Analyzing Trends in Emotional Tone of Song Lyrics and Their Relationship with Mental Health Disorders

Ömer Yildirim, Jianwen Cao, Yuxuan Wang

November 14, 2024

Abstract

This project examines the emotional tone of song lyrics in the United States from 1990 to 2017, with a focused analysis on the period from 2010 to 2017, to assess its potential correlation with mental health disorder trends. By fine-tuning a GPT-2 model on the GoEmotions dataset, we classified lyrics into 28 distinct emotions, grouped into positive, negative, and neutral categories, allowing us to track annual sentiment trends. Additionally, we curated datasets containing yearly U.S. mental health disorder statistics, enabling a comparison of emotional tones in music and mental health prevalence.

Our findings indicate a significant shift toward negative emotional tone in song lyrics over the entire period. In the 2010-2017 period, statistical analysis revealed strong correlations between increased negative emotional tone in lyrics and higher rates of depression, alcohol use disorders, and drug use disorders, suggesting potential links between music sentiment and societal mental health. This research highlights the need for further exploration of the relationship between lyrical content, music accessibility, and public mental health, with future work potentially incorporating additional musical attributes from audio signals.

1 Introduction

This project explores the evolving emotional tone of song lyrics in the United States and its potential connection to trends in mental health disorders from 1990 to 2017, with a focused analysis on the period from 2010 to 2017, when music accessibility expanded via internet tools and streaming platforms. Rooted in the sentiment analysis domain of natural language processing (NLP), the project utilizes a fine-tuned GPT-2 model on the GoEmotions dataset to classify lyrics into 28 distinct emotions, which are then grouped into positive, negative, and neutral categories. This categorization enabled year-by-year sentiment tracking, with song lyrics and release years compiled into a structured dataset (lyrics_and_years.csv), facilitating the analysis of emotional trends in song lyrics.

Additionally, we prepared a mental health dataset encompassing U.S. statistics on disorders such as depression, anxiety, and substance use disorders for the same period. By calculating yearly cumulative emotion scores from song lyrics and comparing them with mental health trends, we conducted statistical analyses to examine potential correlations. Our framework includes both an overall analysis from 1990 to 2017 and a focused study from 2010 to 2017 to assess shifts associated with increased music accessibility through streaming.

2 Datasets

2.1 GoEmotions Dataset

We use the GoEmotions dataset to fine-tune GPT-2[1] for emotion detection, as GPT models require fine-tuning for downstream tasks. GoEmotions is a large-scale, fine-grained dataset with 28 emotion categories and 58k samples, offering a broader taxonomy (12 positive, 11 negative, 4 ambiguous, and 1 neutral emotion) than traditional emotion datasets, which typically contain only 2 to 6 categories. This fine-grained classification allows for intermediate fine-tuning, which involves training on detailed categories before moving to a more coarse-grained classification, often leading to improved model performance. By leveraging GoEmotions, we enhance both classification accuracy and flexibility for various downstream tasks.

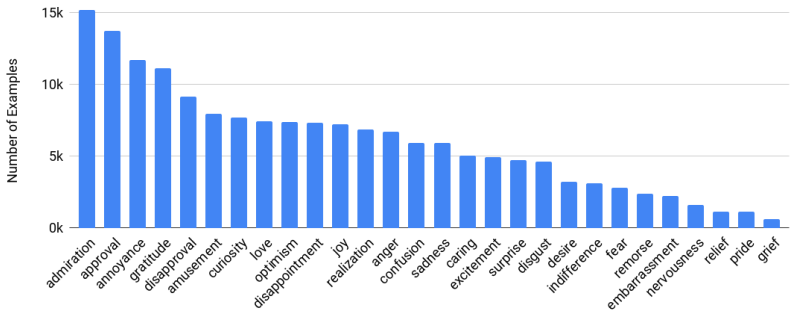

Figure 1: The density distribution of all 27 emotions in GoEmotions[2].

2.2 Song Lyrics and Years Datasets

To create a dataset for evaluating yearly emotion scores in song lyrics, we combined two Kaggle datasets: 160k Spotify Songs from 1921 to 2020[3] and 960K Spotify Songs With Lyrics[4]. The 160k dataset includes musical attributes like release date, artist, and audio features (e.g., acousticness, danceability, valence) but lacks lyrical content. To fill this gap, we incorporated the 960K Spotify Songs With Lyrics dataset, which provides lyrics for many overlapping tracks. Merging these datasets enabled us to compile both release

year and lyrics for each song, allowing for cumulative yearly emotion score calculations and longitudinal sentiment analysis from 1990 to 2017.

2.3 Mental Health Disorders Dataset

We used a Kaggle dataset containing yearly statistics on mental health disorder prevalence in the U.S. from 1990 to 2017[5]. This dataset includes percentages of the population affected by disorders such as schizophrenia, bipolar disorder, anxiety, depression, and substance use. Its comprehensive range aligns with our project's timeline, facilitating the analysis of potential correlations between emotional tone shifts in music and mental health trends. By pairing this dataset with yearly emotion scores from song lyrics, we aim to examine whether changes in musical sentiment correspond with fluctuations in mental health disorder prevalence.

3 Preprocessing

3.1 Preparation of Song Lyrics and Years Dataset for Evaluation

To prepare the song lyrics and years dataset for evaluation, we applied several auxiliary NLP preprocessing steps to ensure the data was appropriately structured and filtered for sentiment analysis. First, we selected and merged relevant columns from two distinct datasets: one containing English song lyrics (id and lyrics columns) and the other providing song metadata (id and year columns). This merge operation on the id column ensured each song entry in the new dataset contained both the lyrics and release year, allowing us to analyze temporal trends in emotional tone. The final dataset, lyrics_and_years.csv, contains 90,647 entries, each including a unique song id, its lyrics, and the year it was released. This dataset is now ready for further NLP tasks to calculate yearly emotion scores.

3.2 Preparation of Mental Health Dataset for Statistical Analysis

To prepare the mental health disorders dataset for analysis, we filtered it to include only U.S. data and converted the Year and disorder columns (e.g., Schizophrenia, Bipolar disorder, Depression) to numeric formats, enabling accurate calculations. Rows with any missing values were removed to ensure completeness, and we focused on data from 1990 to 2017. We then calculated yearly averages for each disorder, creating a consistent, year-by-year dataset. The resulting dataset, saved as mental-health-yearly-score-USA.csv, provides reliable prevalence scores by year, ready for statistical comparison with yearly emotion scores from song lyrics.

4 Model Training

We fine-tuned a GPT2-large model with 36 transformer blocks, 20 attention heads for each attention block and 1280 embedding dimensions. The GPT2 model is a decoder-only transformer model without cross attention. It is stacked with transformer blocks, while in each transformer block there is an attention block with subsequent mlp block[6].

Since GoEmotions is a multi-label classification task which means there could be multiple emotions within a sentence. So, during fine-tuning, for each sentence we make a label of length 28, if the emotion is contained, the corresponding place will be 1, otherwise it is 0. For example, for a sentence with admiration and approval emotion, the label would be [1, 0, 0, 0, 1, 0, 0, 0. . . . . , 0]. However, this will cause most of the positions to be 0, which will lead to the problem of label imbalance. The model will simply classify all samples as 0 to achieve a high accuracy and a small loss. Our investigation found that the field of object detection has also been plagued by this problem for a long time, that

is, the imbalance between the background class and the target class. Therefore, in order to avoid this shortcut solution, we use Focal Loss[7].

$$\mathrm{FL}(p_\mathrm{t}) = -\alpha_\mathrm{t}(1-p_\mathrm{t})^\gamma \log(p_\mathrm{t}).$$

Figure 2: The α and β are factors which control the penalty of high-confidence predictions.

Generally speaking, focal loss assigns high weights for low confident predictions (hard sample) and extremely low weights for high confident predictions. This will allow the model to focus on the real problem instead of the already well learned cases. The experiment shows that it is extremely helpful for improving the performance of our model.

Table 1: The hyperparameters for fine-tuning.

Hyperparameter	**Value**
Is causal	True
Loss function	Focal Loss ($\alpha = 0.75$, $\gamma = 2$)
Dropout rate	0.1
Batch size	16
Epochs	2
Initial learning rate	1×10^{-5}
Final learning rate	0.0
Warmup steps	100
Initial weight decay	0.0
Final weight decay	1×10^{-4}

5 Evaluation

5.1 Evaluation of Fine-Tuned GPT-2 Model

We report our performance on both 28 emotions with respect to F1 score, precision and recall.

Table 2: Model Performance Comparison

	SCRATCH	**GPT2**	**GPT2 (wo focal)**	**GPT2-large**
Precision	0.4928	0.5304	0.6693	0.5290
Recall	0.3017	0.5252	0.3691	0.5757
F1	0.3588	0.5206	0.4568	0.5458

According to the table, we can clearly find out that the performance of vanilla GPT2 and BERT pre-trained is similar (both 0.46 F1), which excludes the possibility that it's the pre-trained model that contributes to the performance gain. The second ablation study shows that compared to learning from scratch, using pre-trained models can largely benefit the learning (from 0.36 to 0.46 F1). The third ablation is about the loss function, we find that the model performance significantly jumps from 0.46 F1 to 0.52 F1 by simply replace cross-entropy loss with focal loss. Finally, to figure out how the model size would contribute to the result, we use a GPT2-large model instead of GPT2-base, it turned out to bring 2.5% performance gain (from 0.52 to 0.545 F1).

As is shown in Figure 3, the training loss of GPT2-large decreases faster and converges to a better position compared to the base model, so does the valid loss. As a matter of fact, due to the large model's strong compacity of fitting, it overfits GoEmotions in 3 epochs.

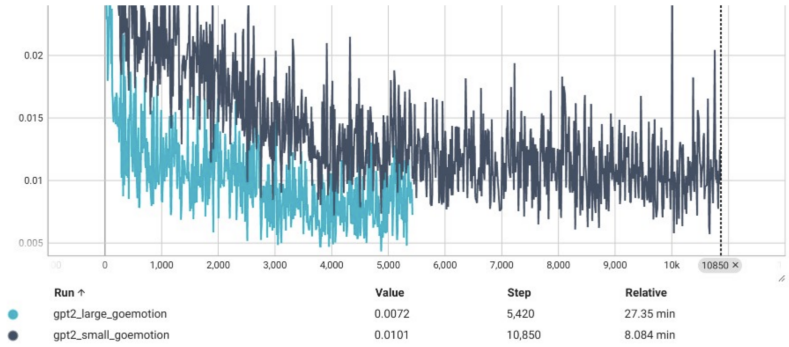

Figure 3: The training loss comparison between GPT2-base and GPT2-large.

6 Statistical Analysis

6.1 Analysis of Song Lyrics and Mental Health Disorders Trends Between 1990 and 2017

Our statistical analysis between the years 1990 and 2017 did not reveal a statistically significant correlation between the increase in negative emotional tone in song lyrics and the prevalence of various mental health disorders, such as depression, anxiety, and substance use disorders. As shown in the Appendix, the trend over this period indicates a shift in song lyrics from a positive emotional tone toward a more negative one. This significant shift in emotional tone highlights a change in the lyrical content of songs, with an increase in expressions of negative emotions relative to positive ones. However, the lack of significant correlation with mental health trends during this period suggests that increased accessibility and exposure to music may not have been a strong influence, possibly due to limited access to streaming services and digital music platforms, which only became widely available after 2010.

Given these results, we decided to conduct a separate analysis for the period from 2010 to 2017, focusing on the impact of growing internet-based music accessibility, including streaming platforms and mobile applications. With these tools, music became more widely and immediately accessible to the public, potentially intensifying its influence on societal attitudes and mental health trends. By analyzing this later period separately, we aim to determine whether increased access to music in the digital age corresponds with any observable relationships between the emotional tone of song lyrics and mental health disorder trends.

6.2 Analysis of Song Lyrics and Mental Health Disorders Trends Between 2010 and 2017

The statistical analysis of song lyrics and mental health disorder trends from 2010 to 2017 revealed several significant findings. Notably, a strong negative correlation was observed between the increase in negative emotional tone in song lyrics and the prevalence of depression, alcohol use disorders, and drug use disorders in the population, with Pearson correlation values of $r = 0.822$, $r = 0.826$, and $r = 0.838$, respectively, all with p-values below 0.05. These results suggest that as song lyrics expressed more negative emotions during this period, there was a corresponding increase in certain mental health disorders, particularly those related to mood and substance use. This supports the hypothesis that the emotional tone of popular music might reflect or even influence societal mental health patterns, especially in the digital age where access to music has significantly expanded.

In addition to these key findings, we observed other statistically meaningful correlations, including between positive emotional tone and schizophrenia ($r = 0.795$, $p = 0.018$), and bipolar disorder ($r = 0.783$, $p = 0.022$). These results may indicate complex rela-

tionships between the emotional content of lyrics and various mental health conditions. Overall, the analysis points to notable associations between lyrical sentiment and mental health prevalence, which warrant further exploration to understand potential causal links or societal implications.

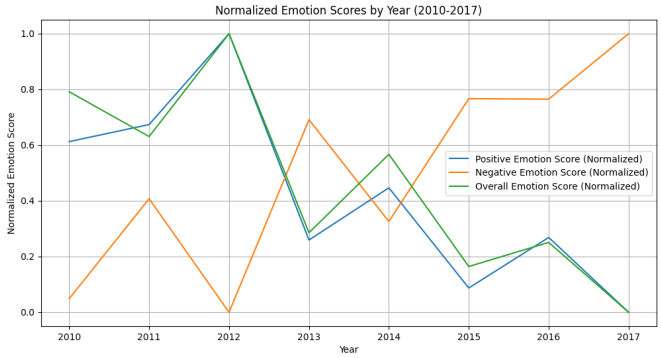

Figure 4: Normalized Emotion Scores by Year (2010-2017).

Figures 4 and 5 illustrate the trends in normalized emotion scores in positive, negative and overall categories over time, and the particular relationship between negative emotional tone and depression.

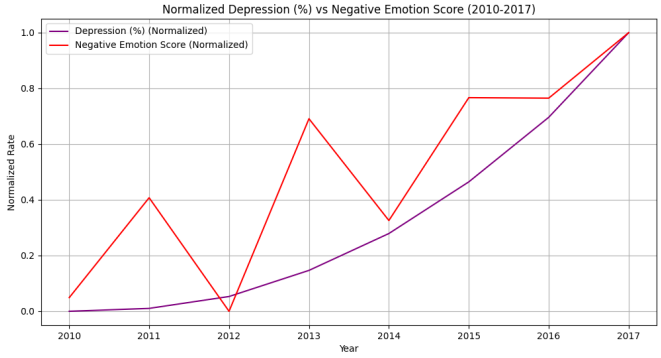

Figure 5: Normalized Depression Rate vs Negative Emotion Score (2010-2017).

Additional visualizations of the negative emotion trends and their alignment with other mental health disorder prevalence rates can be found in the appendix.

7 Future Work

In the future, this work could be extended by investigating the relationship between song lyrics' emotional tone and other musical attributes such as tune, tempo, and style, derived from the audio signals of the songs. Analyzing these associations could reveal whether specific musical styles or audio features, such as tempo and key, are more frequently associated with positive or negative emotional lyrics. Additionally, incorporating audio features alongside lyrical sentiment could provide a more holistic understanding of how music as a whole reflects or influences societal moods and mental health trends.

References

[1] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[2] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547, 2020.

[3] 160k Spotify songs from 1921 to 2020 (Sorted) — kaggle.com. https://www.kaggle.com/datasets/fcpercival/160k-spotify-songs-sorted. [Accessed 14-11-2024].

[4] 960K Spotify Songs With Lyrics data — kaggle.com. https://www.kaggle.com/datasets/bwandowando/spotify-songs-with-attributes-and-lyrics. [Accessed 14-11-2024].

[5] Global Trends in Mental Health Disorder — kaggle.com. https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health-disorder. [Accessed 14-11-2024].

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[7] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017.

A Appendix: Additional Figures and Tables

A.1 Training Loss Plots of GPT2-base and GPT2-large Models

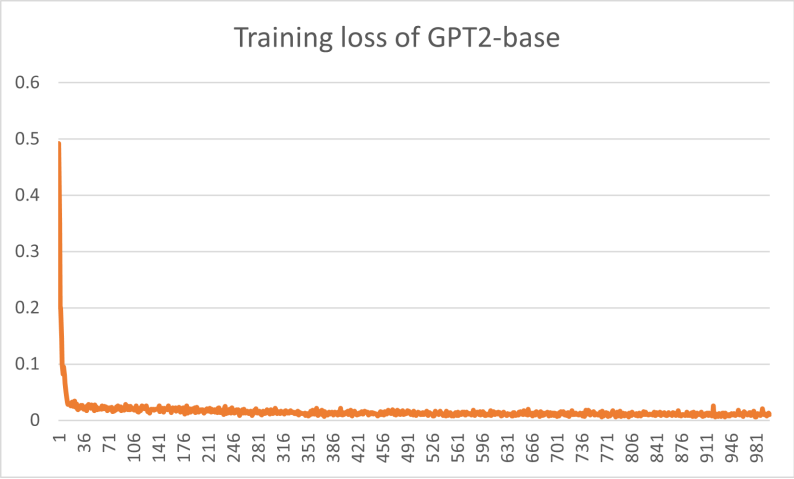

Figure A1: The training loss for GPT2-base GoEmotions fine-tuning.

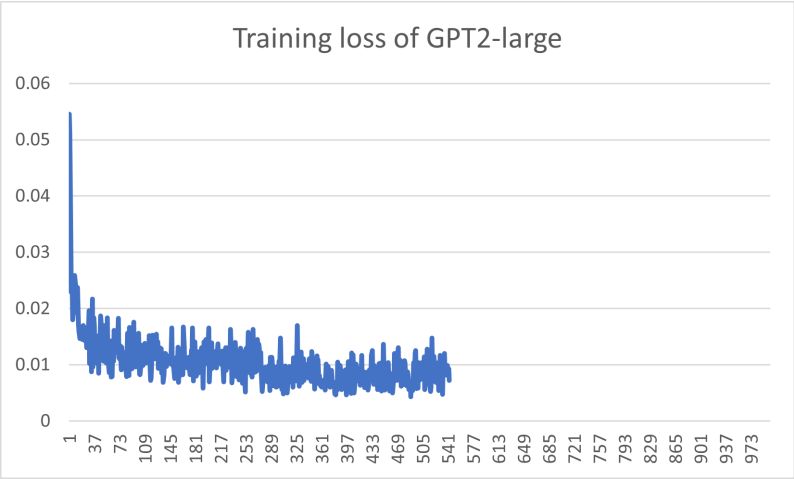

Figure A2: The training loss for GPT2-large GoEmotions fine-tuning.

A.2 Statistical Analysis - Normalized Plots for the Analysis Between 1990 and 2017

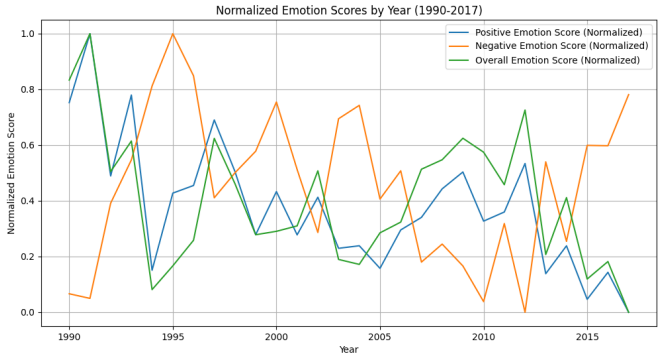

Figure A3: Normalized Emotion Scores by Year (1990-2017)

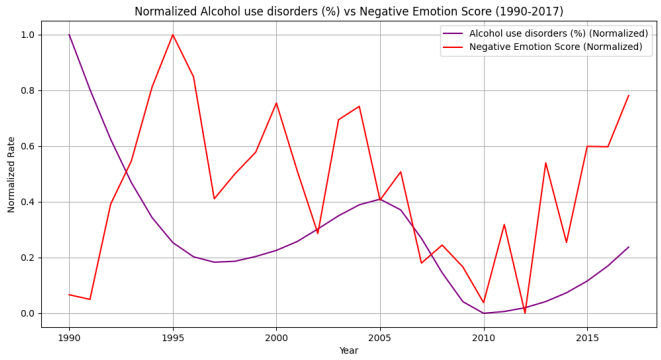

Figure A4: Normalized Alcohol Use Disorders (%) vs Negative Emotion (1990-2017)

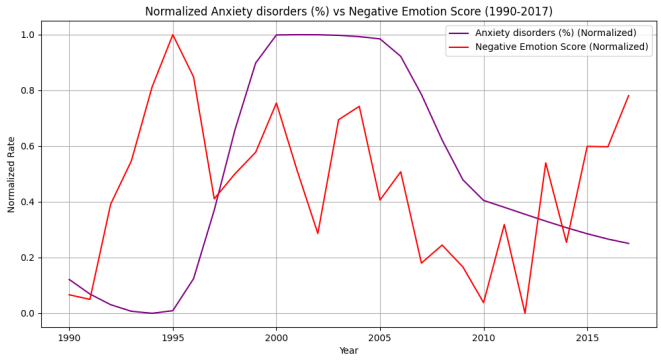

Figure A5: Normalized Anxiety Disorders (%) vs Negative Emotion (1990-2017)

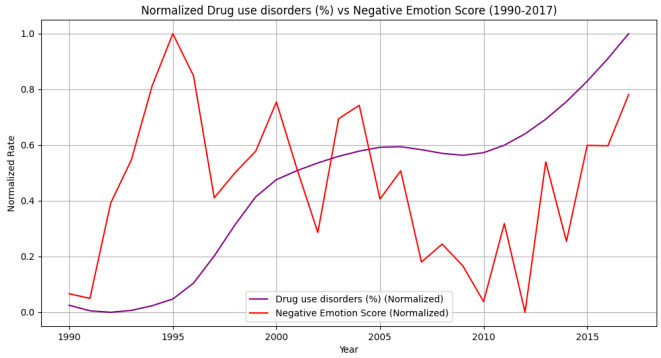

Figure A6: Normalized Drug Use Disorders (%) vs Negative Emotion (1990-2017)

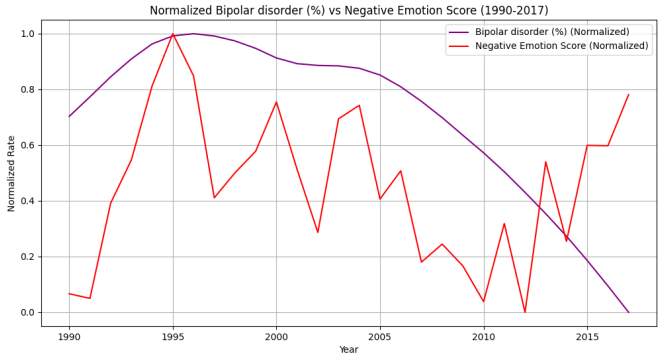

Figure A7: Normalized Bipolar Disorder (%) vs Negative Emotion (1990-2017)

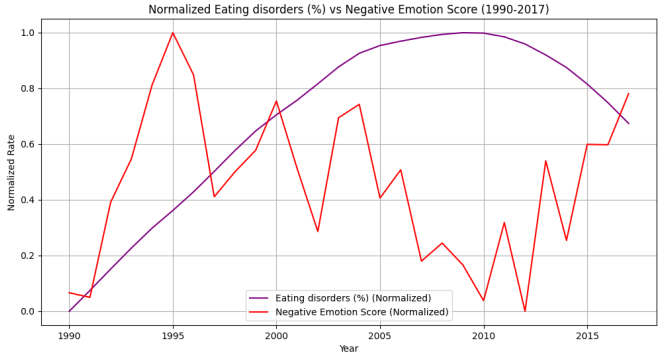

Figure A8: Normalized Eating Disorders (%) vs Negative Emotion (1990-2017)

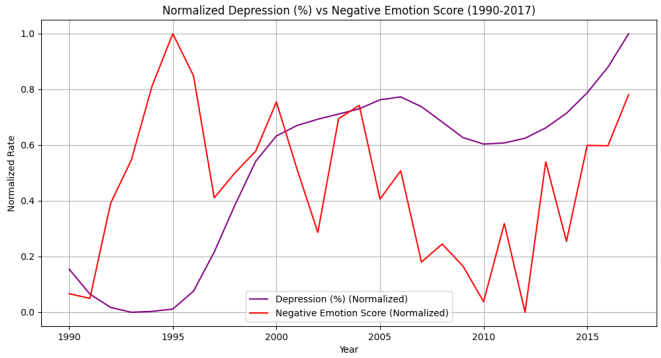

Figure A9: Normalized Depression (%) vs Negative Emotion (1990-2017)

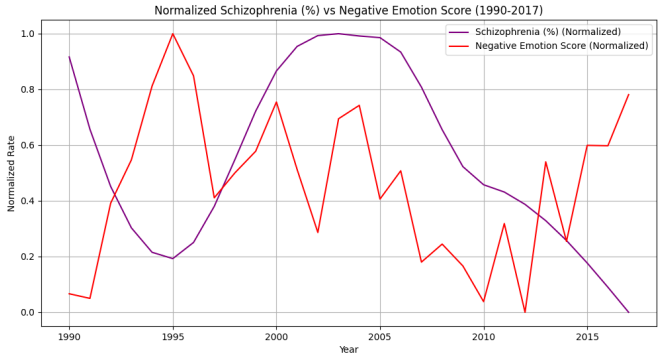

Figure A10: Normalized Schizophrenia (%) vs Negative Emotion (1990-2017)

A.3 Statistical Analysis - Normalized Plots for the Analysis Between 2010 and 2017

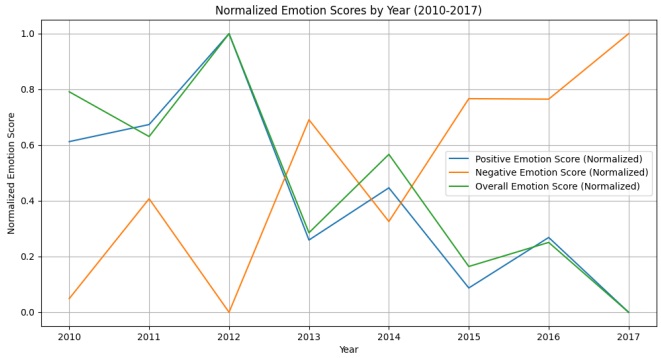

Figure A11: Normalized Emotion Scores by Year (2010-2017)

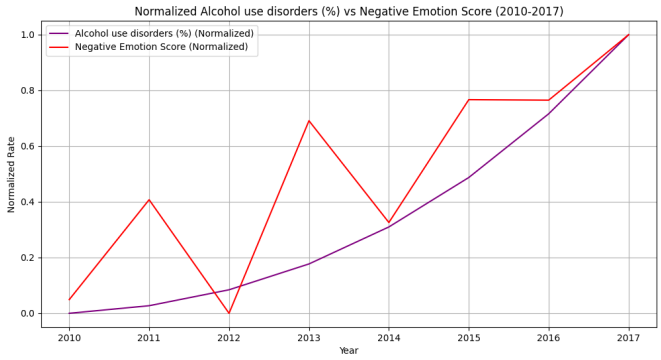

Figure A12: Normalized Alcohol Use Disorders (%) vs Negative Emotion (2010-2017)

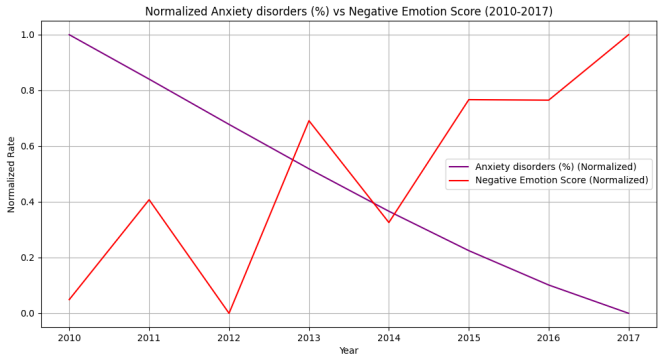

Figure A13: Normalized Anxiety Disorders (%) vs Negative Emotion (2010-2017)

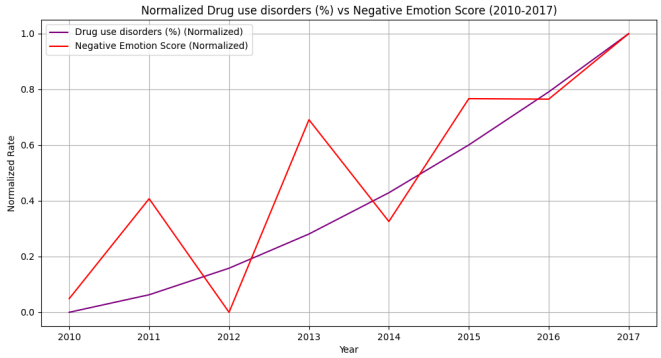

Figure A14: Normalized Drug Use Disorders (%) vs Negative Emotion (2010-2017)

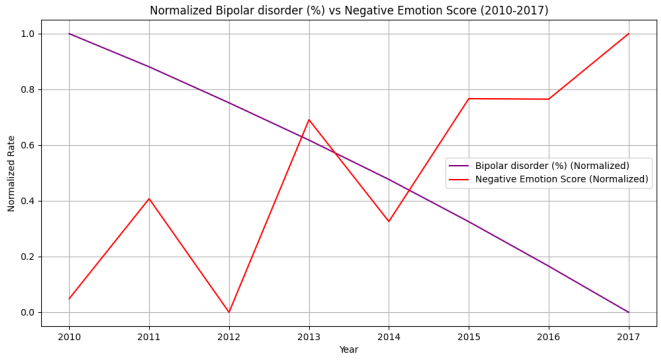

Figure A15: Normalized Bipolar Disorder (%) vs Negative Emotion (2010-2017)

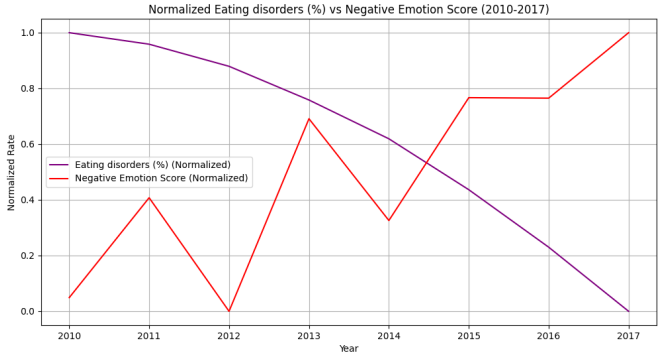

Figure A16: Normalized Eating Disorders (%) vs Negative Emotion (2010-2017)

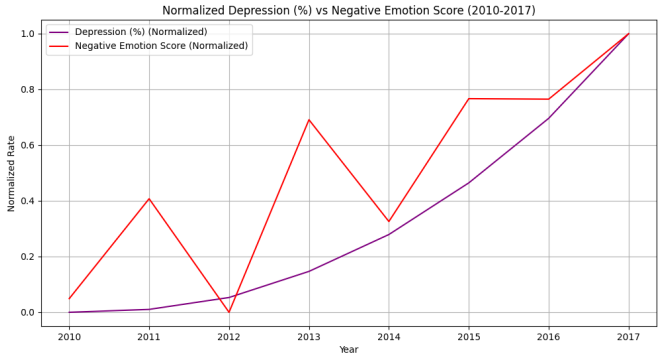

Figure A17: Normalized Depression (%) vs Negative Emotion (2010-2017)

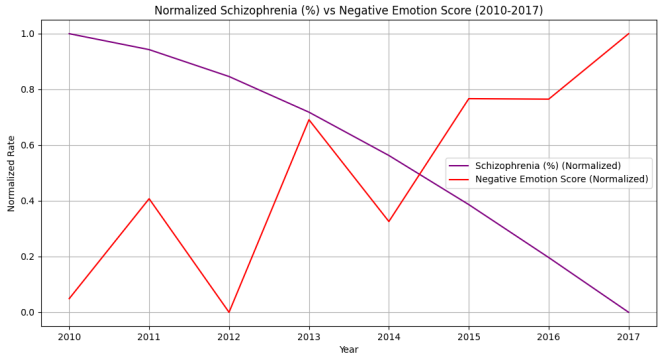

Figure A18: Normalized Schizophrenia (%) vs Negative Emotion (2010-2017)