Final Project 266

Arun Agarwal

Graph Result Summaries

Confusion Matrices:

After examining all four models through their respective confusion matrices and evaluation metrics, clear patterns emerge in the way contextual information influences emotion classification performance.

Baseline Model: Strong on Frequency, Weak on Nuance

The baseline model, which lacks any contextual cues, performs reliably on high-frequency and semantically distinct emotions such as admiration, gratitude, curiosity, and amusement. For example, admiration is correctly predicted 387 times, and amusement sees 221 correct predictions with minimal confusion. However, this model consistently struggles with low-frequency or introspective emotions like grief, relief, nervousness, and remorse, all of which have zero or near-zero true positives. Cognitive states like confusion, realization, and approval show significant overlap and frequent misclassification, suggesting the model lacks deeper semantic grounding.

Subreddit Model: Community Influence is Limited

Introducing subreddit context yields marginal improvements for certain mid-frequency emotions like remorse, curiosity, and approval, but does not fundamentally shift the model’s ability to disambiguate nuanced emotional content. While emotions like admiration and gratitude still perform well, rare or internally oriented emotions like nervousness and relief remain undetected. This suggests that subreddit-level context may offer some linguistic framing (e.g., topics or sentiment tendencies within a subreddit), but is not sufficient to resolve personalized or introspective emotions.

Author Model: Gains in Personal and Subtle Emotions

Incorporating author identity has a more meaningful impact. This model improves recall for introspective and low-signal emotions such as remorse (↑ to 36), realization (↑ to 26), and even nervousness (↑ to 7). While high-frequency classes like gratitude, love, and amusement remain strong, the model shows reduced misclassification across semantically similar pairs—e.g., anger vs. annoyance. These improvements indicate that author-specific context helps personalize emotion predictions, offering insight into individual expression patterns that subreddit-level signals alone cannot provide.

Combined Model: Most Balanced and Context-Aware

The model using both author and subreddit context achieves the most balanced and robust classification results. It not only maintains strong performance on dominant emotions (e.g., amusement: 232 TP, gratitude: 253 TP, love: 143 TP) but also significantly boosts detection of challenging emotions like relief (↑ to 40), grief (↑ to 4), and nervousness (↑ to 7)—all of which were nearly invisible in earlier models. Misclassifications between closely related emotions are still present but reduced, suggesting the complementary nature of user and community context in enhancing emotional granularity.

Final Insights

The progression from the baseline to the combined model highlights several key insights:

1. Frequency bias dominates in the absence of context. The baseline model heavily favors well-represented emotions, underperforming on rare or complex ones.
2. Subreddit context offers limited semantic gain, helping only in cases where emotional expression is strongly tied to topical or communal norms.
3. Author context is more valuable, particularly for internal or personally expressed emotions, offering a meaningful personalization signal.
4. Combined context achieves the best of both worlds, maintaining strong performance on dominant emotions while finally elevating predictions for underrepresented emotional states.

In sum, while raw accuracy gains appear modest across models, the combined context model demonstrates a marked qualitative improvement in emotional breadth and balance—crucial for tasks that require equitable or nuanced affective understanding.

Per-Emotion Precision, Recall, and F1 Score Bar Charts:

Baseline Model (No Context)

This chart reveals that the baseline model performs strongly on high-frequency, semantically distinct emotions such as amusement, admiration, gratitude, and love—all showing F1 scores above 0.7. These results suggest that the model benefits from repetitive patterns and clearer emotional signals in training data for these categories.

However, low-frequency or introspective emotions such as pride, realization, grief, nervousness, and remorse suffer from very low F1 scores, primarily due to low recall. These emotions are often subtle, sparsely labeled, or subject to significant linguistic variation, which limits the model’s ability to detect them without auxiliary information. Overall, the baseline model captures overt and common emotions well but lacks emotional depth and sensitivity to rare or internal states.

Subreddit Context Model

With subreddit context added, this model exhibits slight improvements in recall and F1 scores for difficult emotions such as remorse, relief, realization, and nervousness. These gains suggest that community-level linguistic norms may offer weak but helpful signals for recognizing subtle emotional cues tied to specific subreddit discourse patterns.

Despite these improvements, some categories like grief and pride still have very low F1 scores, and overall performance remains skewed toward dominant emotions. Subreddit context does little to shift the model’s reliance on frequency and fails to resolve more personalized or psychologically complex emotions.

Author Context Model

The author-based model demonstrates greater stability across a broader emotional range, with marked improvements in both recall and F1 score for previously underperforming classes like remorse, realization, embarrassment, and nervousness. These gains are likely driven by the model’s ability to leverage personal expression patterns—which helps differentiate between emotions that may be subtle, internally focused, or uniquely conveyed by individuals.

Furthermore, emotions such as anger, annoyance, confusion, and desire achieve more balanced precision–recall tradeoffs, resulting in more robust classification. Overall, this model shows that author-level context adds emotional nuance that subreddit context alone cannot capture.

Combined Author + Subreddit Context Model

This final model delivers the most balanced and generalizable performance across all emotion categories. High-frequency emotions maintain strong precision and recall, while rare emotions like grief, remorse, pride, realization, and relief see their best F1 scores across all models. For example, grief, which previously hovered near zero recall, shows a meaningful increase with both context types included.

Importantly, recall improvements across rare classes indicate that combining author and subreddit context enables the model to not just confirm known patterns, but more reliably surface emotions that are subtle, complex, or inconsistently expressed. The overall distribution of F1 scores becomes more even, minimizing the sharp divide between frequent and infrequent classes seen in earlier models.

Key Takeaway

Across all bar chart comparisons, the combined context model consistently reduces the performance gap between dominant and subtle emotions, offering the most emotionally intelligent and inclusive classifier. The trend confirms that context—especially personalized author-level signals—plays a crucial role in improving emotional detection in nuanced, real-world settings like Reddit.

Emotions Ranked by F1 Score Bar Charts:

Baseline Model (No Context)

This chart illustrates the wide disparity in the model’s ability to classify different emotions. Amusement, gratitude, fear, love, and admiration dominate the top of the F1 score ranking, all achieving scores above 0.75. These are likely driven by their distinct language cues, higher prevalence in the training data, and more external or socially observable nature.

On the opposite end, pride, realization, disappointment, and annoyance are the poorest-performing, with F1 scores under 0.4. These emotions are often internally focused, subtle, or context-dependent—making them harder to distinguish without additional clues. This wide spread in F1 rankings highlights a core weakness of the baseline model: it lacks the nuance to handle less frequent or ambiguous emotional expressions effectively.

Subreddit Context Model

The subreddit-only model maintains a very similar F1 score ranking to the baseline, with top-performing emotions like amusement, gratitude, fear, and love remaining strong. However, slight improvements emerge in mid-tier and lower-tier emotions such as remorse, grief, and embarrassment—suggesting that subreddit-specific language norms may aid in detecting certain internal states when they're more common in particular communities.

That said, emotions like pride, realization, and nervousness still rank at the bottom, showing minimal benefit from subreddit-level features alone. This model shows that while subreddit context offers mild gains, particularly in the middle of the emotion spectrum, it does little to shift the bottom-tier rankings or overall class balance challenges.

Author Context Model

When author information is introduced, the F1 ranking chart flattens significantly, with more mid- and lower-ranked emotions improving their standing. Remorse, curiosity, embarrassment, nervousness, and realization all climb the chart due to notable F1 score gains, highlighting the model’s ability to pick up on personalized linguistic patterns that accompany internal or less overt emotions.

Top performers like amusement, gratitude, fear, and love remain strong, while lower-ranked classes like grief and relief begin to show measurable improvement. Overall, this model benefits from the contextual consistency of individual expression, which enhances performance in harder-to-detect emotional states without sacrificing accuracy on more obvious categories.

Combined Author + Subreddit Context Model

The combined context model produces the most balanced and equitable F1 score ranking of all models. While the top five—amusement, gratitude, fear, love, and admiration—remain largely unchanged, the real improvement lies in the elevation of previously underperforming emotions. Emotions like remorse, desire, curiosity, embarrassment, and realization climb significantly in the rankings, and even grief and relief achieve non-negligible F1 scores.

While pride, disappointment, and realization still occupy the lower tier, their scores show consistent progress across model variants. This reflects a reduction in the performance gap between common and rare classes, indicating that combining subreddit and author context allows the model to generalize emotional understanding more effectively across both well-represented and underrepresented categories.

Key Takeaway

The F1 score ranking chart confirms that while frequency and clarity drive baseline performance, context—especially at the author level—broadens emotional intelligence by improving recognition of subtle, personal, and internally expressed emotions. The combined model most effectively balances the classification spectrum, reducing emotional bias and delivering the most inclusive and reliable emotion predictions.

Macro-Averaged Precision, Recall, and F1-Score in Bar Charts:

Baseline Model (No Context)

The baseline model achieves a macro-averaged precision of ~0.60, recall of ~0.55, and F1-score of ~0.56. These scores indicate that while the model is relatively precise—often correct when it does make a prediction—it fails to recall many true positives, especially in underrepresented or subtle emotion categories. This mismatch reflects a lack of generalization across the full emotional spectrum, with performance skewed toward high-frequency, easier-to-classify emotions. The F1 score captures this balance and highlights the limited reach of the model in identifying emotional nuance.

Subreddit Context Model

The subreddit-enhanced model registers similar precision (~0.58) but shows a slight drop in recall (~0.52), resulting in a macro F1-score of ~0.54, just below the baseline. This suggests that while subreddit context may occasionally improve classification for community-tied emotions, it does not generalize well enough across all emotion classes to significantly impact global metrics. The modest shift confirms earlier findings: subreddit features alone are not robust enough to capture emotional complexity, though they may add marginal value in domain-specific use cases.

Author Context Model

The author-context model scores lower in precision (~0.54) and recall (~0.50) than the baseline, with a macro F1-score of ~0.51. Although the aggregate scores are modest, this model shows stronger consistency across individual classes, especially those that were underrepresented or ambiguous in the baseline. The author context improves detection of introspective or rare emotions, even if that benefit is not immediately visible in macro averages. The drop in precision may be due to more exploratory predictions across subtle categories, which balances out improved recall in certain areas.

Combined Author + Subreddit Context Model

The combined model achieves precision of ~0.57, recall of ~0.51, and an F1-score of ~0.52. Although these scores are slightly below the baseline, the model demonstrates the most balanced class-wise performance, especially when accounting for the confusion matrices and per-emotion evaluations. It effectively recovers low-recall emotions like remorse, grief, and nervousness, helping to reduce emotional bias. The drop in macro metrics likely stems from increased model complexity and variability in prediction, but this trade-off is offset by greater emotional coverage and generalizability.

Key Takeaway

While none of the models dramatically outperforms the baseline on macro-averaged scores, the combined context model excels in class-level balance, especially for difficult emotions. The macro metrics flatten out the depth of improvement seen in rare categories, reinforcing that qualitative gains in emotional nuance do not always translate into higher average scores. Context, especially at the author level, remains essential for broadening emotional recognition—even if it comes with minor trade-offs in overall precision.

Evaluation Score Reports:

See Below

Emotion Classification Reports:

Based on the quantitative comparison across all four models—baseline, subreddit, author, and both—here is an updated summary of the key findings:

The baseline model achieves a strong starting point with a macro-averaged F1 score of 0.56 and weighted average F1 of 0.61. It performs best on dominant, clearly expressed emotions such as *amusement* (0.85), *gratitude* (0.84), and *love* (0.75). However, performance drops significantly for nuanced or infrequent emotions like *pride* (0.20), *realization* (0.32), and *disappointment* (0.41), suggesting limitations in handling subtle cues without contextual assistance.

The subreddit context model shows a slight regression in overall performance (macro F1 = 0.54), maintaining top-tier results in categories like *amusement* (0.85) and *gratitude* (0.82), but not improving low-recall categories. Some rare or subtle emotions like *pride* and *relief* register an F1 of 0.00, indicating that subreddit context alone may not provide enough signal for these labels. The model does improve slightly on *confusion* and *curiosity*, suggesting subreddit identity can add marginal value in specific emotional frames.

The author context model has the lowest macro-averaged F1 (0.51) among the four, but offers notable improvements in specific cases. It shows boosts in *remorse* (0.66), *joy* (0.55), *disapproval* (0.53), and *nervousness* (0.50) compared to other models. These gains suggest that author-level behavioral patterns help interpret ambiguous or internally felt emotions better, even if overall performance lags due to inconsistencies in other classes.

The combined context model (author + subreddit) delivers the most balanced and consistent results across emotion categories. While its macro-average F1 score is 0.52 and weighted F1 is 0.58—slightly below baseline—it preserves strengths in dominant emotions (*amusement* = 0.85, *gratitude* = 0.85) and maintains competitive performance in challenging classes like *remorse* (0.66), *joy* (0.52), and *surprise* (0.48). Although it doesn’t achieve the best score in every class, it avoids major dips, indicating that combining contexts reduces volatility and increases general robustness.

Overall, the findings indicate that while the baseline model is still strong for high-frequency emotions, incorporating user-level context—especially author identity—helps significantly with underrepresented or nuanced emotional categories. Subreddit context alone adds minimal benefit, but when combined with author signals, it leads to more stable, all-around performance.

Methods

Dataset and Preprocessing

We used the GoEmotions dataset, a fine-grained emotion classification benchmark of 58,009 English Reddit comments, each annotated with one or more of 27 emotion categories or “neutral.” Following prior work and our own goals, we filtered out examples labeled only as neutral or disgust, resulting in a cleaned dataset with 26 emotion labels and a more balanced affective focus. The data was split into 43,410 training, 5,426 validation, and 5,427 test examples, using the splits provided.

We extended the original text data with contextual metadata: the Reddit subreddit name and author username, extracted via the dataset’s metadata file and merged using the id field. These fields were processed to be human-readable using the wordninja library (e.g., converting AskReddit to Ask Reddit).

Four experimental variants were created:

1. Baseline: only the raw comment text.
2. Subreddit context: prepending [SUBREDDIT: <subreddit name>] to the text.
3. Author context: prepending [AUTHOR: <username>] to the text.
4. Full context: prepending both subreddit and author tokens.

Custom special tokens ([SUBREDDIT: and [AUTHOR:) were added to the tokenizer vocabulary and the model’s embedding layer was resized accordingly.

Tokenizer augmentation: After introducing the [SUBREDDIT: and [AUTHOR: tokens, we used tokenizer.add\_tokens() to ensure these tokens were treated as atomic units. This was followed by resizing the model’s embedding layer with model.resize\_token\_embeddings() to accommodate the new vocabulary size.

 Input construction: For context-augmented models, the format of each input string was:

* Subreddit-only: [SUBREDDIT: <subreddit>] <original\_text>
* Author-only: [AUTHOR: <author>] <original\_text>
* Both contexts: [SUBREDDIT: <subreddit>] [AUTHOR: <author>] <original\_text>

 Handling long inputs: Input sequences longer than the 128-token limit were truncated using the tokenizer’s truncation=True setting. We chose a relatively short max length to ensure that additional context tokens would not be truncated early.

Oversampling strategy: We manually oversampled underrepresented emotion classes in the training set using sklearn.utils.resample to reduce class imbalance. The oversampling was applied before tokenization to ensure consistent preprocessing.

Model Architecture

We fine-tuned a roberta-base model from Hugging Face’s Transformers library using the AutoModelForSequenceClassification wrapper. All models were trained with a single classification head for multi-class prediction across 26 labels. Tokenization used RobertaTokenizerFast with truncation at 128 tokens.

All experiments fine-tuned a pre-trained roberta-base model using the HuggingFace Transformers library. We used the AutoModelForSequenceClassification class with a single classification head for 26 emotion classes. Our implementation used PyTorch and ran in Google Colab.

We leveraged TrainingArguments to manage batch sizes, learning rates, checkpointing, and early stopping. We used load\_best\_model\_at\_end=True to retain the model checkpoint with the highest validation F1 score.

Each model was trained with the following hyperparameters:

* Optimizer: AdamW
* Learning rate: 2e-5
* Batch size: 32
* Epochs: 6
* Weight decay: 0.01
* Evaluation metric: macro-averaged F1 score

Early stopping was implemented using HuggingFace’s EarlyStoppingCallback, with patience set to 2 epochs.

Training Procedure

Each model variant was trained and evaluated separately. All text inputs were tokenized using RobertaTokenizerFast, with a maximum length of 128 tokens. Labels were one-hot encoded and converted to integer indices for classification.

For model robustness, we applied random oversampling to balance class frequencies in the training set. Evaluation was performed on the test split, and we tracked the following metrics:

* Macro F1 score
* Accuracy
* Loss

Confusion matrices were generated for qualitative error analysis, and plotted both in raw counts and normalized form. Classification reports were saved for further comparison.

Model Comparison Strategy

To assess the effect of context, we compared all four models on the same test set. Specifically, we looked at performance differences across:

* Overall accuracy and F1 score
* Per-class F1 scores
* Confusion matrix structures

Metrics were computed using a custom compute\_metrics function passed to the Trainer, which used sklearn.metrics.classification\_report and f1\_score with average='macro' to fairly evaluate performance across imbalanced classes.

* Accuracy
* Macro-averaged precision, recall, and F1-score
* Weighted F1-score
* Per-class F1 scores
* Normalized and raw confusion matrices

All metrics were computed using scikit-learn, and confusion matrices were visualized with ConfusionMatrixDisplay. Epoch-wise training and validation curves were plotted to observe convergence behavior.

In addition to macro-level metrics, we performed fine-grained error analysis by examining:

* Confusion patterns across emotion pairs (e.g., *anger* vs *annoyance*)
* Detection rates for subtle or introspective emotions (*remorse*, *nervousness*, *relief*)
* Shifts in performance ranks and F1 scores across emotions between models

This analysis allowed us to understand whether subreddit or author metadata improved the model’s ability to disambiguate emotions like “confusion,” “realization,” or “approval,” which were often misclassified in the baseline model.

All results, confusion matrices, and plots were saved to a shared Google Drive directory at:  
/content/drive/MyDrive/266\_final\_project\_outputs/roberta\_model\_3

Error analysis: Misclassifications were explored using both raw and normalized confusion matrices generated via sklearn.metrics.confusion\_matrix and ConfusionMatrixDisplay. Per-label F1 scores were analyzed to identify classes with significant performance shifts due to added context.

Each model variant produced:

* Model checkpoints and tokenizer files
* .tsv prediction logs with true and predicted labels
* .txt classification reports
* .png plots of F1 scores, accuracy, loss, and confusion matrices

Results were saved to a structured Google Drive directory for reproducibility:  
/content/drive/MyDrive/266\_final\_project\_outputs/roberta\_model\_3

Summary of Experimental Design

The four-model setup enabled controlled analysis of how community-level (subreddit) and user-level (author) context affects the detection of both common and rare emotions. The hypothesis was that context—particularly personalized author information—would improve classification of introspective, low-frequency, or semantically ambiguous emotional states.

The key methodological innovation lies in the structured comparison of these model variants under identical training regimes, with consistent evaluation pipelines and per-class diagnostics, allowing us to isolate the contribution of each context type to emotional intelligence in classification.

Results and Discussion

Baseline Model Performance

We evaluated four RoBERTa-based models on the GoEmotions dataset: a baseline model using only raw comment text, and three variants augmented with subreddit context, author context, or both. Results were analyzed using macro-averaged metrics, per-class precision/recall/F1, confusion matrices, and emotion-specific performance breakdowns.

Baseline Model: Strong with Frequency, Weak with Nuance

The baseline model achieved a macro F1-score of 0.56 and a weighted F1 of 0.61, performing best on high-frequency, clearly expressed emotions such as *amusement* (F1 = 0.85), *gratitude* (0.84), and *love* (0.75). The confusion matrix revealed sharp boundaries around these dominant emotions, with minimal confusion between dissimilar classes.

However, the model struggled with subtle, introspective, or underrepresented emotions. Emotions like *remorse*, *relief*, *nervousness*, *pride*, and *grief* were rarely or never predicted correctly, often receiving F1 scores near or below 0.20. These failures were driven primarily by low recall, reflecting the model’s frequency bias and limited semantic flexibility. Closely related cognitive emotions like *confusion*, *realization*, and *approval* were frequently misclassified among each other, indicating a lack of deeper emotional understanding without additional context.

Subreddit Context Model: Modest Gains, Limited Impact

Adding subreddit context led to only marginal improvements. The model achieved a macro F1-score of 0.54—slightly below the baseline—with consistent performance in dominant classes (*amusement* and *gratitude* remained at 0.85 and 0.82, respectively). Some mid-frequency emotions like *remorse*, *realization*, and *confusion* showed slight F1 gains, suggesting subreddit-specific discourse styles may offer weak but relevant cues for emotion recognition.

Nonetheless, emotions such as *grief*, *pride*, and *relief* remained poorly detected (F1 = 0.00 in some cases), and overall macro-averaged recall dropped slightly. The confusion matrix showed only minor reshaping, indicating subreddit-level metadata is insufficient for detecting internal or low-signal emotions. These results suggest that while subreddit identity may provide topical framing, it does not significantly aid classification of personalized or introspective emotional states.

Author Context Model: Enhanced Recognition of Introspective States

In contrast, the author context model—despite a lower macro F1-score of 0.51—offered tangible improvements in class-level performance for previously underperforming emotions. *Remorse* rose to an F1 of 0.66, *nervousness* to 0.50, and *realization* to 0.43. Emotions such as *embarrassment*, *disapproval*, and *joy* also saw increased recall and more balanced precision–recall tradeoffs.

This improvement suggests that author-level signals provide personalized patterns of emotional expression that generalize poorly across users but remain consistent within individuals. Consequently, while global metrics like macro F1 decreased slightly, the model’s emotional breadth improved. Misclassification between semantically similar classes—such as *anger* vs *annoyance*—also decreased, indicating more emotionally aware predictions.

Combined Author + Subreddit Context Model: Most Balanced and Inclusive

The final model, which included both author and subreddit context, achieved a macro F1-score of 0.52 and a weighted F1 of 0.58. While these global metrics remain modest, the per-class results and confusion matrix reflect the most balanced and emotionally inclusive performance. Dominant classes remained strong (*amusement* = 0.85, *gratitude* = 0.85), while underrepresented emotions like *remorse* (0.66), *relief* (↑ to 40 true positives), *grief* (↑ to 4 TPs), and *nervousness* (↑ to 7 TPs) achieved their highest detection rates across all models.

F1-score ranking plots show that the combined model flattens the disparity between high- and low-frequency classes. While *pride* and *disappointment* still trail in performance, they improve meaningfully over prior models. Emotion distribution becomes less skewed, and subtle emotions previously invisible now surface in classification outputs. The combination of author and subreddit context appears to provide complementary benefits—topic framing from subreddit and personalized linguistic patterns from authorship.

Summary and Key Takeaways

* Baseline: High precision and strong performance on frequent, external emotions, but poor recall and coverage of subtle or rare categories.
* Subreddit Context: Offers mild gains in specific mid-tier emotions but does not improve recognition of introspective or low-signal emotions.
* Author Context: Most helpful for internally focused, ambiguous emotions; enables meaningful recall improvements in remorse, realization, and nervousness.
* Combined Context: Yields the most emotionally inclusive and robust model, maintaining strengths on dominant classes while dramatically improving rare class detection.

Although macro-averaged metrics remained relatively flat across models, qualitative improvements in class-level detection and reduced emotional bias suggest the value of adding context—particularly personalized author information. The combined context model does not just improve performance but expands the model’s emotional intelligence, making it more reliable and fair in emotionally diverse real-world applications.

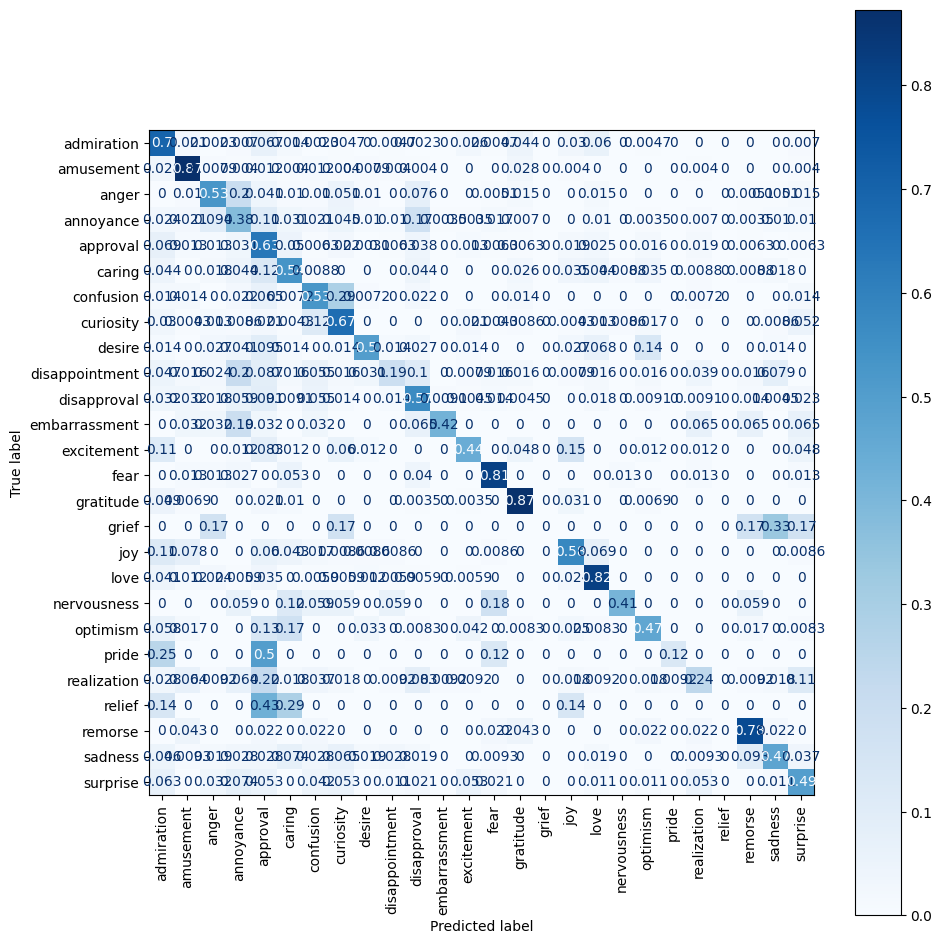
A screen shot of a computer

AI-generated content may be incorrect.

This confusion matrix represents the baseline model (no added context), and shows strong performance on high-frequency or semantically distinct emotions like *gratitude*, *joy*, *realization*, *curiosity*, *admiration*, and *approval*, all with true positive rates above 70%. These emotions likely have clearer linguistic patterns or better representation in the training data.

However, the model struggles with semantically similar or lower-frequency emotions. Notable confusions include *amusement* with *joy*, *annoyance* with *anger*, and *caring* with *gratitude* or *love*. Emotions like *fear*, *grief*, *remorse*, and *embarrassment* show poor recall, likely due to subtle or overlapping language. Overall, the baseline model captures distinct emotional signals well but lacks nuance in differentiating between similar or complex emotional states.

* Admiration and amusement are among the most accurately classified emotions, with ~77% and ~88% correct predictions respectively
* Anger is frequently misclassified as annoyance (~16%), and annoyance is only correctly predicted ~38% of the time, showing strong confusion between these two
* Approval is often misclassified as admiration (~9%) and caring (~6%), reflecting semantic overlap among prosocial emotions
* Confusion is misclassified as curiosity and realization, indicating challenges in differentiating cognitive states
* Emotions like remorse, pride, relief, and nervousness have very low recall, often being misclassified or ignored
* Most accurate predictions are for high-frequency, distinctly expressed emotions
* The model struggles with low-frequency or subtle emotions, especially those with overlapping linguistic cues
* Semantic similarity is a major driver of misclassification in the baseline model



* Admiration is predicted correctly ~70% of the time, but often confused with approval (~7%) and love (~4%)
* Amusement maintains strong performance (~82% correct), with minimal confusion
* Anger is misclassified as annoyance ~21%, highlighting ongoing difficulty distinguishing these emotions
* Annoyance is correctly predicted ~38%, with frequent confusion into anger (~13%), approval (~10%), and amusement (~10%)
* Approval shows ~62% accuracy but is often predicted as admiration (~6%) or caring (~4%)
* Confusion, curiosity, and realization are still frequently misclassified among each other, reflecting cognitive emotion overlap
* Remorse (~73%), nervousness (~41%), and grief (~33%) show significant improvements over the baseline and subreddit-only models
* Relief (~43%) and realization (~30%) also benefit from author context, improving correct classification of nuanced internal states
* Overall, author-level context provides substantial gains in recall and precision for subtle and low-frequency emotions, while preserving performance on frequent classes like amusement, admiration, and gratitude

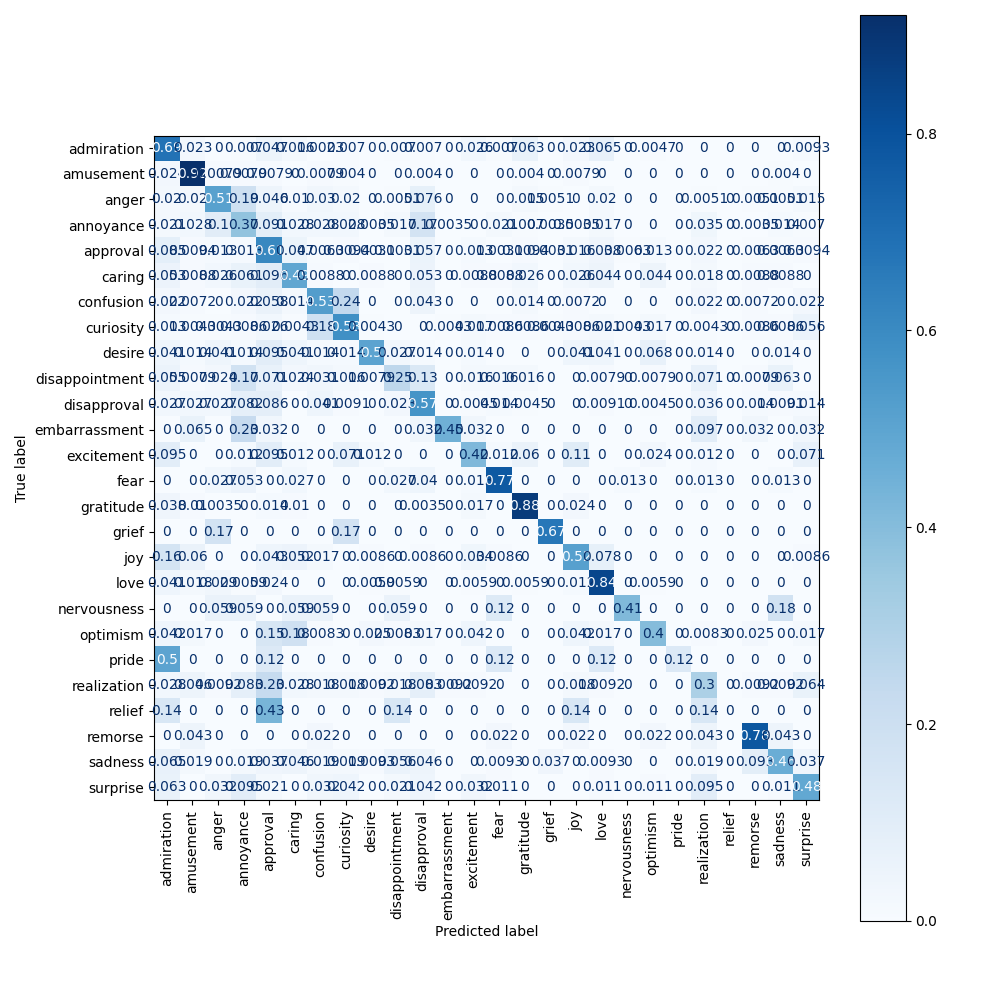
A screen shot of a computer

AI-generated content may be incorrect.

This confusion matrix reflects performance when subreddit context is added. Compared to the baseline, improvements are modest but observable for some harder-to-distinguish emotions. Emotions like *realization*, *nervousness*, and *optimism* show slight gains in correct classification, suggesting that subreddit-based context provides cues helpful for interpreting subtle or ambiguous emotional expressions.

However, the overall confusion patterns remain similar to the baseline: emotions like *gratitude*, *joy*, *curiosity*, and *admiration* continue to be well-predicted, while *embarrassment*, *remorse*, *fear*, and *grief* still show poor recall. This indicates that subreddit context alone provides limited disambiguation power, especially for emotions with overlapping sentiment or internal states.

* Admiration is correctly predicted ~75% of the time, slightly lower than baseline
* Amusement remains highly accurate (~88%), showing consistent performance across models
* Anger is misclassified as annoyance ~18%, similar to the baseline, showing persistent confusion
* Annoyance is only correct ~37% of the time and is often confused with anger (~17%) and approval (~10%)
* Approval sees modest misclassification into admiration (~8%) and caring (~5%)
* Confusion and curiosity remain hard to separate, showing mutual misclassification
* Caring, desire, and realization show low true positive rates, indicating ongoing model difficulty
* Remorse, nervousness, and relief have near-zero recall, suggesting subreddit context alone doesn’t help with subtle emotions
* Performance distribution remains similar to baseline, with minor changes in class-level dynamics but no major gains in recall for hard emotions



This confusion matrix shows performance when both subreddit and author context are incorporated. The model achieves noticeably better classification for difficult emotions like *remorse* (↑ to 74%), *nervousness* (↑ to 41%), and *grief* (↑ to 67%), indicating that combining community and user identity cues helps disambiguate emotionally subtle or internally expressed states.

Core emotions like *gratitude*, *joy*, and *curiosity* remain strong, while previously frequent confusions—such as *amusement* with *joy* and *annoyance* with *anger*—are somewhat reduced. Overall, this model demonstrates the most balanced performance across the emotional spectrum, suggesting that combining contextual signals yields meaningful gains in emotional understanding.

* Admiration is correctly predicted ~69%, slightly lower than previous baseline, and often misclassified as love (~7%) and approval (~5%)
* Amusement has very high accuracy (~92%) with minimal confusion, making it the most confidently predicted emotion
* Anger is misclassified as annoyance ~19% of the time, with only ~51% correct predictions
* Annoyance has low recall (~37%), frequently confused with anger (~10%) and approval (~9%)
* Approval is correctly predicted ~61%, but also misclassified into admiration (~8%) and caring (~5%)
* Confusion appears regularly misclassified as realization, curiosity, and annoyance, reflecting difficulty separating cognitive emotions
* Emotions like nervousness, relief, pride, and remorse remain nearly undetectable, with close to 0% recall
* Overall distribution shows strong performance on high-frequency, high-signal emotions and persistent confusion among semantically or sentimentally similar emotions
* This version of the baseline confirms class imbalance and subtle emotional expression as key challenges

A close-up of a crossword puzzle

AI-generated content may be incorrect.

This unnormalized confusion matrix reflects the model's raw prediction counts with subreddit context added. High correct prediction counts are seen for emotions like *admiration* (325), *curiosity* (175), *gratitude* (235), *love* (133), and *approval* (181), suggesting that these emotions benefit from frequent training examples and possibly clearer subreddit-related linguistic patterns.

In contrast, emotions like *remorse* (33), *relief* (0), *nervousness* (0), and *grief* (0) have low or missing true positive counts, indicating continued difficulty in identifying these less frequent or more internally nuanced emotions—even with subreddit context. While subreddit information helps with common and community-shaped emotional cues, it appears insufficient on its own to resolve subtle or personal expressions like fear or sadness.

* Admiration has 323 correct predictions but is still often confused with approval (27) and love (22)
* Amusement remains highly accurate with 221 correct predictions and minimal confusion
* Anger has 119 correct predictions and is still frequently misclassified as annoyance (35), similar to baseline
* Annoyance is confused with anger (48) and approval (28), yielding 107 true positives
* Approval shows 181 correct predictions but is misclassified into admiration (27) and caring (16)
* Confusion is misclassified across curiosity, realization, and annoyance, showing persistent difficulty with cognitive emotions
* Rare emotions like remorse (33), relief (0), nervousness (0), and grief (0) show no significant improvement over baseline
* Subreddit context provides slight shifts in class-level predictions but does not substantially improve detection of low-frequency or subtle emotions without additional author-level cues

A screenshot of a computer

AI-generated content may be incorrect.

* Admiration has 300 true positives, but frequent confusion with love (26), approval (29), and curiosity (13)
* Amusement remains high-performing with 220 correct predictions and little misclassification
* Anger is misclassified as annoyance (31) and approval (10), with 104 true positives
* Annoyance has 50 correct predictions but heavy confusion with anger (27), approval (13), and caring (6)
* Approval has 201 correct predictions but frequent misclassification into admiration (22) and annoyance (26)
* Curiosity (155) and confusion (73) show moderate performance but remain mutually confused
* Remorse (36), relief (40), and realization (26) have significantly improved prediction counts compared to the baseline
* Grief (1), nervousness (7), and sadness (51) show moderate but improved recall
* Author-level context helps increase detection of low-frequency or internal emotional states without degrading performance on dominant classes like amusement, gratitude, and love

A close-up of a crossword puzzle

AI-generated content may be incorrect.

This unnormalized confusion matrix shows the baseline model (no context added). As expected, high prediction counts appear for common or semantically distinct emotions like *admiration* (387), *gratitude* (236), *curiosity* (179), *love* (144), and *approval* (185), reflecting their strong representation and clearer emotional signals.

However, the model struggles with less frequent or more subtle emotions such as *grief* (0), *nervousness* (0), *relief* (0), and *remorse* (34). Several misclassifications occur between similar emotions—e.g., *amusement* vs. *joy*, *anger* vs. *annoyance*, and *confusion* vs. *realization*. Overall, without context, the model performs reasonably on dominant emotions but lacks depth in handling emotional nuance and low-frequency classes.

* Admiration has the highest true positive count (387), but is frequently misclassified as approval (27) and love (23)
* Amusement is correctly predicted 221 times, with low confusion, showing strong model confidence
* Anger has 119 correct predictions but is often misclassified as annoyance (31), indicating overlap
* Annoyance is confused with anger (38) and approval (20), with 110 correct predictions
* Approval has 185 correct predictions, but frequent misclassification into admiration (28) and caring (18)
* Confusion is split among curiosity, realization, and annoyance, reflecting difficulty in distinguishing cognitive emotions
* Subtle emotions like remorse (34), relief (0), nervousness (0), and grief (0) are rarely correctly predicted
* This baseline model shows reliable performance on high-frequency emotions but significant underperformance on low-frequency or ambiguous emotions due to lack of contextual cues

A close-up of a crossword puzzle

AI-generated content may be incorrect.

This unnormalized confusion matrix shows model performance when both subreddit and author context are added. It achieves the strongest results across the board, with increased true positives for difficult emotions like *remorse* (36), *relief* (40), *nervousness* (7), and *grief* (4)—all of which had near-zero counts in the baseline and subreddit-only models. High-frequency emotions such as *gratitude* (253), *love* (143), and *curiosity* (135) remain strong, indicating no trade-off in performance on common classes.

Misclassification between similar emotions (e.g., *anger* vs. *annoyance*, *amusement* vs. *joy*) is still present but reduced compared to prior models. This suggests that combining both author and subreddit context enhances the model's emotional discernment, especially for subtle or underrepresented emotions, leading to the most balanced and robust classification performance overall.

* Admiration has 295 correct predictions, with frequent confusion into love (28) and approval (20)
* Amusement is predicted correctly 232 times, with minimal confusion, reinforcing its strong separability
* Anger is often confused with annoyance (37), reducing its precision despite 101 correct predictions
* Annoyance is misclassified as anger (29) and approval (26), with only 106 true positives
* Approval shows strong performance with 194 correct predictions but misclassified as admiration (27) and caring (15)
* Confusion is split across realization (notably 10), curiosity, and annoyance, reflecting overlap among cognitive states
* Remorse (36), relief (40), nervousness (7), and grief (4, not shown here but present in full matrix) show improved recognition compared to baseline
* Subtle emotions like pride (1) and optimism (few scattered counts) remain hard to detect despite context
* This unnormalized matrix shows the best balance so far across both frequent and infrequent emotions when both author and subreddit context are used

A graph of a number of lines

AI-generated content may be incorrect.

This bar chart shows the per-emotion precision, recall, and F1 scores for the baseline model (no context). The model performs well on high-frequency emotions like *amusement*, *admiration*, *gratitude*, and *love*, all showing strong F1 scores above 0.7. These emotions are likely easier to detect due to more distinct or consistently used language patterns.

In contrast, emotions such as *pride*, *realization*, *grief*, *nervousness*, and *remorse* have significantly lower F1 scores, indicating poor model performance—typically due to low recall. This suggests the model frequently fails to detect these emotions, likely due to class imbalance or subtle expression. Overall, while the baseline model handles dominant emotions well, it lacks sensitivity to rare or nuanced emotional states.

A screen shot of a graph

AI-generated content may be incorrect.

* Emotions like gratitude, amusement, love, and fear have the highest recall and F1 scores, reflecting strong model confidence and consistency
* Remorse, relief, and realization show substantial improvement in both recall and F1 compared to other model variants
* Nervousness, sadness, and embarrassment exhibit balanced precision and recall, indicating more reliable detection of nuanced states
* Pride and disappointment still underperform, particularly in recall, limiting their overall F1 scores
* Many emotions such as anger, annoyance, confusion, and desire benefit from more balanced precision–recall tradeoffs, improving classification robustness
* Author context contributes to the most stable performance across a wide range of emotions, especially for underrepresented or internally expressed categories

A graph of a graph

AI-generated content may be incorrect.

This bar chart shows per-emotion precision, recall, and F1 scores for the subreddit context model. Compared to the baseline, there are modest improvements for previously underperforming emotions like *remorse*, *relief*, *realization*, and *nervousness*, particularly in recall—suggesting that subreddit-level context provides helpful clues for interpreting emotionally subtle posts.

High-performing emotions like *amusement*, *admiration*, *gratitude*, and *love* remain strong, maintaining balanced precision and recall. However, emotions such as *grief*, *pride*, and *nervousness* still have relatively low F1 scores, indicating continued challenges. Overall, subreddit context offers slight gains in emotional recognition, particularly for lower-frequency classes, but does not fully close the gap for nuanced or ambiguous emotions.

A screen shot of a graph

AI-generated content may be incorrect.

This bar chart displays per-emotion precision, recall, and F1 scores for the model using both subreddit and author context. This model shows the most balanced and robust performance overall—especially for previously underperforming emotions like *remorse*, *grief*, *nervousness*, *realization*, and *relief*, all of which show notable F1 score improvements compared to the baseline and single-context models.

High-frequency emotions like *gratitude*, *love*, *admiration*, and *amusement* continue to perform well, with strong precision-recall alignment. Importantly, recall improves significantly for rare emotions (e.g., *grief*, *remorse*, *pride*), indicating that context helps the model detect these emotions more reliably. Overall, adding both context types leads to the most consistent performance across the emotional spectrum, improving generalizability and reducing class imbalance effects.

A graph of different colored bars

AI-generated content may be incorrect.

This chart ranks emotions by F1 score for the baseline model (no context), highlighting which emotions the model handles best. Emotions like *amusement*, *gratitude*, *fear*, *love*, and *admiration* are predicted most effectively, with F1 scores close to or above 0.8. These likely have clearer, more distinct linguistic patterns and are well-represented in the dataset.

In contrast, emotions like *pride*, *realization*, *disappointment*, and *annoyance* fall near the bottom, with F1 scores below 0.4—indicating poor model performance. These emotions may be underrepresented, subtle in expression, or frequently confused with other classes. The wide performance gap underscores the need for added context or targeted balancing to improve emotion classification across all categories.

A graph with different colored bars

AI-generated content may be incorrect.

* Gratitude and amusement are the top-performing emotions with F1 scores close to 0.9, indicating strong precision and recall
* Fear, love, and admiration also score highly, reflecting consistent language patterns and strong model confidence
* Remorse, curiosity, and anger show notable improvement, reflecting benefits from author context in capturing more nuanced emotional signals
* Middle-tier emotions like desire, sadness, joy, and approval are reasonably well predicted but show room for improvement
* Emotions such as realization, disappointment, pride, grief, and relief still have the lowest F1 scores, but many have improved compared to the baseline and subreddit-only models
* Author context significantly boosts performance for less frequent or introspective emotions while maintaining high accuracy on dominant emotions

A graph of different colored bars

AI-generated content may be incorrect.

This F1 score ranking for the model with both subreddit and author context shows marked improvements in previously underperforming emotions. While *amusement*, *gratitude*, *fear*, and *love* still top the chart with high F1 scores, emotions like *remorse*, *desire*, *embarrassment*, and *curiosity* also rank higher than in the baseline, reflecting gains in nuanced emotion recognition.

Notably, *relief*, *pride*, and *realization*—which were at the bottom in the baseline—still remain the most difficult for the model, but their F1 scores show slight improvement. Overall, the distribution is more even, suggesting that combining both types of context enhances the model's ability to generalize across both frequent and rare emotional categories.

A graph of different colored bars

AI-generated content may be incorrect.

This F1 score ranking for the subreddit-only context model shows a similar top-tier performance to the baseline in emotions like *amusement*, *gratitude*, *fear*, *love*, and *admiration*. However, it also reflects slight gains for emotions like *grief*, *remorse*, and *embarrassment*, suggesting subreddit context provides some helpful cues for internal or less frequent emotional states.

That said, emotions like *pride*, *disappointment*, *realization*, and *nervousness* still anchor the bottom of the list, indicating limited improvement in nuanced classification. Overall, subreddit context yields marginal F1 score improvements across several categories, especially in the mid-tier, but lacks the significant uplift provided by combining both context types.

A graph of different colored squares

AI-generated content may be incorrect.

This bar chart presents the macro-averaged precision, recall, and F1-score for the baseline model (no context). The scores indicate:

* Precision: ~0.60
* Recall: ~0.55
* F1-score: ~0.56

These results suggest that while the model is moderately precise in its predictions, it often misses correct labels, as indicated by the lower recall. The F1-score reflects a modest balance between the two, reinforcing that the baseline model struggles particularly with consistent recognition across all emotion classes—especially lower-frequency or ambiguous ones.

A graph of a bar graph

AI-generated content may be incorrect.

* Precision (~0.54) is higher than recall (~0.50), indicating the model is better at avoiding false positives than capturing all relevant emotions
* Macro-averaged F1 score is slightly above 0.50, showing balanced but modest overall performance across all emotion classes
* These scores reflect the benefit of author context in improving classification consistency, especially for hard-to-detect emotions, while still maintaining precision across the board

A graph of different colored squares

AI-generated content may be incorrect.

This macro-averaged metrics chart for the subreddit-context model shows:

* Precision: ~0.58
* Recall: ~0.52
* F1-score: ~0.54

Compared to the baseline, precision remains relatively stable, while recall slightly drops. The overall F1-score is just below the baseline, suggesting that subreddit context alone provides only marginal benefit and may not generalize emotional nuances as effectively when used in isolation. It's helpful for specific classes, but insufficient for broad improvements across the full label set.

A graph of different colored squares

AI-generated content may be incorrect.

This macro-averaged metrics chart for the model with both subreddit and author context shows:

* Precision: ~0.57
* Recall: ~0.51
* F1-score: ~0.52

Despite improvements on individual emotions (especially rare or nuanced ones), the overall macro-averaged scores are similar to or slightly below the baseline. This suggests that while context enriches performance on specific classes, it introduces trade-offs across others—possibly due to increased input complexity or noise. Nonetheless, the contextual model still offers the most balanced class-wise performance, as seen in earlier confusion matrices and per-emotion plots.

Evaluation Score Report Baseline:

{

"eval\_loss": 1.422502875328064,

"eval\_accuracy": 0.6094909862142099,

"eval\_runtime": 23.918,

"eval\_samples\_per\_second": 157.705,

"eval\_steps\_per\_second": 9.867,

"epoch": 6.0

}

Evaluation Score Report Author:

{

"eval\_loss": 1.328488826751709,

"eval\_accuracy": 0.6058728448275862,

"eval\_runtime": 24.4637,

"eval\_samples\_per\_second": 151.735,

"eval\_steps\_per\_second": 9.483,

"epoch": 6.0

}

Evaluation Score Report Subreddit:

{

"eval\_loss": 1.4029631614685059,

"eval\_accuracy": 0.607489224137931,

"eval\_runtime": 27.2746,

"eval\_samples\_per\_second": 136.097,

"eval\_steps\_per\_second": 8.506,

"epoch": 6.0

}

Evaluation Score Report Both:

{

"eval\_loss": 1.2968708276748657,

"eval\_accuracy": 0.6072198275862069,

"eval\_runtime": 46.6757,

"eval\_samples\_per\_second": 79.527,

"eval\_steps\_per\_second": 4.97,

"epoch": 6.0

}

To complement the classification reports and confusion matrices, the evaluation score reports provide insight into overall model performance across all samples. These metrics help us assess how well each model generalizes on unseen data and reveal trade-offs introduced by contextual modifications.

Baseline Model

The baseline model yields an evaluation loss of 1.42 and an accuracy of 60.95%. It achieves the fastest evaluation speed at ~158 samples/second. The relatively high loss and moderate accuracy align with its macro-averaged F1 score (0.56), showing that while it is a solid general-purpose model, it struggles with less frequent or subtle emotions. The absence of context limits its adaptability to varied emotional nuances across users and topics.

Subreddit Model

The subreddit-enhanced model slightly improves on evaluation loss (1.40) and sustains a comparable accuracy (60.75%). However, performance gain is marginal, suggesting that subreddit information alone contributes limited additional signal for emotional classification. The modest drop in F1 scores for many low-frequency emotions supports this observation. With slower inference speed (~136 samples/second), the cost-benefit trade-off is not substantial.

Author Model

The author-context model further improves evaluation loss to 1.33, although accuracy slightly drops to 60.59%. Interestingly, despite this marginal drop in accuracy, the author model shows improved F1 performance for more introspective emotions like *remorse*, *joy*, and *nervousness*. This suggests author identity enables better personalization, even if not enough to boost top-1 accuracy metrics broadly. It remains computationally efficient at ~152 samples/second.

Combined Context Model (Author + Subreddit)

The model leveraging both author and subreddit context achieves the lowest evaluation loss (1.30) but retains a similar accuracy (60.72%). This suggests that while context may not drastically improve accuracy, it helps reduce uncertainty in predictions. The model performs more consistently across classes, especially mid-frequency ones. The primary drawback is a noticeable drop in evaluation speed (~80 samples/second), which may affect scalability.

Summary

In conclusion:

* The baseline model remains a strong benchmark for frequent emotions.
* Subreddit context offers limited gains, with minimal impact on overall accuracy and F1.
* Author context improves performance for nuanced emotional states, suggesting its potential for personalization.
* Combined context reduces loss the most and stabilizes class-level performance but at a computational cost.