Interim report

Tarpan Rai

BD08: NLP machine learning to better understand public knowledge, perception, and behavior for brain health during the COVID-19 pandemic.

Contents:

Overview:

The pandemic has had a major impact on mental health and wellbeing factors such as brain health, diet, and physical fitness. In this project, we will be using natural language processing (NLP) method and data analytic methods to understand the responses to a questionnaire done by the Centre for public health at QUB.

The data set includes unstructured, and free-text responses by participants where they describe aspect of brain health and diet, as well as other aspects of wellbeing impacted by the COVID-19 pandemic. It has 19 question total divided into 4 different main categories: Demographic variable, General covid questions, covid and brain health, Covid and diet. The dataset has 6818 total difference respondents but only 1124 and 1217 responded to the 2 free text question responses respectively.

Example of question:

- Demographic Variables Question Examples:
    - Age: (16-39, 40-49, 50-65, 66-74, 75+)
    - Country (Northern Ireland, Scotland, England)
    - Gender (Male, Female, Non-binary, I'd prefer not to answer this question)
- General COVID Question Examples:
    - Do you feel impacted by COVID?
    - Have you had to check for COVID?
    - Have you received treatment for COVID-19?
- COVID and Brain Health Question Examples:
    - Has the ongoing pandemic impacted your brain health?
    - Do you feel your brain health has improved or deteriorated?
    - We are interested to know why your brain health might have deteriorated during COVID-19. Please provide a reason for this using the free text box below.
- Covid and Diet Question Examples:
    - Has COVID-19 impacted your dietary habitats?
    - Do you feel that your diet has improved or deteriorated?
    - We are interested to know why your dietary habits might have deteriorated during COVID-19. Please provide a reason for this using the free text box below.

Example of free-text responses:

- Question: We are interested to know why your brain health might have deteriorated during COVID-19.
    - “Being at home, less able to converse with friends, anxiety re being out shopping etc.”
    - “Lack of exercise and more sitting”
    - “I’m now very isolated and getting a bit forgetful, nothing serious”.
    - “I’ve been unable to see family and friends, and this is really difficult and makes me really sad”.
- Question: We are interested to know why your dietary habits might have deteriorated during the COVID-19 pandemic.
    - “Eating a lot more rubbish, i.e., chocolate and biscuits!”
    - “Out of routine with diet and exercise”
    - “I’m at home more, and so treat myself to less healthy foods”.
    - “I’m drinking more alcohol because I am not free to do the things I want to do.

Tools used:

- NumPy: Scientific computing package in python that provides a multidimensional array object, various derived objects, and other operation on array.
- Matplotlib: Data visualization and graphic plotting library for Python.
- Pandas: Python library for data analysis consisting of the previous NumPy and Matplotlib library.
- SciPy: Scientific and mathematical library
- Csv: The files used will be in csv format hence the csv library is included
- Urllib: URL handling library
- PyTorch: PyTorch is a machine learning framework.
- Cuda: Parallel computing platform and programming model developed by NVIDIA that enables the program to use NVIDIA GPU to speed up computing intensive application.
- Hugging-face/Transformer: Python library used for building NLP systems.
- Genism: Open-source python library for representing documents as semantic vectors.

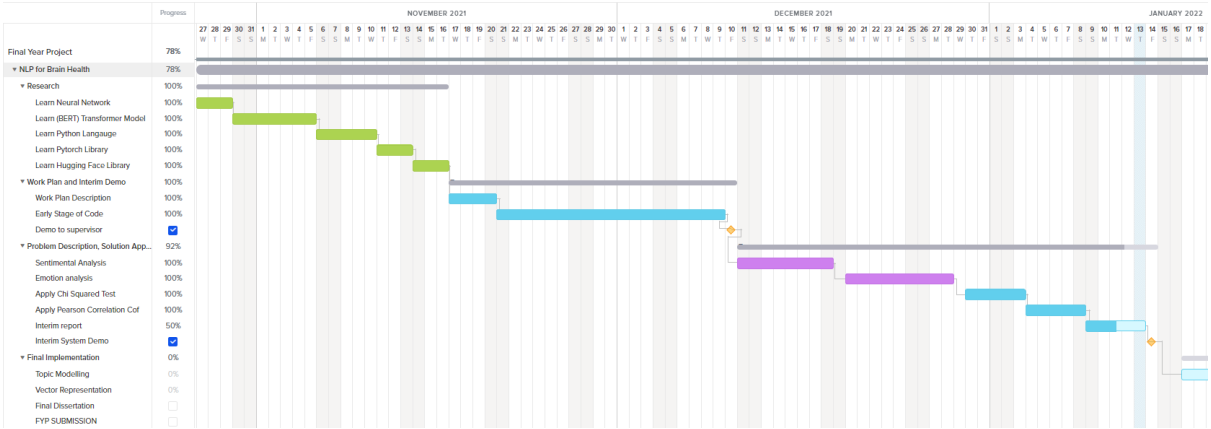


Figure 1: Update Gantt Chart

Progress:

The first few weeks was used to learn the necessary knowledge required to undertake the project:

- Neural Network
- What are transformer models (BERT)
- NLP
- Python and libraries
- PyTorch

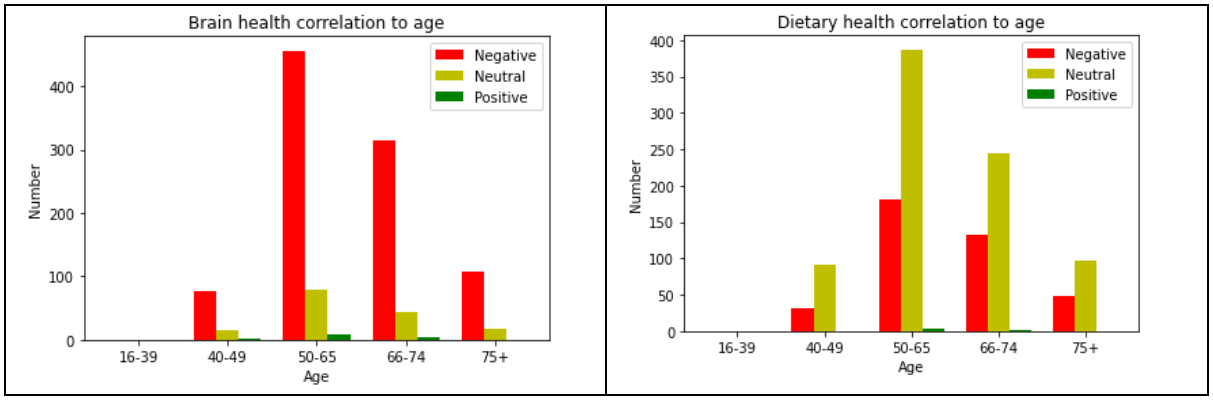
For this project, we will be using the BERT model from Cardiffnlp as stated in minute 3.txt. The original plan was to use the BERT model from Venelin Valkov but due to complications with the model the model was switched to the current one.

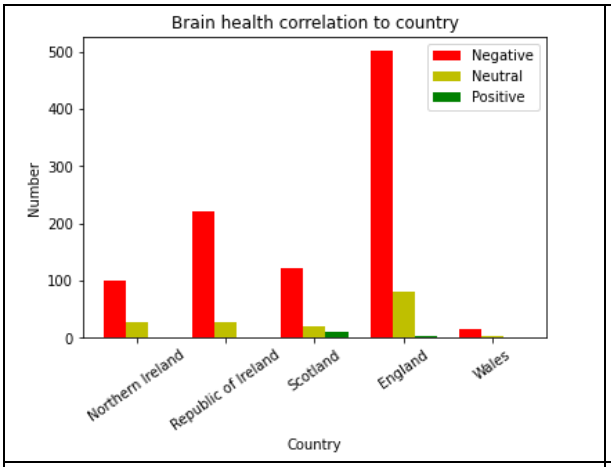
Code:

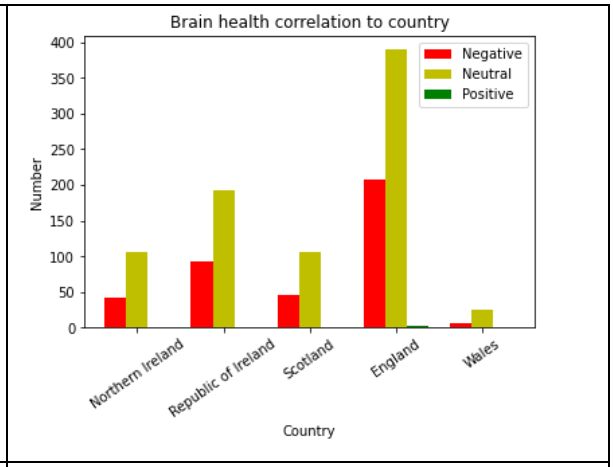
Before putting the file through the model, the .xlsx file is converted to csv to make use of the csv library. The main file consists of 2 different columns of free text responses with different question, it is separated into 2 different files, 'Brain full.csv' and 'Dietary full.csv'. The file then goes through a loop that applies the sentiment/emotion model to each row and count the total number of negative, neutral, and positive for sentiment and anger, joy, sadness, and optimism for the emotion. Using the results obtained, it is then compared to the different demographic variables in the survey e.g., age, country, gender. The results are then plotted into a graph using the matplotlib library in python.

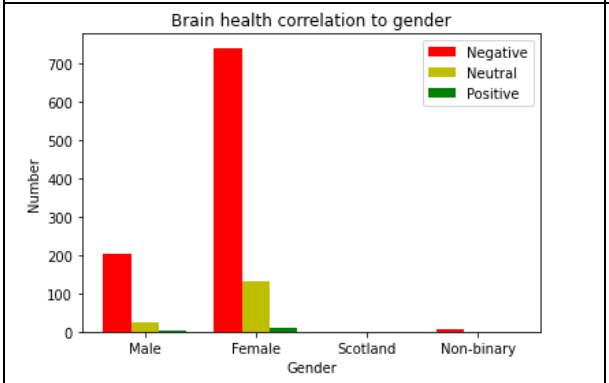
Results:

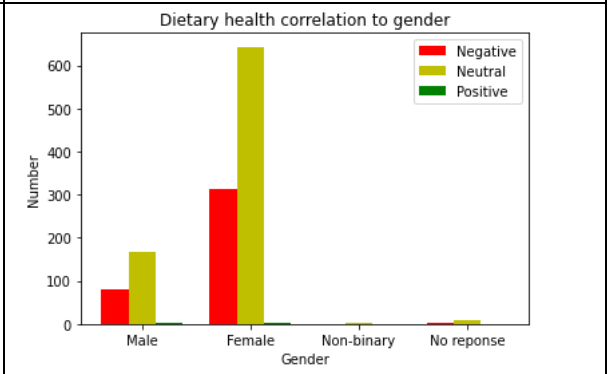
Sentimental analysis in correlation to demographics results

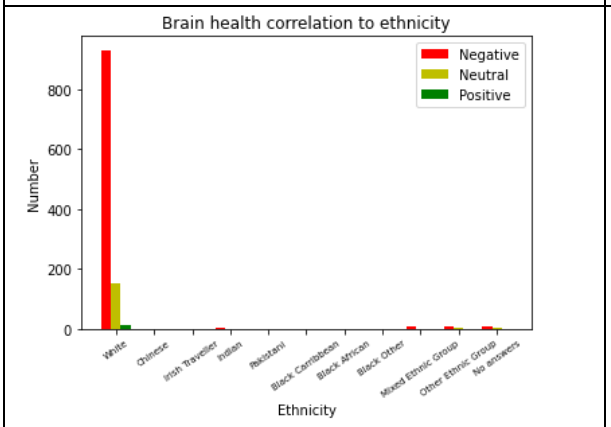


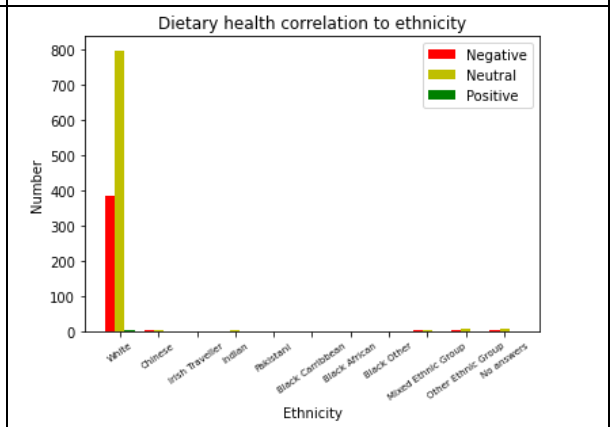

Brain health correlation to country
Negative
Neutral
Positive
Number
Northern Ireland
Republic of Ireland
Scotland
England
Wales
Country

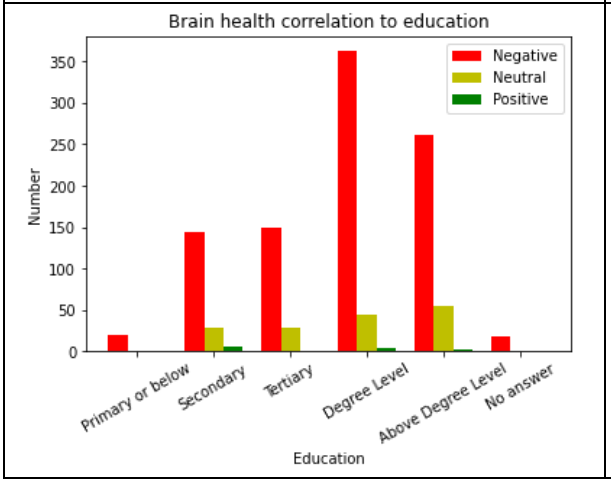

Brain health correlation to country
Negative
Neutral
Positive
Number
Northern Ireland
Republic of Ireland
Scotland
England
Wales
Country

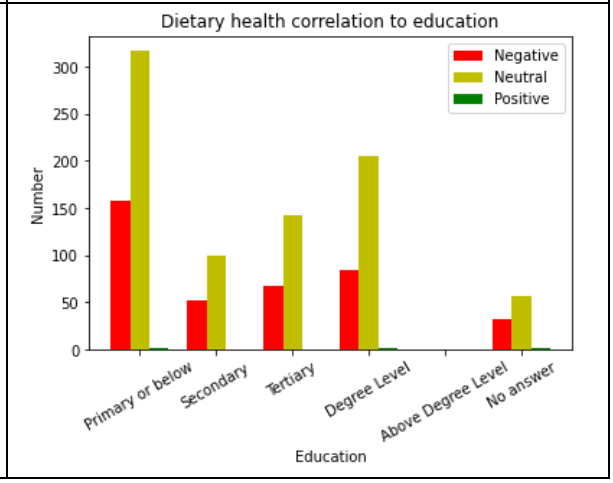

Brain health correlation to gender
Negative
Neutral
Positive
Number
Male
Female
Scotland
Non-binary
Gender


Dietary health correlation to gender
Negative
Neutral
Positive
Number
Male
Female
Non-binary
No reponse
Gender


Brain health correlation to ethnicity
Negative
Neutral
Positive
Number
White
Chinese
Irish Traveller
Indian
Pakistani
Black Carribbean
Black African
Black Other
Mixed Ethnic Group
Other Ethnic Group
No answers
Ethnicity

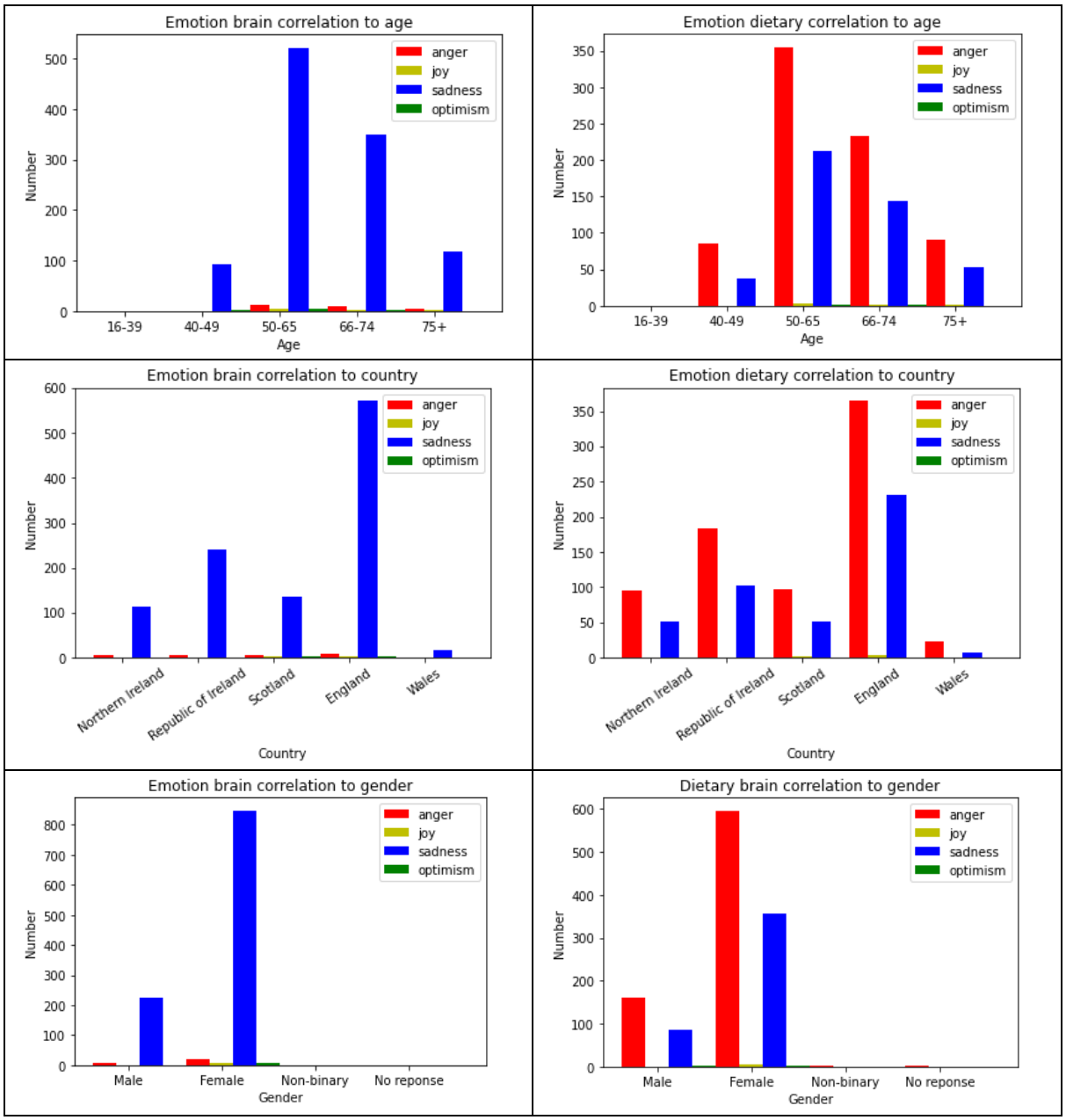

Dietary health correlation to ethnicity
Negative
Neutral
Positive
Number
White
Chinese
Irish Traveller
Indian
Pakistani
Black Carribbean
Black African
Black Other
Mixed Ethnic Group
Other Ethnic Group
No answers
Ethnicity

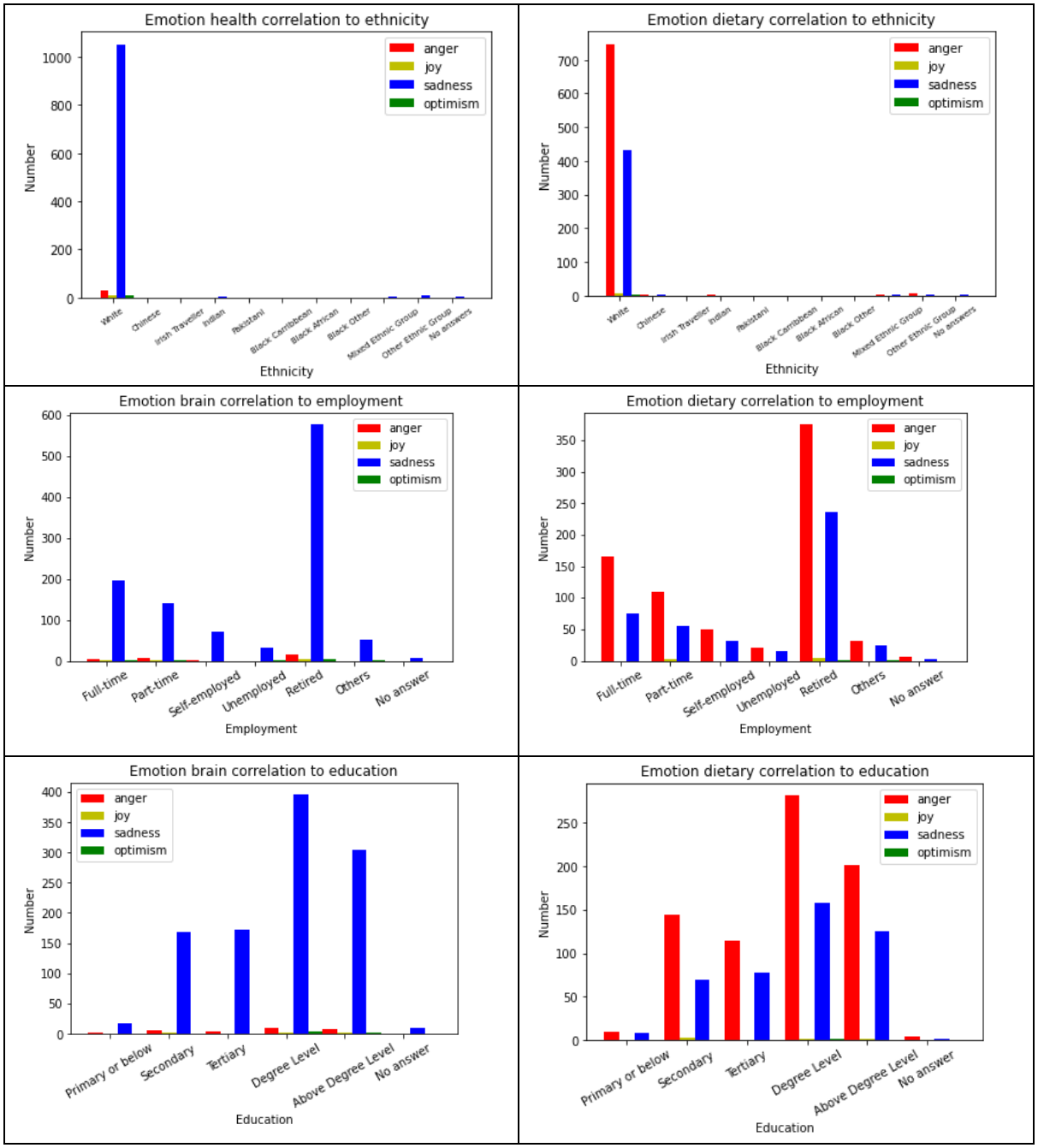

Brain health correlation to education
Negative
Neutral
Positive
Number
Primary or below
Secondary
Tertiary
Degree Level
Above Degree Level
No answer
Education


Dietary health correlation to education
Negative
Neutral
Positive
Number
Primary or below
Secondary
Tertiary
Degree Level
Above Degree Level
No answer
Education

As shown in brain health sentiment correlation, most of the responses were negative, however dietary health sentiment correlation were mostly neutral. There were little to no positive responses to both graphs.

Emotion analysis in correlation to demographics results





As shown in brain health emotion analysis, brain health responses were mostly sadness while dietary health was a mix of anger and sadness. Both had little to no positive emotion such as joy or optimism.

Pearson Correlation Coefficient: To compare the correlation between brain health sentiment and dietary health sentiment the Pearson correlation method is used. For the sentimental analysis the correlation coefficient is 0.116 which implies that there is little to no correlation between brain health sentiment and dietary health sentiment. For the emotion analysis, the correlation coefficient is 0.269 which show little correlation.

Future work:

Train owns sentiment model: Use the obtained sentiment and train/test own model. Cross check accuracy with the model currently used.