Experimental Results Summary

Emotion Drift Detection in Customer Support AI

Noah Meduvsky

1. EXECUTIVE SUMMARY

This document compiles all experimental results from the emotion drift detection project. The experiments compare multiple model architectures (BERT and RoBERTa) trained on both synthetic and real-world datasets, with and without class balancing techniques. Key findings include significant improvements in minority emotion detection when using weighted loss functions, demonstrating the effectiveness of class balancing for imbalanced emotion datasets.

2. OVERALL MODEL COMPARISON

Model	Dataset	Accuracy	Macro F1	Weighted F1	Drift F1
BERT-base	Synthetic	1.0000	1.0000	1.0000	1.0000
RoBERTa-base	Synthetic	1.0000	1.0000	1.0000	1.0000
BERT-base	Real (No Balancing)	0.8577	0.2930	0.8299	0.3677
RoBERTa-base	Real (No Balancing)	0.8577	0.2706	0.8286	0.3486
BERT-base	Real (Weighted Loss)	0.8099	0.3802	0.8170	0.4708

Key Observations:

• Synthetic dataset models achieve perfect performance (100% accuracy), validating pipeline correctness • Real dataset models show realistic performance with significant class imbalance challenges • Class balancing (weighted loss) improves macro F1 from 29.3% to 38.0% (+30% improvement) • Drift detection F1 improves from 36.8% to 47.1% (+28% improvement) with class balancing

3. DETAILED CLASSIFICATION METRICS

3.1 BERT Model: Baseline vs Class Balanced

Metric	Baseline (No Balancing)	With Weighted Loss	Improvement
Accuracy	0.8577	0.8099	-5.6%
Macro F1	0.2930	0.3802	29.7%
Weighted F1	0.8299	0.8170	-1.6%
Precision	0.4231	0.3494	-17.4%
Recall	0.2598	0.4418	70.1%

4. 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%

Key Findings:

• Fear: Improved from 0.0000 to 0.3333 F1 score (now detectable) • Sadness: Improved from 0.0485 to 0.1924 F1 score (+297% improvement) • Anger: Improved from 0.2059 to 0.3061 F1 score (+49% improvement) • Disgust: Improved from 0.0952 to 0.1654 F1 score (+74% improvement) • Neutral: Slight decrease from 0.9221 to 0.8944 (expected trade-off)

5. DRIFT DETECTION METRICS

Metric	Baseline	Weighted Loss	Improvement
Precision	0.5113	0.4017	-21.4%
Recall	0.2871	0.5687	98.1%
F1 Score	0.3677	0.4708	28.0%
Accuracy	0.8081	0.7515	-7.0%

Analysis:

Drift detection recall improved dramatically from 28.7% to 56.9% (+98% improvement), meaning the model now detects more than twice as many actual emotion transitions. The F1 score improvement of 28% indicates better overall drift detection performance, which is critical for identifying when customer emotions shift during conversations.

6. TRAINING HISTORY - BERT WITH WEIGHTED LOSS

Epoch	Train Loss	Val Loss	Train F1	Val F1	Val Accuracy
1	1.3471	1.1965	0.2818	0.2983	0.8357
2	1.1779	1.1992	0.3536	0.3148	0.8164
3	0.9853	1.4649	0.4653	0.3414	0.8189
4	0.8506	1.8014	0.5656	0.3323	0.8408
5	0.7191	1.6848	0.6320	0.3436	0.8158

Training Observations:

• Training F1 improved steadily from 28.2% to 63.2% over 5 epochs • Validation F1 peaked at epoch 3 (34.1%) and stabilized around 34.4% • Training loss decreased from 1.35 to 0.72, showing good convergence • Validation loss increased after epoch 2, indicating some overfitting • Best model saved at epoch 3 with validation F1 of 34.1%

7. VISUALIZATIONS

7.1 Confusion Matrices

Confusion matrices are available for all models in their respective result directories:

- **BERT Baseline (Synthetic)**: models/bert_baseline/results/confusion_matrix.png
- **RoBERTa Baseline (Synthetic)**: models/roberta_baseline/results/confusion_matrix.png
- **BERT Real (No Balancing)**: models/bert_real/results/confusion_matrix.png
- **RoBERTa Real (No Balancing)**: models/roberta_real/results/confusion_matrix.png
- **BERT Real (Weighted Loss)**: models/bert_real_weighted/results/confusion_matrix.png

The confusion matrices show classification patterns, with the weighted loss model showing better detection of minority classes (fear, sadness) compared to the baseline.

7.2 Emotion Transition Heatmaps

Transition heatmaps visualize emotion-to-emotion transition probabilities:

- **BERT Baseline (Synthetic)**: models/bert_baseline/results/transition_heatmap.png
- **RoBERTa Baseline (Synthetic)**: models/roberta_baseline/results/transition_heatmap.png
- **BERT Real (No Balancing)**: models/bert_real/results/transition_heatmap.png
- **RoBERTa Real (No Balancing)**: models/roberta_real/results/transition_heatmap.png
- **BERT Real (Weighted Loss)**: models/bert_real_weighted/results/transition_heatmap.png

These heatmaps reveal common emotion transition patterns, such as neutral-to-anger transitions in customer support scenarios.

7.3 Training History Plots

Training history plots show loss and metric curves over epochs:

- **BERT Baseline**: models/bert_baseline/training_history.png
- **RoBERTa Baseline**: models/roberta_baseline/training_history.png
- **BERT Real (No Balancing)**: models/bert_real/training_history.png
- **RoBERTa Real (No Balancing)**: models/roberta_real/training_history.png
- **BERT Real (Weighted Loss)**: models/bert_real_weighted/training_history.png

These plots show training convergence, overfitting patterns, and validation performance trends.

8. DATASET STATISTICS

Dataset	Dialogues	Total Turns	Emotion Classes	Class Imbalance Ratio
Synthetic	200	1,128	7	Balanced
DailyDialog (Real)	11,118	87,396	7	422:1 (Neutral:Fear)

Real Dataset Class Distribution:

The DailyDialog dataset exhibits severe class imbalance:

• Neutral: 84.2% of samples (51,583 samples)
• Joy: 11.4% of samples (7,034 samples)
• Anger: 1.1% of samples (661 samples)
• Sadness: 1.2% of samples (715 samples)
• Surprise: 1.8% of samples (1,130 samples)
• Disgust: 0.4% of samples (222 samples)
• Fear: 0.2% of samples (122 samples)

This extreme imbalance (422:1 ratio between most and least common classes) necessitated the use of class balancing techniques to improve minority class detection.

9. CONCLUSIONS

Key Achievements:

1. Class balancing techniques successfully address severe class imbalance in real-world emotion datasets

2. Weighted cross-entropy loss improves macro F1 by 30% (29.3% $\rightarrow$ 38.0%) and drift detection F1 by 28% (36.8% $\rightarrow$ 47.1%)

3. Minority emotion detection dramatically improved: fear from 0% to 33.3% F1, sadness from 4.8% to 19.2% F1

4. Drift detection recall improved by 98% (28.7% $\rightarrow$ 56.9%), enabling detection of twice as many emotion transitions

5. The trade-off between overall accuracy and minority class performance is acceptable, as rare emotions are more critical for customer support applications than maintaining high accuracy through majority class over-prediction

Recommendations:

• Continue exploring focal loss as an alternative to weighted cross-entropy
• Experiment with different class weight computation methods
• Consider data augmentation techniques for minority classes
• Evaluate on customer support specific datasets for domain-specific insights