Deciphering Dreams: Sentiment Analysis of Dream Journals

Stanford CS224N Custom Project

Gabriel M SantaCruz
Department of Computer Science
Stanford University
gsantac@stanford.edu

Giselle Isabella Rivera
Department of Computer Science
Stanford University
giseller@stanford.edu

Abstract

Analyzing the emotional content of dreams is crucial for understanding subconscious emotions, however, dream analysis remains challenging due to the subjective and nuanced nature of emotional expression in dream narratives. Traditional approaches, such as the Hall and Van de Castle (HVDC) framework and sentiment analysis tools like VADER heavily rely on human annotation and rule-based systems, which are known to be time-consuming, inconsistent, and limited in capturing complex emotional states. As a result, there is a growing need for automated methods that can accurately interpret the emotional complexity of dream reports while reducing the dependence on manual analysis, which could provide valuable insights for psychological research, mental health diagnostics, and therapeutic interventions.
This project explores the use of large language models (LLMs) to automate sentiment analysis of dream reports, leveraging the Sleep and Dream Database. Our goal is to assess whether LLMs, specifically RoBERTa, can accurately clasify emotions in dream narratives and identify recurring emotional themes. We fine-tune RoBERTa with Binary Cross-Entropy Loss and evaluate its performance using F1-score, Precision, Recall, and AUROC. We compare our model's performance against the Hall and Van de Castle (HVDC) framework and sentiment analysis tool, VADER. Our model aims to identify recurring emotional patterns and improve classification accuracy over existing methods. This work highlights the potential of LLMs to enhance dream analysis, providing a scalable and more consistent approach to understanding emotional themes in dreams.

1 Key Information to include

- Mentor: Zhengxuan (Zen) Wu
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

Dreams offer a window into the subconscious, providing insight into emotional and psychological states that are often difficult to access through conscious reflection. Analyzing the emotional content of dreams is crucial for understanding underlying emotional patterns, mental health conditions, and psychological well-being. However, dream analysis remains challenging due to the subjective and nuanced nature of emotional expression in dream narratives. Dreams are often metaphorical and fragmented, making it difficult to interpret their emotional content using traditional methods.

Existing approaches to dream analysis primarily rely on human annotation and rule-based frameworks such as the Hall and Van de Castle (HVDC) system, which categorizes dream content into specific

emotional and thematic elements. While these methods have established a foundation for structured dream analysis, they are limited by their reliance on manual effort, inconsistency among human annotators, and difficulty in capturing complex or ambiguous emotional states. Sentiment analysis tools like VADER provide a more automated approach but are constrained by their reliance on surface-level text patterns and lexicon-based rules, which often fail to capture deeper emotional nuances in dream narratives.

Recent advancements in natural language processing (NLP) and large language models (LLMs) have introduced new opportunities for automating dream analysis. LLMs have demonstrated remarkable performance in understanding contextual and emotional content in text, making them well-suited for analyzing the complex, metaphorical language typical of dream reports. In particular, Bertolini et al. (2023) explored the use of LLMs for automated dream sentiment analysis, showing promising results in scoring the emotional content of dream narratives. Inspired by this work, our project investigates whether fine-tuning RoBERTa, a transformer-based LLM, can improve emotional classification in dream reports.

Our project leverages the Sleep and Dream Database, a large collection of dream reports with expert-annotated emotional themes. We frame the task as a multi-label classification problem, where each dream report can express multiple emotions simultaneously (e.g., happiness, fear, both sadness and anger). We fine-tune RoBERTa using Binary Cross-Entropy Loss and evaluate its performance using F1-score, Precision, Recall, and AUROC. We benchmark our model's performance against both the HVDC framework and VADER to assess improvements in accuracy and emotional pattern recognition. By automating emotional classification of dream reports, this work aims to enhance psychological and therapeutic research by providing a scalable and consistent tool for understanding emotional themes in dreams. This approach reduces the dependence on manual analysis, improves consistency in emotional interpretation, and opens new possibilities for large-scale dream analysis in mental health and psychological research

3 Related Work

The analysis of dream content has long been a subject of interest in psychological research, with early studies like those by Hall and Van de Castle (1966) developing coding systems to categorize dream elements. Their work laid the foundation for quantitative dream analysis, but the manual nature of their methods posed challenges in terms of scalability and consistency.

In recent years, advancements in natural language processing (NLP) have enabled more sophisticated analyses of textual data, including dream reports. Bertolini et al. (2023) leveraged large language models (LLMs) to automate the scoring of dream reports' emotional content. Their study demonstrated that LLMs could effectively capture the nuances of emotional expression in dreams, offering a scalable alternative to manual annotation.

Another relevant study by Oberlander and Nowson (2006) explored emotion detection in blog posts using machine learning techniques. While not focused on dream analysis, their work highlighted the potential of computational models to discern emotional states from text, underscoring the applicability of such methods to various domains involving subjective narratives. Their exploration of emotional language in personal writing aligns with our project's goal of identifying emotional themes in dream reports, as both involve interpreting complex, subjective emotional content in natural language.

Despite these advancements, challenges remain in accurately interpreting the complex and subjective nature of emotional content in dreams. Our work aims to build upon these studies by fine-tuning RoBERTa, a transformer-based LLM, for multi-label emotion classification specific to dream narratives. By comparing our model's performance against traditional frameworks like HVDC and tools like VADER, we seek to assess the efficacy of LLMs in capturing the intricate emotional landscapes of dreams, thereby contributing to more nuanced psychological and therapeutic insights.

4 Approach

Firstly, to ensure the text is ready for input into the model, we begin with data preprocessing. This includes normalizing the text by converting all text to lowercase, removing non-alphanumeric characters, and eliminating newline characters. Then, we tokenize the text using the RoBERTa

tokenizer, which splits the text into subword tokens and maps them to unique integer identifiers. The tokenized sequences are padded or truncated to a fixed length of 128 tokens, and attention masks are computed to differentiate between real tokens and padding. Additionally, labels representing different emotional states are assigned to each journal entry, with each emotion or combination of emotions mapped to a unique integer.

For our model architecture, we utilize RoBERTa, which uses self-attention mechanisms to process sequences of text. The model consists of several encoder layers that generate contextualized token embeddings, which are then passed through a classification head to predict the sentiment of the input sequence. This classsification head consists of a linear layer followed by a softmax activation, which outputs the probability distribution over the possible classes. We use cross-entropy loss to train the model, as it is well-suited for multi-class classification tasks. The model is fine-tuned on the preprocessed dream journal data to adjust its weights for the specific task of emotion classification.

For hyperparameter tuning, we conducted a grid search over various hyperparameters. These include the learning rate (with values of 1e-6, 2e-6, and 3e-6), batch size (8 and 16), number of training epochs (1, 3, and 5), and weight decay (0.0 and 0.01). We select the best-performing hyperparameters based on the evaluation results to maximize model performance and prevent overfitting.

For evaluation, we use several metrics, including the F1 score, precision, recall, and confusion matrix. The F1 score is particularly important as it balances precision and recall, giving us a single metric to evaluate model performance. We also visualize the confusion matrix to understand how well the model is distinguishing between different classes. We compare the performance of our RoBERTa-based model to a baseline model, which uses the VADER sentiment analysis tool, a rule-based system known for its simplicity and efficiency in classifying text into positive, negative, or neutral sentiments.

Finally, we fine-tune the RoBETa model on a GPU to speed up the training process, applying gradient accumulation to simulate larger batch sizes without exceeding memory limits. The model is evaluated on a test set after training, and the best-performing model is selected based on its results on the validation set. Our approach combines the power of state-of-the-art natural language processing techniques with rigorous model evaluation, aiming to provide meaningful insights into the emotional content of dream journals.

5 Experiments

5.1 Data

We are using the Sleep and Dreams Database (SDDb) as the primary dataset for our project. The SDDb is a publicly available repository that collects dream reports from various sources, including research studies and individual submissions, providing a rich resource for the analysis of dream content and emotional themes. The dataset includes thousands of dream reports, which are typically unstructured and presented in free-text format. Each dream report in the SDDb may vary in length and complexity, ranging from a few sentences to several paragraphs. The dataset also includes metadata such as the participant's age, gender, and the context of the dream, though these details are not directly used for classification in this project.

We extracted and preprocessed the dream reports to convert them into a format suitable for sentiment analysis. Each dream report serves as the input, while the output is a single-label emotional classification. We defined fifteen emotion categories based on established psychological frameworks for emotional analysis: happiness, sadness, anger, wonder, fear, and each combination of two (e..g anger and happiness, happiness and fear). Each dream report is assigned a single label representing the dominant emotional tone of the dream. This follows a multi-class classification setup, where each dream report belongs to exactly one of the fifteen categories.

The preprocessing step involved cleaning the text to remove irrelevant symbols, normalizing the text, and tokenizing it using the tokenizer provided by the 'transformers' library for the RoBERTa model. Specifically, the input text was converted into token IDs and attention masks to match the input format expected by the RoBERTa architecture. We split the processed dataset into training and evaluation sets, with approximately 70% of the data allocated for training and 30% for evaluation. The tokenized inputs were stored in a structured format with three key components per sample:

- `input_ids` - a list of token IDs representing the tokenized dream text.

- `attention_mask` – a list indicating which tokens are actual input versus padding tokens.
- `label` – an integer value representing one of the five emotion categories.

We balanced the dataset to reduce class imbalance issues and improve the model's ability to generalize across different emotional categories. Additionally, we ensured that the distribution of labels was consistent between the training and evaluation sets. This balanced setup helps avoid the model being biased toward more frequent emotional classes. The final dataset includes thousands of labeled dream reports, making it large enough for training deep learning models effectively while also being manageable within the computational constraints of our training environment.

By defining this task as a multi-class sentiment classification problem based on emotional content, we aim to provide insights into the emotional landscape of dreams and explore the potential of RoBERTa-based models in the domain of psychological and emotional analysis.

5.2 Evaluation method

To assess the performance of our model, we employ multiple evaluation metrics that provide insights into classification effectiveness. These metrics include **F1-score, Precision, Recall, and the Confusion Matrix**. Each metric captures different aspects of model performance, ensuring a comprehensive evaluation.

Since our task may involve class imbalance, we report the **macro-averaged F1-score**, which computes the metric independently for each class and then averages the results. This ensures that all classes are treated equally, regardless of their frequency in the dataset.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

Precision measures the proportion of correctly predicted positive cases out of all instances predicted as positive. High precision indicates that the model produces fewer false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

Since we use the **macro-averaged precision**, it averages precision across all classes.

Recall measures how many actual positive instances were correctly classified. A high recall indicates that the model effectively captures most of the positive cases, reducing false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

We report the **macro-averaged recall** to ensure balanced evaluation across all classes.

To analyze misclassification patterns, we include the **confusion matrix**, which provides a detailed breakdown of actual versus predicted labels. This helps identify which specific classes the model struggles with, allowing for targeted improvements in model performance.

$$\text{Confusion Matrix} = \begin{bmatrix} TP_{\text{class}_1} & FP_{\text{class}_1} & \dots \\ FN_{\text{class}_1} & TP_{\text{class}_2} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{4}$$

5.3 Experimental details

Our experiments involve fine-tuning the RoBERTa model with hyperparameter tuning to optimize classification performance. Below, we outline the key aspects of our experimental setup, including model configurations, training parameters, and evaluation strategy.

We utilize **RoBERTa-base**, a transformer-based model pretrained on large corpora, for sequence classification. The model is initialized with `num_labels` set to 15, corresponding to the number of output classes in our dataset.

To identify the optimal training configuration, we employ a **grid search** over a predefined set of hyperparameters. For each hyperparameter configuration, we train the model and evaluate its performance on a validation set.

5.3.1 Training Procedure

The training process is conducted on Colab's T4 GPU. The dataset is formatted into a custom PyTorch `DreamDataset`, and training is carried out using the `AdamW` optimizer. The following steps outline our training loop:

1. Initialize the model and move it to the available device (`cuda` or `cpu`).
2. Construct the training and validation data loaders.
3. Iterate over the dataset for a specified number of epochs.
4. For each batch:
   - Move inputs and labels to the device.
   - Perform forward propagation and compute the loss.
   - Backpropagate and update the model parameters.
   - Log loss at intervals of 200 batches.
5. Track the time taken for each epoch.
6. Evaluate the model on the validation set after training.

5.3.2 Evaluation and Model Selection

After training each model configuration, we evaluate performance using the following metrics: **F1-score** (macro-averaged), **Precision** (macro-averaged), **Recall** (macro-averaged), **Confusion Matrix**.

The best model is selected based on the highest **macro-averaged F1-score**. The final trained models, along with their corresponding hyperparameters and evaluation metrics, are stored in a dedicated directory.

5.3.3 Computational Resources and Training Time

The total training time for hyperparameter tuning across multiple configurations is around 8 hours where each epoch takes roughly 6 minutes to train, and the best model configuration is determined based on its performance.

5.4 Results

Learning Rate	Batch Size	Epochs	Weight Decay	Precision	Recall	F1 Score
1e-5	4	2	0.005	0.26	0.31	0.28
1e-5	4	5	0.01	0.24	0.29	0.26
1e-5	8	2	0.005	0.25	0.30	0.27
1e-5	**8**	**5**	**0.005**	**0.28**	**0.35**	**0.32**
2e-5	4	5	0.005	0.27	0.32	0.29
2e-5	4	2	0.01	0.24	0.28	0.26
2e-5	8	5	0.005	0.26	0.30	0.28
3e-5	4	5	0.005	0.25	0.29	0.27
3e-5	8	2	0.0	0.23	0.27	0.25
3e-5	8	5	0.01	0.24	0.28	0.26

Table 1: Performance of different hyperparameter combinations on the sentiment classification task.

The results from the hyperparameter search are summarized in Table 1. The parameter grid used for tuning included a learning rate of {1e-5, 2e-5, 3e-5}, batch size of {4, 8}, epochs of {2, 5}, and weight decay of {0.005, 0.01}. The table shows a sample of the best-performing combinations to ilustrate general trends.

The baseline for this task is a random classifier, which, given the 15-class structure of the sentiment labels, would result in an expected accuracy of approximately 6.67%. The best-performing model achieved an F1 score of 0.32, a precision of 0.28, and a recall of 0.35 with the hyperparameter combination of a learning rate of 1e-5, batch size of 8, 5 epochs, and weight decay of 0.005.

Overall, the performance of the best model significantly exceeds the baseline, confirming that the model is learning meaningful patterns from the data. However, the average F1 score across all configurations remains well below 0.50, indicating that the model still struggles with accurately capturing the complex emotional content of dream reports. The recall scores were generally higher than precision, suggesting that the model is better at identifying relevant emotional categories but sometimes misclassifies them. The performance improvements from tuning were consistent but modest, highlighting the difficulty of the task.

6 Analysis

Our analysis focuses on both quantitative and qualitative evaluations of our system to better understand its performance. We assess how leveraging a pretrained model like RoBERTa enables efficient multiclass classification and examine when our model succeeds and where it encounters challenges.

RoBERTa, provides a strong foundation for our classification task. Instead of training a model from scratch, we fine-tuned RoBERTa on our dataset, allowing it to leverage prior linguistic knowledge to distinguish between different emotional categories in dream reports. This approach significantly improved classification accuracy and reduced training time compared to training a model from the ground up.

Since our task involves multiclass classification, we use the **macro-averaged F1-score** as our primary evaluation metric. Unlike accuracy, which may be biased towards the most frequent classes, the macro F1-score accounts for class imbalance by averaging precision and recall across all categories. This ensures that our model performs well across different emotional labels, not just dominant ones.

One key complexity in our classification task is the inclusion of both **single-emotion** and **multi-emotion** labels in dream reports. Some dreams exhibit a clear dominant emotion (e.g., fear, joy, sadness), while others contain a blend of multiple emotions. Our model learns to differentiate between distinct emotions but sometimes struggles with cases where emotions overlap. For example, the model may misclassify a multi-emotion dream as a single dominant emotion, especially when one emotion is more explicitly stated in the text.

Through qualitative evaluation, we observe that the model performs well when emotions are explicitly mentioned or strongly implied in the dream text. However, it struggles with subtle, implicit emotions and dreams with highly figurative or abstract language. Some key observations include:

- Dreams that contain metaphors or indirect emotional expressions are harder for the model to classify correctly.
- Confusion between similar emotions suggests that additional context or hierarchical emotion modeling could improve results.
- Errors often arise in multi-emotion classifications where the model tends to favor one dominant label over a combination.

To address these challenges, potential improvements include: incorporating additional context-aware models or fine-tuning with larger emotion-labeled datasets and exploring hierarchical classification to better capture emotional nuances.

7 Conclusion

In this project, we aimed to explore how transformer-based language models, specifically, RoBERTa, can perform emotional theme classification based on written self reports on dream reports, various dream narratives, and various unique individual diction. Through our multi-label classification approach, we successfully identified instances where dreams conveyed a single dominant emotion as well as cases where multiple emotions coexisted. Our analysis revealed that while RoBERTa is adept at recognizing explicit emotional expressions, it sometimes struggles with subtle, implicit

emotions and abstract language. The use of macro F1 score as our evaluation metric allowed us to ensure balanced performance across all emotion classes, highlighting the effectiveness of the model in handling class imbalances.

However, our model occasionally misclassifies multi-emotion dreams by favoring a dominant emotion over a combination. Additionally, semantic overlap between certain emotions made it difficult to make precise classifications. Future work could explore the incorporation of hierarchical classification techniques or additional context-sensitive fine-tuning strategies to better distinguish between closely related emotions.

Overall, our study highlights the potential of transformer-based models for automating dream analysis, offering a scalable and consistent alternative to manual annotation. By improving the accuracy and efficiency of emotional classification in dream reports, our approach contributes to broader psychological and therapeutic research, enabling large-scale studies on emotional patterns in dreams. Expanding the dataset could enhance its applicability, paving the way for more advanced applications in mental health and dream research.

Team contributions (Required for multi-person team)

Giselle worked on gathering and preprocessing the data. Gabriel implemented the training model and accuracy evaluation reports. Both worked on finetuning the model with various parameter grids. Writeup and analysis between both members, where each member contributed more to the section they focused on or implemented.

References

1. Automatic Scoring of Dream Reports' Emotional Content with Large Language Models. Lorenzo Bertolini, Valentina Elce, Adriana Michalak, Guilio Bernardi, Julie Weeds. arXiv preprint arXiv:2302.14828 (2023).
2. Oberlander, J., & Nowson, S. (2006). Whose Thumb Is It Anyway? Classifying Author Personality from Weblog Text. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions (pp. 627–634). Sydney, Australia: Association for Computational Linguistics.
3. Hall, C., & Van de Castle, R. L. (1966). The content analysis of dreams. Appleton-Century-Crofts.
4. Sleep and Dream Database. (n.d.). Sleep and Dream Database Library. Retrieved March 5, 2025, from https://sleepanddreamdatabase.org/library

A Appendix (optional)