Project Report

Emotion Classification Assistant

Written by: Sehr Abrar

CSc 44800: Artificial Intelligence

Professor Erik Grimmelmann

CUNY: The City College of New York

December 18, 2025 - Fall 2025

Table of Contents

Introduction

In recent years, classifying emotions from text has become an important task in artificial intelligence (AI) and natural language processing (NLP). Understanding the emotions in user-generated text enables applications such as sentiment analysis, mental health support, customer feedback analysis, and human-computer interaction. By automatically detecting emotions, AI systems can respond more appropriately to users, improve engagement, and provide useful insights (Demszky et al., 2020).

This project focuses on building an Emotion Classification Assistant, a tool that predicts the dominant emotion expressed in short text comments. Unlike basic sentiment analysis that classifies text simply as positive, negative, or neutral, this project classifies text into multiple fine-grained emotions such as joy, anger, sadness, surprise, admiration, and others. The dataset used, GoEmotions, contains 58,000 English-language Reddit comments labeled with 27 distinct emotion categories plus neutral, making it one of the most comprehensive emotion-labeled corpora available for research and practical applications (Demszky et al., 2020).

The goal of this project is to compare different models and identify the one that performs best in predicting emotions. Three models were evaluated: (1) TF-IDF vectorization with Logistic Regression, (2) TF-IDF vectorization with Linear Support Vector Machine (SVM), and (3) GloVe word embeddings with a small feedforward neural network. Each model was assessed using accuracy, macro F1-score, and confusion matrices to determine how well it predicts each emotion and handles both frequent and rare emotion classes.

The project also includes a simple demonstration interface where users can input short text comments and see the predicted emotion in real time. This highlights practical challenges, such as ambiguous or rare emotional expressions, and allows a qualitative evaluation of the model outputs.

Emotion detection from text has evolved substantially in NLP. Traditional approaches rely on feature-extraction methods like TF-IDF (Term Frequency–Inverse Document Frequency), which convert words into weighted numerical vectors that reflect their informational value in a corpus. Combined with classical classifiers like logistic regression or SVMs, these techniques have been shown to provide reasonable performance in emotion and sentiment classification tasks (Manning et al., 2008). More recent approaches use dense word embeddings, such as GloVe, or contextual embeddings derived from transformer-based models to better capture semantic relationships and contextual nuance (Pennington et al., 2014). Hybrid methods combining embeddings with lexicon-based or heuristic features have also been explored to improve performance on informal, noisy text, as found in social media or online forums.

Given this landscape, this project leverages a large, fine-grained emotion dataset, evaluates multiple textual representations, and compares model architectures to assess trade-offs in accuracy, interpretability, and handling of rare emotions. The work aims not only to measure model performance but also to provide insight into how text representation and model choice affect emotion detection, providing a foundation for practical tools that understand human emotional expression in text.

Data Collection & Preprocessing

The dataset used for this project consists of 38,242 comments labeled with 27 different emotion classes. Comments were chosen to cover a wide spectrum of emotional expressions, including common emotions such as joy, love, and admiration, as well as less frequent emotions like nervousness, relief, pride, and grief. Exploratory analysis showed that the distribution of these emotions is highly imbalanced (Figure A1). On average, comments contain about 21 words, with most being short but a few longer ones (Figure A3). Exploratory analysis also revealed that common words across the dataset include informal expressions such as "im," "just," "like," "love," "dont," "good," "thats," "thanks," "really," and "people" (Figure A4). Additionally, certain words were strongly associated with specific emotions—for example, "happy," "glad," and "enjoy" for joy; "sad," "sorry," and "feel" for sadness; and "!," "hate," and "kull" for anger (Figure A2). Positive emotions tended to appear in slightly longer comments (Figure A6) and dominate the dataset overall (Figure A7), while negative emotions were more varied, with some correlations observed between related emotions, such as sadness and grief or anger and annoyance (Figure A5).

Before training models, the text was preprocessed to ensure it was suitable for machine learning. Preprocessing steps included cleaning the text by removing punctuation, URLs, and extra whitespace, and converting all text to lowercase. Each comment was then tokenized, splitting the text into individual words. The emotion labels were encoded into numeric format to allow models to process them effectively. Finally, the dataset was divided into training and test sets using an 80/20 split to evaluate model performance reliably.

Two main approaches were used to convert the cleaned text into numerical features. The first, Term Frequency–Inverse Document Frequency (TF-IDF), assigns weights to words based on their frequency in a comment relative to their rarity across the dataset. This method was applied for the Logistic Regression and Linear SVM models. The second approach utilized pre-trained GloVe embeddings, which represent each word as a 100-dimensional vector capturing semantic meaning. Each comment was represented as the average of its word vectors, which was then used as input for a small neural network model.

These preprocessing steps, combined with exploratory data analysis, ensured that the raw comments were transformed into formats suitable for machine learning while retaining the semantic and emotional content necessary for accurate emotion classification.

Model Development

The goal of this project was to predict the dominant emotion of each comment from 27 emotion classes. Three different modeling approaches were explored to compare traditional machine learning methods with a small neural network.

All models shared the same preprocessing pipeline, which included cleaning text, converting it to lowercase, removing punctuation and URLs, and splitting the dataset into training (80%) and test (20%) sets. The text was then converted into numerical features either via TF-IDF vectorization or pre-trained GloVe embeddings. Each model was trained to map these features to the corresponding emotion label, and performance was evaluated using accuracy, macro-F1, and confusion matrices.

a. Model A: TF-IDF + Logistic Regression

Model A predicts a comment's emotion by looking at the words it contains. It uses TF-IDF to turn words into numbers that reflect their importance and assigns weights to word combinations to determine which emotions they are most associated with. The steps were as follows:

1. Preprocessed text was vectorized using TF-IDF (uni- & bi-grams, top 10,000 features).
2. Trained a Logistic Regression model with balanced class weights.
3. Predicted the dominant emotion for each comment.

In our results, Model A achieved the highest performance among the three approaches. Accuracy was moderate across frequent emotions such as joy, love, and admiration, while rare emotions

like relief, pride, and grief were predicted less accurately. Overall, this model demonstrated a strong balance of simplicity and predictive capability.

b. Model B: TF-IDF + Linear SVM

Model B also uses TF-IDF to convert words into numerical features but employs a Linear Support Vector Machine (SVM) to find boundaries between emotion classes. The steps were as follows:

- Preprocessed text was vectorized using TF-IDF (top 10,000 features).
- Trained a Linear SVM with balanced class weights.
- Predicted the dominant emotion for each comment.

In our results, performance patterns were similar to Model A, with frequent emotions predicted well and rare emotions predicted poorly. Overall accuracy was slightly lower than Logistic Regression, and the macro-F1 score was reduced, indicating less consistent performance across all emotion classes.

c. Model C: GloVe Embeddings + Neural Network

Model C predicts emotions by representing each word with pre-trained GloVe embeddings and averaging them into a single vector for each comment. A small neural network then learns patterns in these vectors to classify emotions. The steps were as follows:

- Converted each comment into a 100-dimensional vector using pre-trained GloVe embeddings.
- One-hot encoded labels for multi-class classification.

- Trained a feedforward neural network with two hidden layers (64 → 32 neurons) and a softmax output layer.
- Predicted the dominant emotion for each comment.

In our results, the neural network performed poorly relative to the TF-IDF models. While it achieved moderate accuracy for common emotions such as admiration, love, and gratitude, it struggled with rare emotions like relief, pride, and grief. This highlights the challenge of using a small neural network on a relatively small dataset with many classes.

d. Model Selection

Based on the evaluation of all three approaches, TF-IDF combined with Logistic Regression (Model A) was selected as the best-performing model. This decision was supported by its superior accuracy and macro-F1 score, as well as its consistent handling of both frequent and rare emotions. Additionally, a comparison of the normalized confusion matrices (Appendix B) showed that Model A produced the most balanced predictions across emotion categories, with fewer severe misclassifications than the other models. While the Linear SVM model (Model B) displayed similar trends, its overall performance was slightly lower and its confusion matrix indicated more confusion between related emotions. The GloVe-based neural network (Model C) struggled significantly with rare emotions and showed lower accuracy overall, likely due to the modest dataset size and the large number of emotion classes. Overall, Model A offered the most reliable, interpretable, and stable performance for this text-based emotion classification task.

Results & Discussion

The AI Wellness Assistant, powered by the selected TF-IDF + Logistic Regression model (Model A), was tested with a variety of brief user inputs. The model successfully predicted emotions such as joy, love, sadness, surprise, curiosity, pride, disappointment, and confusion. More frequent and explicit emotions, such as joy or love, were reliably detected, while less common or ambiguous emotions occasionally led to repeated or uncertain predictions. Example interactions demonstrating these outcomes are shown in Appendix C.

During testing, the model demonstrated a tendency to overpredict ambiguous emotions like confusion or surprise when input text was short or lacked context. This highlights a limitation of bag-of-words-based models: without deeper semantic understanding, subtle distinctions between similar emotions can be difficult to capture. Nevertheless, the assistant consistently provided reasonable predictions for most inputs, supporting its use as a basic real-time emotion classification tool.

The results also emphasize the trade-offs between simplicity and performance. While the neural network with GloVe embeddings (Model C) showed lower accuracy overall, it offers a foundation that could be improved with larger datasets or more sophisticated architectures. Overall, the demo confirms that the selected model provides a strong baseline for interactive emotion prediction and serves as a practical starting point for future enhancements.

Conclusion & Future Improvements

The project successfully developed an AI Wellness Assistant capable of predicting the dominant emotion in user comments. Among the models evaluated, TF-IDF combined with Logistic Regression (Model A) provided the best balance of accuracy and consistency, particularly for frequent emotions such as joy, love, and admiration. While the model performed less reliably on rare or ambiguous emotions, it demonstrated the practical viability of text-based emotion classification for real-time interaction. The demo confirmed that even a relatively simple machine learning model can provide meaningful predictions and support interactive applications.

Future improvements could focus on enhancing the model's understanding of context and subtle emotional cues. Incorporating transformer-based models such as BERT or fine-tuned pre-trained embeddings could improve performance on rare emotions and ambiguous inputs. Additionally, expanding the dataset with more balanced examples of all 27 emotion classes would help reduce bias toward frequent emotions. Finally, integrating multi-modal inputs, such as voice tone or facial expression analysis, could allow the assistant to make more accurate and nuanced emotion predictions in real-world applications.

Citations

Chollet, F. (2017). Deep learning with Python. Manning Publications.

Demszky, D., Movshovitz-Attias, D., Ko, J., Ravi, S., Fung, P., & Chang, M. (2020). GoEmotions: A dataset of fine-grained emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.372

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Proceedings of the European Conference on Machine Learning. https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

Ma, F., Cai, Y., & Gao, J. (2020). Regularised text logistic regression: Key word detection and sentiment classification for online reviews. arXiv. https://arxiv.org/abs/2009.04591

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. https://nlp.stanford.edu/IR-book/

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). https://nlp.stanford.edu/projects/glove/

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv. https://arxiv.org/abs/1609.04747

Appendix A: Exploratory Data Analysis

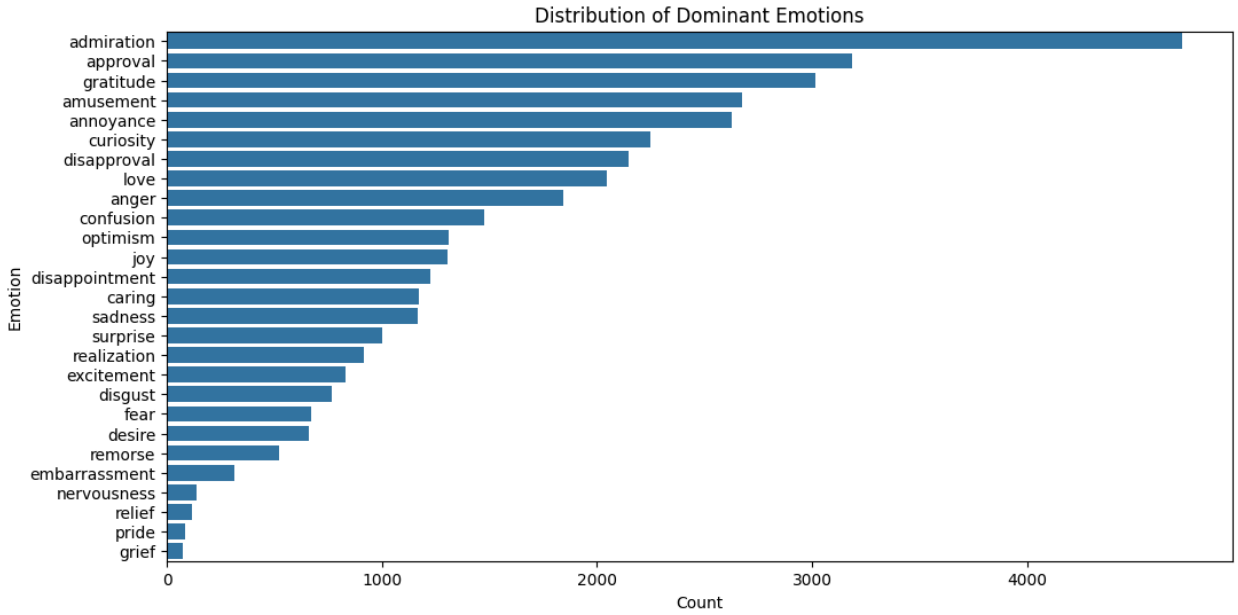

Figure A1. Distribution of dominant emotions in the dataset. Most comments are labeled with frequent emotions like joy, love, and admiration, while rare emotions such as grief, pride, and nervousness appear less frequently.

Examples of 'joy':

	clean_text
26283	i gave the show a chance because i enjoyed bad...
56990	so glad i moved here.
6851	dude, you made my day.

Examples of 'sadness':

	clean_text
21953	it also makes me cry! but because it shows how...
14623	due to work, got to stay legal. edibles are no...
10120	if i died i have always wondered who will be a...

Examples of 'anger':

	clean_text
10703	good for you for shutting that door! you are n...
16355	so now disabled people do not have rights . yo...
7957	i will kill you if you do that again, honey

Figure A2. Sample comments from several emotion classes (joy, sadness, anger, fear) showing typical language patterns associated with each label.

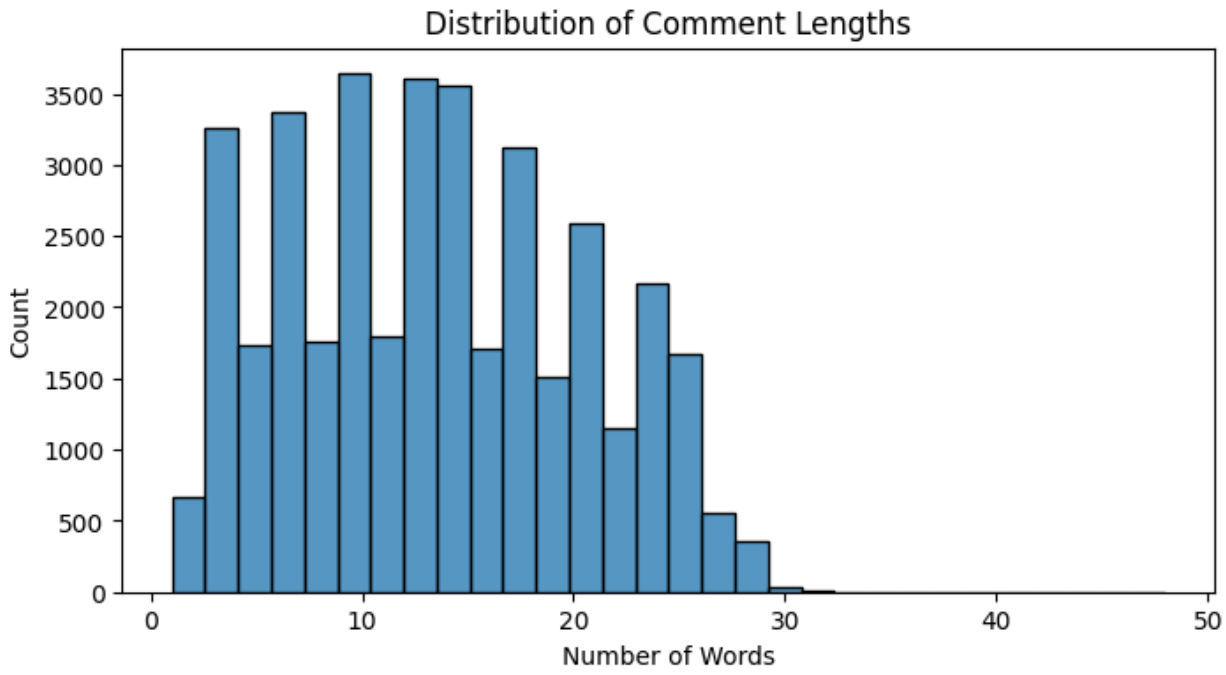

Figure A3. Distribution of comment lengths in the dataset, showing the number of words per comment and the average comment length.

```
[('user', 4599),
 ('just', 2998),
 ('like', 2844),
 ('love', 2193),
 ('num', 2130),
 ('good', 1657),
 ('did', 1524),
 ('really', 1414),
 ('thank', 1331),
 ('people', 1318),
 ('know', 1222),
 ('think', 1201),
 ('thanks', 1094),
 ('lol', 1019),
 ('it.', 968),
 ('going', 941),
 ('got', 928),
 ('does', 922),
 ('hope', 899),
 ('want', 778)]
```

Figure A4. Twenty most common non-stop words in the dataset, showing the frequently used words across all comments.

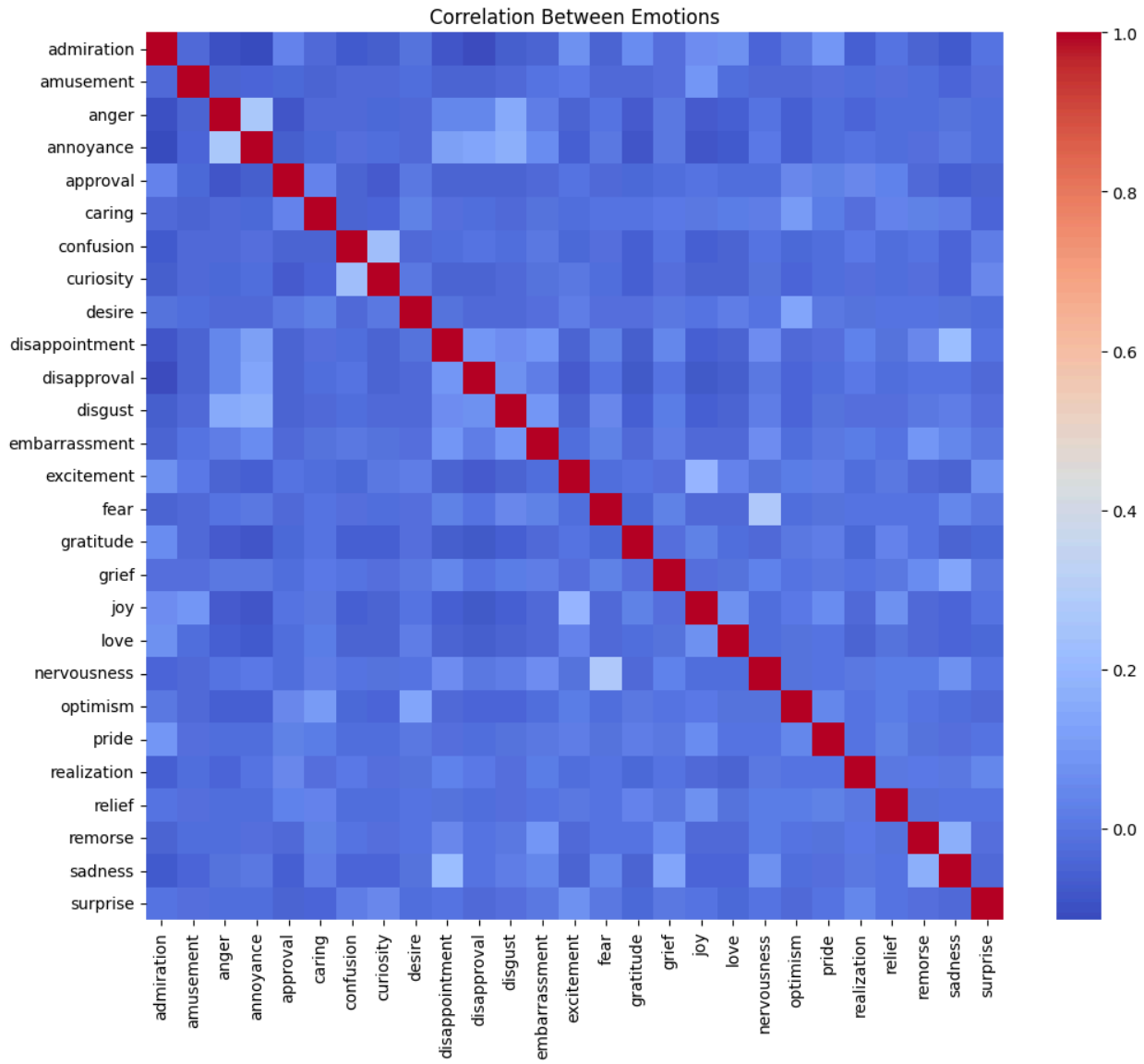

Figure A5. Heatmap showing correlations between different emotions, indicating which emotions tend to co-occur in comments.

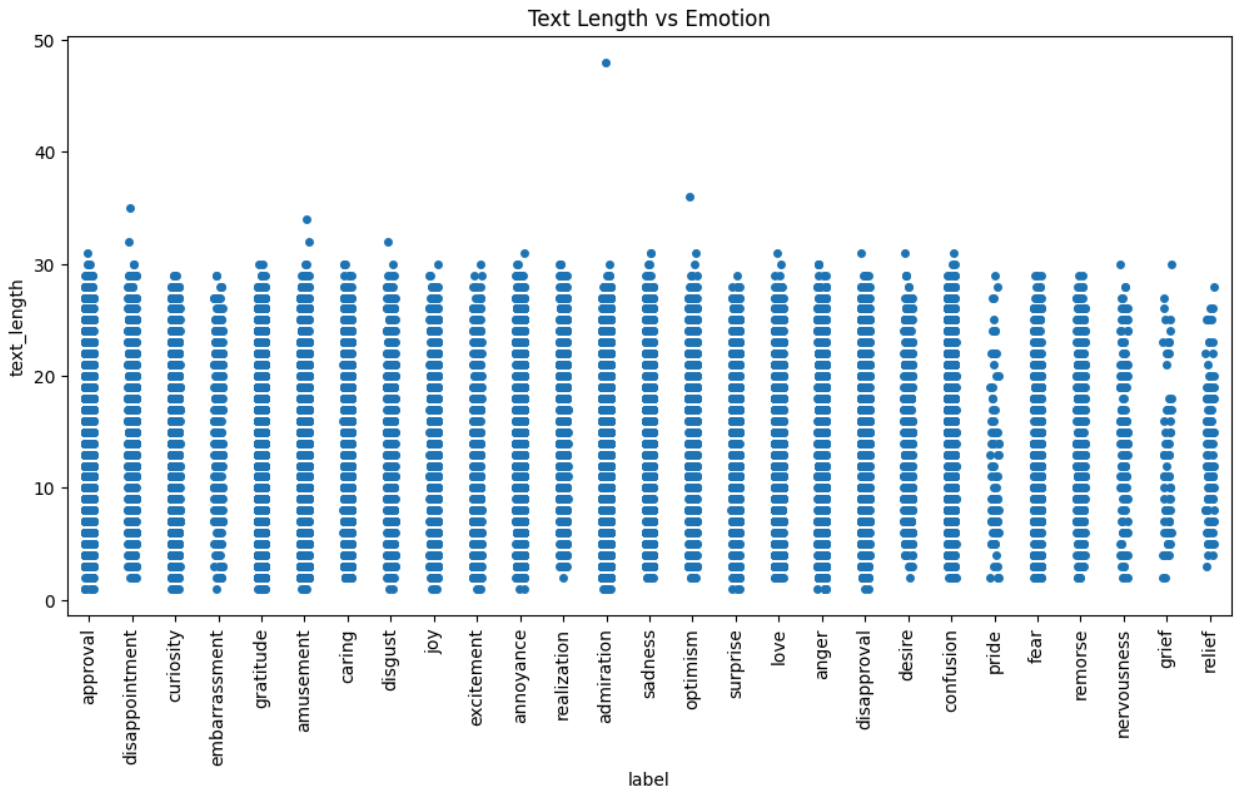

Figure A6. Strip plot showing the distribution of comment lengths for each emotion, highlighting trends in text length across emotions.

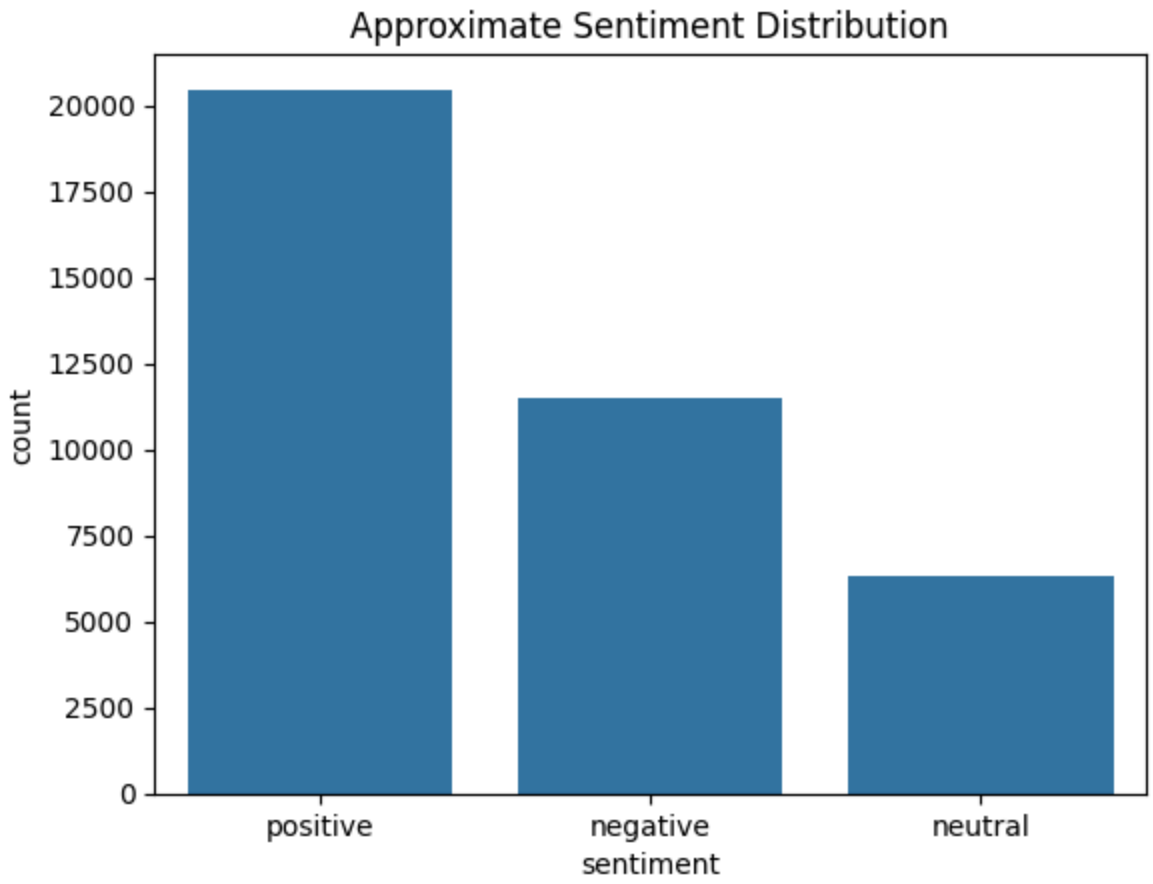

Figure A7. Count plot showing the distribution of approximate sentiment categories (positive, negative, neutral) across all comments.

Appendix B: Confusion Matrices

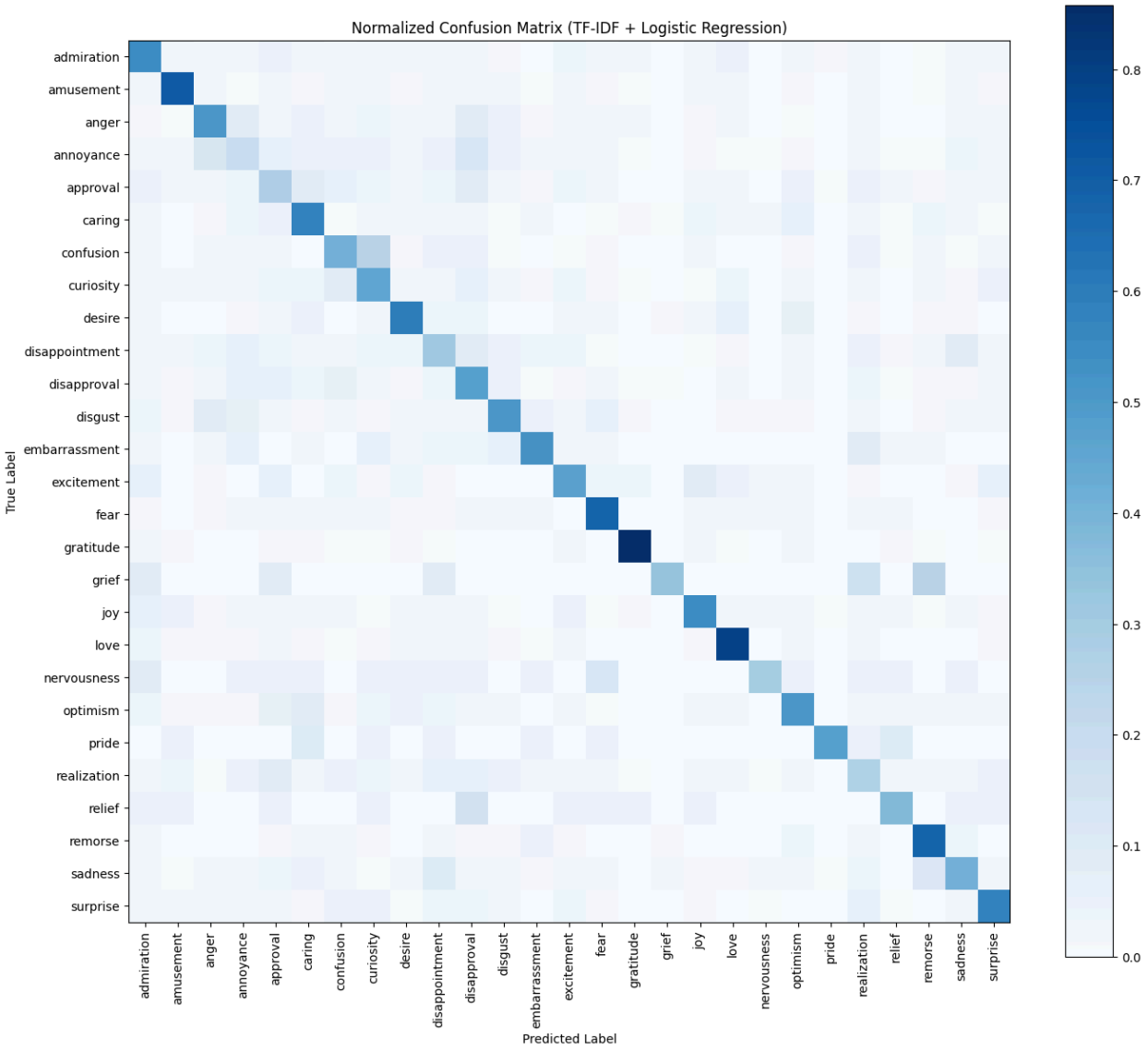

Figure B1. Normalized confusion matrix for Model A (TF-IDF + Logistic Regression), showing the model's performance across all 27 emotion classes.

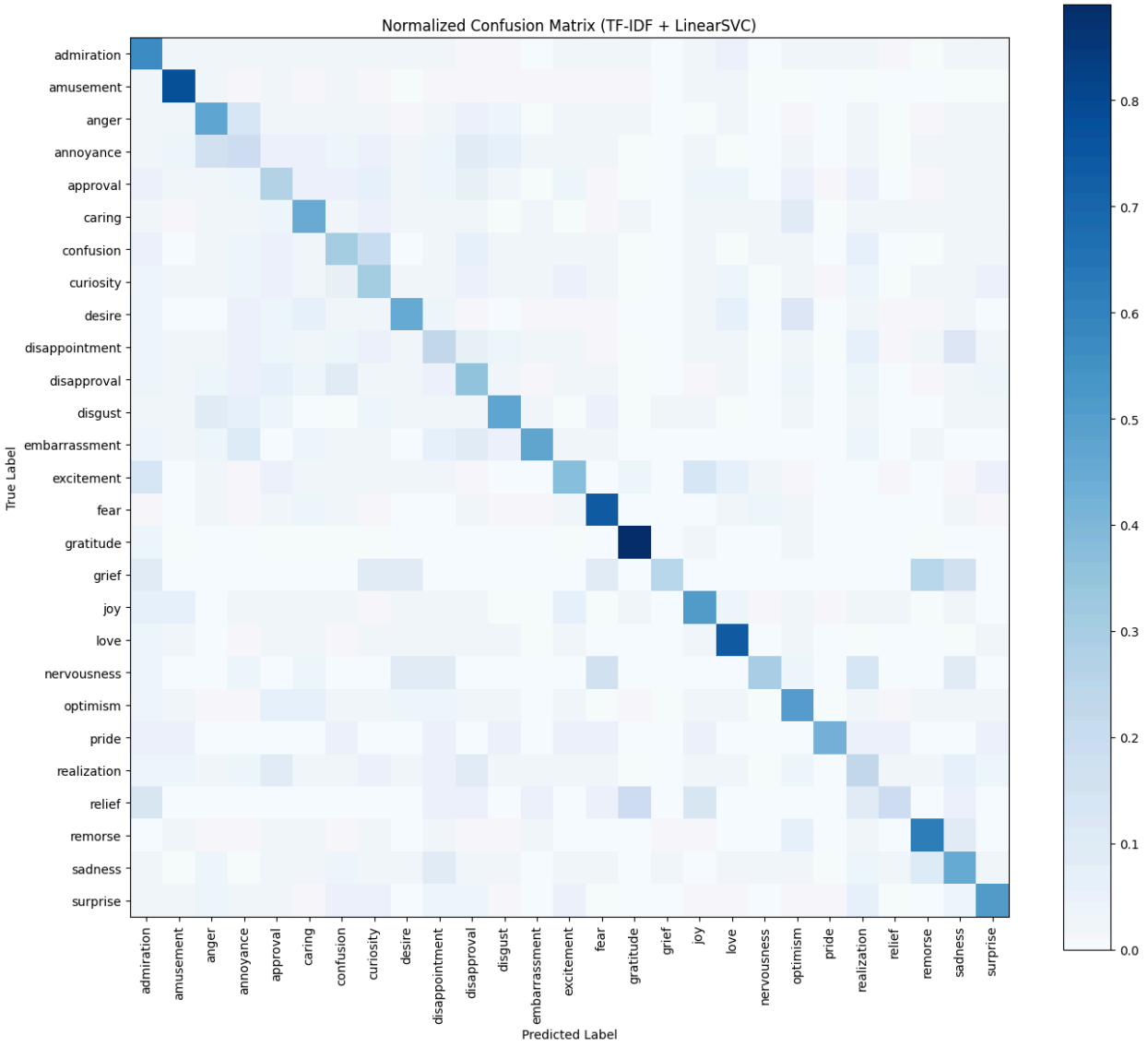

Figure B2. Normalized confusion matrix for Model B (TF-IDF + Linear SVM), illustrating prediction accuracy across all 27 emotion classes.

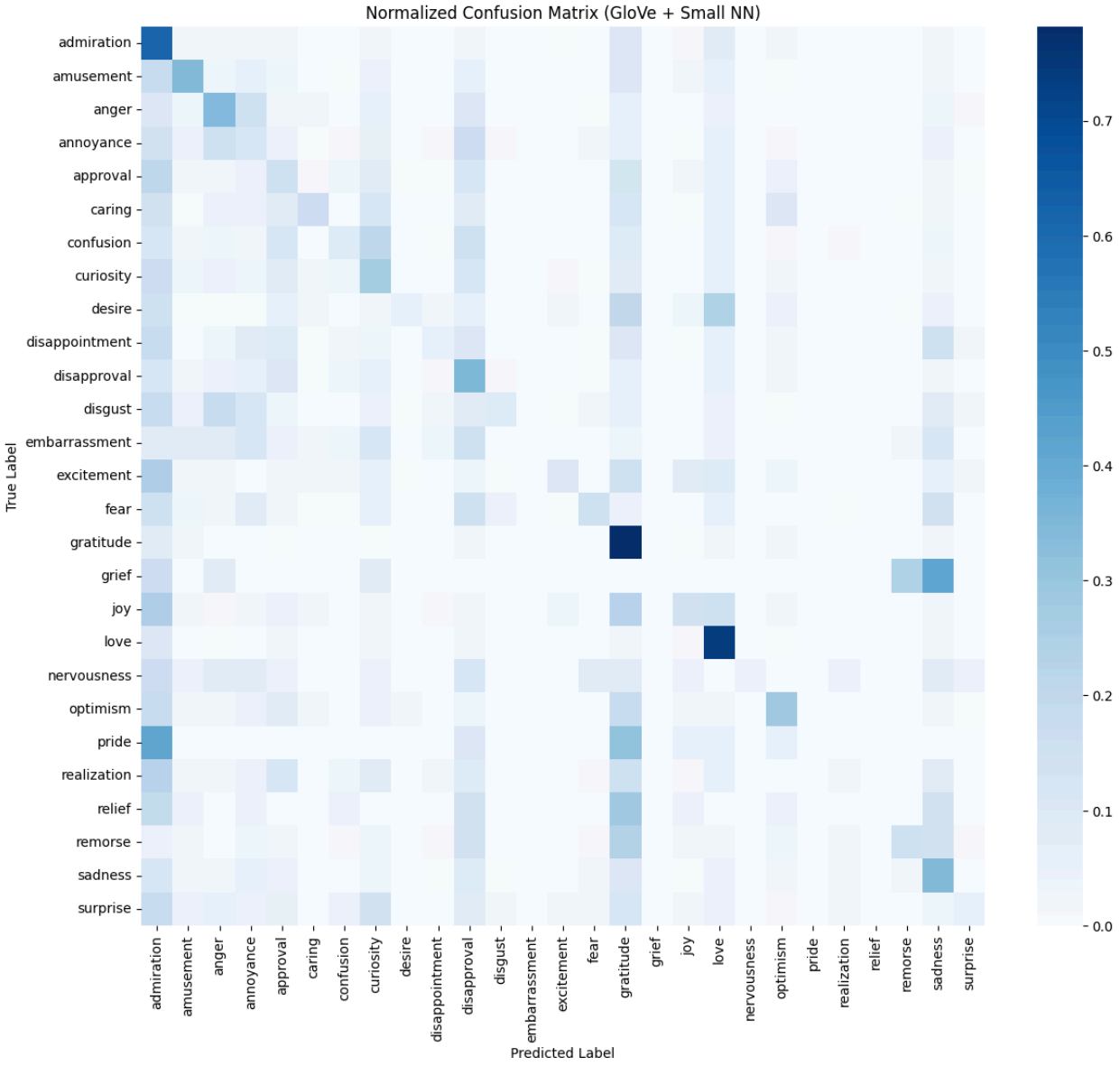

Figure B3. Normalized confusion matrix for Model C (GloVe Embeddings + Small Neural Network), showing prediction performance across all 27 emotion classes.

Appendix C: Sample Input & Output

```
Welcome to your AI Wellness Assistant! Type 'quit' to exit.
[Emotion: joy] AI: I'm sensing you feel joy. How can I help?
[Emotion: nervousness] AI: I'm sensing you feel nervousness. How can I help?
[Emotion: love] AI: I'm sensing you feel love. How can I help?
[Emotion: curiosity] AI: I'm sensing you feel curiosity. How can I help?
[Emotion: pride] AI: I'm sensing you feel pride. How can I help?
[Emotion: sadness] AI: I'm sensing you feel sadness. How can I help?
[Emotion: surprise] AI: I'm sensing you feel surprise. How can I help?
[Emotion: confusion] AI: I'm sensing you feel confusion. How can I help?
[Emotion: surprise] AI: I'm sensing you feel surprise. How can I help?
[Emotion: confusion] AI: I'm sensing you feel confusion. How can I help?
[Emotion: confusion] AI: I'm sensing you feel confusion. How can I help?
[Emotion: joy] AI: I'm sensing you feel joy. How can I help?
[Emotion: sadness] AI: I'm sensing you feel sadness. How can I help?
[Emotion: disappointment] AI: I'm sensing you feel disappointment. How can I help?
[Emotion: realization] AI: I'm sensing you feel realization. How can I help?
[Emotion: desire] AI: I'm sensing you feel desire. How can I help?
[Emotion: annoyance] AI: I'm sensing you feel annoyance. How can I help?
[Emotion: confusion] AI: I'm sensing you feel confusion. How can I help?
You: QUIT
Goodbye! Take care. 😊
```

Figure C1. Example interaction with the AI Wellness Assistant, showing predicted emotions and corresponding responses for various user inputs.