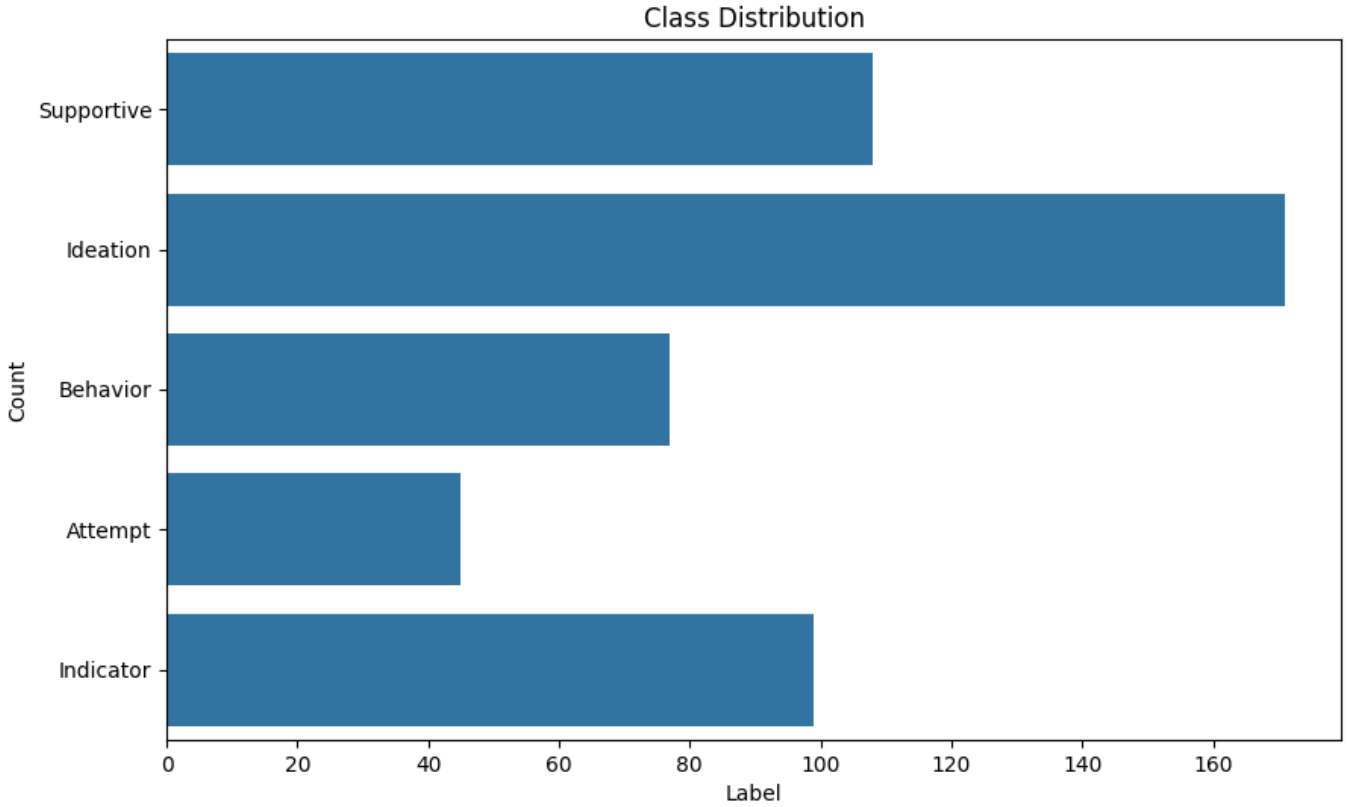
Reddit Suicidal Post Dataset Analysis

Tiernan Lindauer

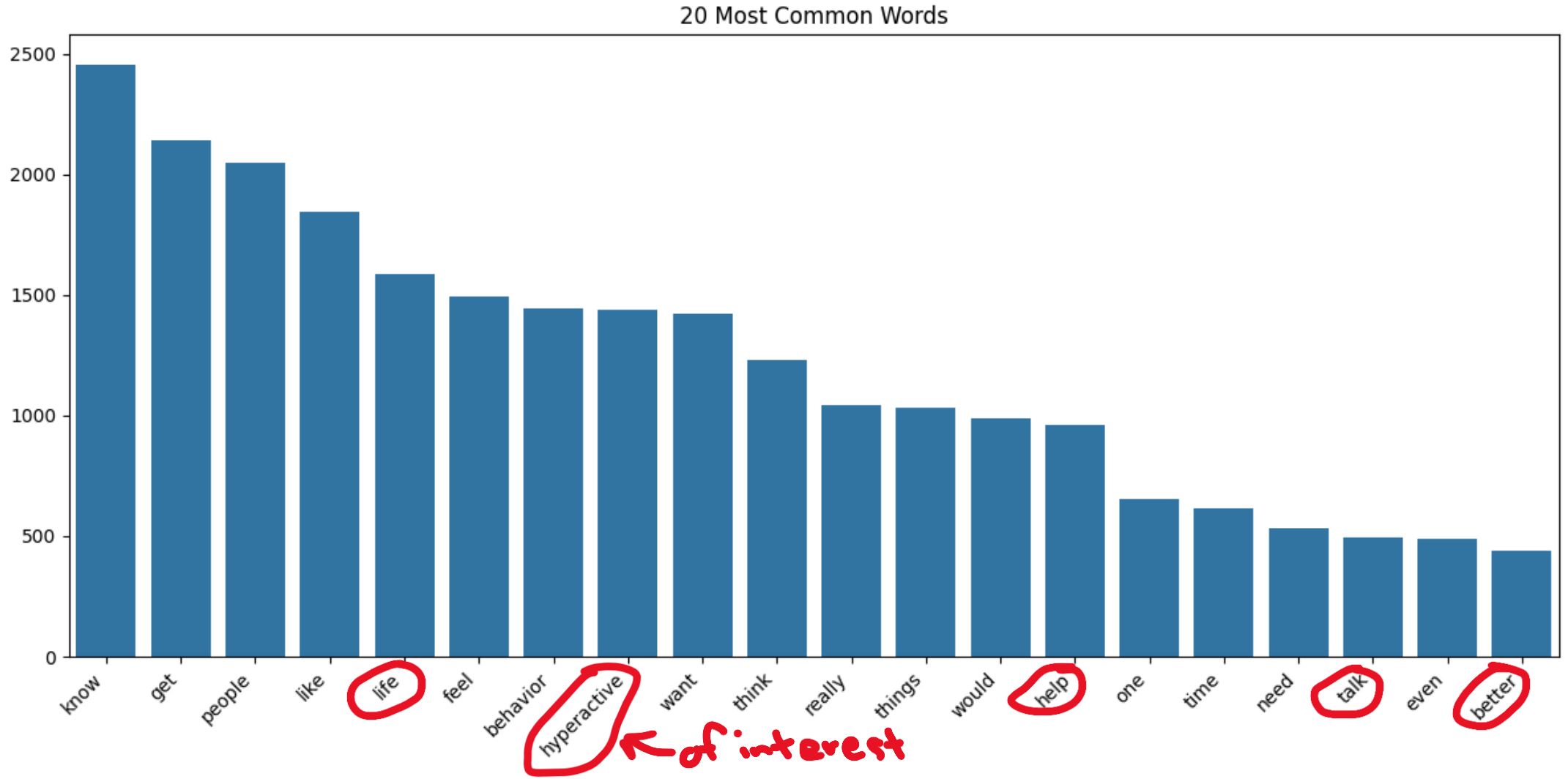
10/5/2024

Dataset Breakdown

- 500 entries total
- Initial dataset needs significant cleaning for use (cleaned version in `/data/500_Reddit_users_posts_labels.csv`)
- Five different classes to predict, ranging in severity
- Our goal is to correctly predict the class of a post in the validation dataset
- To do so, let's perform linguistic (semantic + syntactic analysis) (in `analysis/analysis.py`)

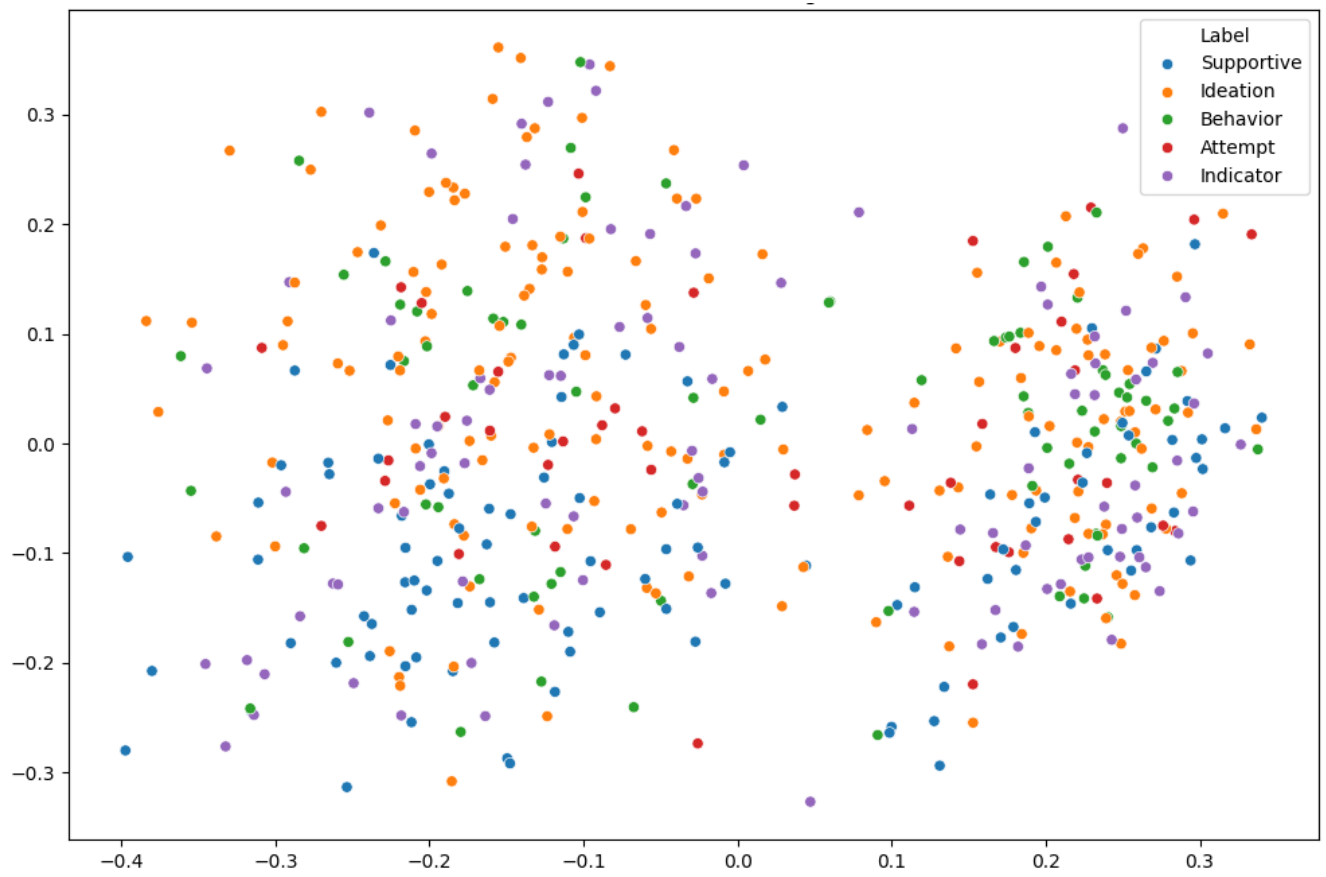


Most Common Words (Stopwords Removed)

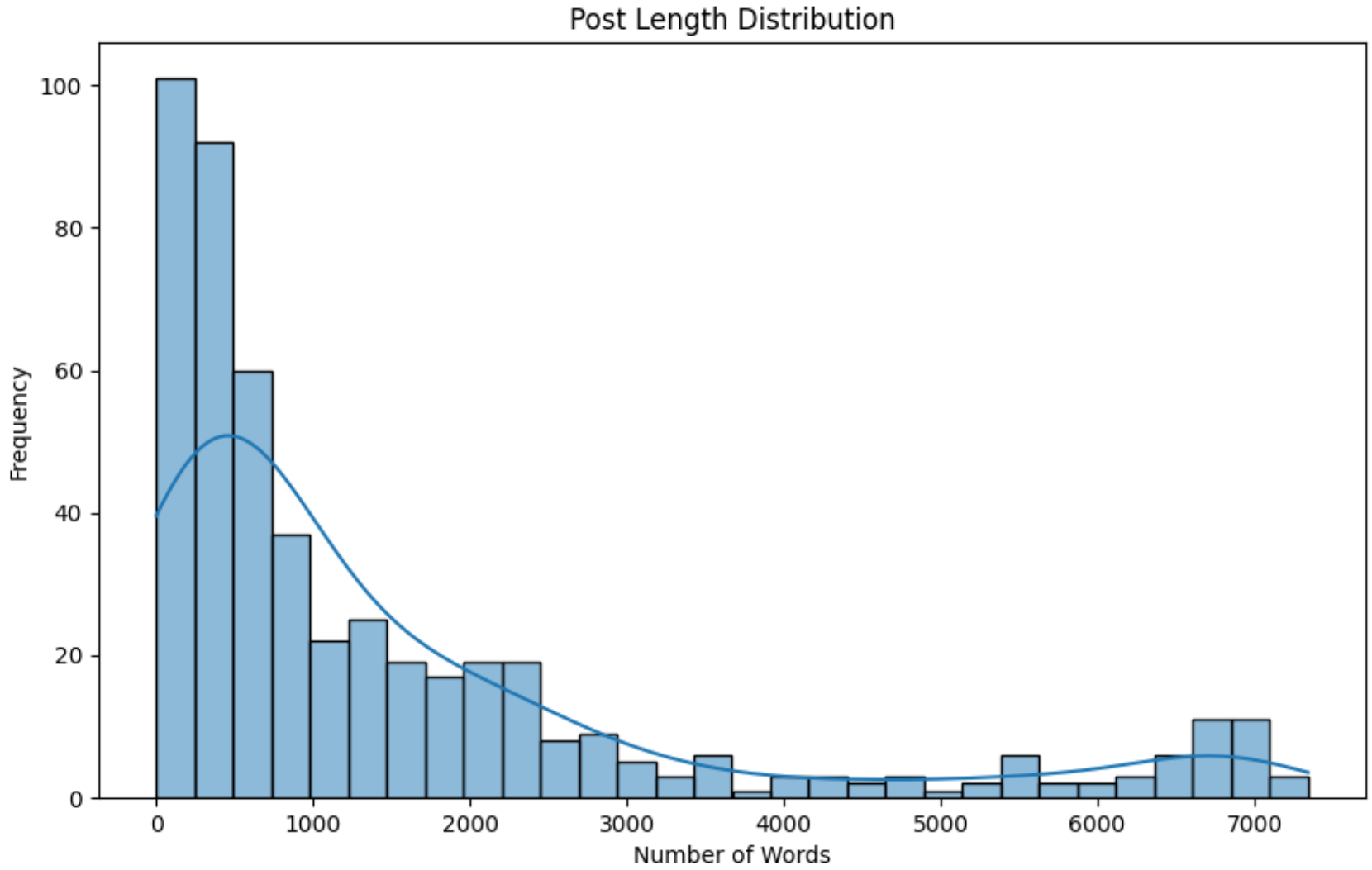


PCA Plot of Post Embeddings

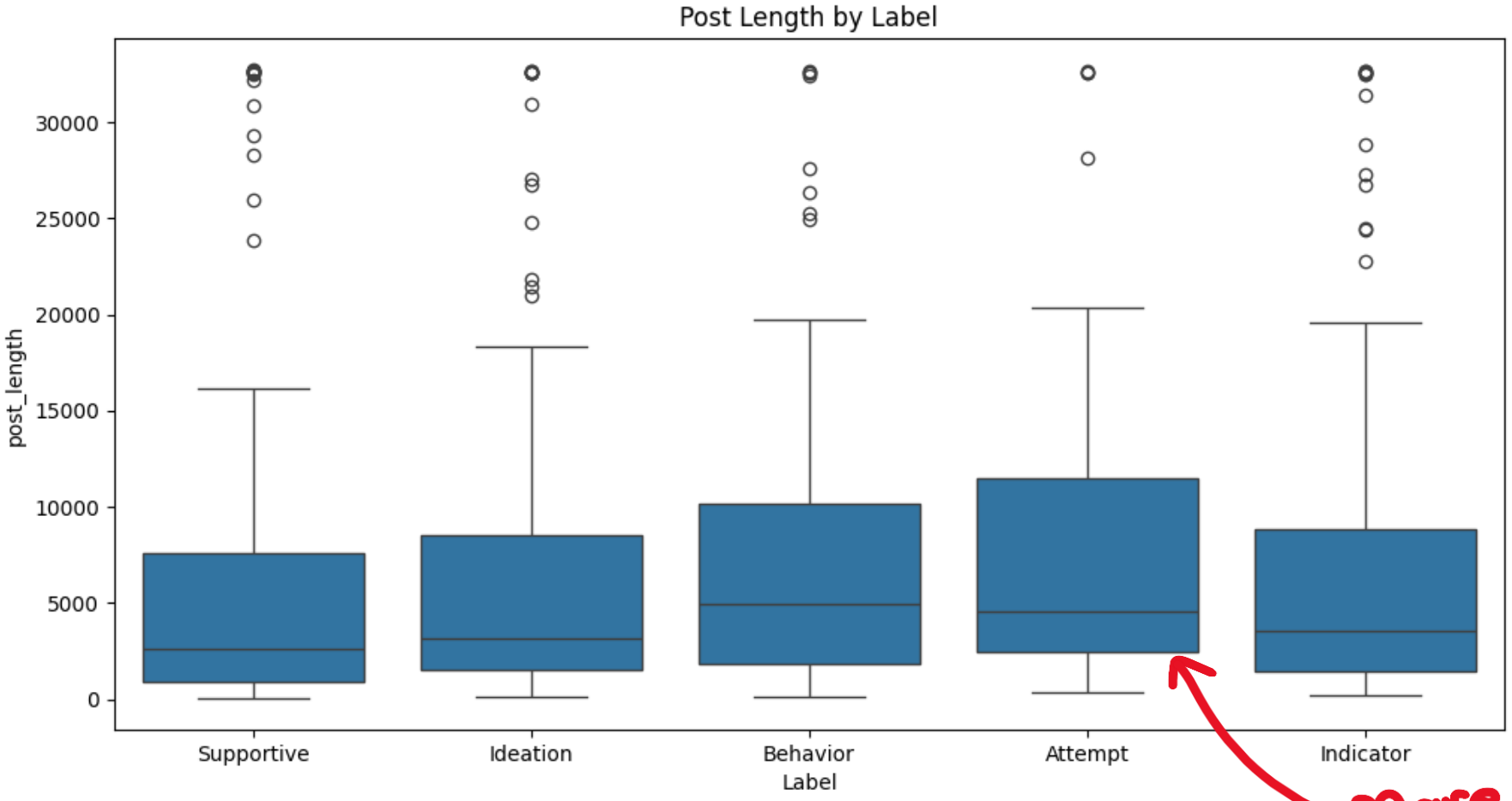
- Used `text-embedding-3-large`
- Little-to-no semantic grouping between categories
- This indicates that vector-based classification isn't a viable approach
- Silhouette score of `-0.02` indicates overlapping clusters (semantic grouping is not very useful)



Post Length Distribution

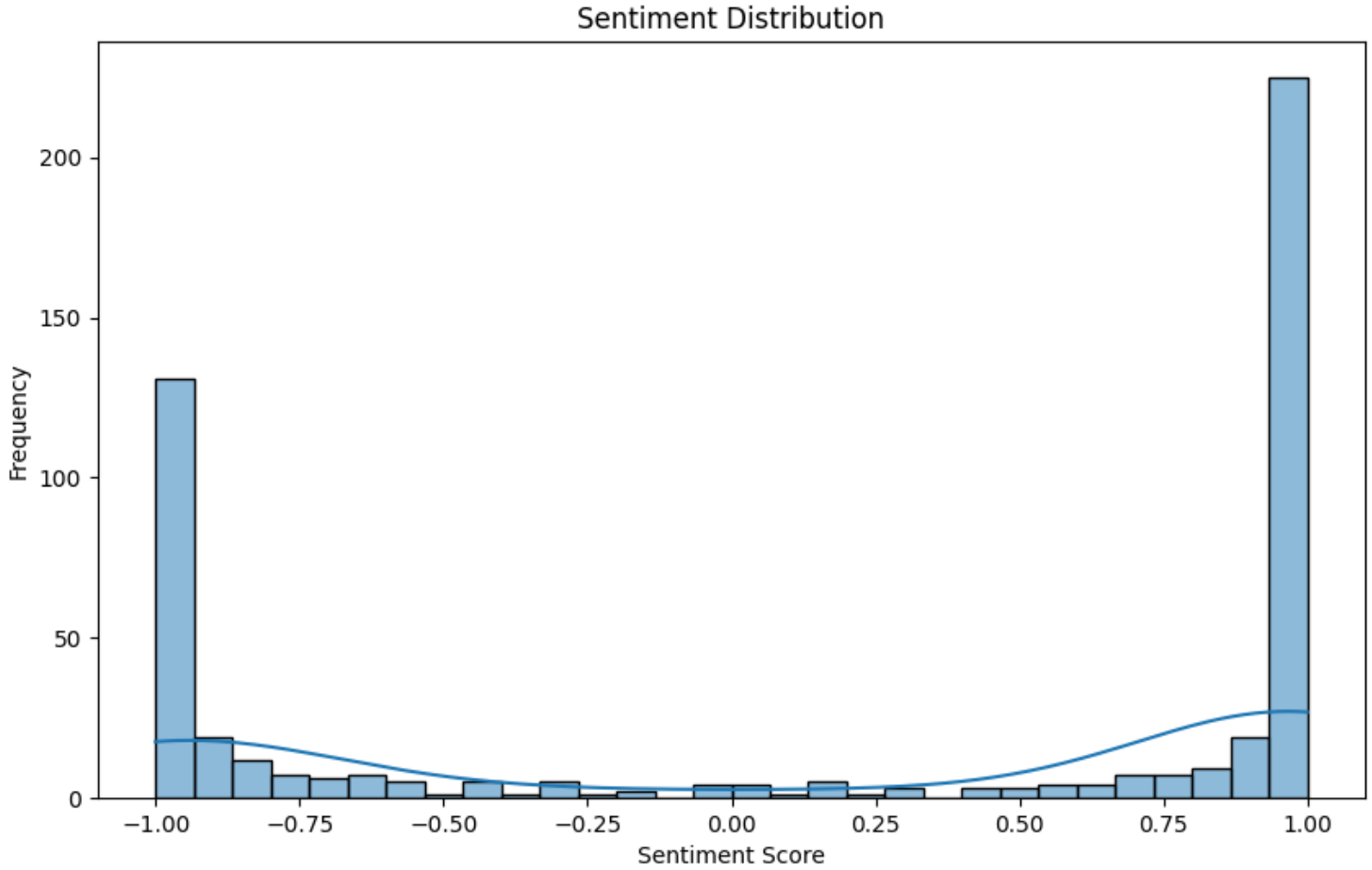


Post Length Distribution By Category



VADER Sentiment Analysis

- -1 is very negative, +1 is very positive
- There many highly negative posts (suicide related) but even more very positive posts (people helping)



VADER Sentiment Analysis By Category