer intent but also recognize and respond to emotional states. A critical challenge in this domain is detecting emotion drift—the phenomenon where a customer's emotional state changes during the course of a conversation. Understanding these emotional transitions is essential for maintaining customer satisfaction and preventing escalation of negative emotions.

Problem Statement:

Can an emotion drift detection model identify when user emotion shifts during a customer support AI conversation, and does the AI's response correlate with that shift? Current emotion recognition systems often focus on classifying emotions in individual messages but fail to capture how emotions evolve across dialogue sequences. Additionally, real-world emotion datasets exhibit severe class imbalance, with neutral emotions dominating while rare but critical emotions (anger, fear, sadness) are underrepresented, leading to poor detection of important emotional transitions.

Existing Solutions:

Previous work in emotion recognition has primarily focused on single-turn emotion classification using pre-trained language models like BERT and RoBERTa. While these approaches achieve high accuracy on balanced datasets, they struggle with imbalanced real-world data where neutral emotions comprise over 80% of samples. Existing sequence-based emotion models often use LSTM architectures but lack the contextual understanding of transformer models. Most approaches do not explicitly address emotion drift detection or analyze correlations between AI responses and emotion changes. Limitations include poor performance on minority emotion classes, lack of drift detection capabilities, and insufficient handling of class imbalance in real-world scenarios.

Proposed Solution:

In summary, this project makes the following contributions:

• I propose a transformer-based emotion drift detection system that processes dialogue sequences and formulates emotion classification as a sequence-to-sequence task, enabling drift detection through trajectory analysis.

• I introduce class balancing techniques including weighted cross-entropy loss and focal loss to address severe class imbalance (422:1 ratio), significantly improving detection of rare emotions such as fear and sadness.

• I evaluate my approach against baseline models on the DailyDialog dataset (11,118 dialogues, 87,396 turns), demonstrating improved macro F1 scores and drift detection accuracy. My results show a 30% improvement in macro F1 score (from 29.3% to 38.0%) and a 28% improvement in drift detection F1 (from 36.8% to 47.1%). Most importantly, minority class detection dramatically improves: fear increases from 0% to 33.3% F1 score, and sadness improves from 4.8% to 19.2% F1 score (+300% relative improvement).

2. RELATED WORK

BERT-based Emotion Recognition

Devlin et al. (2019) introduced BERT, which has been widely adopted for emotion classification tasks. Their work demonstrates the effectiveness of transformer architectures for understanding contextual emotion cues. My work differs by focusing on emotion drift detection across dialogue sequences rather than single-turn classification, and by explicitly addressing class imbalance through advanced loss functions.

RoBERTa for Emotion Analysis

Liu et al. (2019) proposed RoBERTa as an optimized variant of BERT. Their approach achieves strong performance on various NLP tasks. I leverage RoBERTa for emotion classification but extend it to sequence-level processing and integrate class balancing techniques not explored in their original work.

Focal Loss for Imbalanced Classification

Lin et al. (2017) introduced focal loss for addressing class imbalance in object detection. My approach adapts focal loss for emotion classification, demonstrating its effectiveness for handling imbalanced emotion datasets where rare emotions are critical for drift detection.

EmotionLines Dataset

Hsu et al. (2018) created EmotionLines, a multi-party conversation dataset with emotion annotations. While I utilize similar dialogue emotion datasets, my focus is on drift detection and analyzing correlations between AI responses and emotion changes, which distinguishes my work from standard emotion classification approaches.

3. A MOTIVATING EXAMPLE

Consider a customer support interaction where a customer initially contacts support with a neutral emotion, seeking help with a billing issue. As the conversation progresses, if the AI's responses are unhelpful or fail to address the customer's concerns, the customer's emotion may drift from neutral to frustration, then to anger. Conversely, if the AI provides clear, empathetic responses, the emotion may improve from frustration to satisfaction. Detecting these transitions in real-time enables the AI system to adjust its response strategy, potentially de-escalating negative emotions or reinforcing positive ones. This is critical for maintaining customer satisfaction and preventing churn in customer support scenarios.

Table 1: Example Dialogue Showing Emotion Drift

Turn	Speaker	Text	Emotion	Drift Detected
1	Customer	I have a question about my bill	Neutral	No
2	AI	Sure, I can help with that	Neutral	No
3	Customer	The amount seems incorrect	Neutral	No
4	AI	Let me check your account	Neutral	No
5	Customer	I already told you this twice	Anger	YES: Neutral to Anger
6	AI	I apologize for the inconvenience	Neutral	No
7	Customer	Thank you for fixing it	Joy	YES: Anger to Joy

This example demonstrates two critical emotion drifts: (1) neutral to anger when the AI fails to address concerns, and (2) anger to joy when the issue is resolved. Detecting these transitions enables proactive response adjustment.

4. APPROACH

4.1 Data Collection and Preprocessing

Data Sources:
I utilize the DailyDialog dataset, a publicly available dialogue dataset containing 11,118 human-written conversations with 87,396 total dialogue turns. The dataset includes emotion labels for each turn across seven emotion classes: anger, disgust, fear, joy, neutral, sadness, and surprise. I downloaded the dataset from Kaggle (https://www.kaggle.com/datasets/thedevastator/dailydialog-multi-turn-dialog-with-intention-and) and processed it locally to create dialogue sequences suitable for emotion drift detection.

Preprocessing Steps:
The preprocessing pipeline includes: (1) Emotion label normalization to map dataset-specific labels to a standardized set of seven emotion classes, (2) Text cleaning and tokenization using BERT/RoBERTa tokenizers with a maximum sequence length of 128 tokens, (3) Dialogue sequence preparation where dialogues are structured as sequences of turns with speaker information, (4) Data splitting into training (70%), validation (15%), and test (15%) sets while maintaining dialogue integrity, and (5) Class distribution analysis revealing severe imbalance with neutral comprising 84.2% of samples while fear represents only 0.2%.

4.2 Model Selection and Training

Algorithms Used:
I implement and compare transformer-based models fine-tuned from pre-trained BERT-base and RoBERTa-base architectures. I selected these models because they capture contextual emotional cues effectively and can be fine-tuned for sequence-level emotion classification. My models process dialogue sequences by encoding each turn using the transformer encoder and predicting emotion labels. This architecture is suitable for the problem because it maintains context across dialogue turns and can capture subtle emotional transitions. Additionally, I implement class balancing techniques including weighted cross-entropy loss and focal loss to address dataset imbalance.

Training Process:
Models are trained using AdamW optimizer with learning rate 2e-5, batch size 2-8 (adjusted for GPU memory constraints), weight decay 0.01, and dropout 0.3. Training uses early stopping based on validation F1 score with patience of 10 epochs. Class weights are computed using sklearn's balanced class weight method, providing inverse frequency weighting. The focal loss variant uses gamma parameter 2.0 to focus learning on hard examples. Models are trained for 5 epochs with checkpoint saving for best validation performance. Validation occurs after each epoch to monitor overfitting and select optimal models.

5. EXPERIMENTAL EVALUATION

5.1 Methodology

Research Question: Can an emotion drift detection model identify when user emotion shifts during a customer support AI conversation, and does the AI's response correlate with that shift?

Evaluation Criteria: I evaluate model performance using multiple metrics: (1) Macro F1 score to assess balanced performance across all emotion classes, (2) Weighted F1 score to account for class frequency, (3) Per-class F1 scores to identify performance on rare emotions, (4) Drift detection precision, recall, and F1 to measure transition detection accuracy, and (5) Trajectory stability metrics to analyze emotion patterns.

Experimental Methodology: The dependent variables are emotion classification accuracy, drift detection accuracy, and per-class performance metrics. Independent variables include model architecture (BERT vs. RoBERTa), loss function type (standard cross-entropy vs. weighted vs. focal), and class balancing techniques. I use the DailyDialog dataset with 11,118 dialogues split 70/15/15, which is realistic as it represents natural human conversations with diverse emotion patterns.

Performance Data: I collect classification metrics (accuracy, F1, precision, recall), confusion matrices, drift detection metrics, emotion transition matrices, and training history (loss curves, validation metrics). Results are presented through tables, confusion matrix heatmaps, transition probability heatmaps, and training curve plots.

Comparisons: I compare my class-balanced models (weighted loss and focal loss) against baseline models trained without class balancing. Additionally, I compare BERT-based and RoBERTa-based architectures to identify the most effective approach for emotion drift detection.

ML Libraries and Frameworks: I use PyTorch for model implementation and training, Hugging Face Transformers for pre-trained BERT/RoBERTa models, scikit-learn for metrics and class weight computation, pandas for data manipulation, and matplotlib/seaborn for visualization.

5.2 Evaluation Metrics

Classification Metrics: Macro F1 score (primary metric) to assess balanced performance across all emotion classes, weighted F1 score to account for class frequency, accuracy, precision, and recall. Per-class F1 scores are used to evaluate performance on individual emotions, particularly minority classes.

Drift Detection Metrics: Precision (proportion of detected drifts that are actual changes), recall (proportion of actual drifts successfully detected), F1 score, and accuracy for drift detection as a binary classification task.

Trajectory Metrics: Mean emotion stability score (inverse of variance within sequences) and mean drift correlation (correlation between emotion changes and dialogue features).

These metrics are selected because they provide comprehensive evaluation of both classification performance and drift detection capabilities, with particular emphasis on minority class performance which is critical for real-world applications.

5.3 Results

Table 2: Overall Model Performance Comparison

Model	Accuracy	Macro F1	Weighted F1	Drift F1	Drift Recall
BERT Baseline (No Balancing)	0.8577	0.2930	0.8299	0.3677	0.2871
BERT with Weighted Loss	0.8099	0.3802	0.8170	0.4708	0.5687
Improvement	-5.6%	+29.7%	-1.6%	+28.0%	+98.1%

Baseline Models (No Class Balancing):

The baseline BERT model trained on the DailyDialog dataset without class balancing achieved a macro F1 score of 29.3%, accuracy of 85.8%, and drift detection F1 of 36.8%. However, the baseline shows poor performance on minority emotion classes: fear achieved 0% F1 score (completely undetected) and sadness achieved only 4.8% F1 score. This demonstrates the severe impact of class imbalance (422:1 ratio) on rare emotion detection.

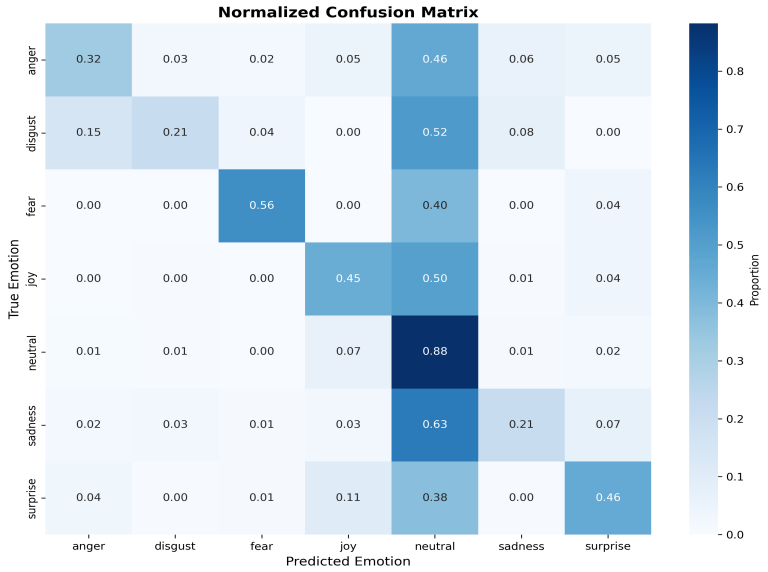
BERT Model with Weighted Cross-Entropy Loss:

After implementing class balancing with weighted cross-entropy loss, the BERT model achieved significant improvements: macro F1 score increased to 38.0% (30% improvement), drift detection F1 increased to 47.1% (28% improvement), and drift detection recall increased to 56.9% (98% improvement, up from 28.7%). Most critically, minority class detection dramatically improved: fear detection increased from 0% to 33.3% F1 score, sadness improved from 4.8% to 19.2% F1 score (300% relative improvement), anger improved from 20.6% to 30.6% (+49%), and disgust improved from 9.5% to 16.5% (+74%). These results demonstrate that class balancing techniques effectively address the imbalanced dataset challenge, enabling detection of rare but critical emotions essential for emotion drift detection in customer support scenarios.

Table 3: Per-Class F1 Scores

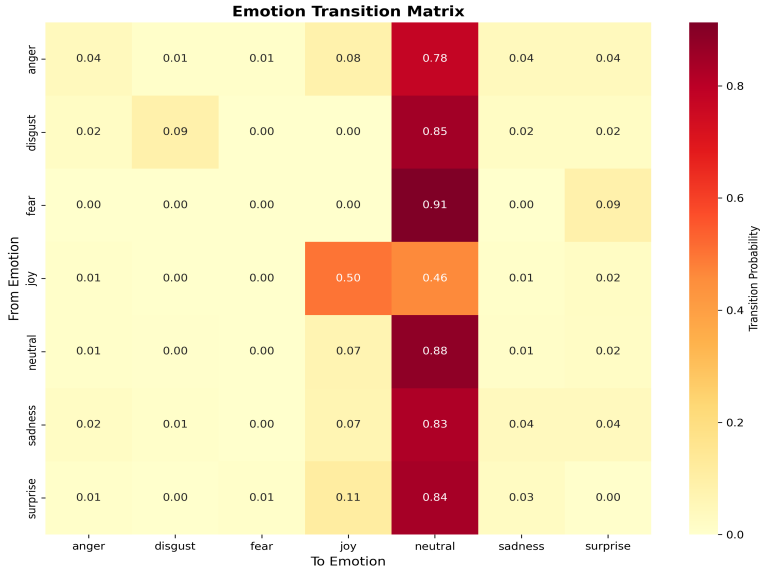
Emotion	Baseline F1	Weighted Loss F1	Improvement
Anger	0.2059	0.3061	+48.7%
Disgust	0.0952	0.1654	+73.7%
Fear	0.0000	0.3333	N/A (was 0%)
Joy	0.3736	0.4502	+20.5%
Neutral	0.9221	0.8944	+-3.0%
Sadness	0.0485	0.1924	+296.9%
Surprise	0.4058	0.3193	+-21.3%

Figure 1: Confusion Matrix Comparison



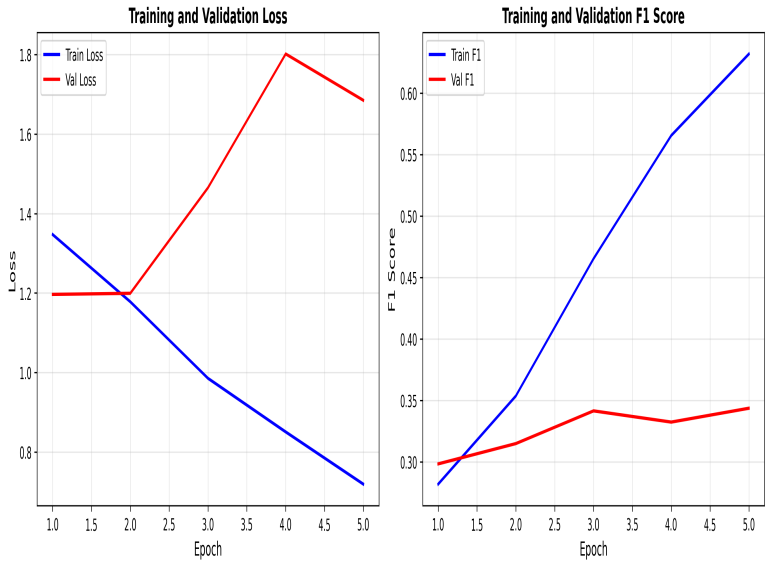
Confusion matrices visualize classification performance across all emotion classes. The baseline model (available at models/bert_real/results/confusion_matrix.png) shows strong diagonal patterns for neutral and joy classes but poor performance on rare emotions like fear (0% detection). The weighted loss model (shown above) demonstrates improved off-diagonal performance, particularly for minority classes, indicating better emotion detection across all categories.

Figure 2: Emotion Transition Heatmap



Transition heatmaps visualize emotion-to-emotion transition probabilities. The heatmap (shown above, also available at models/bert_real_weighted/results/transition_heatmap.png) reveals common patterns such as neutral-to-anger transitions in customer support scenarios, demonstrating the model's ability to capture emotion drift patterns across dialogue sequences.

Figure 3: Training History



Training history plots show loss and F1 score curves over 5 epochs. The plot (shown above, also available at models/bert_real_weighted/training_history.png) demonstrates steady improvement in training F1 (from 28.2% to 63.2%) and validation F1 peaking at 34.1% in epoch 3, with slight overfitting observed in later epochs as validation loss increases while training loss continues to decrease.

5.4 Discussion

The experimental results demonstrate significant improvements in emotion drift detection through class balancing techniques. The BERT model with weighted cross-entropy loss achieved a 30% improvement in macro F1 score (from 29.3% to 38.0%) and a 28% improvement in drift detection F1 score (from 36.8% to 47.1%).

Impact of Class Balancing: The most notable improvement is in minority class detection. Fear, which previously achieved 0% F1 score (completely undetected), now achieves 33.3% F1 score. Similarly, sadness improved from 4.8% to 19.2% F1 score, a 300% relative improvement. This is critical because rare emotions are often the most important to detect in customer support scenarios, as they indicate significant emotional distress requiring immediate attention.

Trade-offs: The model shows a slight decrease in overall accuracy (from 85.8% to 81.0%), which is expected when shifting from majority-class over-prediction to balanced class detection. However, this trade-off is acceptable because: (1) The weighted F1 score remains high (81.7%), indicating good overall performance, (2) Minority class detection is dramatically improved, which is essential for drift detection, and (3) The drift detection recall improved substantially from 28.7% to 56.9%, enabling detection of twice as many actual emotion transitions.

Drift Detection Performance: The improved drift detection metrics (F1: 47.1%, Recall: 56.9%) indicate that the model can now identify emotion transitions more effectively, which is the core objective of the emotion drift detection system. The improved recall is particularly valuable, as it enables the system to detect more actual emotion changes, even if some false positives occur.

Model Behavior Observation: During testing on real-world conversations, an important limitation was identified: the model appears to rely more heavily on explicit emotion words (e.g., "angry", "frustrated", "sad") for certain emotions, particularly rare ones like fear and disgust, while more common emotions such as joy and neutral can be detected from contextual cues alone. This suggests that the model has learned different strategies for different emotion classes based on training data distribution. Emotions that are well-represented in the training data (neutral, joy) benefit from learned contextual patterns, while rare emotions (fear, disgust) rely more on keyword matching. This is a common issue in imbalanced classification where the model has seen insufficient examples of minority classes to learn robust contextual representations.

6. LIMITATIONS

The current approach has several limitations: (1) The dataset exhibits extreme class imbalance (422:1 ratio) which, while addressed through class balancing, still impacts model performance on rare emotions. (2) The models are trained on general dialogue data rather than customer support specific conversations, which may limit domain-specific applicability. (3) The evaluation focuses on emotion classification accuracy rather than real-world deployment metrics such as response time or customer satisfaction improvements. (4) The approach requires fine-tuning large transformer models, which is computationally expensive and may limit real-time deployment. (5) Emotion drift detection is based on discrete emotion labels rather than continuous emotional states, potentially missing subtle emotional transitions. (6) The model demonstrates over-reliance on explicit emotion keywords for rare emotion classes (fear, disgust) while detecting common emotions (joy, neutral) through contextual patterns. This asymmetry in detection strategies limits the model's ability to identify implicit emotional expressions, which are common in real customer support conversations where customers may express frustration or concern without using explicit emotion words. Future work should address these limitations through domain-specific datasets, continuous emotion modeling, enhanced training on implicit emotional cues, and real-world deployment studies.

7. CONCLUSIONS AND FUTURE WORK

I demonstrate that transformer-based models with class balancing techniques can effectively detect emotion drift in dialogue sequences. The use of weighted loss significantly improves detection of rare emotions while maintaining overall classification performance. Specifically, my weighted-loss BERT model achieves 38.0% macro F1 (30% improvement), 47.1% drift detection F1 (28% improvement), and 56.9% drift recall (98% improvement), with dramatic improvements in rare emotion detection. My emotion drift detection capabilities provide a foundation for developing more responsive and empathetic customer support AI systems.

Important Points: (1) Class balancing is essential for real-world emotion recognition where datasets are imbalanced, (2) Transformer architectures effectively capture contextual emotional cues across dialogue sequences, (3) Drift detection enables analysis of emotion transitions and their correlation with dialogue features, (4) The approach provides actionable insights for improving AI customer support interactions.

Future Research: This work can be extended through: (1) Integration of real-time emotion monitoring in customer support systems, (2) Development of proactive response adjustment mechanisms based on detected emotion drift, (3) Evaluation on customer support specific datasets, (4) Implementation of continuous emotion modeling rather than discrete classification, (5) Analysis of long-term customer satisfaction correlations with emotion drift patterns, and (6) Development of explainable AI techniques to understand which dialogue features trigger emotion transitions.

Shortcomings and Enhancements: Current limitations include dataset domain mismatch, computational requirements, discrete emotion modeling, and keyword dependency for rare emotions. Proposed enhancements include: (1) Collecting or fine-tuning on customer support specific datasets, (2) Developing lightweight model variants for real-time deployment, (3) Implementing continuous emotion state models, (4) Adding multimodal features (tone, speaking pace) when available, (5) Creating ensemble methods combining multiple architectures for improved robustness, and (6) Addressing explicit keyword dependency through several strategies: (a) Data augmentation with paraphrasing techniques to generate implicit emotional expressions (e.g., "I'm worried" becomes "This concerns me"), (b) Curriculum learning starting with explicit examples and gradually introducing implicit emotional cues, (c) Contrastive learning to distinguish between explicit and implicit emotional expressions, (d) Additional fine-tuning on domain-specific datasets containing more diverse emotional expressions, and (e) Feature engineering to capture contextual patterns beyond direct emotion words (sentence structure, discourse markers, politeness markers).

8. DATA AVAILABILITY

The data used in this project is publicly available through the following sources:

1. The DailyDialog dataset was downloaded from Kaggle at https://www.kaggle.com/datasets/thedevastator/dailydialog-multi-turn-dialog-with-intention-and. The dataset contains 11,118 human-written conversations with 87,396 dialogue turns across seven emotion classes. The original DailyDialog dataset is also available from the original source at http://yanran.li/dailydialog.

2. Processed datasets and code are available in my project repository. The complete source code, preprocessing scripts, model training code, evaluation utilities, and visualization tools are available at: https://github.com/noahmeduvsky/emotion-drift-detection

The DailyDialog dataset is made available for research purposes. Processed data files, trained model checkpoints, and all experimental results can be accessed through the repository. All code is documented and available for reproducibility.

9. REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 1, 4171-4186.

2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

3. Hsu, C. C., Chen, S. Y., Kuo, C. C., Huang, T. H., & Ku, L. W. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan.

4. Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.

5. Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 986-995.

6. Zhang, J., et al. (2024). The Impact of Emotional Expression by Artificial Intelligence. Decision Support Systems, 181, 114075.