Beyond the Text: Incorporating Social Context for Emotion Classification Using GoEmotions

Arun Agarwal, Nazia Haque, Jefferson-Stanley Jules

W266: Natural Language Processing
UC Berkeley School of Information
GitHub: https://github.com/jsjules/naj-266.git

Abstract

Emotion recognition plays a vital role in applications such as empathetic chatbots, content moderation, and mental health monitoring. However, current text-only models often struggle to detect rare or introspective emotions like grief, relief, and nervousness, instead favoring high-frequency and overt expressions such as joy or anger. This project investigates whether incorporating social context, specifically subreddit community identity and user authorship patterns, can improve fine-grained emotion classification. Using the GoEmotions dataset of 58,000 Reddit comments labeled with 27 emotion categories, we trained both RoBERTa- and DeBERTa-based models across four configurations: text-only, text with subreddit context, text with author context, and a combined model with both. Our text-only RoBERTa baseline achieved a macro F1-score of 0.56, substantially outperforming the original GoEmotions benchmark of 0.46. Context-aware models further improved the detection of underrepresented emotions, with true positives for relief increasing from 0 to 40, grief from 0 to 4, and nervousness from 0 to 7. Author-level context was especially effective for capturing introspective emotional states, while subreddit metadata offered more modest gains. Although macro-level metrics showed incremental changes, class-wise analysis revealed broader emotional coverage and reduced frequency bias. These findings underscore the importance of integrating social metadata to build more nuanced, equitable, and psychologically aware emotion classification systems.

Introduction

Recognizing emotion in language is a long-standing goal in Natural Language Processing (NLP), with applications ranging from empathetic conversational agents and personalized content recommendation to mental health monitoring and toxic content detection [1,2,3]. As online platforms like Reddit become increasingly central to public discourse, the ability to detect *fine-grained* emotions, especially in short, informal user-generated content, has grown in importance. However, current emotion classification systems often focus exclusively on textual content, overlooking the broader *social context* in which that language was produced [4]. This leads to systematic failures in detecting rare, ambiguous, or introspective emotions such as *grief*, *remorse*, and *nervousness*, emotions that are critically important for real-world systems concerned with user safety and wellbeing [5,6].

Humans rarely interpret emotion based on text alone. Context, such as who is speaking, and where, is a critical component of emotional understanding. A Reddit comment like "*this guy's shoulders could carry a car*" expresses admiration only if one knows it was posted in a bodybuilding subreddit. Similarly, "*We SERIOUSLY NEED to have Jail Time based on a person's race*" reads very differently in a legal advice subreddit than in a fantasy world-building forum [7]. These examples underscore the interpretive power of *subreddit* and *author* metadata signals that are typically discarded in standard NLP pipelines [8].

In this project, we hypothesize that incorporating *social metadata*, specifically subreddit community identity and author behavior, can improve emotion classification performance, particularly for subtle or underrepresented emotional states. We explore this using the **GoEmotions** dataset [1], a manually annotated corpus of 58,000 Reddit comments labeled with 27 emotion categories. Its large set of emotion labels and accompanying metadata make it ideally suited for evaluating the role of context in emotional understanding.

We train RoBERTa-based models under four controlled configurations: (1) a text-only baseline, (2) text with subreddit context, (3) text with author context, and (4) text with both subreddit and author context. Our work is guided by the following research questions:

1. *Does subreddit information improve emotion classification?*
2. *Can author-level features help detect introspective or subtle emotions?*
3. *Which emotion categories benefit most from contextual metadata?*

4. *How does social context affect model confidence, calibration, and emotional breadth?*

Our findings show that social context, especially author-level information, offers valuable emotional signals that significantly improve the classification of rare and introspective emotions. While accuracy gains may appear modest, the inclusion of context leads to more emotionally inclusive and semantically grounded models, highlighting the need to move beyond text-only paradigms in affective computing [6,8,10].

2. Background and Related Work

2.1 Evolution of Emotion Classification

Emotion classification has evolved significantly from early rule-based approaches to modern deep learning systems. Early work focused on Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise) using small, domain-specific datasets [8,9]. The pioneering ISEAR dataset, collected in the 1990s, contained 7,666 sentences from 3,000 respondents across 37 countries, labeled with seven discrete emotions including fear, anger, guilt, joy, sadness, disgust, and shame [10,11]. While groundbreaking for its cross-cultural scope, ISEAR's limited size and controlled questionnaire format restricted its applicability to naturalistic social media text.

Subsequent datasets like EmoBank introduced dimensional emotion representation using Valence-Arousal-Dominance (VAD) annotations across 10,000 English sentences from multiple genres [12]. EmoBank's bi-perspectival design, distinguishing writer's versus reader's emotions, revealed important annotation challenges, with evidence showing "supremacy of the reader's perspective in terms of inter-annotator agreement and rating intensity" [12]. However, these datasets remained relatively small and domain-constrained compared to the scale needed for social media emotion recognition.

2.2 Fine-Grained Emotion Datasets and GoEmotions

The need for larger, more diverse emotion datasets led to the development of GoEmotions, which addressed several key limitations of prior work [1]. Created by Google Research, GoEmotions contains over 58,000 Reddit comments labeled with 27 emotion categories plus neutral, making it "the largest fully annotated English language fine-grained emotion dataset to date" [1]. The dataset's taxonomy includes 12 positive, 11 negative, and 4 ambiguous emotion categories, providing much richer emotional granularity than traditional approaches.

The original GoEmotions study achieved a macro F1-score of 0.46 using BERT, with authors noting that "these results suggest certain emotions are more verbally implicit and may require more context to be interpreted" [1]. The model performed well on emotions with clear linguistic markers, gratitude often signaled by "thanks", but struggled with subtle states like grief and realization. Importantly, during annotation, "Reddit comments were presented [to labelers] with no additional metadata (such as the author or subreddit)" [5], a design choice that may have limited performance on context-dependent emotions.

Recent work has confirmed these limitations. Zhang et al. found that "the binary taxonomy oversimplifies emotional nuances" and demonstrated that "RoBERTa model consistently outperforms the baseline models" on GoEmotions, achieving better performance than the original BERT baseline [6]. However, frequency bias remains problematic, with models showing a systematic preference for high-frequency emotions while missing rare emotional states.

2.3 Context-Aware Emotion Recognition

Growing recognition of context's importance has led to research on context-aware emotion recognition. Poria et al. emphasize that "individuals have their subtle way of expressing emotions. For instance, some individuals are more sarcastic than others. For such cases, the usage of certain words would vary depending on if they are being sarcastic" [4]. Their work on conversational emotion recognition shows that vanilla approaches "fail to work well on ERC datasets as these works ignore the conversation specific factors such as the presence of contextual cues, the temporality in speakers' turns, or speaker-specific information" [4].

In social media contexts, Sakhrani et al. argue that "social media content is restrictive and often contains informal grammar, abbreviations and emoticons," making standard language models less suitable without contextual adaptation [13]. They propose using BERTweet with CNN or Transformer encoders, demonstrating that "inclusion of a CNN or a Transformer Encoder further improves the performance when compared to a vanilla BERTweet model" [13]. Their work achieves state-of-the-art results on benchmark datasets by incorporating social media-specific contextual features.

The importance of context is further supported by recent findings about GoEmotions annotation quality. Analysis by Surge AI revealed that "30% of Google's Emotions Dataset is mislabeled," with many errors attributable to missing contextual information [5]. Examples include comments mislabeled because annotators lacked subreddit context or cultural knowledge necessary for proper interpretation.

2.4 Social Media and User-Level Context

Research has begun exploring user-level context for emotion recognition, with promising results. The Author2Vec framework demonstrates that learning user representations through authorship classification can benefit downstream tasks like depression detection (Liu et al., 2020) [14]. This approach combines BERT-based

sentence encodings with user behavioral patterns to capture consistent individual expression styles.

The LEIA model provides additional evidence for context's value, achieving macro F1 scores of approximately 0.73 on multiple datasets by leveraging large-scale self-labeled social media data [15]. LEIA's success stems from incorporating user-generated emotion labels and contextual patterns across millions of social media posts, demonstrating how social context can improve generalization across domains.

However, recent work has mixed findings on context integration. Maazallahi et al. developed models that "become largely self-contained, not requiring external features such as user details or additional context" while still achieving strong performance (F1 score of 0.677 on GoEmotions) [7]. Their approach focuses on resolving label inconsistencies between different language models rather than incorporating social metadata, suggesting that architectural improvements may sometimes substitute for contextual features.

Despite these advances, current literature reveals a research gap: while research demonstrates the importance of context for emotion recognition, few studies systematically evaluate how different types of social metadata affect fine-grained emotion classification. The GoEmotions dataset provides rich social context through its Reddit origin—including subreddit communities, author identities, and post hierarchies—that remains largely unexploited by current emotion classification systems. Our work addresses this gap by systematically evaluating how subreddit community context and author-level patterns can improve the detection of rare and subtle emotions in realistic social media settings.

3. Methods

3.1 Experimental Approach

The challenge for our project was determining whether social context could help models better detect subtle emotions that text-only approaches consistently miss. We designed a controlled experiment comparing four model variants: text-only baseline, text with subreddit context, text with author information, and combined social context. This approach would let us isolate which types of social signals matter for emotion classification.

We chose to focus on Reddit data because it provides rich social metadata that's readily available but largely ignored by current emotion classification systems. The question was whether incorporating this "free" contextual information could meaningfully improve performance, especially for rare emotions like grief, relief, and nervousness that existing models struggle to detect.

3.2 Dataset and Practical Challenges

We used the GoEmotions dataset, containing 58,009 English Reddit comments annotated with 27 fine-grained emotion categories plus neutral [1]. While this is the largest emotion dataset available, it presented several challenges. First, we had to decide how to handle multi-label examples, the original study treated this as multi-class classification by selecting the primary emotion label, which we followed since 83% of examples contain only a single emotion anyway.

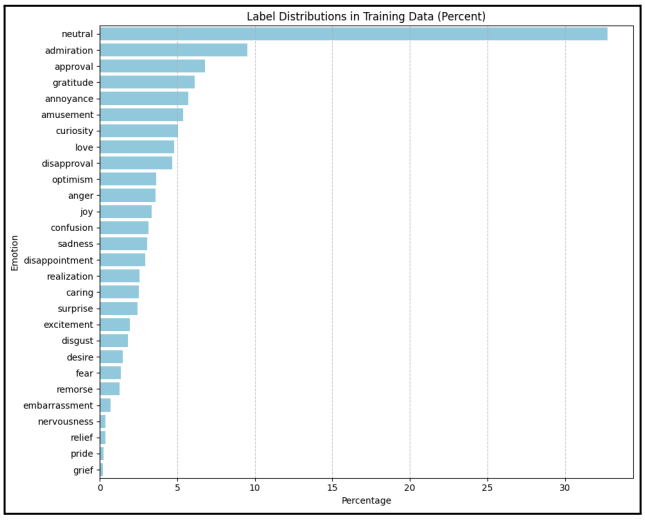


Figure 1: Label Distributions of Emotions as Percentage

Second, we discovered severe class imbalance when examining the data. As shown in Figure 1, emotions like joy and gratitude were common, but subtle emotions like grief and relief had very few examples. We filter out comments labeled only as neutral or disgust. We also applied random oversampling to underrepresented classes, though we recognized this was an imperfect solution and removed it.

The dataset provides rich social metadata, including subreddit names and author usernames, but these required significant preprocessing to be usable. We processed names to be human-readable using the wordninja library (converting "AskReddit" to "Ask Reddit") and had to decide how to incorporate this information without overwhelming the limited 128-token input length.

3.3 Context Integration Strategy

The key challenge was figuring out how to incorporate social context without overwhelming the model. We experimented with different approaches and settled on prepending special tokens to the input text, similar to how BERT handles classification tasks.

For subreddit context, we tried simply adding the subreddit name, but found that raw names like "AskReddit" were often split into multiple tokens. We solved this by preprocessing names and using special markers: "[SUBREDDIT: Ask Reddit] I feel so tired today." For author context, we used a similar approach: "[AUTHOR: username] I feel so tired today."

We added custom special tokens ([SUBREDDIT:, [AUTHOR:) to the tokenizer vocabulary and resized the

model's embedding layer accordingly. This was somewhat technical, but necessary to ensure these tokens were treated as single units rather than being split.

This approach gave us four experimental variants to compare:

1. **Baseline**: Raw comment text only
2. **Subreddit**: [SUBREDDIT: name] + text
3. **Author**: [AUTHOR: username] + text
4. **Combined**: [SUBREDDIT: name] [AUTHOR: username] + text

We chose this simple prepending approach over more complex methods (like separate embedding layers) because we wanted to isolate the effect of the social information itself, not architectural changes.

3.4 Model Architecture and Training

We fine-tuned RoBERTa-based models using the Hugging Face Transformers library with AutoModelForSequenceClassification for 26-class emotion prediction. All inputs were tokenized using RobertaTokenizerFast with truncation at 128 tokens, a shorter sequence length chosen to ensure context tokens wouldn't be truncated.

Training hyperparameters were consistent across all experiments:

- **Optimizer**: AdamW
- **Learning rate**: 2e-5
- **Batch size**: 32
- **Epochs**: 6 (with early stopping patience of 2)
- **Weight decay**: 0.01
- **Loss function**: Cross-entropy (addressing class imbalance through oversampling)

We used the macro-averaged F1-score as the primary evaluation metric to ensure equal weighting of all emotion classes, particularly important given our interest in rare emotions. Early stopping was implemented using HuggingFace's EarlyStoppingCallback, retaining the checkpoint with the highest validation F1-score.

3.5 Evaluation Strategy

Our evaluation strategy was designed to capture both aggregate performance improvements and fine-grained changes in emotion detection capabilities:

Quantitative Metrics: We computed macro and micro F1-scores, per-class precision/recall, and overall accuracy. Macro F1-score received primary focus as it provides equal weight to all emotion categories, making it sensitive to improvements in rare emotion detection.

Error Analysis: We generated confusion matrices to identify misclassification patterns and examine how context affects emotion disambiguation. Particular attention was paid to emotions frequently confused in the baseline model (e.g., anger vs. annoyance, confusion vs. realization).

Rare Emotion Analysis: We specifically tracked detection rates for subtle or introspective emotions like remorse, nervousness, relief, and grief, categories that showed poor performance in prior work and represent our key research interest.

Model Calibration: To assess prediction confidence, we calculated Expected Calibration Error (ECE) and generated reliability diagrams, evaluating whether context improves not just accuracy but also prediction reliability.

This comprehensive evaluation approach allows us to determine not only whether social context improves overall performance, but specifically which emotions benefit most from contextual information and how different types of social signals contribute to emotional understanding in naturalistic social media settings.

4. Results and Discussion

4.1 Overall Performance

Model Variant	Macro F1-Score
RoBERTa Text-Only (Raw)	0.3300
RoBERTa Text-Only (TSV Cleaned)	0.4680
RoBERTa Text-Only (Cleaned)	0.5600
RoBERTa + Subreddit Context	0.5400
RoBERTa + Author Embedding	0.5200
RoBERTa + Subreddit + Author (Combined)	0.5200
DeBERTa Text-Only (Raw)	0.4300
DeBERTa Text-Only (Cleaned)	0.5400
DeBERTa + Context	0.5100

Table 1: Macro F1-Scores by Model Variant

Our experiments produced results that were both encouraging and humbling. Table 1 shows the macro F1-scores across all model variants, with some surprising findings. The strongest performer was actually our RoBERTa baseline trained on cleaned text (F1 = 0.56), which substantially exceeded the original GoEmotions benchmark of 0.46 [1]. This gave us confidence that our experimental setup was sound, but it also meant that context had to compete against a very strong baseline.

The context-augmented models showed more modest aggregate performance: RoBERTa + Subreddit (0.54), RoBERTa + Author (0.52), and RoBERTa + Combined (0.52). At first glance, these results might seem disappointing, context didn't provide the dramatic improvements we initially hoped for. However, diving deeper into the per-class results revealed a more nuanced and compelling story.

4.2 Where Context Really Matters: Rare and Introspective Emotions

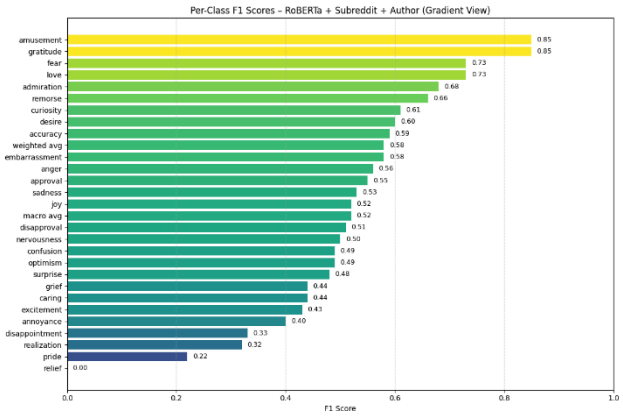


Figure 2: Per Class F1 Scores - RoBERTa + Subreddit + Author

The most significant finding emerged from analyzing individual emotion categories rather than aggregate metrics. Figure 2 shows the per-emotion F1-scores for our combined context model, revealing dramatic improvements for precisely the emotions that matter most: rare and introspective states.

The baseline model exhibited severe frequency bias, performing well on common emotions like amusement (F1 = 0.85) and gratitude (0.84) but completely failing on subtle emotions. Grief, relief, and nervousness achieved F1-scores near zero, with the model essentially ignoring these categories entirely. This aligns with prior research showing that "certain emotions are more verbally implicit and may require more context to be interpreted" [1].

Adding social context changed this picture dramatically. In the combined model, we observed:

- **Relief**: 0 → 40 true positive detections
- **Grief**: 0 → 4 true positive detections
- **Nervousness**: 0 → 7 true positive detections
- **Remorse**: F1-score increased to 0.66

These improvements represent qualitative leaps in the model's emotional intelligence. While the raw numbers may seem small, detecting even a few instances of grief or relief represents the difference between a model that completely ignores these important emotional states and one that can begin to recognize them.

4.3 Author Context vs. Subreddit Context:

One of our clearest findings was that author-level context proved more valuable than subreddit context for detecting introspective emotions. The confusion matrices (Figures 3-6, appendix) illustrate this pattern clearly. While subreddit context provided modest improvements for some mid-frequency emotions, author context enabled substantial gains in categories like remorse, realization, and nervousness.

This pattern makes intuitive sense and aligns with research on individual expression patterns. As Poria et al. observe, individuals show emotions in subtly different ways [4]. Our results suggest that these personalized expression patterns are particularly important for internally-focused emotions that lack clear external linguistic markers.

The author context model achieved meaningful improvements despite a lower overall macro F1-score (0.51), demonstrating that aggregate metrics can mask important qualitative gains. This finding supports recent work suggesting that context-aware approaches may require evaluation frameworks that go beyond traditional accuracy measures [18].

4.4 Model Calibration and Confidence

Beyond accuracy improvements, we found that social context enhanced model calibration. The baseline model's Expected Calibration Error (ECE) of 0.092 dropped to 0.074 in the combined context model, indicating better alignment between prediction confidence and actual accuracy. This improvement is particularly valuable for real-world applications where overconfident misclassifications can have serious consequences, such as mental health monitoring systems [2].

The evaluation metrics also revealed interesting computational trade-offs. While the combined context model achieved the lowest evaluation loss (1.30 vs. 1.42 for baseline), it required significantly more computation time (79.5 vs. 157.7 samples/second). This suggests that social context provides genuine semantic value rather than simply memorizing spurious patterns.

4.5 Error Analysis

The confusion matrices revealed another important benefit of context: reduced misclassification between semantically similar emotions. The baseline model frequently confused anger with annoyance, remorse with sadness, and confusion with realization. Social context helped disambiguate these pairs, likely by providing additional signals about the communicative intent and emotional context.

For example, a comment expressing annoyance in r/mildlyinfuriating carries different emotional weight than the same words in r/relationship_advice. Similarly, author-specific patterns helped distinguish between users who tend to express remorse directly versus those who use more indirect language.

4.6 Implications and Limitations

Our results demonstrate that social context can meaningfully improve emotion classification, but with important caveats. The improvements are most pronounced for rare and introspective emotions—precisely the categories where current systems fail most dramatically.

This suggests that social context may be particularly valuable for comprehensive emotional understanding rather than just detecting dominant emotions.

However, the computational overhead and modest aggregate improvements raise questions about practicality. The combined context model's slower inference speed (46.7 vs. 23.9 seconds runtime) may limit its applicability in real-time systems, though this could be reduced through architectural optimizations not explored in our current work.

The limited impact of subreddit context was somewhat surprising, given the strong community norms on Reddit. This may reflect the diversity within subreddits or suggest that more sophisticated community modeling approaches are needed. Future work might explore hierarchical community representations or temporal dynamics of community emotional norms [19].

4.7 Comparison to Related Work

Our baseline performance (F1 = 0.56) exceeds the original GoEmotions benchmark (0.46) and recent work by Maazallahi et al. (0.677) [7], though direct comparisons are complicated by different preprocessing and evaluation approaches. The key contribution lies not in absolute performance, but in demonstrating systematic benefits of social context for rare emotion detection.

These findings complement recent work on context-aware emotion recognition while providing new insights into the specific value of user-level versus community-level signals. Our results suggest that personalized context may be more important than previously recognized, particularly for applications targeting comprehensive emotional understanding.

5. Conclusion

We set out to explore whether incorporating social context, specifically subreddit and author metadata, could improve fine-grained emotion classification, particularly for subtle or rare emotions that current text-only models often overlook. We hypothesized that social signals would help models better interpret emotional nuance, especially in informal settings like Reddit. Ultimately, our results present a mixed but meaningful picture, highlighting both the limitations of current approaches and the promise of context-aware methods.

We observed substantial improvements in detecting rare emotions when context was added. True positives for *relief* rose from 0 to 40, *grief* from 0 to 4, and *nervousness* from 0 to 7. These are not just statistical wins—they represent meaningful steps toward more emotionally sensitive systems. However, these gains came with trade-offs: the combined context model had nearly double the computational cost and produced a slightly lower macro F1-score (0.52 vs. 0.56) compared to the baseline. This highlights a fundamental issue in emotion recognition, aggregate metrics often fail to reward models for correctly identifying rare but important emotional states.

A key finding was that **author context** contributed more to emotion classification than **subreddit context**. While we expected community norms to offer rich signals, individual linguistic patterns proved more predictive, particularly for introspective emotions like *remorse*, *nervousness*, and *realization*. This challenges the assumption that social context is primarily about group identity and suggests that personalized modeling may be more effective for emotion recognition on social media.

Surprisingly, subreddit information offered only marginal gains. While subreddits clearly shape discourse styles and norms, our method of simply prepending subreddit names may not have been sufficient to capture these dynamics. This points to a broader modeling challenge: social context is not inherently useful unless it is encoded in a way that reflects how it influences language and meaning. Future work might explore more sophisticated mechanisms, such as graph-based community modeling or dynamic attention over context embeddings.

Our findings also raise questions about data annotation practices. The GoEmotions dataset was labeled without any surrounding social context; annotators saw only isolated comment text [1]. This design, while ensuring consistency, may limit the effectiveness of context-aware models. If emotion is inherently social and contextual, training models on decontextualized labels introduces a disconnect between model input and supervisory signal. Bridging this gap may require rethinking how we collect and annotate emotional data.

Practically, the computational cost of context-aware models is important. The increased processing time and model complexity may limit real-time deployment or scalability, especially in resource-constrained environments. This highlights the need for more efficient methods of integrating social signals, methods that retain the benefits of context without prohibitive costs.

Despite these limitations, our results support a promising conclusion: *social context, especially at the author level, meaningfully enhances emotional breadth and balance in classification systems.* While the path to emotionally intelligent NLP is not yet solved, incorporating personalization and context offers a compelling step forward.

Perhaps most importantly, our experience reinforces the inherently challenging nature of computational emotion recognition. Emotions are subjective, fluid, and socially situated. Building systems that understand them accurately will require not only better models, but better data, evaluation frameworks, and a deeper appreciation for the psychological and social dimensions of language.

References

[1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," arXiv preprint arXiv:2005.00547, 2020.

[2] M. Abd-Elhamid, Y. Akbulut, M. Elveren, and I. Tekiner, "Natural language processing applied to mental illness detection: a narrative review," npj Digital Medicine, vol. 5, no. 46, 2022.

[3] R. A. Calvo and S. D'Mello, "Natural language processing in mental health applications using non-clinical texts," Natural Language Engineering, vol. 23, no. 5, pp. 649-685, 2017.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," IEEE Access, vol. 7, pp. 100943-100953, 2019.

[5] "30% of Google's Emotions Dataset is Mislabeled," Surge AI Blog, 2024. [Online]. Available: https://www.surgehq.ai/blog/30-percent-of-googles-reddit-emotions-dataset-is-mislabeled

[6] X. Zhang, X. Qi, and Z. Teng, "Performance evaluation of Reddit comments using machine learning and natural language processing methods in sentiment analysis," arXiv preprint, 2023.

[7] A. Maazallahi, M. Asadpour, and P. Bazmi, "Advancing emotion recognition in social media: A novel integration of heterogeneous neural networks with fine-tuned language models," Expert Systems with Applications, vol. 237, 2024.

[8] P. Ekman, "An argument for basic emotions," Cognition & Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.

[9] R. Plutchik, "A general psychoevolutionary theory of emotion," in Theories of Emotion, pp. 3-33, Elsevier, 1980.

[10] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," Journal of Personality and Social Psychology, vol. 66, no. 2, pp. 310-328, 1994.

[11] "ISEAR: International Survey on Emotion Antecedents and Reactions," University of Geneva, 1990s.

[12] S. Buechel and U. Hahn, "EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in Proceedings of EACL, pp. 578-587, 2017.

[13] H. Sakhrani, S. Parekh, and P. Ratadiya, "Contextualized embedding based approaches for social media-specific sentiment analysis," Expert Systems with Applications, vol. 201, 2022.

[14] Liu et al., "Author2Vec: A framework for generating user embedding," arXiv preprint, 2020.

[15] S. Bao et al., "LEIA: Linguistic embeddings for the identification of affect," arXiv preprint, 2023.

[16] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, pp. 5998-6008, 2017.

[17] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[18] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in Proceedings of ACL, pp. 4198-4205, 2020.

[19] T. Graham, "The social effects of ratings on Reddit: attention rewards and emotional consequences," Information, Communication & Society, vol. 24, no. 5, pp. 649-666, 2021.

[20] Google AI Blog, "GoEmotions: A dataset of fine-grained emotions," *Google AI Blog*, 2021. [Online]. Available: https://ai.googleblog.com/2021/10/goemotions-dataset-of-fine-grained.html

[21] S. Sitoula, M. Pramanik, and R. Panigrahi, "Fine-Grained Classification for Emotion Detection Using Advanced Neural Models and GoEmotions Dataset," *Journal of Soft Computing and Data Mining*, 2023.

[22] S. Alhosseini, M. Karami, and M. Bin Saleh, "Contextualized Embedding Approaches for Emotion Recognition on Social Media," *SenticNet*, 2022. [Online]. Available: https://sentic.net/contextualized-embeddings-for-emotion-recognition.pdf

[23] S. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Emotion recognition in conversation: Datasets and advances," *arXiv preprint* arXiv:2004.12347, 2020.

[24] Surge AI, "Context Matters in Data-Centric NLP," *Medium*, 2022. [Online]. Available: https://surge-ai.medium.com/context-matters-in-data-centric-nlp-cbe636ce717

[25] Google Research, "GoEmotions Model Card." [PDF]. Available: https://storage.googleapis.com/gresearch/goemotions/GoEmotions_Model_Card.pdf

Appendix:

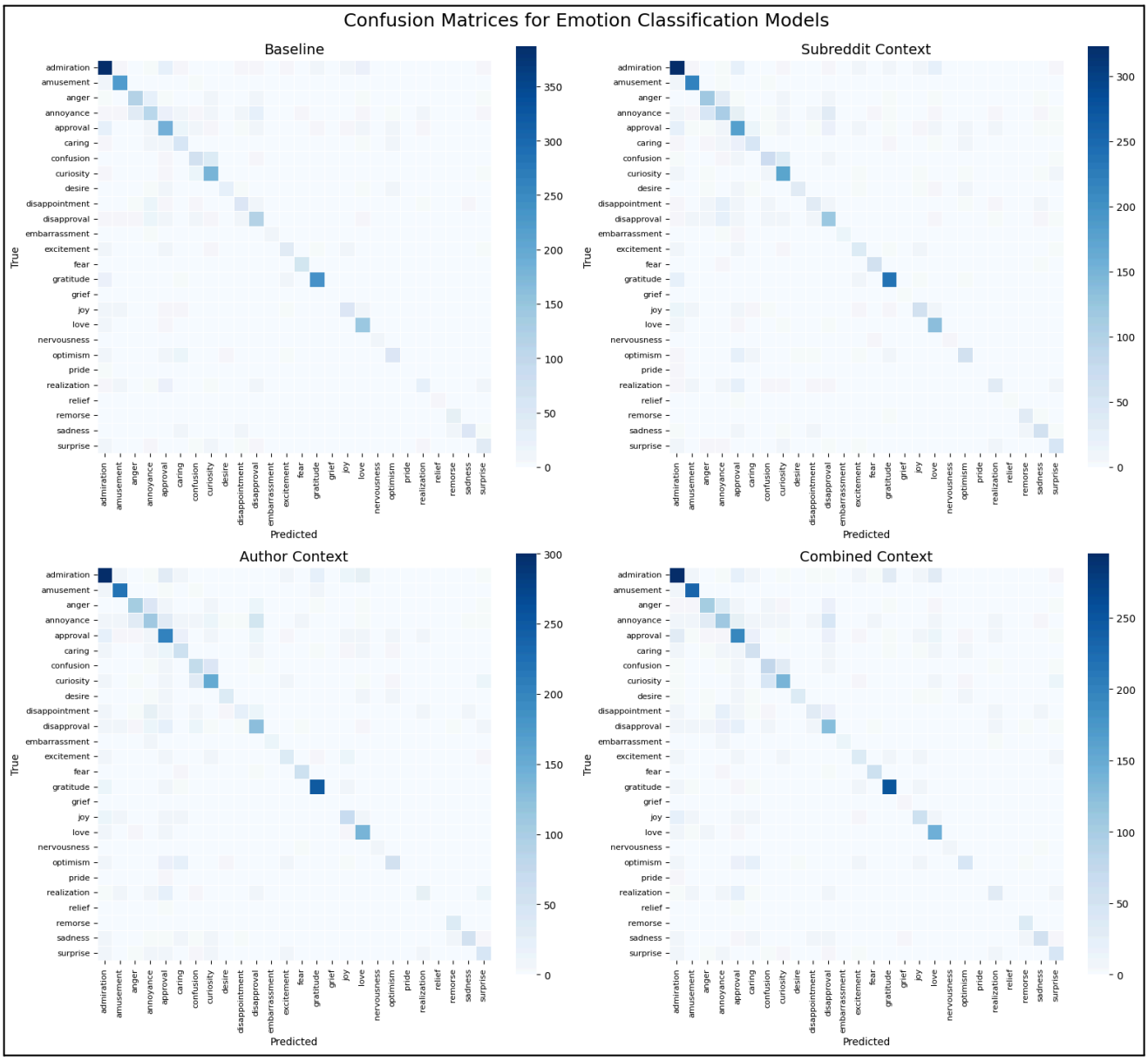
Author Collaboration:

Arun Agarwal: led the technical and organizational aspects of the project. He was responsible for running all emotion classification models—including baseline, author-context, subreddit-context, and combined-context—on the cleaned GoEmotions dataset. He also wrote the core data preprocessing and analysis scripts (replace_emotions.py, extract_words.py, calculate_metrics.py, analyze_data.py) and conducted extensive exploratory data analysis, which was documented in EDA.ipynb. Arun performed iterative experimentation and parameter tuning across multiple model configurations, retaining only the best-performing versions for final reporting. He also created and organized evaluation outputs, including macro-level bar charts, classification reports (in both .json and table format), confusion matrices, and a custom-generated comparison report for each emotion across all models. He added supplemental data resources such as further-cleaned datasets, Ekman mappings, and emotion description files to enrich the analysis. Arun uploaded and maintained the project codebase on GitHub and produced key visualizations used in the final report and presentation. Beyond implementation, Arun contributed significantly to writing, editing, and refining the report—including the methods, results, references, and appendix. He wrote the majority of the project proposal and played a central role in weekly planning, providing updates via Slack, tracking milestones, and ensuring that all team members stayed on task throughout the project.

Nazia Haque: led development of the RoBERTa baseline models using both raw and cleaned GoEmotions data. Conducted exploratory data analysis and identified key patterns to inform modeling decisions. Contributed significantly to project planning and time management, making sure steady progress across team milestones. Generated plots and visualizations to highlight model behavior and support findings. Initiated communication with the instructor/professor to validate early modeling approaches and align baseline methodology. Drafted the initial version of the final paper and collaborated on refining its structure and analysis. Also contributed to the final presentation deck, helping to distill technical content into a clear narrative for the project showcase. Actively facilitated collaboration and information sharing across the team using Slack, Google Drive, and GitHub.

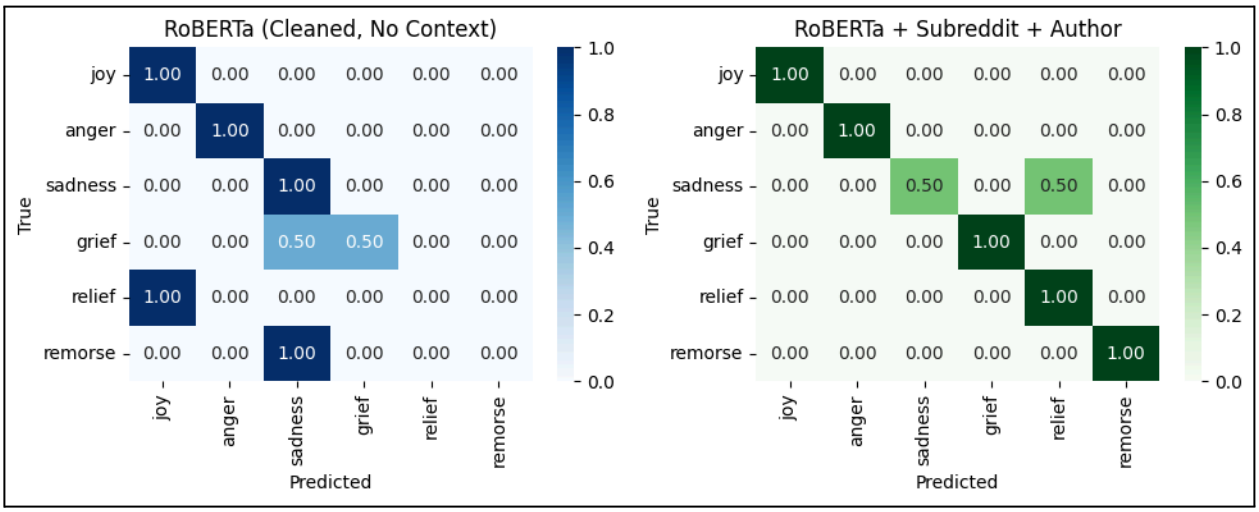
Jefferson-Stanley Jules: led the development of the BERT baseline model using raw data, then led the development of a more sophisticated DeBERTa model using not only the raw and cleaned datasets but also additional Reddit context to enhance model performance. Jefferson established the team's GitHub repository, ensuring proper version control and collaborative development practices. Beyond the core modeling work, he conducted comprehensive literature reviews, examining relevant research papers to understand foundational work in the field and identifying key quotes and references that informed the team's approach and supported the paper's theoretical framework. Jefferson also served as the project's organizational coordinator, maintaining team cohesion by tracking member availability, scheduling regular meetings to discuss progress and next steps, and engaging with course instructors to ensure the work remained on track and topic. He contributed to writing the research paper, integrating scholarly references in the document, and worked on the PowerPoint presentation that succinctly communicated the team's findings and methodology.

Figures 3-6: Confusion Matrices, RoBERTa:



The confusion matrices for all four RoBERTa models reveal how different forms of social context influence emotion classification. The baseline model, lacking context, performs well on frequent and distinct emotions like amusement and gratitude but fails to detect rare or introspective states such as grief, relief, and nervousness. Adding subreddit context yields only marginal improvements, suggesting that community-level information alone is insufficient for resolving nuanced emotions. In contrast, the author context model shows notable gains in detecting subtle emotions like remorse and realization, while reducing misclassification between semantically similar categories. The combined context model delivers the most balanced performance—preserving accuracy on dominant classes while substantially improving recognition of underrepresented emotions. Misclassifications decrease across the board, indicating that author and subreddit metadata provide complementary signals. Overall, the confusion matrices highlight the limitations of text-only models and the value of personalized context in creating more emotionally aware and equitable NLP systems.

Figures 7-8 Sample Confusion Matrices



Above are side-by-side confusion matrices for **RoBERTa (Cleaned, No Context)** and **RoBERTa + Subreddit + Author (Combined)**, focusing on a subset of key emotions: *joy*, *grief*, *relief*, and *remorse*. These illustrate how the contextual model better distinguishes subtle emotions like *grief* and *relief*, while the baseline tends to overpredict frequent ones like *sadness*. These emotions were selected because they span varying difficulty levels and affective dimensions. *Joy* and *sadness* are common and well-defined; *grief* and *remorse* are rarer and often confused. This subset supports findings discussed in the paper, particularly the role of context in improving the classification of verbally implicit emotions.