Emotion Analysis in Journal Entries : Classification and Regression

INF385T- INTRO TO MACHINE LEARNING (27844)
Taught by Professor Jyothi Vinjumur
Jiwon Park, Hitanshi Dhaktode

***Abstract*—This project explores emotion analysis in journal entries, predicting six emotions (joy, sadness, anger, fear, disgust, and surprise) on a 1–5 scale. Two approaches were tested: BERT for classification and RoBERTa for regression. While effective in identifying dominant emotions, the models struggled with capturing nuanced interplay of multiple emotions and lacked accuracy in predicting negative emotions due to data imbalance. These challenges highlight the need for improved methods and richer datasets for fine-grained emotion prediction.**

Keywords—journal entries, bert, emotion prediction, sentiment classification, regression for emotions

1. Introduction

Emotion analysis has become a critical area of research, with applications spanning various fields such as mental health, user experience design, and personalized recommendations. In the domain of mental health, emotion analysis can assist individuals in recognizing and recording their emotions through journaling, fostering self-awareness, and emotional regulation. Recent studies in this area have focused on sentiment detection and fine-grained emotion classification, aiming to provide actionable insights for analytical purposes. This project explores emotion analysis in journal entries by predicting six core emotions—joy, sadness, anger, fear, disgust, and surprise—on an integer scale from 1 to 5. Predicting emotion intensity is particularly valuable as it enables personalized analysis, such as tracking emotional trends over time or identifying triggers that provoke specific emotional responses. Such insights can empower individuals to better understand their emotional patterns and take proactive measures to enhance their well-being.

A. Problem Definition

This project aims to predict and analyze emotion intensity in journal entries. Specifically, the problem involves:
1. Predicting six core emotions—joy, sadness, anger, fear, disgust, and surprise—that appear in diary entries.
2. Quantifying the intensity of each emotion on a scale from 1 (very weak) to 5 (very strong).

The objective is to assess the accuracy of two distinct modeling approaches for emotion intensity prediction: classification using BERT (Devlin et al., 2019) and regression using RoBERTa (Liu et al., 2019). By comparing these approaches, the study seeks to evaluate the strengths and weaknesses of classification versus regression models in the context of emotion prediction. This research aims to provide insights into how emotion intensity can be effectively predicted and applied to support emotional well-being.

2. Related Work

Emotion analysis, as discussed by Cambria et al. (2013), has been extensively studied using traditional and deep learning-based approaches. Early methods, such as those based on the NRC Emotion Lexicon (Mohammad & Turney, 2013), relied on word-level associations with emotions, which provided foundational insights but struggled with context-dependent expressions. Recent advancements have introduced models like EmoBERTa, which fine-tunes RoBERTa specifically for emotion classification, achieving state-of-the-art performance by better understanding nuanced emotional contexts.

In addition to text-based models, Google's GoEmotions (Demszky et al., 2020) dataset has significantly expanded the scope of emotion classification, offering 27 fine-grained emotion categories derived from over 58,000 Reddit comments. Its use has become a standard benchmark for evaluating the robustness of emotion analysis models.

A notable trend in recent research is the shift toward multimodal emotion analysis, combining text with visual and audio data to capture emotions more comprehensively. This approach acknowledges the limitations of text-only systems in fully representing human emotional expressions, paving the way for more holistic solutions.

This study focuses on comparing the performance of BERT and RoBERTa in predicting six core emotions from psychology—**joy**, **sadness**, **anger**, **fear**, **disgust**, and **surprise**. These models are evaluated to understand their effectiveness in emotion intensity prediction and their potential applications in personalized emotion tracking systems.

3. Data Description

The dataset used in this study was derived from multiple sources and meticulously processed for emotion intensity prediction. A total of 1,472 diary entries were used for fine-tuning, excluding emotion-labeled data from Kaggle's "Journal Entries with Labeled Emotions" (Malhotra, 2022). Each diary entry, consisting of 250 to 400 characters, was labeled with six core emotions—joy, sadness, anger, fear, disgust, and surprise—on a scale from 1 to 5. Here, 1 indicates "very weak," 2 "weak," 3 "moderate," 4 "strong," and 5 "very strong." The labeling process utilized OpenAI's GPT-4o model for initial scoring, followed by human review and adjustments to ensure label accuracy and consistency.

A. Class Distribution

The class distribution for the six emotions is as follows :

Emotion	1	2	3	4	5
Joy	17	9	15	72	35
Sadness	100	36	8	4	0
Anger	132	10	3	3	0
Disgust	128	15	4	1	0
Surprise	79	48	20	1	0

4. Methodology

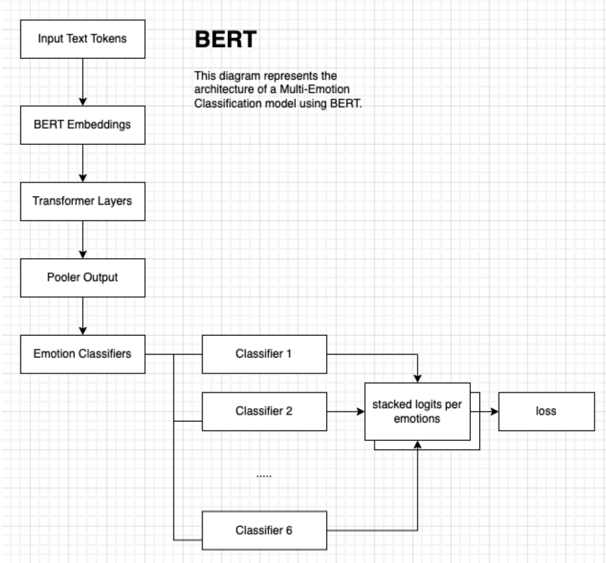
The BERT-based classification model was fine-tuned to predict emotion intensities, leveraging its architecture for multi-class classification tasks. In parallel, the RoBERTa-based regression model was employed to predict the intensity of emotions on a continuous scale, which was then converted into an integer scale. The following methodology outlines the steps taken for each approach:

A. BERT

A1. Preprocessing

Diary entries were tokenized using BertTokenizer with truncation and padding to a maximum sequence length of 128 tokens. Labels were mapped to a range of 0 to 4 for compatibility with the CrossEntropyLoss function.

A2. Model Architecture



The base model used in this study was **bert-base-uncased**, enhanced with additional layers to classify emotion intensities. To address imbalances in emotion intensity distributions, a custom model was developed incorporating class weights. These weights were manually adjusted to prioritize higher intensity levels, which were underrepresented in the dataset.

A3. Training Configuration

Learning Rate	1e-5
Batch size	16
Epochs	20
Optimizer	AdamW with weight decay
Evaluation Strategy	Per epoch
Early Stopping	After e epochs without improvement

A4. Evaluation Metrics

MAE, MSE, RMSE, and Pearson correlations were computed for each emotion to provide detailed insights into the model's performance.

Metric	Purpose
Mean Absolute Error (MAE)	Measures regression accuracy by averaging absolute differences between predicted and true values.
Mean Squared Error (MSE)	Similar to MAE but penalizes larger errors more heavily due to its quadratic nature.
Root Mean Squared Error (RMSE)	Penalizes larger errors and provides an interpretable scale for error magnitude.

Pearson Correlation Coefficient	Assesses the strength and direction of the linear relationship between predicted and true values.

Instead of using accuracy, Mean Absolute Error (MAE) was selected as the evaluation metric due to its ability to quantify the magnitude of error between predicted and true values, as discussed by Willmott & Matsuura (2005). Specifically, MAE considers the difference in severity, such as a prediction of 2 when the true value is 1 being less severe than a prediction of 5 when the true value is 1, thereby providing a more nuanced assessment of prediction quality.

B. RoBERTa

B1. Preprocessing

Preprocessing the data was a critical step in preparing the input for the RoBERTa model. Each diary entry was tokenized using the RoBERTa tokenizer, which converted the text into token IDs, added special tokens (such as [CLS] and [SEP]), truncated longer entries to a maximum length of 512 tokens, and padded shorter entries to ensure consistent input sizes. The labels representing emotion intensities were normalized to a 0–1 range to align with the regression training objective. These normalized labels were then converted into PyTorch tensors for compatibility with the Hugging Face training pipeline. The dataset was split into training, validation, and testing subsets in an 80:10:10 ratio, ensuring a balanced distribution for evaluation purposes. This preprocessing pipeline ensured that the textual and numerical data were in a format optimized for the RoBERTa model, enabling efficient training and prediction while maintaining the integrity of the original emotion labels.

B2. Model Architecture

The RoBERTa-based model architecture leveraged a pre-trained RoBERTa-base transformer for multi-label regression, fine-tuned to predict the intensity of six emotions—joy, sadness, anger, fear, disgust, and surprise. The model's architecture included 12 transformer layers, each capable of capturing contextual relationships within the text. The tokenizer's outputs (input_ids and attention_mask) served as inputs to the RoBERTa model, which generated high-dimensional contextual embeddings for each token in the text. These embeddings were passed through a regression head specifically added for this task, which included a fully connected layer to map the embeddings to six continuous outputs corresponding to the normalized emotion intensities. The Mean Squared Error (MSE) loss function was employed during training to minimize the difference between predicted and true emotion scores. This architecture effectively utilized RoBERTa's contextual understanding of language to capture subtle emotional cues embedded in diary entries.

B3. Training Configuration

Learning Rate	2e-5
Batch size	8
Epochs	Up to 20 (Best at 3 with early stopping)
Optimizer	AdamW with weight decay
Evaluation Strategy	Per epoch
Early Stopping	After 3 epochs without improvement

B4. Evaluation Metrics

MAE, MSE, RMSE, and Pearson correlations were computed for each emotion to provide detailed insights into the model's performance.

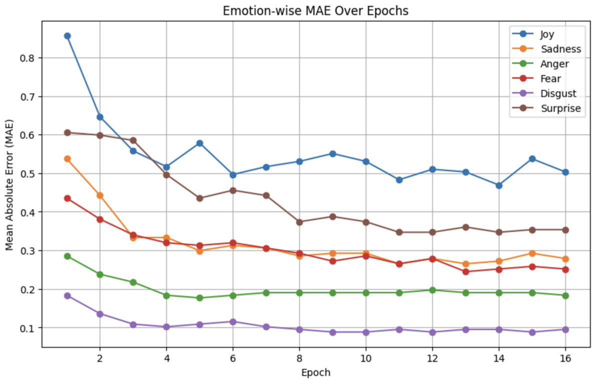
Metric	Purpose
Mean Absolute Error (MAE)	Measures regression accuracy by averaging absolute differences between predicted and true values.
Mean Squared Error (MSE)	Similar to MAE but penalizes larger errors more heavily due to its quadratic nature.

5. Results

A. BERT

A1. Fine-tuning Result

Early stopping occurred at Epoch 15. Overall, MAE decreased consistently during initial epochs, showing improved performance. Disgust and Anger had the lowest MAE, likely due to their imbalance in the dataset, where levels 4 and 5 were scarce. This suggests the model may achieve low error by predicting these emotions as mostly absent. Conversely, Joy and Surprise had higher MAE, indicating the model struggles with these emotions and requires further optimization.



MAE for each emotion at Epoch 15 is as follows:

joy	sadness	anger	fear	disgust	surprise
0.497	0.299	0.184	0.272	0.102	0.374

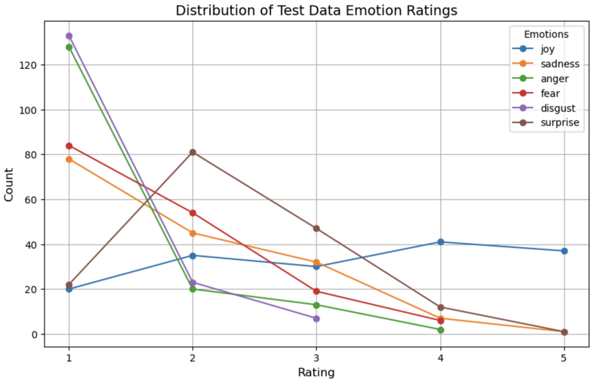
A2. Test Result

The following are the results of testing conducted on 164 test data samples.

MAE for each emotion with 164 test dataset:

joy	sadness	anger	fear	disgust	surprise
0.577	0.387	0.209	0.497	0.184	0.681

When analyzing the data distribution of the test dataset, it was observed that the dataset, which was collected arbitrarily, lacked a significant number of cases with negative emotions.



A3. Test Data Modified to Reflect Negative Tone

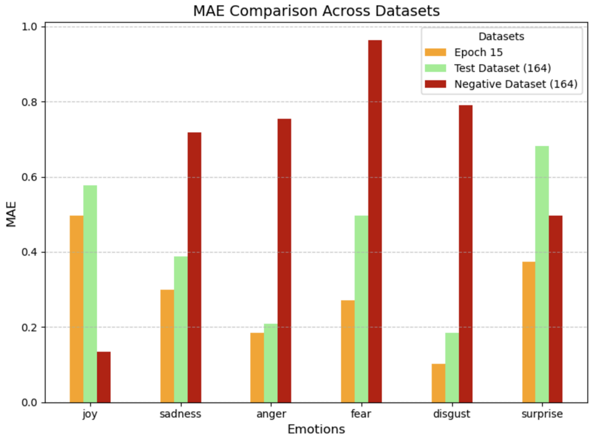
To experiment with how the model responds to negative data, 164 entries were converted to a negative tone using GPT-4, and the answer labels were reassigned. The results were then compared with those of the BERT classification model.

MAE for each emotion with 164 negative dataset:

joy	sadness	anger	fear	disgust	surprise
0.135	0.718	0.755	0.963	0.791	0.497

A4. Summary

Positive data, particularly emotions related to joy, are evaluated with high accuracy, indicating the model's strong performance in predicting positive emotions. However, when the data is converted to a negative tone, the MAE for emotions such as Fear (0.963), Anger (0.755), Sadness (0.718), and Disgust (0.791) increases significantly. This demonstrates that the model struggles to accurately predict these emotions in a negative context, highlighting its limitations in handling negatively toned data.



B. RoBERTa

B1. MAE and MSE for Each Emotion

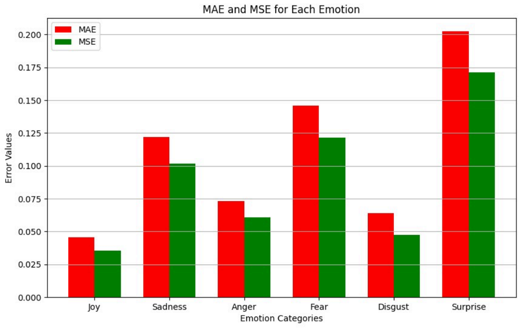


fig. 5.B.1(a)

The bar chart above (fig. 5.B.1(a)) depicting the Mean Absolute Error (MAE) and Mean Squared Error (MSE) across emotions reveals distinct trends in the model's predictive performance. The model performed best for “Disgust” and “Anger,” achieving the lowest MAE values of 0.0753 and 0.0793, respectively. These results suggest that these emotions are easier to predict due to their more explicit linguistic patterns in diary entries. However, “Surprise” and “Sadness” exhibited the highest MAE values of 0.1728 and 0.1235, respectively, indicating the model’s struggle to capture the subtlety and variability of these emotions. The higher MSE values for “Surprise” further emphasize that the model penalizes larger prediction errors more significantly for this emotion.

B2. F1-Score and Accuracy Over Epochs

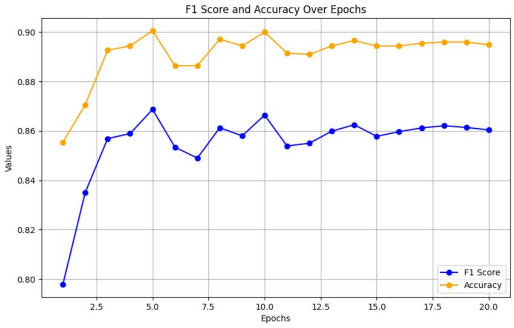


fig 5.B.2(a)

The line graph above (fig 5.B.2(a)) tracking F1-score and accuracy over epochs provides insights into the model’s training dynamics. Accuracy steadily increased, stabilizing at around **90%** by the fifth epoch, after which it plateaued. The F1-score demonstrated a similar trend, peaking at **86.8%** around the same epoch. This suggests that the model reached optimal performance early in training due to effective parameter tuning and the application of early stopping. The consistency of these metrics in later epochs indicates that the model avoided overfitting, validating the robustness of the training process.

B3. Test Results

The following are the MAE results of testing conducted on 164 test data samples for RoBERT:

joy	sadness	anger	fear	disgust	surprise
1.0859	1.7362	0.6871	1.6871	0.6810	1.6442

The Mean Absolute Error (MAE) values highlight the model’s predictive performance across six emotions. The lowest MAE values were observed for **anger (0.6871)** and **disgust (0.6810)**, indicating strong performance likely due to distinct linguistic markers associated with these emotions. In contrast, higher MAE values for **sadness (1.7362)** and **surprise (1.6442)** reflect challenges in accurately predicting these emotions, likely due to their subtle and context-dependent nature. Moderate performance was observed for **joy (1.0859)** and **fear (1.6871)**, suggesting room for improvement in capturing the nuanced expressions of these emotions.

C. BERT and RoBERTa Comparison

The following diary data was analyzed using four different models, and their respective results were compared.

Diary Entry: I felt like I was soaring, bursting with joy so intense it almost hurt—but then, out of nowhere, it was ripped away, leaving me drowning in a pit of despair so dark I couldn’t breathe. It’s like my heart is being torn in two—one half screaming with happiness, the other crumbling under the weight of unbearable pain.

	BERT	RoBERT	Human	GPT-4
Joy	1	3	5	5
Sadness	3	5	5	5
Anger	2	4	2	3
Fear	2	5	4	4
Disgust	2	3	1	2
Surprise	1	4	4	4

The comparison highlights distinct differences in emotion prediction across the models. While GPT-4 and human annotators showed strong agreement in identifying the intensity of emotions like joy and sadness (both rated 5), BERT significantly underestimated the intensity of joy (1) and failed to fully capture the depth of sadness (3). RoBERTa, although closer to human evaluations than BERT, tended to overestimate emotions like fear (5 compared to 4 by humans) and surprise (4 compared to 1 by BERT and 4 by humans). Notably, BERT struggled with more intense or conflicting emotions, as seen in its low ratings for both joy and sadness in a single entry, whereas RoBERTa better captured the nuances of mixed emotions but still deviated in certain cases, such as disgust. These results underscore the limitations of current models in handling complex emotional interplay, with GPT-4 demonstrating the most human-like judgment, though not flawless. Fine-tuning with more nuanced data may help bridge this gap.

6. Conclusion

This study explored emotion analysis in journal entries using BERT for classification and RoBERTa for regression. Both models demonstrated their potential in predicting emotion intensities, with RoBERTa showing a slight edge in handling nuanced emotional data. However, the challenges posed by data imbalance and the complexity of mixed emotions highlight the need for more robust and context-aware models. Future work should focus on augmenting datasets and integrating multimodal data to enhance the accuracy and applicability of emotion prediction models in real-world scenarios.

7. Limitations and Future Work

A. BERT

While BERT demonstrated strong performance in identifying dominant emotions, it struggled with negative emotions such as sadness, anger, and fear. This limitation is likely due to the data imbalance, where positive emotions were more prevalent. To address this, future work will involve augmenting the dataset with additional entries emphasizing negative emotional tones, ensuring a more balanced representation.

Another challenge observed was the model's inability to accurately identify complex and conflicting emotions. For example, entries containing equally strong opposing emotions (e.g., simultaneous joy and despair) often resulted in misclassification. Even human annotators faced difficulties in these cases, and the model's performance aligned more closely with GPT's outputs, which also showed limitations. This suggests that both humans and advanced models struggle with the intricacies of nuanced, mixed-emotion data.

Future efforts should explore techniques to enhance the model's capacity to discern these subtleties. Incorporating methods such as integrating multimodal data (e.g., text, visuals, or audio) could provide a more holistic approach to understanding human emotions. Expanding the study to include more diverse datasets, capturing a broader spectrum of emotional expressions, and fine-tuning with specifically curated mixed-emotion data will also be key priorities.

B. RoBERTa

While the RoBERTa-based emotion prediction model demonstrated promising results, there are several areas for improvement that could enhance its accuracy and generalizability. One of the primary challenges was the imbalanced dataset, where certain emotions, such as “Surprise,” were underrepresented. Addressing this issue through data augmentation or oversampling techniques like SMOTE could help the model better learn underrepresented emotional patterns. Additionally, expanding the dataset with more diverse diary entries would improve the model’s ability to generalize to unseen data.

Another area of improvement lies in exploring alternative loss functions tailored to the ordinal nature of emotion intensities, such as ordinal regression loss, which could help the model better align its predictions with the scale-based labels. Fine-tuning the hyperparameters further, including batch size, learning rate, and regularization strategies, could also optimize the training process and improve results.

From a model architecture perspective, experimenting with larger or domain-specific transformer models, such as RoBERTa-large or models pre-trained on emotional text datasets, may enhance performance. Incorporating ensemble techniques that combine predictions from multiple models could also provide more robust and accurate results.

Lastly, deploying the model in real-world applications such as sentiment analysis tools, mental health monitoring platforms, or personalized recommendation systems could provide valuable insights into its practical utility. Future work should focus on these aspects to refine the model and expand its applicability in real-world scenarios.

8. References

1. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019)**:
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

2. **Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019)**:
RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

3. **Willmott, C. J., & Matsuura, K. (2005)**:
Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance. Climate Research, 30, 79–82. https://doi.org/10.3354/cr030079

4. **Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013)**:
New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems, 28(2), 15–21. https://doi.org/10.1109/MIS.2013.30

5. **Alm, C. O., Roth, D., & Sproat, R. (2005)**:
Emotions from Text: Machine Learning for Text-based Emotion Prediction. Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), 579–586. https://doi.org/10.3115/1220575.1220650

6. **Malhotra, M. (2022)**:
Kaggle: Journal Entries with Labelled Emotions. https://www.kaggle.com/datasets/madhavmalhotra/journal-entries-with-labelled-emotions

7. **Mohammad, S. M., & Turney, P. D. (2013):**
Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence, 29*(3), 436–465.
https://doi.org/10.1111/j.1467-8640.2012.00460.x

8. **Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020):**
GoEmotions: A Dataset of Fine-Grained Emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4040–4054.
https://doi.org/10.18653/v1/2020.acl-main.372