The Influence of Social Factors on Mental Health and Wellbeing during the COVID-19 Pandemic

Alex Sartain
Lancaster University
Lancaster, United Kingdom
https://orcid.org/0000-0003-3720-1298

Mahmoud El-Haj
Lancaster University
Lancaster, United Kingdom
https://orcid.org/0000-0002-6136-3898

***Abstract*— This study uses Natural Language Processing and Machine Learning techniques to understand the effect of the COVID-19 pandemic on mental wellbeing. We considered different user groups and locations in the USA to analyze the influence contrasting social factors, such as political stance, had on wellbeing. We measured the mental wellbeing of the social media users through understanding negative sentiment and mental health topic discussion in** *Twitter posts added by users from the top 10 Democrat and top 10 Republican cities in the USA.* **To measure the topic discussion, we used a mental health keyword list and developed machine learning models to classify the topic of a tweet. The primary findings include the similarity of the effect the pandemic had on Republican and Democrat cities when considering a timeline of tweets, whilst an increase in 'Anxiety' was discussed across different user groups and cities. Enforcement strategies had an influence on mental wellbeing with results differing for Republican and Democrat cities. An accurate text classifier was developed and used to categorize tweets into different mental health topics. The results showed how medical and unemployed users discussed topics like 'anxiety' and 'depression' more than a control set of users. The best machine learning model was developed using a Decision Tree algorithm which achieved an accuracy of 87% on unseen data.**

I. Introduction

This study aims to understand how mental wellbeing has been impacted by COVID-19 and assess the influence of different social factors. Investigating the social factors builds a better understanding of how mental wellbeing may be impacted in the future in a similar scenario such as an epidemic. The social factors considered in this study are:

- Political stance of users in a geo-location.
- Enforcement Strategy (State of Emergency, Lockdown order, Mask advice, Mask requirement, Phase 1 Re-opening, Phase 2 Re-opening, Phase 3 Re-opening).
- Employment Status (recently unemployed, medical user, control user). The control users are users selected without using a specific search term.

Since we use Natural Language Processing (NLP), completing the study would become more complicated when working with tweets of different languages across different nations. To allow the comparison of different enforcement strategies in a common language, the USA was selected as the location. Enforcement was distinct across states and cities whilst 95% of the population speak English (USA Census Bureau [1]). Politically, the USA has two dominant political parties allowing for a simplified comparison of political influence.

The work is important for mental health in the USA as before and during COVID-19 mental illness within adults has been increasing [2]. Reports show 70% of people stated 'isolation' and 'loneliness' have impacted their mental health, a cause of some COVID-19 enforcement strategies. A 2016 study comparing the USA with high-income countries showed emotional distress such as 'anxiety' was found in 26% of participants, nearly 10% more than the UK at 17% [3]. The study also showed the USA had a consistently increasing suicide rate which was the highest of the 11 high-income countries sampled.

The COVID-19 pandemic has had unprecedented effects on the global population. In the USA some control was delegated to states and cities allowing for discussion of enforcement strategies across locations. The same cities can be used to compare political stances on mental wellbeing based upon their voting records from the 2016 election. To consider the impact of COVID-19 generally on wellbeing, datasets from before the pandemic were collected as baselines.

Different user groups were also found to compare employment status, with a control user group used for comparison.

It is important to consider positivity in the language used and the discussion of mental health topics for measuring well-being. To measure positivity, sentiment analysis was used. For the discussion of mental health topics, a mental health keyword list was used initially before a machine learning classifier was built to improve accuracy and speed. For analysis, patterns were found and compared with external pandemic data such as COVID-19 case rates and unemployment rates.

The data-gathering stage considers large user groups and a wide range of data to find and acknowledge new patterns of mental wellbeing. We avoid finding data from small subsets of users to analyze the effect of the COVID-19 pandemic on a wider population. The development of a mental health topic classifier is used for finding topic discussion in this study and can be used in future work too. The findings and analysis of the research assess the effect of different social factors on mental wellbeing and can be used for future planning and decisions made by policymakers. In this work, we explore whether the political stance of a city influences the mental wellbeing of the population during a global pandemic using NLP and machine learning. We also explore how a global pandemic has affected the wellbeing of people using NLP and machine learning techniques. Through the use of NLP and Machine Learning, the study helps in understanding how different enforcement strategies have affected the wellbeing of a population during the global pandemic. Finally, we explore the different mental health topics discussed during the global pandemic to understand how people's wellbeing has specifically been affected using NLP and a machine learning classifier and how employment status affects people's mental health during a pandemic.

II. Related Work

Previous research related to COVID-19 by Low et al [4] investigated how users with different types of mental health problems were impacted as COVID-19 cases rose. They used pre-classified data (based on subreddits) which avoids finding out how mental health topic discussion has entered mainstream discussion, which we investigate in the study.

It was expected and has also been shown that the mental health impact of COVID-19 has led to increased uncertainty and potential symptoms of mental health conditions as shown by Pan et al. [5]. The study explored mental health symptoms through questionnaires during the pandemic which made it difficult to consider how mental health symptoms or wellbeing changed from pre-pandemic. Our study avoids using a questionnaire to gather data from previous years and consider the change in mental wellbeing during the pandemic.

Study [4] explored how sentiment analysis provides quick and useful insight when considering mental wellbeing using social media data, therefore in this study sentiment analysis was used.

Within the COVID-19 pandemic domain, Alamoodi et al [6] considered epidemics and COVID-19 in relation to sentiment analysis. They highlighted the benefit of information dissemination and public broadcasting in reducing the spread of fake news and panic. The opinions and broadcasts from high-status people such as Donald Trump, therefore, are likely to affect the result data.

When considering mental wellbeing, mental health is a major topic to consider. Cohan et al [7] explored Reddit posts from users identifying with a mental health condition against a control group to examine and identify the users with a mental health condition. The study used a keyword list of mental health conditions and terms which was used in our study to find tweets related to mental health within the datasets collected. This study avoids longer social media text and looks at shorter text using Twitter data. Research by Coppersmith et al [8] and Jaidka et al [9] used Twitter data and highlighted the usefulness of Twitter as a data source.

Twitter data was also used by McClellan et al [10] as they monitored the trends of mental health topics. Their research found the volume of mental health tweets spiked around a behavioral health event or unexpected event. The spike would drop and return to a regular level after around two days. The COVID-19 pandemic was unexpected but has lasted a lot longer than two days therefore the pattern is expected to be different and worth analyzing in this study.

Our study aims to find mental health discussions without using search terms to analyze changes to mental health topic discussion at different points during the pandemic.

Using NLP, there has been difficulty classifying mental health topics. An example is with Fink et al [11] who had to separate disruptive terms from the dataset collection. This would be useful for the training of a classifier but would produce incorrect results in practice. Within the context of COVID-19, the terms mentioned in [12] such as the 'Great Depression' are common due to the increased unemployment. This allowed us to find and consider exceptional cases when training the classifier and measuring mental health topic discussion.

When collecting and labelling the mental health topics, keywords are used. This is a common task used in social media data mining [13, 14] and allows for detection of events or indicates the post context, in this study mental wellbeing.

Our data collection also selects user groups at times, such as medical users, using self-reporting techniques first suggested in 2014 [15] to select related users quickly.

III. Techniques

For the experimental work, we used text classifiers combined with a mental health keyword list. Fig. 1 shows the general workflow of the entire study in gathering and analyzing the tweets.

The snscrape[1] library was used to gather tweets. The collection of data is discussed in more detail in the Data Collection section.

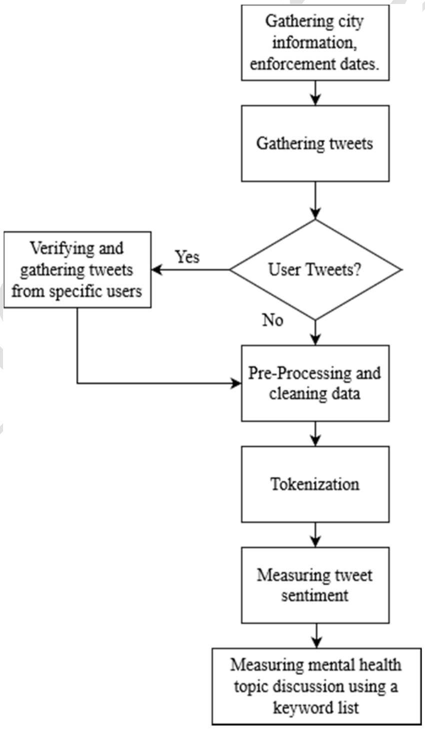

Fig. 1
The general workflow in gathering and analyzing tweets

The tweets were pre-processed and cleaned removing noisy data like URLs, emoji's and usernames with the hashtag symbols being removed too. The SpaCy[2] library was then used for tokenization, splitting the tweets into individual words for analysis. The TextBlob[3] library was used for sentiment analysis to give a polarity score of -1 to 1 which can be separated into: negative, neutral, and positive classifications. The negative tweets associated with lower mental wellbeing were considered for this study. The presence of a mental health topic was then checked for each tokenized tweet to measure mental health topic discussion.

[1] https://pypi.org/project/snscrape/

[2] https://spacy.io/

[3] https://textblob.readthedocs.io/en/dev/

The Fig. 2 shows the workflow for producing the mental health topic classifier. A key stage is the removal of stop words, common words in text data that have a low influence on the classification. The NLTK[4] corpus was used since it provides a pre-defined list of stop words.

When training the model, the text needs to be vectorized, for this the Bag-Of-Words model is often used. The BOW model only considers the words and no context, useful within the classification for mental health keywords like 'depression'. To improve accuracy, a TF-IDF rating can be produced, which considers a percentage of the frequency of a term in a piece of text (tweet) compared with the frequency of a term in the dataset corpus. This allows words to carry weights and highlight the relevance of a word in a tweet and whether it is relevant to a specific category such as 'depression'.

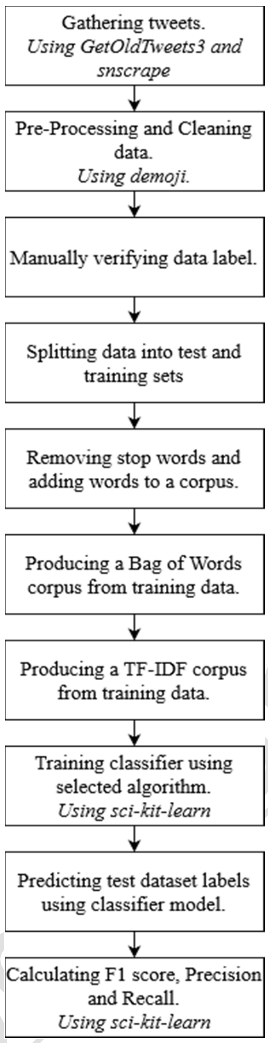

Fig. 2
The general workflow in training the classifier

Lemmatization is a method used to reduce related forms of a word to a single form and was beneficial to produce a classifier. Within the domain of this study and data collected it can relate to the words: depress, depression, depressed all being converted to the base form: depress. This was completed when the words were added to the corpus as shown in Fig. 2.

For topic classification, there are many machine learning algorithms used in NLP which provide useful solutions. Research into the 'best' algorithm is often flawed since the dataset affects how the algorithm works. Therefore, for the training of the classifier in this study it is worth considering a range of algorithms. For any machine learning algorithm selected, the Sci-Kit-learn[5] library was used as it provides a simple way to train a set of data and produce a machine learning model within Python and quickly train the data on a different algorithm.

Word Clouds were also used in the study for analysis. They provided additional information for mental health topic discussion and understanding the training datasets.

IV. Data Collection

A. Twitter

To get a large amount of data voicing the opinions of users that would be readily available and easy to filter microblogging data was chosen.

Twitter was chosen as the selected platform since it was simple to collect similar data using different search terms as the length of every tweet is limited to 280 characters. Twitter also allowed for geotagging, accurately determining the location of a user's tweet. This was important when considering political stance, where the location of a tweet was measured.

A number of 20 cities were selected to analyze the influence of political stance. This allows for comparison between Democrat and Republican cities by choosing the 10 cities who voted Democrat and Republican in the 2016 election with the largest populations respectively. Cities with the largest populations were considered since there would be more data available when filtering by location on Twitter. Table 1 shows the cities alongside the political party they represented and tweets collected for some of the data collected.

B. What data was collected?

For the analysis of sentiment and mental health topic discussion in tweets, data needed to be collected during and before the pandemic. A set of data also needed to be collected to allow for the training of a classifier that contained mental health keywords and a set of generic control tweets. The control tweets were collected without using a search term whilst COVID-related tweets used COVID-19 search terms. It was important to collect control data to allow effects brought on by the pandemic to be analyzed and ensure that tweets unrelated to mental health topic discussion were ignored. To measure employment status, self-classification techniques were used to find users of varying employment statuses. The technique used is taken from study [15] when selecting users with mental health conditions. The initial user collection tweets were human analyzed to ensure the tweet was a valid statement related to the data collected. When considering a medical user, the user needed to explicitly label themselves a nurse/doctor in the tweet to be valid, whilst not labelling themselves as retired. When considering a user who

[4] https://www.nltk.org/

[5] https://scikit-learn.org/sTable/

had recently become unemployed, the tweet needed to be a valid statement to suggest the tweet author had lost their job at the time of the tweet.

Tweets were then collected during the pandemic for medical and control users whilst for unemployed users, tweets were collected 3 months before and after announcing the loss of their job.

Collecting data for enforcement strategies required the use of government websites to get the correct dates for collecting the data. The date of the announcement was used, rather than the date something was put in place since reactions were discussed when an announcement was made.

The enforcement strategies were as follows:

- Lockdown: A "Stay-at-home" order placed by the governing body.
- Mask Guidance: When a government body recommended wearing a face covering.
- Mask Enforcement: When a government body enforced wearing of masks such as through fines.
- Phase Re-opening: When a government used a traffic light tier system of re-opening or if such a system wasn't used: 3 dates restrictions were lifted incrementally.

After collecting the data, the raw data needed to be transformed into something useful for analysis.

City	Party	#Tweets
Austin	Democrat	19,350
Chicago	Democrat	32,391
Dallas	Democrat	31,416
Houston	Democrat	29,492
Los Angeles	Democrat	18,095
New York	Democrat	21,345
Philadelphia	Democrat	34,420
Phoenix	Democrat	22,676
San Francisco	Democrat	14,985
San Jose	Democrat	38,253
Colorado Springs	Republican	33,000
El Paso	Republican	4,172
Fort Worth	Republican	31,350
Fresno	Republican	12,920
Jacksonville	Republican	3,674
Mesa	Republican	22,737
Miami	Republican	33,038
Oklahoma City	Republican	9,345
Omaha	Republican	4,512
San Diego	Republican	10,095

Table 1

Cities used in the study with the number of COVID-19 tweets collected from January to March 2020.

V. Methodology.

A. Sentiment Analysis

For the analysis of the data, only negative tweet sentiment was used to understand how mental wellbeing was affected due to the pandemic. As sentiment reflects an opinion, negative sentiment highlights negative opinion which corresponds to lower mental wellbeing which is being investigated in this study. The sentiment was collected for each tweet and the percentage of tweets with a negative sentiment were found across the datasets.

B. Mental health topic analysis

When considering the analysis of topics in tweets, a keyword list provides a simple way of finding the discussion of certain keywords. In the case of mental health topics, this was useful since keywords such as depression or anxiety reference a mental health topic in a tweet.

Using a keyword list [7] makes it simple to analyze the dataset with high accuracy since only tweets containing the relevant keyword will be classified. All words in the tweets and keyword list were converted to lower case so capitalization in the tweet did not affect the presence of a keyword. The list contained misspellings to account for the social media dataset which is often informal.

There are difficulties with using a keyword list as considering the appearance of a single word in a tweet does not give the full context or meaning. Two common examples are with the discussion of depression and bipolar, some users tweeting about depression were referencing the 'Great Depression' or an 'Economic Depression' whilst users tweeting about bipolar were often relating to the weather which were common exceptions found in [11].

Avoidance of the exceptions was attempted through checking the immediate words surrounding the words 'depression' and 'bipolar'. Removing exceptions increased the time taken to complete the analysis whilst some tweets were still mislabeled since only the immediate surrounding words were checked. This led to the production of a mental health topic classifier to improve the speed and accuracy of topic detection.

C. Mental Health Classifier

For the consideration of mental health topic discussion during the pandemic, the aim was to understand the changes in mental wellbeing. Mental health topics such as 'autism' were of no relevance since they appear from an early age whilst topics such as 'anxiety' or 'depression' can be onset from any age and situation [16]. The mental health topics chosen for the classifier were: Anxiety, Bipolar, Depression, PTSD, and Suicide.

After analysis of the datasets was complete, an additional topic: 'Economic' was added to reduce inaccuracy associated with tweets referencing economic depression or recession.

The data was then split into training and test sets. When considering the quantity of each topic to add to the classification model, it is important to have as much training data as possible for more accurate machine learning models to be produced. A number of 150 tweets from each topic were added into a test dataset since the full dataset originally contained nearly 1,500 tweets on average for each topic (1,463) and this allowed for ~90% of the data to be used for training.

The keyword list used for measuring topic discussion was modified for use in the production of the classifier as shown in Table 2. Keywords were also added for the 'Economic' class like 'Great' or 'Recession' to ensure the 'depression' and 'economic' tweets were accurately separated.

Anxiety	Bi polar	Bipola
Bipolar2	Bipolari	bipolarii
Depresion	depressed	Depressiion
Ptsr	Ptss	Socialphobic
weather	Recession	economy
Bipolar	Bi-polar	Bipolar 1
Bipolars	Bipolat	Bi-poler
depressio	depressiom	Mdd
Suicidal	Suicide	Suicides
Bipolar 2	Bipolar1	Ptsd
Depression	Depres	Great
Pstd	Economic	

Table 2

Keyword list for the mental health classifier

The training dataset was shortened further from 90% of all tweets using the sci-kit-learn library to shuffle the dataset and allow 10% of the tweets for evaluation of the model. The 10% evaluation was used to understand if the training dataset and model had a high accuracy before making predictions. A dataset previously collected relating to 2020 COVID-19 El Paso tweets were used for validating the classifier to avoid over/underfitting with the location selected at random.

The tweets were machine labelled as they were collected, based upon the keyword search term. The tweets were then human-verified to ensure there were no errors made. For some tweets, more than one mental health topic was mentioned which disrupted the labelling. In this scenario, the label selected when collecting the tweets was used to avoid adding bias where more than one mental health topic was discussed. This means the accuracy is not expected to be perfect when testing and verifying the classifier too.

After the tweets were labelled duplicates were removed to avoid any bias in the training data. Therefore, the training dataset contained 8,631 tweets. A breakdown of the number of tweets for each label is shown in Table 3.

The quantities of tweets for different labels are not perfectly weighted, this is so as much data as possible could be trained. The Results section tested the usefulness of the training dataset and ensured that no adjustments were required.

As highlighted in the techniques section, the bag of words model was initially used for vectorization. There were some additional design considerations such as vector size. Allowing too few words would result in difficulty making predictions whilst allowing too many words would increase the time spent training the model and include words that have little relevance. For this reason, two corpora were created, one containing a refined keyword list and one corpus containing words of a frequency above 15. The value of 15 was selected since this was around 1% of the training tweet set for a specific label and suggested that if 1% of tweets for each label contained the word then the word was important. The mental health keyword list was also appended to the frequency corpus to avoid ignoring keywords.

The dataset was trained on unigram words, which considers individual words only for the vector, so only where there was a single word entry in the keyword list would the keyword be appended to the corpus.

In this study, it is highly useful to use TF-IDF for vectorization. This gives the mental health keywords a larger weight for the classification decision.

Label	**Quantity of tweets**
Anxiety	993
Bipolar	768
Control	1586
Depression	1689
Economic	1084
PTSD	1022
Suicide	1487
Total	**8631**

Table 3

Quantity of tweets for each topic class in the training dataset.

D. Measuring the classifier effectiveness

To analyze the effectiveness of each model: Precision, Recall and the F1-Score were calculated for each corpus and algorithm.

The recall of the control label was especially important for the classifier since one aim was to reduce the pitfalls of the keyword list method. If the recall of the control label was high, there would be fewer non-mental health tweets labelled as a mental health topic.

The F1 score and precision values were useful for comparing the corpora and algorithms to make a distinct decision when performance was similar. If they were low across all corpora or algorithms, the training dataset itself would then be modified.

When searching for high accuracy it is important to realize that there are limits within a dataset and a stopping point needs to be reached, for this experiment the validation set highlights the stop point. The range of F1 scores for different models on the 2020 COVID-19 validation dataset was 57 - 87%. The middle of the range becomes 72% to suggest that an F1 score above 72% can be considered effective for making predictions.

E. Topic classifier experiment

As highlighted in the Methodology, the test dataset contained the same 150 tweets for each label. Both the test and training datasets were split 5 times to understand how the accuracy of the model changed as more labels were added. This verified whether there was a limit to the number of labels or a label that enabled too much confusion with the dataset.

The first labels compared were: Depression, Economic and Control. This was to understand if ‘economic’ and ‘depression’ tweets could be distinguished. Including a ‘control’ dataset was especially important to ensure the classifier didn’t place every tweet into a mental health topic. Following this, the labels added were: Anxiety, PTSD, Suicide and Bipolar with no order considered.

Both the keyword and frequency corpora were initially tested using Random Forest as the algorithm since it was quick and simple to test and validate.

After the keyword corpus performed better it was selected. Different algorithms were then tested using this corpus: Naïve Bayes, Random Decision Forest, Decision Tree and SVM. Comparing algorithms is useful to find the algorithm best suited to the data but also for analyzing the training dataset by ensuring accuracy is relatively high for each algorithm.

Deep learning techniques such as Convolutional Neural Networks[17], tend to require a large amount of data, whilst the machine learning dataset in this paper is small so the dataset was more suited to a less complex algorithm. [18]. With a larger dataset, the use of the keyword list corpora could also have been removed as this may have introduced some bias.

VI. Results

The following sections highlight the training of the mental health topic classifier alongside the results. The data was collected for this study, therefore there are no other studies to make comparisons with. Two baselines were created to compare the different algorithms and datasets used. Baseline 1 uses the control class to see the accuracy of the control group in relation to a separate mental health label. Baseline 2 uses the accuracy of all labels with the naïve bayes algorithm.

Label	Accuracy	Precision	Recall	F1-Score
Control	0.95	0.82	1	0.90

Table 4
Accuracy of baseline 1 comparing the control class with a mental health topic class.

Label	Model	Precision	Recall	F1 Score	Support
Anxiety	Naïve Bayes	1	0.88	0.94	150
Bipolar	Naïve Bayes	0.97	0.88	0.92	150
Control	Naïve Bayes	0.05	0.01	0.02	150
Depression	Naïve Bayes	0.37	0.91	0.53	150
Economic	Naïve Bayes	0.96	0.66	0.78	150
PTSD	Naïve Bayes	0.99	0.89	0.94	150
Suicide	Naïve Bayes	0.99	0.93	0.96	150
Macro Average	**Naïve Bayes**	**0.76**	**0.74**	**0.73**	**1050**

Table 5
Accuracy of baseline 2 comparing the control and mental health labels using the Naïve Bayes algorithm.

Baseline 1 has a high accuracy; however the precision is fairly low suggesting that mental health topics are placed in the control label. In this study, it is also important to measure individual mental health topics therefore the second baseline was developed. Baseline 2 uses the Naïve Bayes algorithm to assess a higher number of labels, as the precision and recall are low for certain classes more algorithms were later tested.

A. Analysing the Results

Firstly, as labels are added, the macro-average F1 score improves from 0.85 to 0.90 and 0.87 to 0.90 in the keyword and frequency corpus respectively. This suggests no labels need to be removed and there is high accuracy for each label in the dataset.

Table 5 highlights the lower precision for the 'Depression' class and lower recall for the 'Economic' class set compared with the other classes. The initial concern is the 'depression' precision being low to suggest many tweets are labelled as 'depression' when they are not. Meanwhile, 'economic' tweets are being labelled another class due to the lower recall. It would be a valid assumption to suggest this is since both classes are closely related and 'economic' tweets are often mis-labelled into the 'depression' class.

The frequency corpus doesn't improve the results much, a precision of 0.70 for the 'depression' class only improves to 0.72 and 0.68 recall for the 'economic' class only rises to 0.70. This issue may be explored individually in further work. After reviewing the word clouds for the 'depression' and 'economic' training data, both word clouds were mostly made up of keywords. This can suggest why the frequency corpus doesn't improve the accuracy much since both labels are mostly dependent on the keyword.

B. Validating the classifiers

When comparing both corpora, the validation set of tweets was also used. This dataset considers how useful the classifier would be in predictions. There was an overwhelming amount of 'control' tweets in relation to mental health topics, which would be expected in a general dataset.

Despite similarities in the test dataset, Table 8 shows the keyword corpus performs much better than the frequency corpus on the validation set, with an 87% F1 score compared with 35%. The keyword corpus has difficulty only with classifying 'economic' and 'control' tweets which in the context of a mental health topic classifier is irrelevant.

Once the keyword corpus was found to be the most usable in the task context, different algorithms: Random Forest, SVM, Naïve Bayes and Decision Tree were compared. These comparisons are shown in Table 7. As accuracy was very similar on the test dataset, the validation set was again used to make a distinct selection.

The Random Forest and Decision Tree algorithms performed best, with the Decision Tree algorithm performing slightly better at classifying control and economic tweets correctly. Either of the models could've been selected, however as mentioned in the Methodology, a higher recall in the control class is important when selecting the best classifier. Since the Decision Tree has a perfect recall of 1 compared with 0.99 with the Random Forest algorithm, the Decision Tree model was selected.

Label	Corpus	Precision	Recall	F1 Score	Support
Macro Average	Keyword	0.86	1	0.87	7580
Macro Average	Frequency	0.29	0.92	0.35	7580

Table 6
Comparison of the Keyword and Frequency corpora on the validation dataset.

Label	Model	Precision	Recall	F1 Score	Support
Macro Average	Random Forest	0.86	1	0.87	7580
Macro Average	SVM	0.77	1	0.79	7580
Macro Average	Naïve Bayes	0.72	0.84	0.57	7580
Macro Average	**Decision Tree**	**0.87**	**1**	**0.87**	**7580**

Table 7
Comparison of machine learning algorithms corpora on the validation dataset.

Keyword Corpus					
Label	Model	Precision	Recall	F1 Score	Support
Anxiety	Random Forest	1	0.92	0.96	150
Bipolar	Random Forest	0.98	0.91	0.94	150
Control	Random Forest	0.82	1	0.90	150
Depression	Random Forest	0.70	0.91	0.79	150
Economic	Random Forest	0.94	0.68	0.79	150
PTSD	Random Forest	0.99	0.95	0.97	150
Suicide	Random Forest	0.99	0.93	0.96	150
Macro Average	**Random Forest**	**0.92**	**0.90**	**0.90**	**1050**

Table 8
Keyword corpus performance on the test dataset with all labels.

The Results section has proved that the classifier built is accurate and useful for this data study. This allowed the classifier to be used for predictions for analysis in the next section. The next section also discusses the result data for sentiment analysis and topic discussion.

VII. Findings and Discussion

The following discussion attempts to find results to the aims outlined in the introduction. Throughout the discussion, the term "Mental Wellbeing" will be used of which the definition in the case of this study is outlined in Table 9.

A. *Influence of political stance of area*

Hypothesis: Users from Democrat voting cities have lower mental wellbeing than Republican users.

The negative sentiment of users from Democrat and Republican areas was very similar in tweets from 2019, 2020 and the enforcement stages. This may suggest opinion is shared equally during the pandemic. A slight difference can be observed in February 2020 as Democrat negative sentiment was higher than Republican cities. This correlates with President Donald Trump downplaying the COVID-19 risk [19] and links to related work suggesting that information dissemination affects mental wellbeing, [10].

There is an observable difference when looking at mental health topic discussion. Democrat areas have an average quantity of mental health keyword tweets around 0.006%, double that of Republican areas during the pandemic as shown in Fig. 4.

Mental Wellbeing	Negative Sentiment Quantity	Mental Health Topic Discussion Quantity
Lower	High	High
Higher	Low	Low

Table 9
Table of keys representing what lower and higher mental wellbeing stands for in the findings section

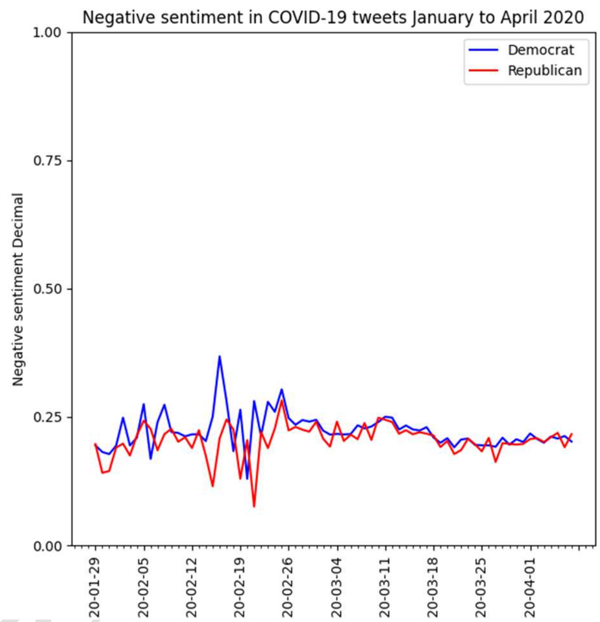

Figure. 3
Negative sentiment in COVID-19 tweets from January to April 2020.

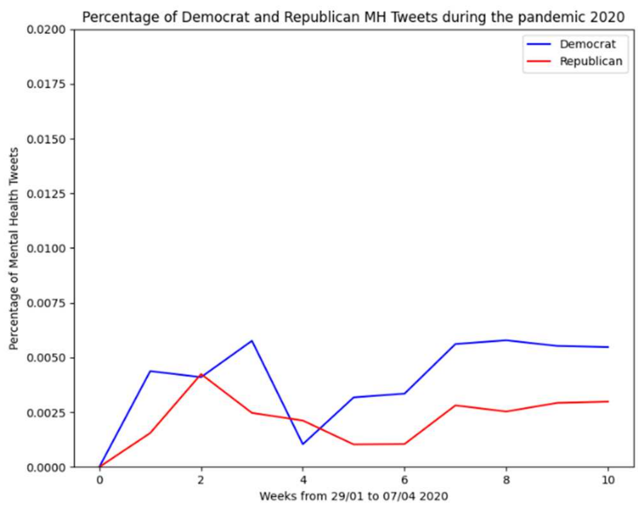

Figure. 4
Mental health topic discussion in COVID-19 tweets from January to April 2020.

In 2019 there is also more mental health topic discussion from Democrat cities rather than Republican. The word clouds of the two stances are dissimilar in 2019 but become more homogenized in 2020. One reason is the large amount of 'anxiety' discussion by Democrat areas in 2019. This can suggest that political stance has limited influence on mental wellbeing in a pandemic scenario, but it may be found through extra work that Democrat areas have higher anxiety than Republican areas.

B. Mental health topic discussion across user groups

This study aims to observe expected patterns such as medical and recently unemployed users experiencing lower mental wellbeing than control users. Unemployed users are at a higher risk of suffering poorer mental wellbeing as found by the charity Mind [20].

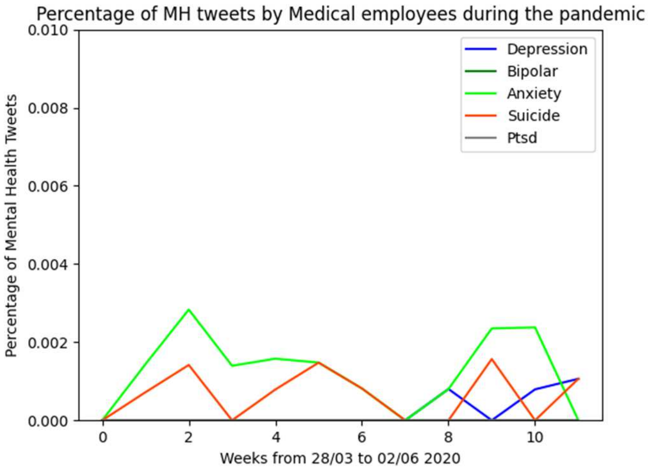

Figure. 5
Mental health topic discussion in Medical User tweets from March to June 2020.

Hypothesis: There will be a higher amount of mental health topic discussion in doctors & nurses (medical users).

Medical users discuss 'anxiety' and 'suicide' more than other mental health topics, as found by the classifier shown in Fig. 5. The pattern is observed in the COVID-19 case rates [21] to suggest that the case rates link to 'anxiety' in medical employees. PTSD was expected to be prominent, however the data was collected at the beginning of the pandemic where this is less likely to be discussed.

The control users (Fig.6) discuss mental health topics at distinct dates, linked likely to events in the pandemic, whilst medical users discuss mental health topics more regularly. This shows how common it was for medical employees to discuss mental health topics, in which discussion increased as COVID-19 case rates increased [21].

Hypothesis: There will be a higher amount of mental health topic discussion in recently unemployed users

Job loss user data was collected at the point of an 'I have lost my job' statement. Figures 7 and 8 represent mental health topic discussion for users before and after losing their job respectively. At week 10 before a user loses their job, 'anxiety' rates begin to increase which reflects the unemployment rates rising at a rapid rate [22]. Rates of 'depression' also increase which remain high after the user announces the loss of their job. There is then a drop in 'depression', 'anxiety' and 'suicide' discussion after 8 weeks which may highlight a cool-off period to feelings felt after losing their job. It may also highlight possible new employment opportunities as there is a drop-in unemployment from the peak unemployment rates observed in the USA in March.

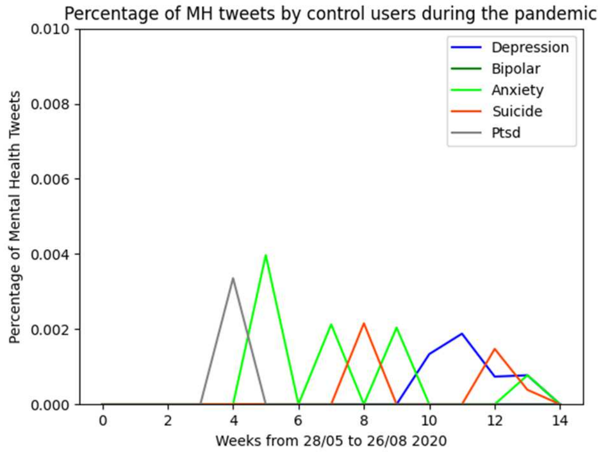

Figure. 6
Mental health topic discussion in Control User tweets from May to August 2020.

C. The unemployed users discuss mental health topics often, unlike the control users. A comparison between unemployed and medical users is the increased prominence of 'depression' discussion with the unemployed user data and 'anxiety' in the medical user data: Mental wellbeing when discussing COVID-19.

Hypothesis: There will be lower mental wellbeing in tweets when discussing COVID-19.

When comparing 2019 and 2020 mental wellbeing, a baseline with no search term was collected for both January to April 2019 and 2020. The baselines have a lower negative sentiment compared with the COVID-19 tweets. The same pattern is observed for mental health topic discussion too.

Within the COVID-19 tweets, it is important to link back to previous research [12] which suggested a 2-day drop after large events with mental health topic discussion. It is clear from the result data such as Fig. 4 that there is no 2-day drop showing how the global pandemic is distinct from other events which trigger an increase in mental health topic discussion.

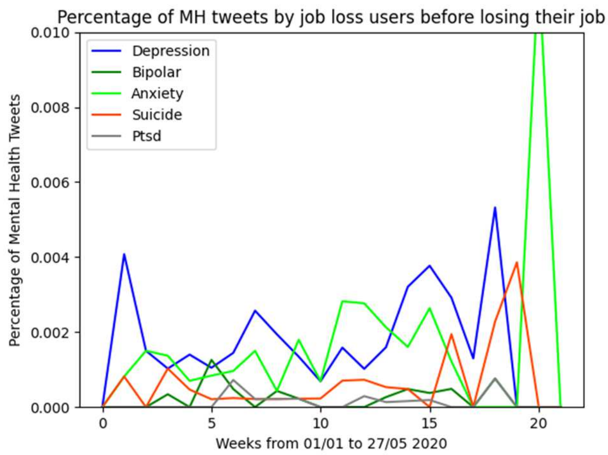

Figure. 7
Mental health topic discussion in Job Loss User tweets before losing their job from January to May 2020.

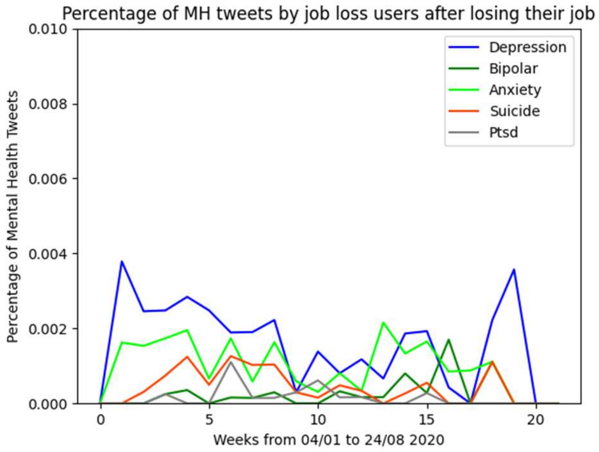

Figure. 8
Mental health topic discussion in Job Loss User tweets after losing their job from March to August 2020

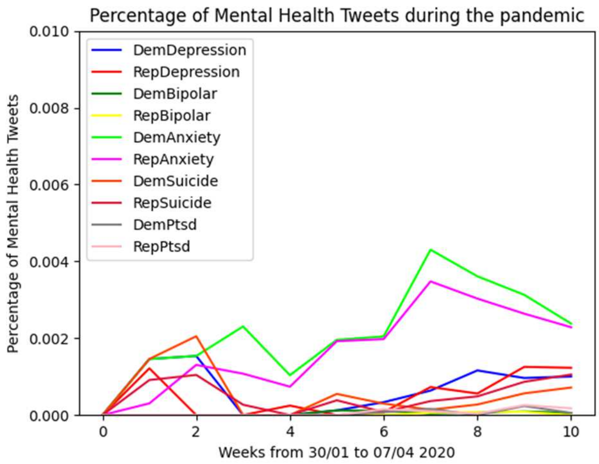

Figure. 9
Mental health topic discussion in COVID-19 tweets using the classifier from January to April 2020.

D. Discussion of Anxiety during the COVID-19 pandemic

***Hypothesis: There will be more anxiety discussion than the other mental health topics when using the classifier*.**

The classifier was used to find distinct topic discussions in COVID-19 tweets. The discussion of 'anxiety' is at a 3-4x higher number than 'depression' highlighting the prominence of the topic from February to March 2020.

It can clearly be stated that the initial rise in COVID-19 cases and unemployment rates severely impacted, or increased, anxiety discussion for both Republican and Democrat areas. From March to August, anxiety becomes a distinctly common topic in weeks 15-25 as COVID-19 cases rose (June to August). These findings further highlight the link between the COVID-19 case increase and an increase in anxiety discussion first observed in medical user tweets. Other mental health topics such as 'depression' and 'suicide' increase in discussion during March to August whilst 'bipolar' and 'PTSD' aren't discussed much throughout the datasets.

E. Influence of face coverings on mental wellbeing

Hypothesis: There will be lower mental wellbeing around the discussion of face coverings/masks.

This hypothesis is considered due to the discussion of mask anxiety noted on many mental health sites such as the charity "Mind". It is, therefore, worthwhile to assume that there will be an increase in mental health topic discussion when masks are enforced legally in the USA.

There is a large amount of mental health discussion about the requirement of face coverings independent of political stance, highlighting the wearing of masks as a nationwide issue. For both Republicans and Democrats, negative sentiment is highest for the enforcement of a mask requirement.

Hypothesis: There will be lower mental wellbeing when the lockdown / stay-at-home orders are enforced compared with when the orders are lifted

Another expectation surrounding enforcement strategies is that there is more mental health topic discussion and negative sentiment around the enforcement of lockdown which would reduce as the Phase 1, 2 and 3 measures are introduced.

Democrat users discuss mental health topics more in re-opening with a high discussion at Phase 2 compared with lockdown enforcement. Republican users follow the expected pattern of reducing topic discussion from lockdown to Phases 1-3. The negative sentiment, however, was lowest when the lockdown was enforced and higher as the re-opening stages were announced. Lower wellbeing with re-opening was found in other studies as articles from the BBC and mental health experts [23] found that there was anxiety about life after lockdown and restriction lifting. This can suggest that the reduction of lockdown can have a higher impact on mental wellbeing for some people. As Republican users show in Fig.10, other users are the opposite and have higher mental wellbeing as lockdown orders are reduced. When looking at the political differences, it is important to note that City Mayors made the majority of enforcement decisions and the political differences may be purely based on how a certain city or state responded to the pandemic.

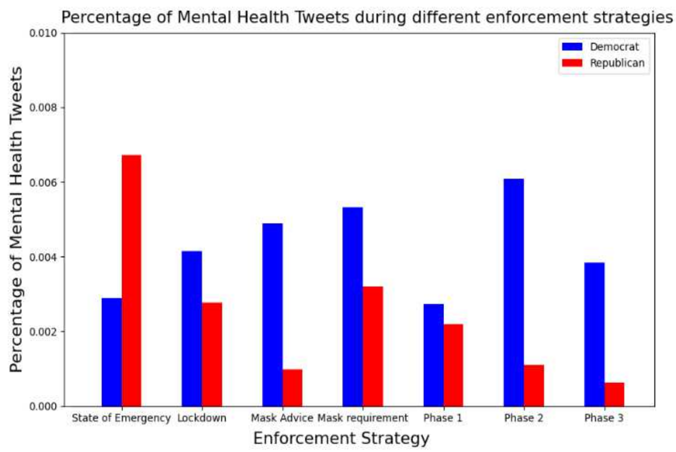

Figure. 10
Mental health topic discussion in COVID-19 tweets at different enforcement strategies using the classifier.

VIII. Conclusion

In the study, we found 'anxiety' was the most discussed mental health topic on Twitter throughout the pandemic. This was collected and analyzed through the development of an accurate mental health topic classifier which can be used in future studies to find mental health topics in tweets. The study also highlights the decrease in mental wellbeing across Twitter users due to a higher negative sentiment and higher mental health topic discussion in tweets during the pandemic.

When considering individual users, those who had lost their jobs and users working in the medical industry had higher rates of 'anxiety' and 'depression' compared with a set of control users. This can be further investigated in another study where questionnaires and a larger dataset of users can be found to identify similar patterns.

The comparison of Democrat and Republican users found few distinct differences during the pandemic to suggest most people in the USA were affected similarly. However, since location alone was used to classify a tweet into a political group dataset the political stance findings could've been stronger with more filtering. Another study could use questionnaires or self-classifying techniques to identify a user as belonging to a political group rather than an assumption based on the recent election voting history of an area.

A common pattern in the study was a rise in COVID-19 case rates correlating with lower mental wellbeing in the different datasets. Since this study uses tweets early in the pandemic due to the unknown foresight of how long the pandemic would continue, a future study could make the effort of collecting data across a larger timeframe. The larger timeframe would act as a support for suggestions made in this study and could uncover new patterns.

The use of Twitter data and NLP to find a useful result dataset and produce an accurate mental health classifier should also be observed. Future studies should take social media data into account for gathering opinions and discussion across a population and scenario such as a pandemic.

References:

[1] United States Census Bureau. (2015). *Age by language spoken at home by ability to speak English for the population 5 years and over.* Available at: https://data.census.gov/cedsci/Table?q=ACSDT5YAIAN2015.B16004&g=0100000US&tid=ACSDT5YAIAN2015.B16004 (Accessed 05 March 2021).

[2] Tikkanen, R., Fields, K., D.Williams II, R., K.Abrams, M. (2020). *Mental Health Conditions and Substance Use: Comparing U.S Needs and Treatment Capacity with Those in Other High-Income Countries.* Available at: https://www.commonwealthfund.org/publications/issue-briefs/2020/may/mental-health-conditions-substance-use-comparing-us-other-countries [Accessed 10 June 2021]

[3] Mental Health America. (2021) *The State of Mental Health in America*. Available at: https://mhanational.org/issues/state-mental-health-america [Accessed 10 June 2021]

[4] Low, D.M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., Ghosh, S.S. (2020, Dec). "Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study". *J Med Internet Res* [*Online].* vol. 22, issue 10. Available: https://www.jmir.org/2020/10/e22635/

[5] Pan, K.Y., Kok, A.A.L., Eikelenboom, M., Horsfall, M., Jörg, F., Lutejin, R.A., Rhebergen, D., van Oppen, P., Giltay, E.J. and Penninx, B.W.J.H. (2020) The mental health impact of the COVID-19 pandemic on people with and without depressive, anxiety, or obsessive-compulsive disorders: a longitudinal study of three Dutch case-control cohors. *The Lancet Psychiaty.* [online], 8(2), 121-129, doi: 10.1016/S2215-0366(20)30491-0.

[6] Alamoodi, A., Zaidan, B.B., Zaidan, A.A, Albahri, O.S., Mohammed, K.I., Malik, R.Q., Almahdi, E.M., Chyad, M.H., Tareq, Z., Albahri A.S., Hammed, H. and Alaa M. (2020). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Syst Appl.* [online], 167(1), 1-32, doi: 10.1016/j.eswa.2020.114155.

[7] Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S. and Goharian N. (2018). *SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions.* In: Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA, August, 2018. ACL: Association for Computational Linguistics.

[8] Coppersmith, G., Dredze, M., Harman, C. and Hollingshead, K. (2015). *From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses*. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Denver, Colorado, USA, June 5-5, 2015. ACL: W15-1201.

[9] Jaidka, K., Giorgi, S., Schwartz, H.A., Kern, M.L., Ungar, L.H. and Eichstaedt, J.C. (2020). *Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods.* Proceedings of the National Academy of Sciences, 117(19), 10165-10171.

[10] McClellan, C., Ali, M.M., Mutter, R., Kroutil, L. and Landwehr, J. (2016). Using social media to monitor mental health discussions – evidence from Twitter. *Journal of the American Medical Informatics Association.* [online], 24(3), 496-502, doi: 10.1093/jamia/ocw133.

[11] Fink, D.S., Santaella-Tenori, J. and Keyes, K.M. (2018). Increase in suicides the months after the death of Robin Williams in the US. *Plos One*. [online], doi:10.1371/journal.pone.0191405.

[12] Vioulès, M.J., Moulahi, B., Azé, J. and Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*. [online], 62(1), 1-7, doi: 10.1147/JRD.2017.2768678.

[13] Massoudi, K., Tsagkias, M., De Rijke, M. and Weerkamp, W., 2011, April. *Incorporating query expansion and quality indicators in searching microblog posts.* In European Conference on Information Retrieval (pp. 362-367). Springer, Berlin, Heidelberg.

[14] Khandpur, R.P., Ji, T., Jan, S., Wang, G., Lu, C.T. and Ramakrishnan, N., (2017). *Crowdsourcing cybersecurity: Cyber attack detection using social media*. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1049-1057).

[15] Coppersmith, G., Dredze, M. and Harman, C. (2014). *Quantifying Mental Health Signals in Twitter*. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. Baltimore, Maryland, USA, Jun 2014. ACL: W14-3207.

[16] NHS UK. (2019). What is autism? Available at: https://www.nhs.uk/conditions/autism/what-is-autism [Accessed 10 February 2021].

[17] Kim, P. (2017). *Convolutional Neural Network.* In: MATLAB Deep Learning. Apress, Berkely, CA. doi: 10.1007/978-1-4842-2845-6_6

[18] Kalchbrenner, N., Grefenstette, E., Blunsom, P. (2014) *A Convolutional Neural Network for Modelling Sentences.* arXiv e-prints. [online]. arXiv:1404.2188

[19] Bredemeier, K. (2020). *Trump Downplayed Coronavirus Threat Throughout February.* Available at: https://www.voanews.com/science-health/coronavirus-outbreak/trump-downplayed-coronavirus-threat-throughout-february [Accessed 10 February 2021].

[20] Mind. (2020). *How to be mentally healthy at work.* Available at: https://www.mind.org.uk/information-support/tips-for-everyday-living/how-to-be-mentally-healthy-at-work/work-and-mental-health/ [Accessed 12 January 2020].

[21] NY Times. (2021). *Coronavirus in the U.S: Latest Map and Case Count.* Available at: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html [Accessed 10 February 2021].

[22] U.S Department of Labour, (2020). *The Employment Situation – August 2020.* Washington, D.C. pp. 1-41.

[23] Butterly, A. (2020). *Coronavirus anxiety: How to cope with life after lockdown.* Available at: https://www.bbc.co.uk/news/health-52443108 [Accessed 10 February 2021]