Project 2: Final Draft

Econ 1680: Machine Learning, Text Analysis, and Economics

Anushka Kataruka

May 12, 2024

1 Introduction

Music, like many other art forms, is an expression of emotions, for both creators and consumers. The lyrics, rhythm, tone, pitch, and other features of a track can very often capture emotions of the song's artists and its listeners. Preferences in music tastes of a population can reflect on the socioeconomic/environmental conditions of a population, as seen in the study "The Language of Lyrics", Pettijohn and Sacco (2009).
A small handful of studies have tried to explore the relationship between emotions in music and the contemporary socioeconomic conditions, including a study on lyrics listened to during the COVID-19 pandemic (Putter et al. (2022)), reflection of unemployment trends in lyric emotions(Qiu et al. (2021)), and representations of the economy in hip-hop (REHN and SKÖLD (2005)). Economic conditions impact people's lived experiences, which they can express in multiple ways including the production and consumption of art. As such, both the production and consumption of certain kinds of music can be influenced by economic conditions of a time and place.

Examining popular music lyrics and their emotions, we can answer questions like:Do people listen to more "happy" music during times of relative economic prosperity, or conversely do they listen to more sad or angry music when their economic circumstances are worsening?

Previous studies pertaining to the impact of socioeconomic circumstances on lyrics have largely been undertaken in the fields of psychology and cultural studies, but not economics itself. This project will attempt to build on past literature, and inform on the relationship between cultural phenomenon/art (studied here as popular music lyrics) and economic conditions. I analyse data on lyrics and audio features of the monthly most popular songs in the US, Canada, and the UK, correlating these temporal trends in "emotions" with trends in unemployment, using monthly data for the years 2017-2021. Songs were classified using text classification methods to generate "emotions" of a song's lyrics. This was then put through a multi-layer neural network, which generated probabilities to predict emotions of a track given economic circumstances (mainly unemployment) and audio features of the track. We find that higher unemployment is associated with more negative emotions in popular lyrics.

2 Data Sources and Descriptions

My data on the most popular songs comes from a Kaggle dataset developed by Dhruvil Dave (Dave (2021)), which contains all the daily "Top 200" published globally by Spotify, between the dates January 1, 2017, to December 31, 2021. I aggregated this data to reflect the top 5 songs per month between 2017-2021 in the US, ordering songs by the frequency of their appearance in the top 10 of every day of the month. I chose this dataset as Spotify as it is the biggest player in the music streaming industry, dominating 31% of the market as of 2021 (Forde (2022)), and thus its charts are a good representation of music listening trends. My other music-related data sources are Spotify API and RapidAPI, to extract a song's audio features and lyrics.

My data on US monthly unemployment rates, for the same time period, comes from the US Bureau of Labor Statistics. Canada's monthly unemployment was also taken from Statistics Canada, Canada's government data agency. My data on monthly unemployment in the United Kingdom was taken from the UK's Office for National Statistics.

More information on the processing and generation of the data is available in the Github Repository of the project. As mentioned earlier, the analysis only uses monthly data between 2017-2021, for the US, UK, and Canada. I chose these three countries because they are largely English-speaking countries, which makes the process of cleaning and analysing lyrics much simpler. Appendix I contains detailed descriptions for each variable, and Table 1 shows descriptive statistics for all our data.

3 Method

For my **Text Analysis**, after extracting the lyrics of each song from RapidAPI, I cleaned the lyrics and removed all time signatures and duplicate lines present in each lyric. The lyrics were translated using Google Cloud's translation API. This ensured that songs in other languages (like Spanish) were still evaluated correctly. Following this, a text classification task was performed on each song lyrics using the Hugging Face model, avaliable here, which generates for each text, 28 probability values for 28 different emotions. For each song lyric, I extracted the emotion which had the highest probability value, which I further classified into "positive", "negative" and "neutral" emotions, represented in the variable *emotions_id* (more information available below, in Appendix I). A value of 1 in *emotions_id* corresponds to a "negative" emotion, and 0 corresponds to "positive". I also extracted the probabilities of *anger*, *love*, and *sadness* (the three most common emotions in songs in our dataset), to examine trends in those emotions as well. *emotions_id*, *anger*, *love*, and *sadness* are my dependent variables. The main parameter of interest is *unemployment*.

I one-hot encoded the country data, to incorporate country fixed effects, (c_i), and then split the data into training and testing sets. I ran an initial OLS for each of the dependent variables, generating coefficients on unemployment and other control variables:

$$Y_i = \beta_0 + \beta_1 unemployment_i + \beta X_i + \gamma c_i + \varepsilon_i$$

where, Y_i is *emotions_id*, *anger*, *love*, or *sadness*, and X_i represents the rest of the control variables (audio features). I also ran a multi-layer neural network, again for each of my

dependent variables. For assessing *emotions_id*, used a multilayer classifier (as *emotions_id* is a classification between 0 and 1/"positive" and "negative"), while a multilayer regressor was used for the rest of the dependent variables (as they had continuous values). Each of the neural networks had a hidden layer size of 100, with logistic activation functions. A neural network is better at evaluating non-linear relationships, and so more useful in quantifying and visualizing the relationship between emotions in lyrics and unemployment overall.

4 Results or Expected Results

The OLS results are available in Table 2 in Appendix II, with coefficients visualized in Figures 1-4. We only obtain statistically significant results for *emotions_id*, and significant results for *anger* (at the 10% significance level). We see a substantial positive relationship between *emotions_id* and unemployment, indicating that an increase in unemployment is associated with more negative emotion. This fits with our hypothesis that higher unemployment negatively impact people's emotions, which can be reflected in the lyrics they listen to. This relationship is also seen in Figure 5 , where we see a generally positive relationship between our dependent and main independent variable.

After training the neural network separately for each dependent variable, I evaluated its performance on the testing data, and also generated partial dependence plots for each dependent variable, to visualize the interaction between that variable and *unemployment*. The results of our neural network for each of the dependent variables is the following:
1. *anger*: After training, the MLP regressor yields an R-squared of 0.264 on the testing data, showing that our variables explain 26.4% of the variability in *anger*. The partial dependence plot for *anger* (Figure 6) shows a negative relationship between unemployment and anger, similar to the OLS prediction, which is very interesting.
2. *love*: After training, the MLP Regressor yields an R-squared of 0.515 on the testing data, showing that our variables explain 51.5% of the variability in *love*. The partial dependence plot for *love* (Figure 7) seems to show only a slight negative relationship between unemployment and love in lyrics, implying that higher unemployment is associated with people listening to fewer love songs.
3. *sadness*: After training, the MLP Regressor yields an R-squared of 0.032 on the testing data, showing that our variables explain only 3.2% of the variability in *love*. The partial dependence plot for *love* (Figure 8) shows a positive relationship between unemployment and sadness in lyrics, implying that higher unemployment is associated with people listening to more songs with sad lyrics.
4. *emotions_id*: Here, the trained MLP classifier, predicts a positive or negative emotion with 91% accuracy on the testing dataset. The confusion matrix for the results of the classifier can be seen in Figure 9. We see that the model is better at predicting negative emotions (with 84% accuracy) than positive emotions. The partial dependence plot (Figure 10) also shows a positive relationship between unemployment and negative emotions in lyrics, implying that higher unemployment is associated with people listening to more songs with negative emotions in its lyrics.

5 Conclusion

Our results show that increase in unemployment is associated with lyrics with more "negative" emotions being more popular. However, it seems that these negative emotions are more sad than angry, as anger seems to be negatively correlated with unemployment. Our analysis also has a significant limitation: On examining some of its emotion classification predictions, I discovered that the model was somewhat biased. If there are "bad words" or curse words used in a text, the model has a tendency to report more negative emotions like anger and annoyance. As most of the popular songs analysed do use some curse words, this biases our own "*emotions_id*" variable to reflect more negative emotions than there actually may be. This could be a bias in our overall analysis as well, explaining why our neural network is better at predicting negative emotions and not positive. I think this study can lead to more research in the cross-section of culture and economics, helping us learn more about how economic factors shape people's lives and culture.

References

Dave, Dhruvil (2021) "Spotify Charts," 10.34740/KAGGLE/DS/1265407.

Forde, Eamonn (2022) "Spotify Comfortably Remains The Biggest Streaming Service Despite Its Market Share Being Eaten Into."

Pettijohn, Terry F. and Donald F. Sacco (2009) "The Language of Lyrics," *SAGE*, https://doi.org/10.1177/0261927X09335259.

Putter, Kaila C, Amanda E Krause, and Adrian C North (2022) "Popular music lyrics and the COVID-19 pandemic," *Psychology of Music*, https://doi.org/10.1177/03057356211045114.

Qiu, Lin, Sarah Hian May Chan, Kenichi Ito, and Joyce Yan Ting Sam (2021) "Unemployment Rate Predicts Anger in Popular Music Lyrics: Evidence From Top 10 Songs in the United States and Germany From 1980 to 2017," *Psychology of Popular Media*, https://doi.org/10.1037/ppm0000282.

REHN, ALF and DAVID SKÖLD (2005) "'I Love The Dough': Rap Lyrics as a Minor Economic Literature," *Culture and Organization*, https://doi.org/10.1080/14759550500062268.

6 Appendix I

Following is the description of each variable, including its source:
1. *acoustic*: Extracted from Spotify's API for each song. A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
2. *dance*: Extracted from Spotify's API for each song. This describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
3. *energy*: Extracted from Spotify's API for each song. It is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
4. *instrumental*: Extracted from Spotify's API for each song. This describes how much of the song is instrumental versus has more vocals. Approaching 1.0 means that the track is mostlyt instrumental.
5. *loudness*: Extracted from Spotify's API for each song. The overall loudness of a track in decibels (dB).
6. *mode*: Extracted from Spotify's API for each song. Mode indicates the modality (major or minor) of a track Major is represented by 1 and minor is 0.
7. *tempo*: Extracted from Spotify's API for each song. The overall estimated tempo of a track in beats per minute (BPM).
8. *valence*: Extracted from Spotify's API for each song. A measure from 0.0 to 1.0 describing the musical "positiveness" conveyed by a track.
9. *emotions_id*: This is our independent variable. It can attain 3 possible values: 0, 1, and 2. 0 indicates "positive" emotions, as identified by the model, including admiration, amusement, approval, caring, excitement, gratitude, joy, love, optimism, and pride. 1 includes "negative" emotions, including anger, annoyance, disappointment, disapproval, confusion, disgust, embarrassment, fear, grief, remorse, and sadness. Finally, 2 indicates "neutral" emotions, which are hard to classify, including neutral, curiosity, desire, nervousness, realization, relief, and surprise. All data points with $emotions\_id = 2$ were excluded from this analysis This variable was generated using lyrics of each song, which itself was retrieved using the lyrics endpoint of RapidAPI.
10. *anger*: The probability of anger being present in the song lyrics, also generated by our sentiment analysis of each song. It is a measure from 0 to 1. It represents the maximum probability value generated by the text classification model of *anger* and *annoyance*
11. *love*: The probability of love being present in the song lyrics, also generated by our sentiment analysis of each song. It is a measure from 0 to 1. It represents the maximum probability value generated by the text classification model of *love*, *desire* and *caring*.
12. *sadness*: The probability of sadness being present in the song lyrics, also generated by our sentiment analysis of each song. It is a measure from 0 to 1. It represents the maximum probability value generated by the text classification model of *sadness* and *grief*
13. *unemployment*: This is our primary dependent variable. This indicates the monthly unemployment rate in the US, for all workers aged 16 and over. This was sourced from the Bureau of Labor Statistics

7 Appendix II

Table 1: Summary Statistics

	acoustic	dance	energy	instrumental	loudness	mode
Count	899	899	899	899	899	899
Mean	0.21	0.72	0.625	0.002	-6.196	0.568
Std	0.207	0.13	0.14	0.015	2.032	0.495
Min	0.0002	0.234	0.214	0.0	-14.505	0.0
Median	0.163	0.737	0.626	0.0	-6.062	1.0
Max	0.908	0.975	0.913	0.162	-2.253	1.0

	tempo	valence	emotions_id	anger	love	sadness	unemployment
Count	899	899	899	899	899	899	899
Mean	121.83	0.53	0.554	0.179	0.255	0.089	5.495
Std	28.58	0.223	0.548	0.202	0.341	0.165	2.108
Min	62.948	0.0605	0.0	0.003	0.002	0.007	3.5
Median	120.04	0.511	1.00	0.102	0.062	0.012	4.70
max	202.899	0.966	2.0	1.69	1.73	0.896	14.8

Note: *emotions_id*, *anger*, *love*, and *sadness* are our dependent variables, while the rest are the regressors for our model. *unemplyoment* is our main parameter of interest, while the rest are controls. There are a total of 9 indicators which will be used to predict a popular song's "emotion" given its audio features and the prevalent economic conditions

Table 2: OLS Results for anger, love, sadness, and emotions_id

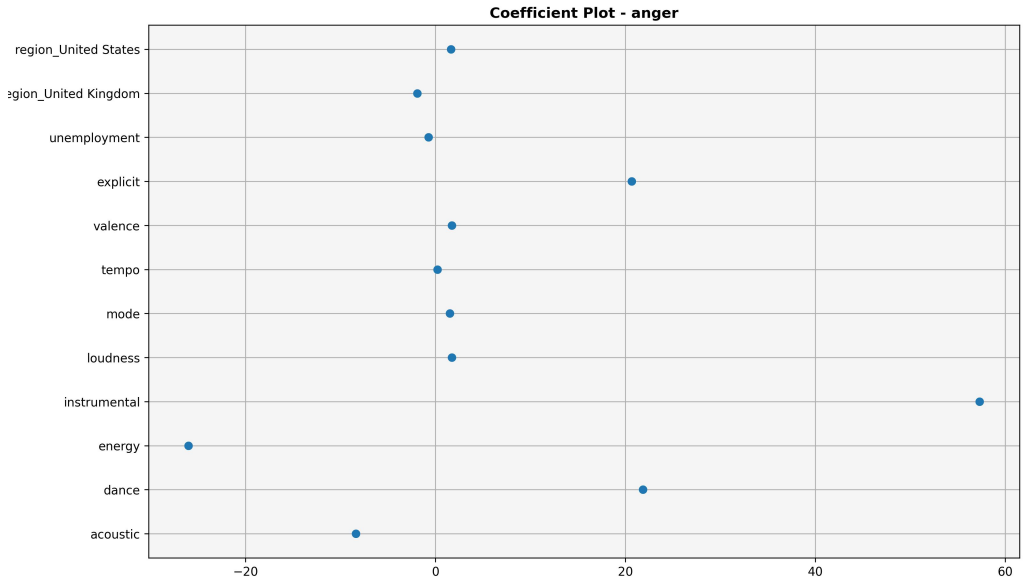
	anger	love	sadness	emotions_id
unemployment	-0.74*	-0.75	0.01	2.35**
	(0.39)	(0.70)	(0.36)	(1.12)
acoustic	-8.37**	23.43***	17.86***	-12.91
	(3.67)	(6.58)	(3.44)	(9.65)
dance	21.86***	-19.35**	-11.20**	48.30***
	(5.33)	(9.56)	(4.99)	(14.36)
energy	-26.01***	38.37***	14.73***	-97.48***
	(5.90)	(10.58)	(5.53)	(16.28)
instrumental	57.31	39.32	-112.43**	-249.33
	(53.67)	(96.21)	(50.27)	(159.78)
loudness	1.73***	-2.62***	-0.98**	5.67***
	(0.41)	(0.74)	(0.39)	(1.10)
mode	1.52	-4.40	0.28	11.17***
	(1.51)	(2.71)	(1.41)	(4.03)
tempo	0.21***	-0.00	0.08***	1.00***
	(0.03)	(0.05)	(0.03)	(0.08)
valence	1.72	28.35***	-18.63***	-47.05***
	(3.69)	(6.61)	(3.45)	(9.85)
explicit	20.68***	-12.09***	-0.15	15.34***
	(1.66)	(2.98)	(1.56)	(4.44)
R-squared	0.62	0.46	0.31	0.61
R-squared Adj.	0.61	0.45	0.30	0.60

Standard errors in parentheses.

* $p < .1$, ** $p < .05$, ***$p < .01$

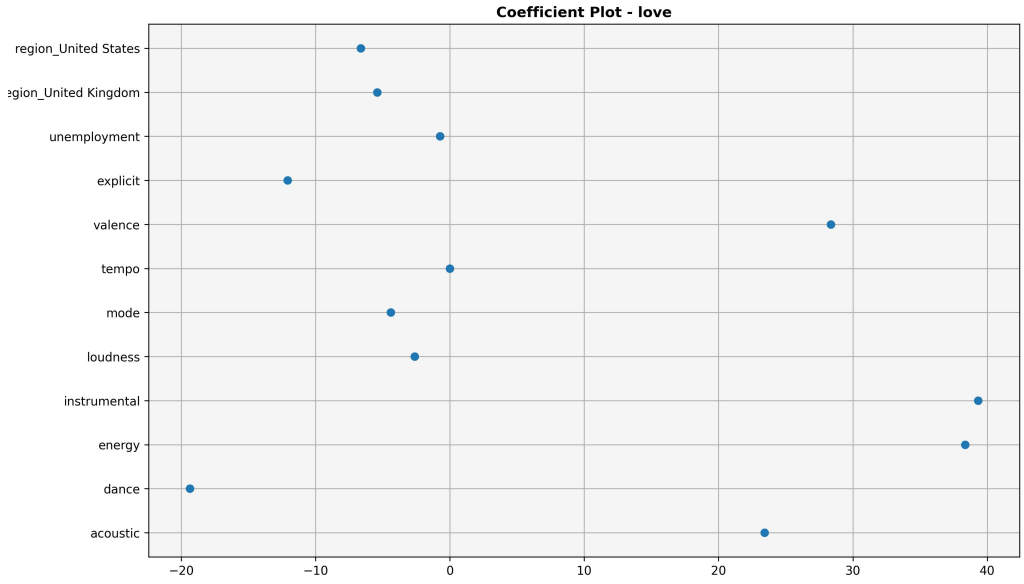
Note: The columns represent the different dependent variables of our analysis. Coefficients for country dummies (country fixed effects) have been excluded for clarity

Figure 1: OLS Coefficient Plot - Anger



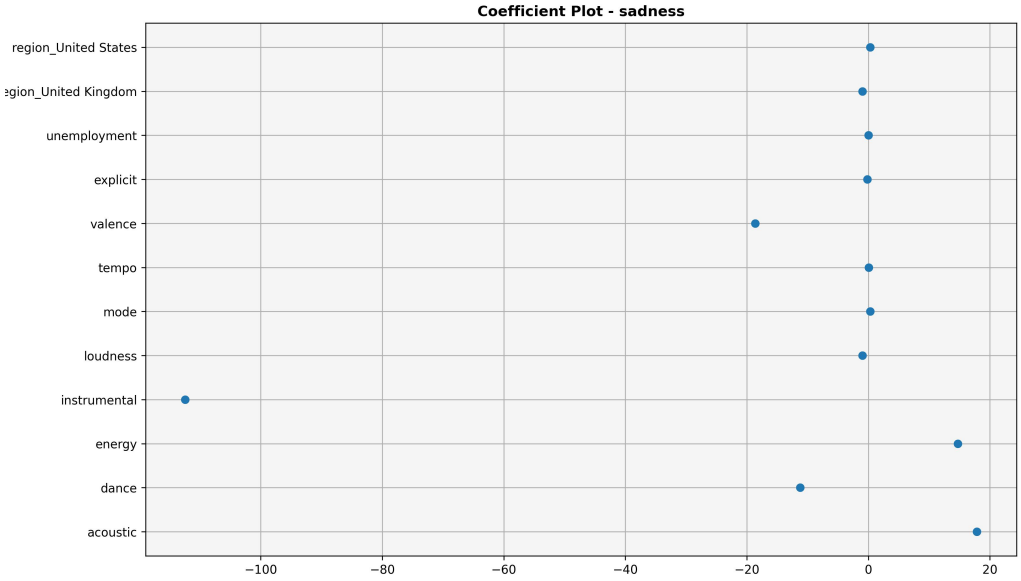
Coefficients for individual countries have been excluded. The coefficient *unemployment* is significant at the 10% significance level.

Figure 2: OLS Coefficient Plot - Love



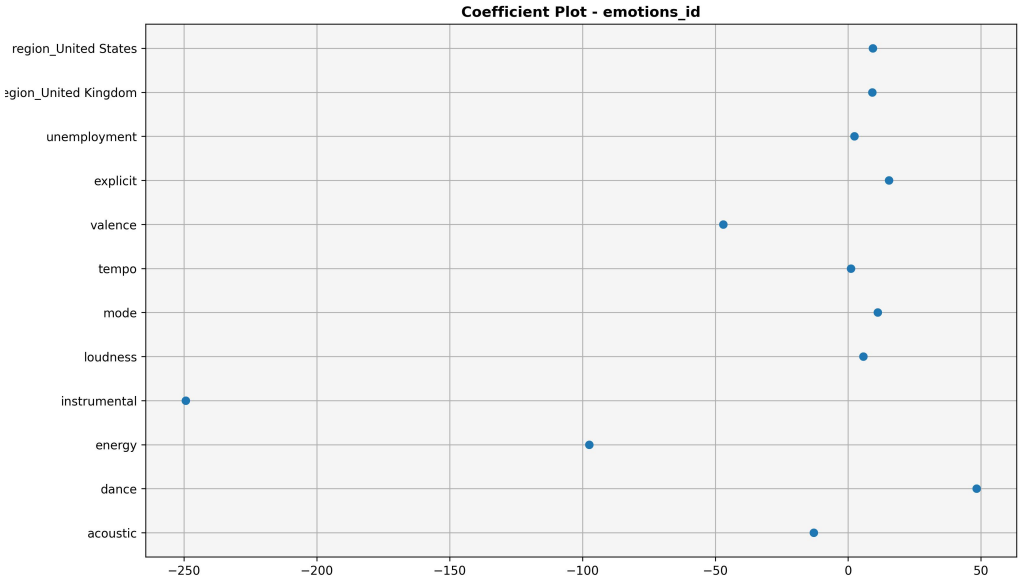
Coefficients for individual countries have been excluded. The coefficient *unemployment* is not statistically significant.

Figure 3: OLS Coefficient Plot - Sadness



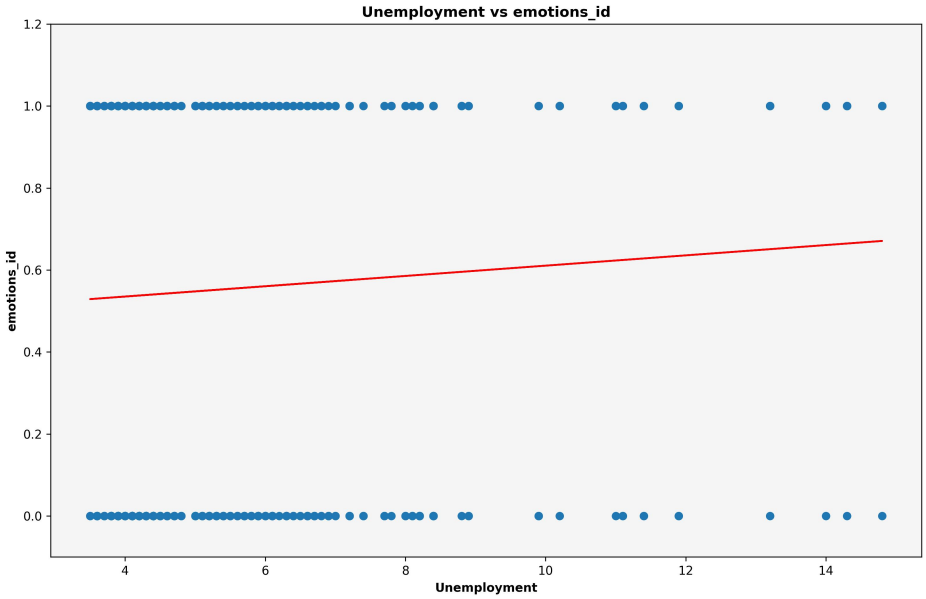
Coefficients for individual countries have been excluded. The coefficient *unemployment* is not statistically significant.

Figure 4: OLS Coefficient Plot - emotions_id



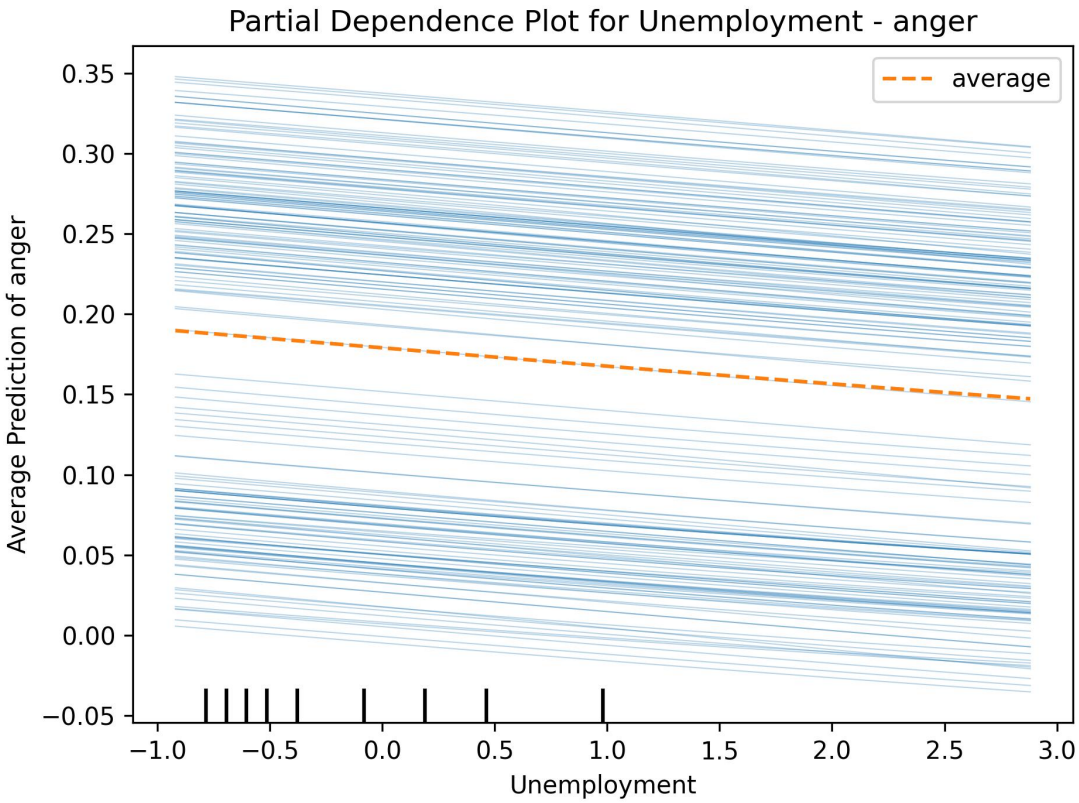
Coefficients for individual countries have been excluded. The coefficient *unemployment* is significant at the 5% significance level, and is slightly positive

Figure 5: Emotions_id and Unemployment



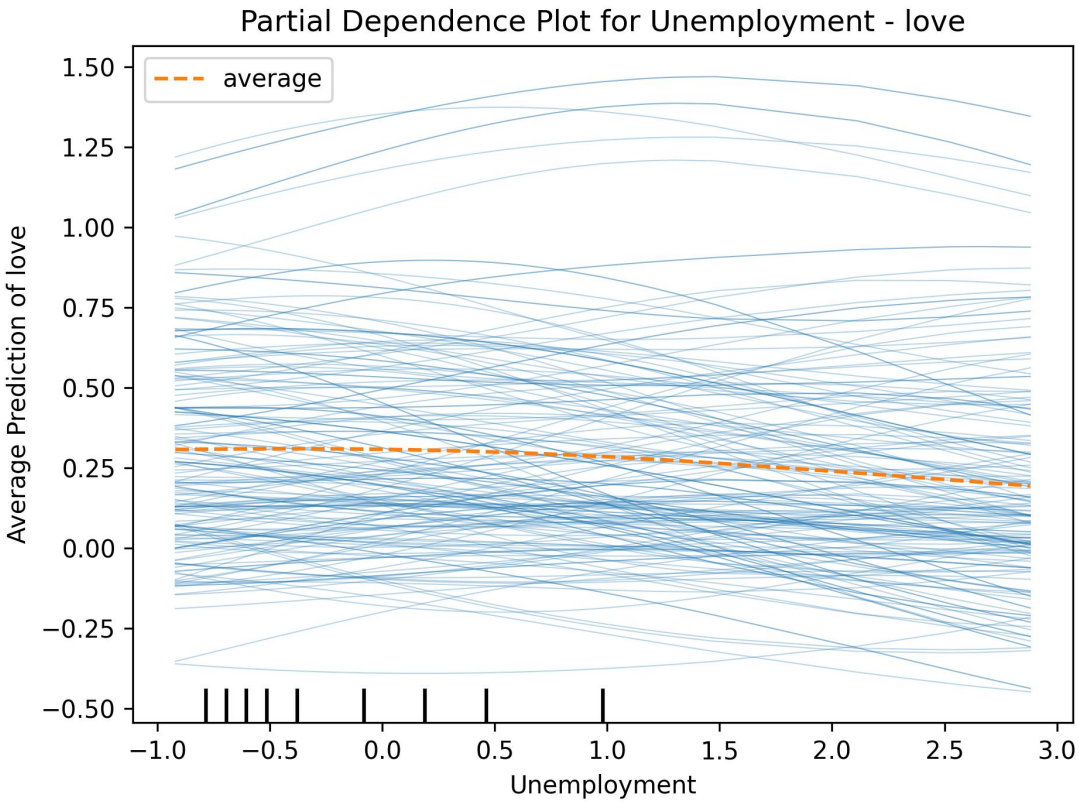
Relationship between unemployment and *emotions_id* in the dataset The red line is the line of best fit, with each scatter point representing a singular data point in our testing dataset. This shows a positive relationship. Statistically significant in OLS.

Figure 6: Partial Dependence Plot - Anger



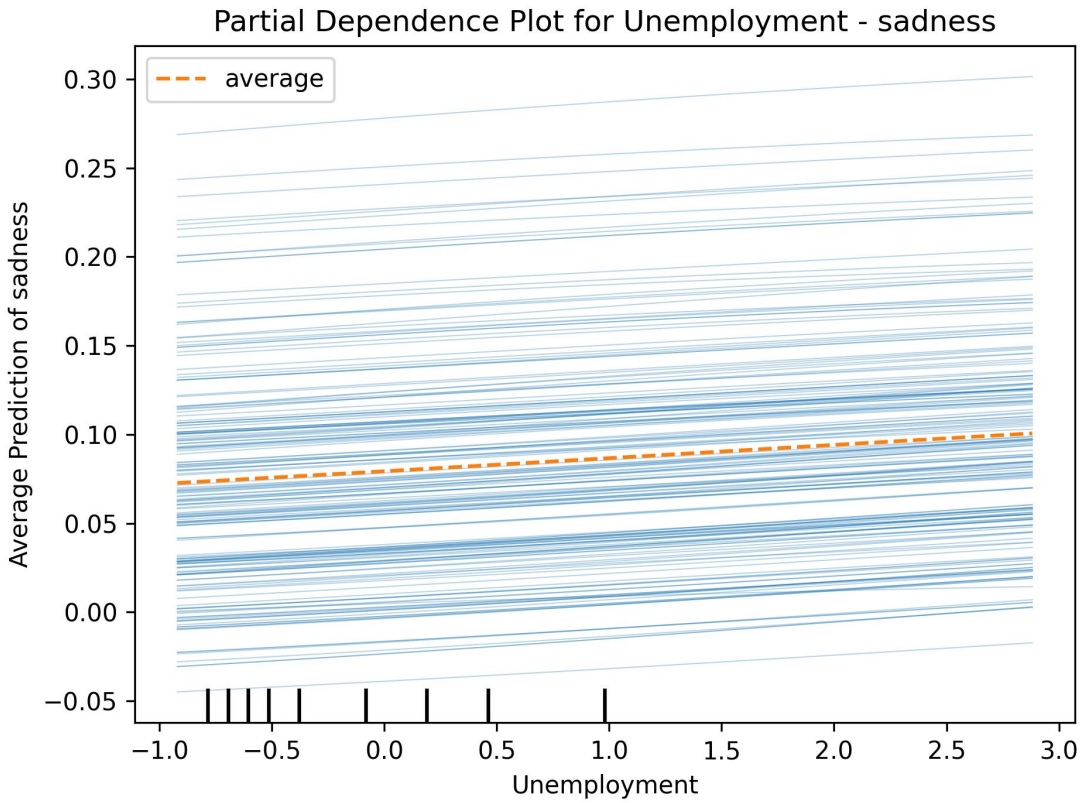
The interaction between *unemployment* and *anger*, in the multilayer regressor model. A negative relationship is observed.

Figure 7: Partial Dependence Plot - Love



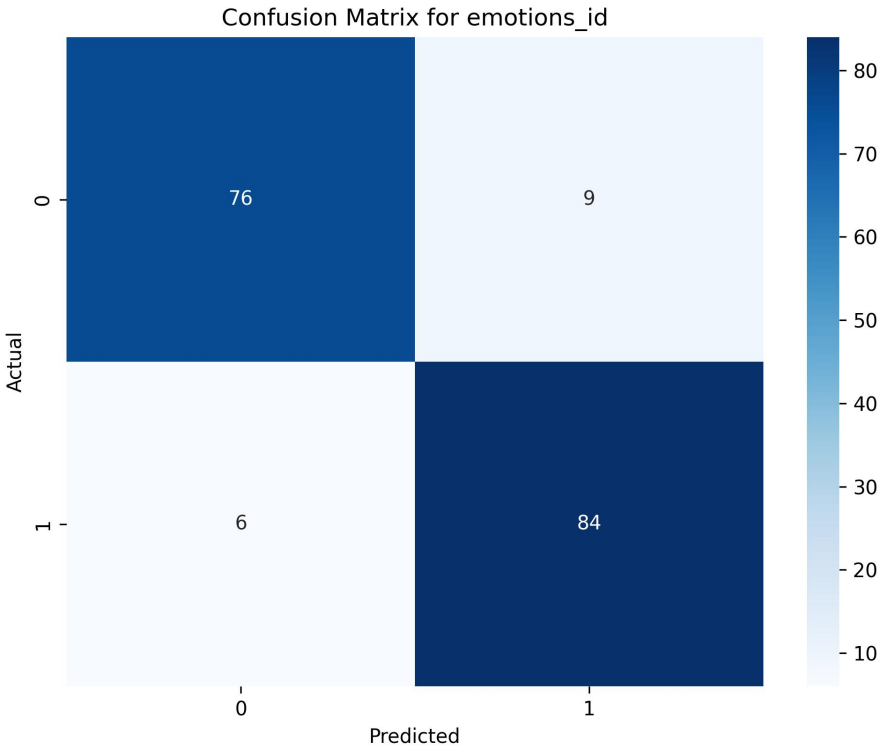
The interaction between *unemployment* and *love*, in the multilayer regressor model. A slight negative relationship is observed.

Figure 8: Partial Dependence Plot - Sadness



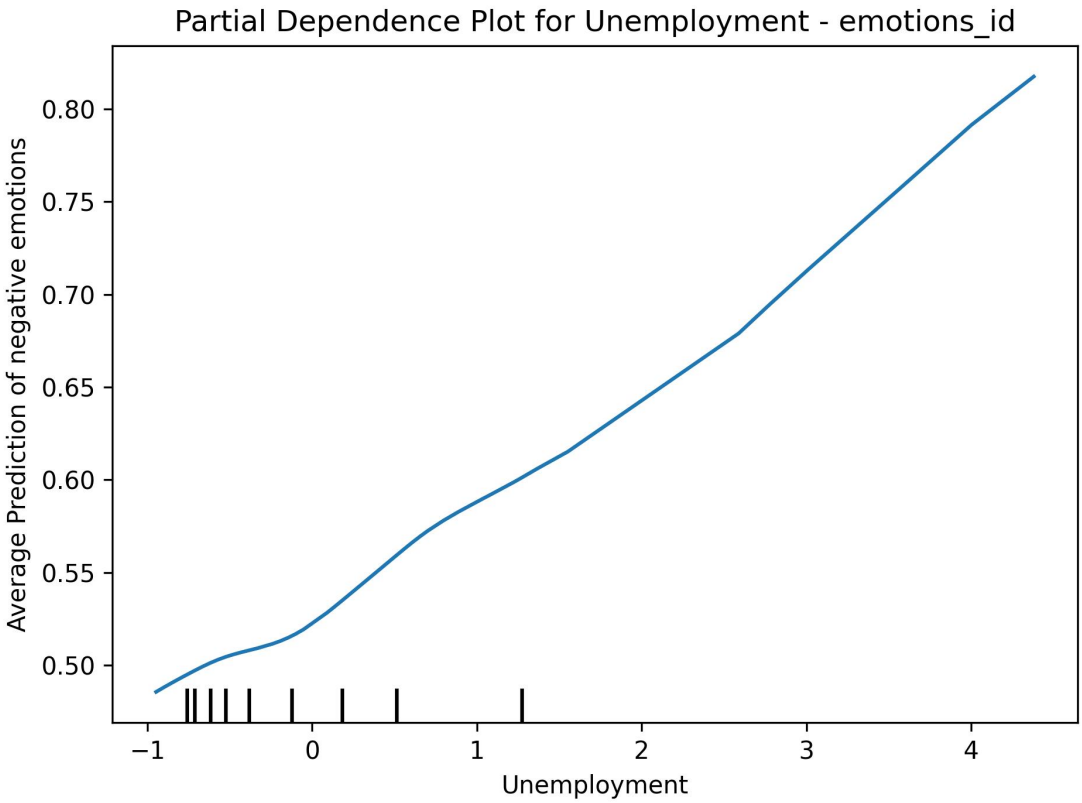
The interaction between *unemployment* and *sadness*, in the multilayer regressor model. A positive relationship is observed.

Figure 9: Confusion Matrix - MLP Classifier for *emotions_id*



This model has overall 91% accuracy. Predictions for negative emotions seem to be more accurate than that of positive emotions.

Figure 10: Partial Dependence Plot - *emotions_id*



The interaction between *unemployment* and *emotions_id*, in the multilayer classification model. A strictly positive relationship is observed.