Deciphering Dreams: Sentiment Analysis of Dream Journals

Giselle Rivera[1], Gabriel M. SantaCruz[1]

Department of Computer Science, Stanford University

Stanford
Computer Science

Project Overview

- **Background:**
  - Dreams are a valuable source of insight into emotional and psychological states.
  - Analyzing dream content may help uncover patterns in emotional processing and mental health.
- **Past Research**
  - Hall and Van De Castle (HVDC): Manual coding of dream content based on categories of characters, interactions, and emotions.
  - VADER: Rule-based sentiment analysis tool effective for social media text but limited for complex narratives like dreams.
- **Problem:**
  - Current dream analysis methods, like HVDC framework, rely on manual coding - which is subjective and time consuming.
  - Existing sentiment analysis tools (e.g. VADER) are not well suited for the unique language and emotional complexity of dream reports.
- **Goal:**
  - Develop an automated model to classify emotional content in dream reports into 15 categories: happiness, sadness, wonder/confusion, fear, with the additional 10 emotions being a combination pair of the five listed emotions (e.g. anger and sadness)

Datasets & Metrics

- ***Sleep and Dream Database*** **(SDDb)** - Contains thousands of manually annotated dream reports with emotional labels.
  - Preprocessing included tokenization using the RoBERTa tokenizer, lowercasing, and removing stopwords.

- F1, Precision, Recall - Evaluate classification performance

$$\text{Precision} = \frac{TP}{TP + FP}$$

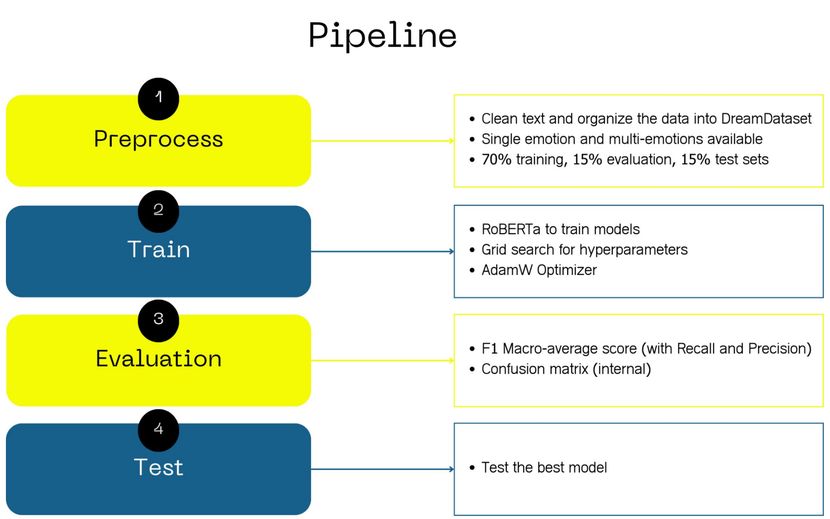
$$\text{Recall} = \frac{TP}{TP + FN}$$

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

- Cross-Validation - Ensure generalization across different data splits

Methods & Experiments

- **Preprocessing:** The reports are preprocessed and tokenized using RoBERTa for sentiment classification into fifteen emotion categories in a multi-class classification setup.
- **Training:** We fine-tune the RoBERTa-base model for sequence classification, setting the number of labels to 15. Training is performed on a Colab T4 GPU using the AdamW optimizer. The dataset is formatted into a custom PyTorch DreamDataset.
- **Evaluation**: We use F1-score, Precision, Recall, and a confusion matrix to analyze performance across all classes. Our experimental setup involves fine-tuning RoBERTa with a grid search for hyperparameter tuning, where the best model is selected based on the highest F1-score.



Discussions & Future Research

Discussions:

- **Model Limitations**: Challenges in distinguishing subtle emotional nuances, especially on dream reports with multi-emotional labeling, and handling ambiguous dream content.

Future Research:

- Explore multimodal approaches (e.g. combining text and physiological data) and expand emotional categories for more granular analysis.

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] Hall, C., & Van de Castle, R. L. (1966). The content analysis of dreams. Appleton-Century- Crofts.

[3] Sleep and Dream Database. (n.d.). Sleep and Dream Database Library. Retrieved March 5, 2025, from https://sleepanddreamdatabase.org/library

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.

- **Performance Overview:**
  - RoBERTa achieved the highest F1-score (0.32) with a learning rate of 2e-5 and a batch size of 16, suggesting that smaller batch sizes and moderate learning rate help improve convergence and generalization.
  - Best-performing model achieved an F1-score significantly exceeding the baseline, confirming that the model is learning meaningful patterns from the data.
  - All configurations per emotion remain below 0.50, indicating the model may still struggle with accurately capturing complex emotional content to distinguish multi-label and single-label classifications (e.g. *happiness* vs. *happiness and wonder).*

- **Emotion Classification Patterns:**
  - Model performed best in detecting *happiness* and *fear*.
  - Lower F1-scores for multi-label classifications may stem from data imbalance and/or less explicit emotional language to distinguish from the single-emotional labels it may overlap.

- **Challenges and Improvement Areas:**
  - Common misclassifications involved mixed emotions or subtle tone shifts, suggesting that further improvement could come from multi-label classification or emotion intensity score.
  - Fine-tuning with additional emotionally rich or balanced data may further improve performance across underperforming categories.
  - Perform a more detailed ablation study of the key parameters to better understand their individual contributions to performance and suggest a more target approach for improving the model's classification capabilities
    - Data Augmentation
    - Fine-tuning Parameters (e.g., epochs, learning rate)
    - Model Architecture (e.g., additional layers, more advanced attention mechanisms)