



**This electronic thesis or dissertation has been
downloaded from Explore Bristol Research,
<http://research-information.bristol.ac.uk>**

Author:
Di Cara, Nina H

Title:
Mental health data science in rich longitudinal cohort studies

General rights

Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>. This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy

Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

- Your contact details
- Bibliographic details for the item, including a URL
- An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

Mental Health Data Science in Rich Longitudinal Cohort Studies

Nina H Di Cara
MRC Integrative Epidemiology Unit
Bristol Medical School
Faculty of Health Sciences
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Population Health Sciences in the Faculty of Health Sciences

Bristol Medical School, April 2022, Word Count: 64,080

Abstract

Data science for mental health using social media may allow us to derive digital phenotypes that present new avenues for our understanding, measurement and treatment of mental health outcomes. The timeliness and scale of these data are significant advantages over traditional survey methods. However, to ensure new technologies are developed safely and responsibly we need to use high quality ground truth that ensures that they are as robust as they can be. In this thesis I explore how population birth cohorts, specifically the Avon Longitudinal Study of Parents and Children (ALSPAC), could provide this high quality evidence through social media data linkage programmes. I use interdisciplinary methods to analyse questionnaires, focus groups and linked Twitter data from the ALSPAC cohort to understand who the populations are that use social media and how acceptable social media data linkage is to participants. I then assess the quality of the literature on mental health inference from social media, and go on to use the linked data to see whether this form of data linkage can provide new information for digital phenotyping for mental health. Overall, I find that cohort participants are generally accepting of social media data linkage, and that the linked sample broadly represents the population of people who use Twitter. Using this linked data I illustrate that population level inference of mental health outcomes of depression, anxiety and well-being are feasible at a population-level, and that using data from a well-specified sample allows us to explore model error in more detail. Future work conducting social media data linkage in other cohort samples is recommended to allow for comparisons across ages and geographies. The involvement of potential users in future research is also encouraged. Ultimately, access to higher quality of ground truth measurement will lead to safer and more reliable technologies.

Acknowledgements

There are so many people that I want to thank for supporting and encouraging me throughout my PhD. First, Claire and Oliver. Thank you for all the thought and time you have given me, as well as the space and encouragement to follow my own curiosity in my research. You are both brilliant scientists and I have learned so much from working with you. Second, thank you to Luke. I have really appreciated our in-depth conversations about my research, and all the advice you have given me about work and beyond. Thank you all for being such kind and thoughtful supervisors throughout my PhD, especially given all the challenges of the last few years. This PhD has allowed me to achieve so many things, and your confidence in me has made all the difference.

To Valerio and Al, I consider myself very lucky to have had you both as colleagues and friends. I have learned a huge amount from you both, which was helped by you always being such patient and enthusiastic teachers. Thank you both. Thanks also to Lizzy, Jacks, Benji, Chris and Jess for being the kind of colleagues that are a joy to get to work with. Having a warm group of PhD students to be part of made the whole process, and the move to Bristol, much easier, and I am so glad to know you all.

Outside of work I have been very lucky to have friends and family who have supported me through all my ups and downs. To Tiff, thanks for always being one message away, for chatting data with me and, of course, for moving to Bristol. To my brother Luke, thanks for making me laugh and for providing late night maths support. I am very proud of you. To Ben, you didn't hesitate to agree to move our lives across the country so I could do this. You've been there for me every day as sounding board, head technician and supporter. I am so grateful for you. Lastly, to my Mum and Dad. From every school pick-up and drop-off, to proof reading this whole thesis, I cannot even begin to thank you for everything you have done that has enabled me to do this. The determination you both have to work hard and achieve what you set your minds to has always inspired me to do the same, and played a big part in helping me to get this done.

Finally, my thanks must go to the participants and staff of ALSPAC. The data I have used in this thesis represents hundreds of hours of participant time that has been donated to science, which has made this research possible. I am also very grateful to the GW4 Biomed MRC DTP who funded my PhD. It was supported by grant MR/N0137941/1 awarded to the Universities of Bath, Bristol, Cardiff and Exeter from the Medical Research Council.

Publications

Di Cara NH, Winstone L, Sloan L et al. (In Press) The mental health and well-being profile of young people using social media. *npj Mental Health Research*. ¹

Di Cara NH, Zelenka N, Day H et al. (2022) Data Ethics Club: creating a collaborative space to discuss data ethics. *Patterns*, 3(7). doi: [10.1016/j.patter.2022.100537](https://doi.org/10.1016/j.patter.2022.100537)

Woolf B, **Di Cara NH**, Moreno-Stokoe C et al. (2022) Investigating the transparency of reporting in two-sample summary data Mendelian randomization studies. *International Journal of Epidemiology*. doi: [10.1093/ije/dyac074](https://doi.org/10.1093/ije/dyac074)

Shiells K, **Di Cara NH**, Skatova A et al. (2022) Participant acceptability of digital footprint data collection strategies: an exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. *International Journal of Population Data Science*, 5(3). doi: [10.23889/ijpds.v5i3.1728](https://doi.org/10.23889/ijpds.v5i3.1728)

Di Cara NH, Song J, Maggio V et al. (2021) Mapping Population Vulnerability and Community Support During COVID-19: a case study from Wales. *International Journal of Population Data Science*, 5(4). doi: [10.23889/ijpds.v5i4.1409](https://doi.org/10.23889/ijpds.v5i4.1409)

Di Cara NH, Boyd A, Tanner AR et al. (2020) Views on social media and its linkage to longitudinal data from two generations of a UK cohort study. *Wellcome Open Research*, 5(44). doi: [10.12688/wellcomeopenres.15755.2](https://doi.org/10.12688/wellcomeopenres.15755.2) ²

¹Forms Chapter 2 of this thesis. The contribution statement given in this chapter.

²Forms Chapter 3 of this thesis. The contribution statement given in this chapter.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed Nina Heather Di Cara

Dated 8th April 2022

Table of Contents

Chapter 1: Introduction	1
1.1 Thesis Overview	1
1.1.1 Chapters	2
1.1.2 Positionality and motivation	3
1.1.3 Reproducibility	5
1.2 Data science for mental health	5
1.3 Social media and mental health	10
1.3.1 Deriving digital phenotypes	10
1.3.2 Perspectives on the relationship between social media and mental health	16
1.3.3 Current challenges	18
1.3.4 Twitter	21
1.3.5 Summary	24
1.4 Questions of ethics	24
1.5 Data linkage with social media	27
1.6 The case for cohort studies	29
1.6.1 High quality ground truth data	30
1.6.2 Ethical data collection and sharing	31
1.6.3 The Avon Longitudinal Study of Parents and Children	32
1.7 Summary	33
Part A: Establishing the use of digital phenotypes in ALSPAC	35
Chapter 2: The mental health and well-being profile of social media users	37
Abstract	38
Aims	38
2.1 Introduction	39
2.2 Methods	42
2.2.1 Sample Description	42
2.2.2 Measures	43
2.2.3 Ethics	46
2.2.4 Data and code	46
2.3 Results	48
2.3.1 Demographics	48
2.3.2 Mental Health and Well-being	50

2.4	Discussion	55
Chapter 3: Participant views on social media and its linkage to longitudinal data		63
	Abstract	64
	Aims	64
3.1	Introduction	65
3.2	Methods	67
	3.2.1 Sample and Recruitment	67
	3.2.2 Ethics	68
	3.2.3 Data collection	68
	3.2.4 Analysis	74
3.3	Results	74
	3.3.1 Personal views on social media in the UK today	75
	3.3.2 Views on using social media data for research	81
3.4	Discussion	88
3.5	Conclusions	91
Chapter 4: Twitter data linkage: features of consenting participants and their data . . .		93
	Abstract	93
	Aims	94
4.1	Introduction	95
4.2	Methods	97
	4.2.1 Cohort description	97
	4.2.2 Twitter Data Linkage Programme	98
	4.2.3 Datasets and measures	101
	4.2.4 Analysis	103
4.3	Results	104
	4.3.1 Features of linked participants	104
	4.3.2 Features of linked Twitter data	106
4.4	Discussion	110
	4.4.1 Consent and successful linkage rates	110
	4.4.2 Characteristics of the linked participants	111
	4.4.3 Characteristics of the linked data	112
	4.4.4 Strengths and limitations	113
4.5	Conclusion	114
Part B: Mental health data science using Twitter		115
Chapter 5: A review of methodologies for monitoring mental health on Twitter		117
	Abstract	117
	Aims	118
5.1	Introduction	119
	5.1.1 Themes from previous reviews	119
	5.1.2 The purpose of the present study	122
5.2	Methods	123
	5.2.1 Search Methodology	123

5.2.2	Screening Methodology	124
5.2.3	Data Collection	124
5.3	Results	125
5.3.1	Mental health outcomes predicted	126
5.3.2	Datasets	128
5.3.3	Modelling workflows	132
5.3.4	Ethics	137
5.3.5	Replicability	138
5.4	Discussion	138
5.4.1	Principal Results	138
5.4.2	Recommendations	142
5.4.3	Limitations	144
5.5	Conclusion	144
Chapter 6: Modelling mental health using linked Twitter data		147
	Abstract	147
	Aims	148
6.1	Introduction	149
6.1.1	Sentiment in population mental health inference	150
6.1.2	Temporality in mental health inference	152
6.1.3	The present study	154
6.2	Methods	155
6.2.1	Sample	155
6.2.2	Measures	157
6.2.3	Analysis	160
6.2.4	Software and code	164
6.2.5	Ethics	164
6.3	Data Description	166
6.3.1	Mental health outcomes across the samples	166
6.3.2	Twitter features	166
6.3.3	Temporal patterns	170
6.4	Results	175
6.4.1	How well do patterns of life and codings of emotion predict mental health?	175
6.4.2	Is prediction improved by using a larger number of linguistic categories?	177
6.4.3	What is the effect of changing the window and weightings of data?	185
6.4.4	How do predictions perform over time?	186
6.5	Discussion	189
6.5.1	How well do patterns of life and codings of emotion predict mental health?	189
6.5.2	Is prediction improved by using a larger number of linguistic categories?	191
6.5.3	What is the effect of changing the window and weightings of data?	192
6.5.4	How do the predictions perform over time?	193
6.5.5	Strengths and limitations	194
6.5.6	Future directions	195
6.6	Conclusion	197
Chapter 7: Discussion		199

7.1	Commentary on thesis themes	202
7.1.1	Contributions of cohorts	202
7.1.2	Digital phenotyping for mental health	205
7.1.3	Ethics and acceptability	207
7.2	Future directions	208
7.3	Conclusion	211
Appendix A: Chapter 2		213
A.1	Additional sample information	213
A.2	Descriptive data on mental health and well-being outcomes	214
A.3	Outcomes by platform for all social media users	217
Appendix B: Chapter 4		219
B.1	Mental health sample comparisons by sex	219
Appendix C: Chapter 5		221
C.1	Systematic search key terms	221
Appendix D: Chapter 6		223
D.1	Data Hazards Analysis	223
D.2	Descriptive mental health data by sex and generation	228
D.3	Correlations between model features	230
D.4	Associations for disaggregated data	233
D.5	Tables of Results Accompanying Figures	236
D.6	Prediction error	238
References		241
Abbreviations		289

List of Tables

1.1	The differing purposes, data and methodologies used by the two alternative approaches to studying the relationship between social media and mental health. . . .	17
1.2	A summary of different types of interaction available on Twitter and the particular meaning and nomenclature associated with these interaction types.	23
2.1	The number of participants in each of the two samples used in this study, subset by demographic characteristics.	43
2.2	The percentage of each demographic group by their self-reported frequency of using any social media each day.	48
2.3	The percentage of each demographic group who indicated that they had an account on each of the social media platforms considered.	49
3.1	Definitions of levels of social media use as given by NatCen.	70
3.2	The options presented to participants to fill each 'blank' in the statements in Figure 1.	72
4.1	Demographic split of the index cohort with linked Twitter data	104
4.2	For the index cohort (G1) only: demographic characteristics of the full index cohort, those who said that they had a Twitter account at age 24, and those who agreed to link their Twitter data.	105
4.3	Weekly tweet frequency for the most recent year of data, split by sex and generation.	109
5.1	The number of papers included in the review that were published each year.	126
5.2	Overview of the different methods used to annotate datasets with ground truth labels.	130
5.3	Overview of feature categories, the number of studies that used at least one feature from each category and a description of the types of features it contains.	135
5.4	The number of studies using each type of algorithm for at least one model.	136
6.1	Summary of key features of the linked Twitter sample against the sample who tweeted at least twice and completed Survey 1 (N=151), and those who tweeted at least twice and completed Survey 2 (N=136)	167
6.2	This table displays the variance accounted for by each sentiment variable with $p < 0.05$ when regressed against depression, anxiety or well-being, and adjusting for sex and generation. The 'Effect' column gives the direction of the coefficient. p-values in orange are < 0.001 and those in teal are < 0.01	178

6.3	Summary of the linear model of the 11 chosen sentiment variables and number of tweets against depression measured at Survey 1. The lower half of the table gives summary information for the model overall.	180
6.4	Summary of the linear model of the 8 chosen sentiment variables and number of tweets against anxiety measured at Survey 1. The lower half of the table gives summary information for the model overall.	182
6.5	Summary of the linear model of the 11 chosen sentiment variables and number of tweets against general well-being measured at Survey 1. The lower half of the table gives summary information for the model overall.	184
6.6	Table containing the Root Mean Squared Error (RMSE) for each sub-group of those predicted at the second survey time point (Survey 2), using the model trained on data from Survey 1. The sub-groups include sex, generation and whether or not the individual being predicted was part of the original sample at Survey 1. P-values were calculated on the original errors using Welch's Test.	187
7.1	A summary of the main findings and subsequent implications from each empirical chapter of this thesis.	200
A.1	Percentage of the users of each social media site by use-frequency and demographics reported.	213
A.2	The percentage of the sample who had experienced each of the four categorical mental health outcomes.	214
A.3	Summary statistics for well-being outcomes, all measured in the sub-sample (N=2,862).	215
A.4	Contingency table of suicidality and disordered eating (N=4,083).	215
A.5	Contingency table of suicidality and self-harm (N=4,083).	215
A.6	Contingency table of disordered eating and self-harm (N=4,083).	216
D.1	An analysis of different potential ethical hazards that might be presented by this project, using the Data Hazard labels.	224
D.2	Summary of numbers of tweets and the mental health outcomes between men and women at Survey 1 and Survey 2.	229
D.3	Summary of numbers of tweets and the mental health outcomes between G0 and G1 at Survey 1 and Survey 2.	229
D.4	This table displays the variance accounted for by each of the sentiment variables that align with standard codings of emotion, as well as patterns of life, when regressed against depression, anxiety or well-being, and adjusted for sex and generation.	234
D.5	This table displays the variance accounted for by each sentiment variable with $p < 0.001$ when regressed against depression, anxiety or well-being, and adjusting for sex and generation with no aggregation by individual.	235
D.6	Table of the results displayed in Chapter 6, Figure 6.11. The results of regression models of each sentiment variable against depression, anxiety and general well-being, adjusted for sex and generation. The percentage variance explained after sex and generation have been accounted for is given.	236

D.7 This table represents the results from Chapter 6, Figure 12. The mean R Squared value obtained from 10 x 5-fold cross validation plotted for the outcomes of depression, anxiety and well-being. The input data contains an increasing number of weeks, and difference weighting functions were used on each number of weeks of data. 237

List of Figures

2.1	Percentage of participants using each of Facebook, Twitter, Instagram, Snapchat and YouTube stratified by the frequency of using that platform, and sex.	50
2.2	Percentage of participants who reported disordered eating, self-harm or suicidal thoughts in the past year, or who met the threshold for depression, differentiated by sex and frequency of any social media use with 95% confidence intervals. . . .	51
2.3	Mean scores for seven well-being measures, stratified by sex and overall frequency of using any social media platform, with 95% confidence intervals.	52
2.4	Percentage of participants who reported disordered eating, self-harm or suicidal thoughts in the past year, or who met the threshold for depression, differentiated by sex for daily users of each social media platform, with 95% confidence intervals.	53
2.5	Mean scores for seven well-being measures for daily users of each platform, stratified by sex, with 95% confidence intervals.	54
3.1	The template that participants filled with options from Table 2 to provide discussion points.	71
3.2	An example of a completed template from the situational exercise.	73
4.1	The number of participants included at each stage of the process for identifying participants for linkage, as well as the reasons for not being included in the final sample.	99
4.2	A diagram illustrating data flow between the ALSPAC participants, data safe-haven, data management team and researchers.	100
4.3	A comparison of the distributions of participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.	106
4.4	Monthly counts of original tweets and retweets collected for all participants. The original tweet and retweet values are layered to fill to the total number of tweets for each month.	107
4.5	Proportions of total tweets that were retweets, split by participants who are in each of the quantiles of total number of tweets. The dashed grey line indicates the median proportion per quartile.	108
4.6	Histogram of the number of tweets per person in the most recent year of data (31st Oct 2019 to 31st Oct 2020), with tweets per person transformed with the binary logarithm.	109

5.1	PRISMA flow diagram of inclusion and exclusion figures for the literature search .	125
5.2	Network digram showing which mental health disorder (pink) each paper (blue) attempted to infer. ³ Depression and suicidality have been the most popular, with most papers attempting to predict a single outcome.	126
5.3	The number of studies considering each mental health disorder published each year. To be presented in this figure the mental health disorder needed to be included in more than two studies. ⁴	127
5.4	The proportion of studies that reported each of the stages of modelling that I considered, split by those published before 2020 (N=90) and those published in 2020 or later (N=75).	133
6.1	The windows of data collection for Survey 1 and Survey 2 where the psychological outcomes were measured.	156
6.2	The number of linked participants with two tweets in the two weeks leading up to the survey date for Surveys 1 and 2, and how many participants are unique to each survey or overlap across both.	157
6.3	The five options of weighting functions	163
6.4	A correlation plot of the relationships between sentiment variables that relate to codings of emotion or positive and negative sentiment, without aggregation by individual. Correlations are calculated using Spearman's Rank, with colour representing the correlation coefficient.	169
6.5	Histogram of tweet frequency over the two weeks leading up to Survey 1, for those who have tweeted at least twice Tweets per person are transformed with the binary logarithm.	170
6.6	In the top figure, VADER compound is split by sex and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of each sex group differentiated by color and stacked on top of each other.	171
6.7	In the top figure, VADER compound is split by generation and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of each generational group differentiated by color and stacked on top of each other.	171
6.8	In the top figure, VADER compound is split by whether or not a tweet is a retweet and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of retweets and original tweets differentiated by color and stacked on top of each other.	172
6.9	VADER compound across the day over 12 months. Warmer months are coloured towards yellow, and colder months towards blue.	173
6.10	The ten most common LIWC values over 12 months (without 'function' which has a mean of approximately 0.28)	174
6.11	The results of regression models of each sentiment variable against depression, anxiety and general well-being, adjusted for sex and generation. The percentage variance explained after sex and generation have been accounted for is given by the size of each point, the sign of the coefficient is given by the colour, and the p-value is represented by the transparency of the point.	176

6.12	The mean R Squared value obtained from 10 x 5-fold cross validation plotted for the outcomes of depression, anxiety and well-being. The input data contains an increasing number of weeks for each graph from left to right, and within each graph the y-axis represents the different weighting functions used on that number of weeks of data.	185
6.13	Each line illustrates the mean predicted values for depression, anxiety and general well-being as predicted by the linear models for each outcome. Predictions were made on data aggregated by individual over each two week period between the 1st January 2020 and 31st October 2020. The figure is annotated with key national events over the same time-period, particularly those regarding the COVID-19 pandemic, with grey horizontal bars indicating the two-week prediction period that each event happened within.	188
A.1	A correlation matrix for all continuous mental health and well-being variables using Spearman's Rank coefficient (all $p < 0.000$).	216
A.2	Percentage of participants who reported disordered eating, self-harm, depression, or suicidal thoughts in the past year, differentiated by sex for all users of each platform, with 95% confidence intervals.	217
A.3	Mean scores for seven well-being measures for all users of each platform, stratified by sex, with 95% confidence intervals.	218
B.1	A comparison of the distributions of female participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.	219
B.2	A comparison of the distributions of male participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.	220
D.1	Correlations between the variables which were best associated with depression . . .	230
D.2	Correlations between the variables which were best associated with general anxiety	231
D.3	Correlation between the variables which were best associated with general well-being	232
D.4	Histogram of the residual errors from predictions of Survey 2 outcomes, using the linear models trained on Survey 1. Residuals for depression and anxiety were calculated after exponentiating the estimate, which is predicted on a log scale. . . .	238
D.5	The graph of predictions made by the models for depression, anxiety and general well-being over the pandemic period using the final models trained on 2 weeks of Twitter data. This version contains the prediction intervals estimated for each fortnightly prediction.	239

Chapter 1

Introduction

This thesis aims to explore what can be learned from the process of linking digital phenotypes from social media, specifically Twitter, in a UK birth cohort study. It illustrates what we can gain from data linkage in cohorts with respect to social media use, the understanding of the acceptability of this form of data linkage for cohort participants, and shows how these data can be used for the purpose of better understanding mental health.

1.1 Thesis Overview

In this thesis I aim to answer two main questions. The first is whether linking social media data in a birth cohort is an effective means of developing high-quality datasets for mental health data science. The second is, given this novel dataset, can we use it to better understand how textual sentiment is related to mental health and well-being. The two parts of this thesis focus on each of these questions respectively.

In the first part, ‘Establishing the use of digital phenotypes in ALSPAC’, I illustrate how social media is already being used in the Avon Longitudinal Study of Parents and Children (ALSPAC) based on questionnaire data in *Chapter 2*, present the results of focus groups with the cohort participants which considers the acceptability of the proposed data linkage in *Chapter 3*, and then I describe the results of linking to Twitter in the ALSPAC cohort, with an overview of the data obtained in the linkage process in *Chapter 4*.

The second part, ‘Mental health data science using Twitter’ then focuses on conducting mental health data science with the linked Twitter data from the cohort. First in *Chapter 5* I report the results of a scoping review of current methodologies in the literature for predicting mental health disorders from Twitter. In *Chapter 6* I go on to use the linked Twitter data from ALSPAC to establish the accuracy of sentiment analysis for inferring changes in mental health.

Since this thesis takes an interdisciplinary approach each chapter has its own unique approach and methodology. As such, this overall introduction aims to give a broad overview of the field of mental health data science, and where cohort data linkage sits within it. The specific literature and methodology relevant to each chapter will then be presented in each of the empirical chapters.

1.1.1 Chapters

Part A: Establishing the use of digital phenotypes in ALSPAC

Chapter 2: The mental health and well-being profile of social media users

This chapter gives a descriptive overview of social media users in ALSPAC, and asks how the populations of users of different social media platforms (namely Facebook, Instagram, Twitter, Snapchat and YouTube) differ in their demographic features, and in their mental health and well-being. This has implications for understanding whether we can consider social media users to be representative of the general population in terms of their mental health, and what this might mean for research conducted using these platforms.

Chapter 3: Participant views on social media and its linkage to longitudinal data

I conducted a qualitative exploration into data from focus groups with the G1 and G0 cohorts that aimed to explore their views towards linking their social media data to the data already held about them by the cohort. These focus groups asked which types of data might be more or less acceptable, and under which conditions participants would consent to them being shared with other researchers. These findings were used to inform the data collection strategy used for linking participant Twitter data.

Chapter 4: Twitter data linkage: features of consenting participants and their data

With Twitter data successfully linked in the cohort I give a brief overview of the linkage framework, and then present data on who agreed to link their data in ALSPAC. This asks what the similarities

and differences are of those who consented to data linkage, including their mental health profiles and descriptive features of their Twitter data. I discuss the implications of these findings for the future use and applications of these data.

Part B: Mental health data science using Twitter

Chapter 5: A review of methodologies for monitoring mental health on Twitter

This chapter presents the results of a review of 165 published papers that attempted to monitor or detect mental illness from Twitter. It asks what their approach to collecting data was, how mental illness was defined and the features that were used for modelling. Finally it discusses the ethical and practical ways forward for the field and presents a series of recommendations for future research.

Chapter 6: Modelling mental health using linked Twitter data

Using the data described in Chapter 4 I go on to use measures taken in 2020 on depression, anxiety and general well-being in the ALSPAC cohort to consider how accurate sentiment can be as a signal of each of these mental health outcomes. I also test the impact of changing the length and weightings of training data time periods to see if different mental health outcomes are predicted more effectively with differently specified training data.

1.1.2 Positionality and motivation

It is common, and usually expected, in qualitative research for the researcher to present a positionality statement. This statement gives an overview of the researcher's understanding of themselves in relation to their topic and their participants in order to frame the position from which they approached their research [1]. In data science, the relevance of positionality and reflexivity are only recently beginning to gain traction [2], but especially when examining social phenomena like mental health it is likely that our positionality has an influence on our research questions and processes. As such I believe it is valuable to be explicit about this by including a brief positionality statement in my thesis, which also touches on my motivations for completing this research.

By way of introduction, I am a White woman who grew up in Kent, UK from a background best described as middle class. By heritage I am Irish-Italian, with my mother and paternal grandparents having moved to England from their respective countries. After studying for an undergraduate

degree in Mathematics, I went on to train as a social worker. This training involved spending two years working directly with children and families who were experiencing mental health challenges in varying contexts of addiction, domestic abuse and trauma. It provided direct experience of how structural barriers prevent people from receiving mental health care, and the effect this has on their lives. Partly as a result of my training I am inclined to view mental health disorders as being socially constructed categorisations of groups of symptoms that are inter-connected and tend to co-occur, rather than being distinct illnesses that cause a group of symptoms. My mixed technical and social care background has also given me a great interest in, and enjoyment of, interdisciplinarity which I hope is evident in this thesis.

In studying the participants of ALSPAC I have at times felt both like an insider and an outsider in relation to them [3]. I am only a couple of years younger than the participants, meaning that I experienced similar exposure to the emergence of social media at the same ages as they did, and that the age of my parents is in the same range as theirs. At times this had made it challenging not to make assumptions about how questions might have been interpreted, or to assume similarities with the ways that I know social media is used between myself, my family and my peers. At the same time, I am a relative outsider to the city of Bristol where the majority of the participants grew up and many continue to live, which has provided some distance between our experiences.

I was originally motivated to study digital phenotyping for mental health in the knowledge of how challenging it can be for individuals to see their own mental health declining, and as a means to support self-management of mental health disorders. As I have spent more time on the topic I have become additionally motivated by concerns that the development of mental health inference technologies could do the most harm to those who are most psychologically vulnerable, coupled with the potential for overstretched statutory systems to put their trust in such technologies, which may serve to maintain the structural inequalities that cause poor mental health in the first place. This journey of my perspective on my research topic has left me now highly invested in building robust evidence that allows us to truly understand what works and for whom, and in finding ways to communicate this clearly when thinking of applying this research.

1.1.3 Reproducibility

This thesis is written using RMarkdown (Version 2.11), meaning that the textual content and code to produce all data-based display items such as tables and graphs are contained within the same source document. The original source for this thesis is available on the Open Science Framework (doi: [10.17605/OSF.IO/HYD9G](https://doi.org/10.17605/OSF.IO/HYD9G)), where the reader may also find other digital supplementary materials which are signposted from the relevant chapters. Given the original ALSPAC datasets, which can be requested from ALSPAC, this document is fully reproducible.

1.2 Data science for mental health

Mental health is a global issue. The World Health Organisation (WHO) estimate that approximately 19% of years lived with disease globally can be attributed to mental health [4]. The mental illnesses which most represent this burden are depression, anxiety, schizophrenia, and bipolar disorder, and suicide is currently the fourth leading cause of death among 15-29 year olds worldwide [5]. In the United Kingdom (UK) the last Adult Psychiatric Morbidity Survey, which is distributed every seven years, estimated that 17% of over 16s in the UK were suffering from a mental health condition, and that nearly half think that they have had a mental health condition at some point in their lives [6]. It also showed that just over a third of people who self-identified as having a mental health condition had never been diagnosed by a professional, and that up to 75% of people with a mental health condition may not get the treatment that they need [6].

This picture of mental health need across the UK population, and across the globe, impresses the importance of understanding how to best prevent, treat and define mental health. Though the statistics are concerning, there has been admirable progress towards this aim over the past 80 years. The rapid development of the field of psychiatry in the twentieth century and the introduction of a diagnostic system for mental illness has aided the recognition and understanding of mental disorders [7]. Our subsequent development of therapeutic techniques for treatment, alongside pharmaceuticals, continues [8]. Up until recently, mental health diagnosis has relied on the observation and interpretation of behaviour by another human, in this case a trained medical professional. As we have entered the internet age however, we have uncovered new opportunities for understanding mental health from the explosion of recorded digital data about our health, well-being and daily

lives, as well as developing the computational power and capabilities for analysing it [9]. This process of datafication of our lives and health is somewhat cynically of great benefit to commercial companies for generating profit from our datafied selves, but its benefits can also be harnessed in the best interests of research into our mental health and well-being.

The sensemaking of this digital data for the benefit of mental health falls under several umbrella terms. It can be known as *mental health data science* [8, 10], *digital epidemiology* [11] or *digital psychiatry* [12], which all tend to sit within the broader field of *digital health*. For simplicity I will defer to the term *mental health data science*. The potentials of this emerging field are broad, and in theory cover any application of data science to any type of mental health related data [10]. This could mean using operational research to optimise treatment queuing systems [8], using deep learning for the interpretation of neuroimaging data [13], or predicting the best matches between individuals and medications [14]. Another application of data science to mental health is the passive sensing of mental health signals using high frequency digital data, such as smartwatches, phones, and pertinent to this thesis, social media [9]. These sources of personal digital data are known as *digital phenotypes*, defined by Torous and colleagues in their seminal work on the topic as the “moment-by-moment quantification of the individual-level human phenotype in-situ using data from smartphones and other personal digital devices” [15]. As such, digital phenotypes may be derived from the language people use, the way they move, metrics related to social connectedness or the times of day they are active [9].

These data have great potential to be used as a mental health “smoke alarm” [16] that could alert to concerning changes in an individual’s mental health, and also to capture ‘genuine’ lived experiences rather than those reported to clinicians which can be influenced by impression management or simply variable recollection [15]. The ability to understand mental health in this way paves significant new avenues for research, as feedback on longitudinal mental health changes may help us to better understand the course and onset of mental disorders in ways that are not possible, or ethical, with traditional research methods [8]. Other attractive incentives for developing these technologies are their potential to scale and their cost effectiveness for use in clinical care. A report from the United States (US) Behavioural Workforce Administration in 2020 estimated the need for mental health treatment in the US alone would require an additional 4 million mental health professionals, which would mean more than quadrupling the current estimated workforce

[10, 17]. Digital methodologies may provide a means to support the provision of mental health care by more efficiently allocating existing resources based on risk, or more quickly identifying the best treatment for an individual from a range of options, and so relieve some pressure from already overstretched services.

As well as making a contribution towards the potentials of personalised medicine, these technologies have significant applications in population health too. Collation of population-level digital phenotypes can be used to provide an understanding of how populations are responding to events such as the COVID-19 pandemic [18], as well as how the impacts may be felt differently over geographic areas. They have also been explored for use in specific sub populations such as students [19]. A digital approach to population health monitoring also has the benefit of timeliness, which traditional surveys do not provide, although this is often at the expense of data which can be noisy and from a biased sample of internet users [20].

Several key reports and voices in the field have been keen to impress the need for these exciting opportunities presented by our digital mental health revolution to be tempered with a responsible approach to innovation, which should comprise an appropriate understanding of the risk that these technologies pose societally, ethically and practically [9, 21–24]. This is perhaps best put by the National Institute for Mental Health (NIMH) in their statement that “the promise is enticing, but there are still many unanswered questions about effectiveness, concerns about privacy, and challenges for regulation of these nascent technologies” [25]. As with many technological advances, there is a balance to be found between deploying technologies as quickly as possible in order to allow people to benefit from them, and waiting to deploy anything until we can be totally satisfied on their stability and safety [21, 22]. Due to the inability of regulation to effectively keep up with the pace of technological development, this balance is largely left to us as a field to find, guided by developing research and existing laws or regulation surrounding the use of data, human rights and research ethics [21]. We are currently in the process of feeling out these acceptable limits in the field of mental health data science, but the nature of digital innovation means that we are essentially laying the tracks in front of the train. Calls for careful next steps [24] are reinforced by a 2018 review by the National Health Service (NHS) which estimated that digital mental health prediction will start to be operationalised within two to five years [26], which brings us to the time of writing. As anticipated, in the commer-

cial sector mental health prediction is already being researched or applied in technologies that are part of our day to day lives. Spotify, a music streaming service, has filed a patent to use mood sensing to enhance personalised music prediction [27]. The fitness watch company Fit-Bit have recently released an upgrade that tracks stress in people using their smartwatches, and natural language processing (NLP) technologies for mental health are being developed by digital counselling providers such as Sondermind (<https://www.sondermind.com/>) and Ieso (<https://www.iesohealth.com/>).

The regulatory environment in particular becomes more challenging when we consider the distinction between so-called *emotion AI* [28], which seeks to infer emotional states, and mental health inference. Whilst emotion and mental health share similar characteristics, and their fields share similar ethical quandaries [29], mental health inference is specific to classifying what would be considered mental health disorders, rather than momentary emotional states, so as to find the boundaries between these and ‘healthy’ functioning. This distinction is important, particularly as mental illness (as an umbrella term for any combination of mental health disorders) is a protected characteristic under UK law [30], and in many other countries, whereas emotion inference is not specifically protected [31]. This means that discrimination on the basis of one’s mental health is illegal, and so inferring information about this attribute, like other protected characteristics such as gender, age or ethnicity, must be considered carefully. However, as the boundaries between what we call *emotion* and what we call *mental health* blur, we reach more difficult questions about what mental health even is. For instance, classifying mood is not the same as classifying a mental health status, but mental health status may be inferred with relatively high sensitivity from particular patterns of moods if the mood measurements were accurate. This opens up further ethical challenges in this area, since this information could be among that shared with health insurance companies without breaching current laws in multiple countries, or perhaps more insidiously used to monitor employees or school-children, which is currently occurring [32].

Outside of concerns about the capabilities of these technologies for surveillance and insurance risk prediction, there are also ongoing challenges in the field of AI and machine learning with respect to the quality and representativeness of training data. We already know that training data collected from digital technologies or Internet of Things (IoT) devices show similar biases as seen in other sectors due to inconsistencies in accuracy of readings from wearable devices in those with darker

skin [33], or divergences from the patterns of English that most of our natural language processing tools have been designed around [34, 35], or simply differences in patterns of life across heterogeneous groups [36].¹ We can foresee that these conditions create a perfect storm for algorithmic bias, through which mental health technologies may become another route for perpetuating systemic inequalities that inevitably arise from these data. This reinforces the importance of high quality data needed to drive these technologies, so that we have robust evidence of what works, and for who. This evidence is currently in short supply in digital phenotyping studies, due to the overabundance of studies focused on small clinical populations, rather than larger epidemiological samples [24]. This high quality data is also needed to help to develop NLP-based tools that can account for individual differences in language use, which is an ongoing challenge in the field [37]. Similar issues of data quality arise from the use of social media data for population health purposes. Without a clear understanding of who the population is that is using social media sites, we cannot yet make decisions using the inferences we are generating from social media when we do not yet understand how our data might be biased. Without accurate ground truth data we also cannot know whether differences that are found in textual content, patterns or sentiment actually relate to changes in the overall mental health of the population, or translate to increased service demand for health care services.

As well as the understanding of ground truth demographic characteristics of samples, the quality of ground truth mental health measurement across studies is also of concern. *Ground truth* refers to information about a sample that is assumed or known to be true, which is usually the outcome that a model is being trained to predict in supervised learning tasks, like a particular mental health disorder [38]. The quality of what is considered to be ground truth in machine learning tasks can vary, and the term *gold standard* ground truth may be used to signify ground truth measurements that are generally agreed to be highly reliable at representing the outcome that is being predicted, such as validated scales or user self-report [38], though what is considered to be high quality is likely to be context dependent. A recent systematic review by Kim and colleagues sought to identify studies that predicted depression from Instagram, Facebook and Twitter, with inclusion criteria that stipulated that the mental health outcome must have been measured using a validated scale, they only found 15 studies to review [40], whereas a review with similar criteria but no

¹In fact, a review into racial bias in medical devices was announced by the UK Health Secretary in November 2021 <https://www.bbc.co.uk/news/uk-59363544>

requirement for the use of validated scales [41] found 40 studies on the same platforms.

In summary, mental health data science is still a relatively new field and one with huge potential to support the increasing demand for mental health services and support that is being felt across the world. However, we can also see that the stakes are high in this field; mental health is one of our most personal human attributes and, as with any research that eventually intends to impact human lives, we need to ensure we can implement robust technical models along with an awareness of their societal implications. In order to successfully harness this potential we need to retain a focus on high-quality measurement in order to develop useful training data and models that we can test for bias and fairness.

1.3 Social media and mental health

As discussed, one aspect of mental health data science is digital phenotyping. Digital phenotyping can take many forms, and given the range of digital data available about us from our smart watches, phones and online lives there are many potential sources of passive sensor data to draw on [9]. This thesis will concentrate largely on the study of social media data for the purpose of monitoring and predicting mental health, and so it seems prudent to begin by defining exactly what I mean by *social media* and how it can be used to derive digital phenotypes. I then go on to discuss current research on the relationship between social media use and mental health, before identifying the current challenges that are facing this field.

1.3.1 Deriving digital phenotypes

Whilst most of us could name several social media sites, it is much more challenging to pin down the definition of what makes these examples, say Facebook, Instagram or Twitter, social media sites. Indeed, in the literature social media are defined in multiple ways, which tend to share themes around user generated content, and allowing people to connect through the internet [42]. Carr and colleagues [42] reviewed these definitions to offer one of the more thoughtful definitions available of what characterises a social media site. As well as being specific about how we define social media sites, the definition covers the facets of social media use that allow us to understand the reasons that social media has so much value to mental health researchers:

"Social media are Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others"

Selectively self-present: A means of understanding how people want to be seen by the world, in the ways they set up their profiles, the ways they talk and the types of content they share.

Real time or asynchronously: The timings with which people use social media give strong indications of their daily routines, or disruptions to them, which has been repeatedly found to be an important factor in predicting mental ill health. This can also give clues to sleep routines, which are strongly related to mental health outcomes.

Broad and narrow audiences: We can consider the online communities that people belong to, and how these reflect a person's interests and beliefs. This may also include which subset of people this person interacts with, how large this group is, to what extent the interactions are reciprocal and how frequently they interact.

Internet-based: Lastly all this data is collected and stored. This gives us access to all the above data throughout the history of that person's interaction with the platform, objectively timestamped so that we can look at it both longitudinally and in real-time.

There is an attractive opportunity to use the volumes of data produced by individuals on social media as a low cost, high time frequency and (in theory) non-invasive method of monitoring individual and population health. This methodology has been widely researched across a number of social media sites, where there has been some success in identifying a wide range of mental disorders [40, 41]. However, there are also concerns about the quality of data in this field - most crucially an imbalance in the availability of digital footprint data versus the accessibility of genuine ground truth information against which to train models and assess the effectiveness of any algorithms or methodologies developed [24, 41].

In order to be specific about the role of social media in mental health measurement and monitoring I will briefly outline four of the mechanisms by which social media can function as a digital phenotype, by drawing connections between the features of data measured on social media platforms and

their links to presentations of mental health disorders. These mechanisms are language, timing, connectivity and visual content.

Language

The text that people share on social media is one of the main contributions of social media to mental health inference. Social media is one of the only sources of digital phenotype data that contains self expression through language [9], outside of voice recording/microphone data. There are two facets of relevance between the language used by individuals and their mental health. The first is what people talk about, for instance the events that they recollect or the topics they choose to focus on; this speaks to the link between someone's immediate interests and concerns and how they are feeling. The second is the way in which they talk about these topics, such as the types of words used to describe them or the sentiment of the text, which is connected more to cognitive processes that are hypothesised to differ between those with or without current mental health disorders [43]. This link between language, usually through speech, and mental health appears mostly in the literature around depression, and has been known and researched long before social media provided a new avenue for its exploration [44, 45]. However, the availability of written natural language produced through social media is a unique addition to this area. In the past written language was difficult to use for statistical analysis since a) it was hand written, and so not easy to process in large volumes or at speed, and b) aside from letters between friends or diaries, the majority of written information was in the form of formal letters [46]. As such, social media presents an opportunity to study language with a huge volume of first-person data that has not been available previously.

The task then is to make sense of this written data, in order to be able to use it with quantitative methodologies. This is broadly known as *Natural Language Processing* (NLP). NLP is a wide field that covers many different approaches to working with textual data, from machine translation to automatic summarisation. A particularly popular technique when using social media for mental health analysis is sentiment analysis, which is a means of understanding the overall feeling of a piece of written text. There have been many algorithms developed for this purpose which range in complexity [47], but in general they categorise text into, or generate a score that summarises the level of positivity, negativity or overall valence of a piece of text. Sentiment features have been found in several studies to be a useful predictive feature in mental health inference algorithms [41],

although this has more recently been contested [48] and is discussed in more depth in Chapter 6. Parts-of-speech tagging is another popular a method that is used to identify the grammatical part of speech that each word in a piece of text matches, such as pronouns, verbs or adjectives [49]. Research into the links between mental health and speech using this technique have found that increased use of first-person pronouns is associated with, or predicts, depression [50]. Many more approaches, such as topic modelling or word importance, can also be used to extract meaning from social media text.

There are obvious challenges to using NLP in the context of social media. Internet language, and by reflection the slang used by young people, changes rapidly and words can take on whole new meanings in different contexts [46]. For instance the word “sick” could refer to being unwell, or as a slang term for something that is very impressive. Additionally, the multi-media format of social media can mean that text is interpreted differently if it is accompanied by an image, emoji or GIF; this aligns with similar concerns about computational methods not correctly interpreting irony [51]. An over reliance on language data also impacts on the ease with which such research can be applied to non-English text, and also has the potential to systematically misclassify or misrepresent those who speak English in a form that is not considered typical by NLP systems based on their training data [34]. Lastly, social media text can be very short, with some platforms like Twitter limiting users to a certain number of characters for each post. This limits the amount of information available to make a reliable prediction, compared to a book or letter, and in particular some sentiment based methods are optimised for use on longer length texts so may perform differently on social media data [52].

In summary, language has the potential to play an important role in mental health inference, and there are a variety of language-based methods available for extracting meaning from social media data. These allow us to make use of the unique presence of language data in social media as a digital phenotype.

Timing

One of the main benefits of objectively measured data like social media posts is that it is stored with exact timestamps, usually with the precision of seconds, meaning that we can easily obtain precise measures of when users are actively posting or interacting on social media sites. This is

useful for a range of reasons. Firstly, timings of data generation can lead to an understanding of someone's *patterns of life* which indicate typical daily routines [53] and, importantly, the timings of sleep [54]. Disrupted sleep or insomnia is thought to be causally related to multiple mental health disorders [55], and is also a risk factor for suicidality [56]. Therefore, changes in sleep patterns have strong potential to be important early indicators of declining mental health. Timings of post generation or interactions also allow researchers to access information about the frequency of using social media [41, 57]. Lastly, because this data is stored and available historically we can use the precise timings stored to construct longitudinal information about social media use that allows us to explore how patterns of behaviour have changed over time. This is especially important for the development of Just-In-Time Adaptive Interventions (JITAI) [58], or research that explores how use patterns change during poor mental health episodes.

Of course, there are potential limitations to timings too, such as life circumstances that may easily explain changes or differences in timings of use; someone may do shift work that changes between days and nights, or may recently have become a new parent. Timings of data generation also do not allow us to measure the times that someone was online and scrolling a social media site, though this could be approximated using data such as the times that someone interacts with content on the site. This so-called passive social media use has been identified as potentially being an important signal of declining mental health [59].

Using time-based data from social media provides another dimension through which we can understand social media users and how they change over time, and allows to approach mental health as a changing and moveable concept rather than a static classification.

Connectivity and interaction

Relationships, communication and willingness or unwillingness to interact with others is another common theme in the symptoms of mental health disorders [7]. By definition, social media serves a social purpose and there are several means of assessing social connection online. This may involve the number of 'friends' or connections someone has on the site, how many times they interact with them and how, and whether they contact specific people directly or broadcast messages to specific groups. This information about digital social connections can be used as a feature of prediction algorithms for mental health disorders. For instance, recent research has found that

mental health can be predicted with modest accuracy from the content generated by the users that people follow on Twitter [60]. This topic was also the subject of an infamous experiment by Facebook in 2014 which explored emotional contagion through social media based on the content that people observed their friends posting [61]. There have also been findings relating to friendships on Facebook, where the number of friends and the connectivity of the friendship network were associated with depression [40].

The nature of online connectivity is likely to be highly subjective, and it cannot be assumed that online relationships reflect somebody's offline interactions. However it is clear that they can be used a useful indicator none-the-less, and aligns with the requirement for diagnosis of a mental health disorder of some impairment in an individual's social functioning based on their symptoms [7].

Visual content

Social media generally allows for the sharing of visual context, as well as textual updates. This could include still images like photographs and graphics, videos and animations or short moving clips known as GIFs. As well as adding layers of meaning to the textual content they often accompany, visual content is the primary communication format of some social media sites such as Instagram, Snapchat and TikTok. However, visual content is comparatively less well understood than other forms of data as a digital phenotype [62] despite the fact that studies attempting to use images to predict mental health have been relatively successful so far. For instance, Reece et al. [63] employed computer vision techniques to extract the hue, saturation and brightness of images and found that these could be used to predict depression in young people. It is likely that the full potential of visual content will be achieved by using it as an input feature alongside the other features already described such as textual content and other metadata

Visual content is an important part of online communication today, and despite being less well-understood in mental health inference it has strong potential to represent valuable information that improves the quality of inference.

Summary

The features language, timing, connectivity and interaction, and visual content are all useful data-driven markers of behaviour change, and as alluded to, the best use of them tends to be from using different types of features together to build a more thorough picture. For instance, combining features of images and textual information to provide better context for images [64, 65], or using both the timing of social media use and text to identify when meaningful changes occur [66, 67].

1.3.2 Perspectives on the relationship between social media and mental health

Having outlined the ways that social media data can give signals of mental health and well-being, I will now outline the two fairly distinct perspectives that existing research has taken on the relationship between social media and mental health. These different perspectives essentially imply two causal directions, with one asking how the way people use social media affects their mental health and well-being, and the other asking how people's mental health and well-being affects the way that they use social media. The differences in the approaches taken between these fields in terms of their purpose, data and methodologies is outlined in Table 1.1. Whilst they are currently operating as largely separate fields recent research into longitudinal associations between social media use and mental health indicates that these two causal directions are actually likely to be reciprocal over time [68], though longitudinal causal research in this area is still in its very early phases [69–71]. A more detailed review of previous research across these areas is presented in Chapter 2. As outlined, these two fields do have slightly different purposes. Understanding the way that social media use affects the mental health and well-being of young people has been of particular interest given the considerable policy and research interest in recent years regarding the impact that social media use has on the mental health and well-being of young people [72], with particular focus on the dose-response relationship between social media use (usually measured by screen time) and declining mental health. However, more recent research has recognised the likelihood that the way social media is being used by the individual is likely to be more significant than just time spent [70].

Although they are currently largely separate, with different venues, methodologies and audiences, both directions of research have strong potential to inform one another. The relative success so far of predicting mental health signals from social media data indicates that those experiencing mental

Table 1.1: The differing purposes, data and methodologies used by the two alternative approaches to studying the relationship between social media and mental health.

	How does the way that people use social media affect their mental health and well-being?	How does people's mental health and well-being affect the way that they use social media?
Purpose	Improving guidance for safe social media use, improving controls on social media companies.	Improving early detection of mental disorders, improving understanding of the longitudinal development and changes to individual and population mental health
Typical Data	Data tends to be related to which sites participants have used, how long they have used them for and what they were doing whilst on those sites.	Data on the use of a specific platform/s with considerations of the language used, the timings of data generation, the social networks engaged with and the types of content shared.
Typical Methodologies	Regression modelling and structural equation modelling. Results tend to be communicated as effect sizes, model fit and significance testing.	Prediction and inference tasks that tend to use machine learning methods. Results are measured with prediction evaluations such as accuracy, precision and recall.

health disorders do use social media differently to those that are considered ‘healthy’. This has been observed using a range of features that include interpersonal connections, language, frequency and types of content being engaged with [40, 41, 57]. However, research about the impact of social media on mental health largely relies on attempts at capturing dose-response relationships that focus on the amount of time spent on social media sites, which is only one aspect of social media use as a digital phenotype [70]. It is likely that research into the relationship between social media use and mental health will benefit from more detailed digital phenotypes, such as those that are beginning to use techniques like screen recordings which then capture passive use which is likely to be very important in understanding mental health [73]. Similarly, identifying mental health signals from social media data would benefit from the more detailed measurement of mood such as Ecological Momentary Assessment (EMA) that are increasingly being used by researchers [74–77].

1.3.3 Current challenges

Whilst I have described the ways in which social media data can contribute to our current understanding of mental health, there are lots of challenges that we still need to understand in order to develop a well-grounded theory for the effective modelling of mental health from these data. Here I discuss three of the technical challenges that currently exist: the lack of ground truth data, the unstructured nature of the data collected and the ability to access digital footprint data for research. I reserve the ethical challenges for a fuller discussion in Section 3.

Lack of ground truth

The lack of high quality ground truth data in mental health inference from social media is arguably the greatest challenge to the field at present. Concerns about data quality have persisted across several years [41, 57], but so far little attention appears to be paid to resolving them. A lack of understanding of who is in the samples we are drawing poses significant risks to the safety and ethics of the systems that these data are used to train. Without knowing the sample we cannot understand who our methodologies do and do not work for, and therefore are unable to test for systematic biases embedded in algorithms. The lack of ground truth also extends to our understanding of the mental health outcomes that are being measured. There have been some solutions to this problem,

mostly relying on collecting survey data from online crowdsourcing platforms such as Prolific or Mechanical Turk [41], however these are generally limited to a single time point of ground truth data collection, and do not tend to be from population samples.

Other attempts at inferring ground truth in the literature tend to be guessing characteristics from the data itself. For instance, demographic attributes that are important for understanding variations in mental health, like gender, socio-economic status and age, are generally not available alongside social media data being collected online. In an effort to establish some understanding of the demographics of their samples many researchers have used algorithmic methods that estimate these characteristics from the data itself [78]. This might include guessing gender from account names and profile pictures [79], or age from styles of language use [80]. However, these can only ever achieve a certain threshold of accuracy, and reinforce gendered stereotypes that contribute to the systematic and ongoing bias against those who are LGBTQIA+ and genderqueer [81]. Many studies also derive their mental health labels from their input data, for example by searching for people who have stated “I am depressed” on Twitter to find positive cases of those who are depressed [41]. This has the potential to introduce several forms of bias, namely that ‘positive’ cases might not actually be depressed, that they may only represent a specific subset of depressed people on Twitter, and that training data is then implicitly linked to the outcome before we have even attempted to use any machine learning techniques. This issue is discussed further as part of the review in Chapter 5. Some studies also use psychiatry or psychology professionals to review and label the data with mental health outcomes based on their professional expertise [67, 82]

Having the ability to test the fairness and bias in algorithms, especially those with social implications, is vital to the development of technologies that behave fairly for everyone.

Unstructured and Highly Variable Data

The nature of social media means that the volume and type of data produced is entirely within the control of the social media account user. Unlike traditional data collection where the researcher has designed what data they will collect and how many questions they will ask, social media has no defined frequency or quantity [83], which also means that we cannot apply traditional approaches to missing data in those who produce very little data. The consequence of this is that the volumes of data that can be collected are highly variable. So, whilst social media presents the potential for

high-resolution understanding of mental health we also need to take into account the effect that the huge variability in training data brings to algorithm effectiveness [83]. This is likely to mean conducting sensitivity analyses to see how prediction accuracy is impacted by variable rates of data production. The variability in social media data is also likely to include individual differences in which social media sites people use for which occasions or feelings [84], with a full high-resolution understanding of social media behaviour only likely to be achieved if we are able to collect data from multiple social media sites for the same individual.

Data Access

One of the core limitations of using social media data for research is the constraints imposed by the commercial companies who own social media platforms, and ultimately control access to the data [12]. Data from digital phenotypes are usually generated via third-party apps like Facebook, Instagram, or Reddit, and non-social media sites too like Strava for health monitoring or Spotify for music streaming. Access to these data is largely controlled by the companies who own them and if they are accessible to researchers then they are made available in a controlled manner through an application programming interface (API). An API is a programmatic system for requesting and receiving data from a central provider, like a company. They usually require short passcodes called API keys in order to be accessed, which can ensure that those accessing the API do not make too many data requests, or that data is not delivered that the requester does not have permissions for.

Approaches to making data available varies across social media companies, and are currently a major limitation in cross-platform social media research [12, 85]. A recent report by the Royal College of Psychiatrists (UK) into the potential harms and benefits of social media for young people recommended the “regulator to urgently review and establish a protocol for the sharing of data from social media companies with universities for research into benefits and harms on children and young people” [72]. At the time of writing Facebook (now *Meta*) owns the sites Facebook and Instagram, which are two of the most popular social media sites in the world. Facebook does not allow any research access to its data and APIs outside of a handful of selected projects, which was a change made largely in response to the Cambridge Analytical scandal [85]. This has led to some researchers adopting a web-scraping strategy (automatically capturing all the content of

a webpage) as an alternative way of accessing the data, which has in some cases led to threats of prosecution by Facebook [86]. It is possible, as some sites do, to make user-level data accessible to applications if the user gives consent for this through a process known as Open Authentication (OAuth), which would allow for users to share their data with research studies that they had consented to take part in. Alternatively sites like Twitter have made their public API highly accessible, and in 2021 released an API specifically for academic researchers which gives full access to the whole of Twitter's history of public data, allowing verified researchers on their platform to access up to 10 million tweets per month [87].

While access to the data of commercial entities is one concern, the other is the level of transparency about how user data may be influenced by personalised engagement algorithms that prioritise the content that people see on their 'timelines' or 'newsfeeds', such that they do not genuinely represent the time-ordered content being produced by others [88]. This is commercially sensitive information, and so is unlikely to ever be shared, but has the potential to highly influence the types of user engagement data we receive to construct digital phenotypes. This may include the content users interact with, the political leaning of the content being shown to users, or the overall emotional valence of the content being displayed. Changes in the algorithms used by Google's search function have been attributed to the lack of replicability of the highly influential piece on the prediction of flu trends using data from Google Trends in 2009 [89].

Summary

The current limitations of social media data available for mental health research all present challenges to the use of these data for making robust inferences. Some of these challenges, such as data quality, may be more straightforward for the research community to address than the possibility of influencing the corporations that generate the data for our use. However, it is important that we consider how the data that we use may impact the effectiveness of the digital phenotypes we are attempting to generate.

1.3.4 Twitter

Having considered the different data types available from social media that can be useful for inferring mental health from social media, as well as the limitations that researchers currently face

in attempting to use these methodologies, I will now also give a more detailed description of the specific social media site that this thesis focusses on, Twitter.

Twitter was publicly launched in July 2006. It is based around the concept of ‘micro-blogging’, where users can use their accounts to broadcast short messages to other users; these messages are known as *tweets*. At the time of writing tweets are limited to 280 characters in length, which was increased from 140 in 2017. Tweets will be displayed on the personal feed of those who follow the user that created them. Tweets were originally text based, but now can also include attachments in the form of images, videos or links. There are various ways that users can interact on Twitter which are outlined in Table 1.2. All of replies, likes and retweets/quote tweets come under what Twitter calls ‘public metrics’ and the numbers of each are displayed underneath every tweet [87]. As well as individual interactions Twitter also aims to summarise what is ‘trending’ on the site by automatically generating lists of tweets that refer to particular topics or hashtags that are popular in real time. User connections on Twitter can be represented by a directed network, in that a user may *follow* another user so that tweets of the followed user will appear on the feed of the user that followed them, but not vice-versa unless they are followed in return. This is unlike platforms such as Facebook where a connection with another person creates a reciprocal link between the two accounts. This feature of Twitter means that accounts can follow very high numbers of people, whilst being followed by very few people and reciprocally accounts can be followed by very high numbers of people and follow few. It is possible for Twitter accounts to be made private, which means that the user must approve all requests to be followed, and their tweets will only be viewed by approved followers. This also means their tweets cannot be retweeted or included in collections of tweets containing specific hashtags. Tweets from private accounts will also not appear on Twitter’s public Application Programming Interface (API) [87]. An additional privacy feature was introduced in 2020 that allows users to restrict who can reply to their tweets, even if they have a public account, so that only people they follow or only people mentioned in the tweet can reply. This feature was introduced to combat ongoing issues of targeted abuse experienced by many users [90].

All of these data described, from tweet text to social interactions to public metrics of tweets, are provided by Twitter’s public API, with Twitter making a specific Academic API available early in 2021 which allows those conducting verified research projects to access up to ten million tweets

Table 1.2: A summary of different types of interaction available on Twitter and the particular meaning and nomenclature associated with these interaction types.

Interaction type	Meaning
Like	Any user can like another's tweet. A user's liked tweets are stored in a list on their profile.
Reply	Users can reply to another user's tweet (or their own), which are then presented as a chain of tweets representing a conversation. A reply is a tweet itself, and generally includes an @-mention at the beginning of the user/s being replied to.
Retweet	Users can reshare a tweet from another user to their own followers. The original tweet data is shared with the tweet, including the total number of likes, replies and retweets it has received. This is considered to be a type of tweet on a users profile, and so is returned in any API searches for a user's tweets unless filtered out.
Quote	Quotes were introduced in 2020, and are retweets with the added functionality of allowing users to reshare tweets with their own comments attached to them. Similarly to retweets these are considered their own types of tweet.
@-mention	This refers to when another user is referenced in a user's tweet. This is done by writing an '@' symbol followed by the username of the other user, and creates a hyperlink to that user's profile in published tweet. Users are notified of any mentions.
Hashtag	A hashtag is written into a tweet by a user, and uses the symbol ('#') followed by a word or phase (with no spaces). This is used to link tweets which are concerned with a particular topic, theme or event.

per month [87]. The Academic API also accounts for previous concerns about Twitter data, which were that only 3,200 tweets were available from each individual's history, and that when collecting historic data it was not clear whether the full set of tweets were being shared or only a pseudo-random sub-sample. Now, the new API guarantees all data is shared that matches the parameters set by a query, provided the payload is within the overall monthly limit. These features make Twitter one of the most accessible social media sites available to conduct research with.

The most recently published review of the use of social media for mental health inference showed that Twitter is by far the most used platform for this type of research, with the review finding 30 studies using Twitter data compared to six and four for Facebook and Instagram respectively [41]. This is likely to be due at least partly to the accessibility of Twitter data, as discussed in Section 1.3.3.

1.3.5 Summary

Overall this section has summarised what social media is, and why it can be a useful source of digital footprint data when attempting to derive digital phenotypes for mental health. This is because social media contains features like language, datetime data, information on social connections and also visual content, all of which have been shown to be useful in the prediction of mental health outcomes at both tweet and user level.

1.4 Questions of ethics

Alongside our attempts to approach the practice and technical challenges we face in social media research, the ethical use of internet-based data for research is a matter of ongoing discussion and development [91]. Research ethics is a pillar of the responsible research process, and has been for decades [21, 92, 93]. However, the many years that academics have spent forming boundaries for acceptable research in the medical and psychological sciences does not always neatly translate to new forms of research that take place primarily online, even though we can still frame these through the lens of the basic ethical principles of respect for persons, beneficence and justice set out in the Belmont Report [92].

Internet-based data gives researchers an opportunity like never before to understand the lives of

thousands of people with minimal data collection effort. Given this ready access to the data of such huge volumes of people one of the primary ethical concerns about the use of internet-based data is the lack of informed consent from those whose data is being collected and used [94]. Ethical questions often tread a difficult line between what is acceptable and what is legal, and in the earlier stages of internet-based research it was more common to consider data that was publicly available on the internet as implying consent for its use by anyone who could access it [94, 95]. This does not explicitly undermine research ethics procedures, and also at the time did not go against guidance since it did not require any interaction with human participants. However, as time has gone on we have developed a more sophisticated approach to the complex public/private nature of internet data [94, 95]. We are now more aware of the fact that data that is publicly available is not necessarily intended to be public by those who produced it [91, 96] and that even though terms of service for social media platforms stipulate that data may be shared this is not broadly understood by users [97]. For instance, in a study by Williams and colleagues in 2017 [98] around 80% of Twitter users expected that researchers would ask for their consent to use their data in published outputs. Similar results from a separate study in 2018 [99] found that most Twitter users would expect to be aware that they were ‘participating’ in a study, and that their decision on whether or not the use of their data was ethical was highly contextual. Williams et al. [98] also found that there were different reactions from participants depending on the identity of the user and whether the organisation using their public Twitter data was the state, a commercial entity or a research institution. For instance, those identifying as LGBTQ+ or female had higher concern about the state collecting and using their data, and were more likely to expect to give informed consent [98].

The issue of open research and reproducibility also becomes particularly contentious with internet-based data, since openly sharing a research dataset might mean sharing inferences about individuals’ mental health (accurate or not) that have been made to construct training data, alongside their username or the text of their tweets which can easily be traced back to their profile, and sometimes their real-world identities [94, 95]. Researchers also need to take care of how they reproduce tweets in research papers, even when care is taken to reword direct quotes so that they cannot be traced back, as this also risks disclosing the identities of individuals [96]. Another issue is the inability of researchers to monitor participant ‘participation’ in their studies; in traditional designs it is generally clear when someone has withdrawn from a study, or intends to, but on the internet this may

look like people later deleting content that has already been collected by the research team [94]. If this data is stored longer term for re-use or reproducibility it creates a permanence which goes against the decision of the original content creator to delete it.

Today there are a variety of ethical guidelines available from organisations such as the British Psychological Society [94], and the Association for Internet Researchers [95] which specifically tackle these issues in internet based research. These guidelines recognise the additional risk that is generated by researchers interacting with and analysing data, deriving insights that were not previously transparent from the data, the need for careful consideration of the anonymity of those whose data is included, and also the potentially high exposure to distressing content online that could adversely affect researchers [94, 95]. As it currently stands, many social media or web-scraping studies do not require ethical approval by research institutions, and so the responsibility for behaving ethically (and the definition of what this means) is largely down to individual researchers and research groups. Even if a study does reach the threshold for ethical approval, those on the boards of research ethics committees do not always have a nuanced awareness of the ethical risks of internet-mediated research [91].

Ethical concerns about mental health inference from social media also inherit dilemmas from the field of machine learning and artificial intelligence (AI), which are currently grappling with the concerning lack of representation in digital data, and the direct impact of this on the fairness² of algorithmic systems [100]. Sampling bias in social media-based research is likely to impact on the potentials of systems to work effectively, both in terms of those who do and do not use the internet, and also how the samples we are using to train these systems might poorly represent the general population [78]. There are also ongoing concerns about the autonomy of those who may eventually use these systems, and how algorithmic inferences will be taken into account in the contexts they are used; this is especially important in the case of children and vulnerable adults [101]. Lastly, there continue to be significant questions raised about the applications of these tools in the real world. There is no question that such tools could be used with malicious intent, and this risk rises when they are solely based on open-source data and so could be reproduced and applied outside of their intended settings which especially applies to social media sites like Twitter and Reddit [102].

²It is worth acknowledging that there is not one single definition of *fairness* in algorithmic systems, much like there are varying definitions of what it means to be ethical.

As mentioned in Section 1.2, similar technologies in emotion AI are already being deployed and used for surveillance purposes [32].

In summary, there are significant ethical concerns in this field at present, with informed consent and the fairness and accountability of developed systems often at the core. These are concerns that can only be fully addressed by having a thorough understanding of, and consent from, our underlying samples. In order for this to happen without slowing the progress of the field we need to find methods for dataset creation and sharing that are safe, secure and align similar measures. Other fields, most notably genetics, have faced and addressed similar concerns by forming international consortia that allow for researchers to collaborate with much larger samples than individual research groups had access to [103]. A similar approach could be the answer for digital phenotyping too.

1.5 Data linkage with social media

So far I have discussed some of the current challenges that we face in making the best use of social media as a digital phenotype, such as the lack of effective ground truth data, challenges accessing the data at all or with an appropriate level of informed consent. The best way to address these challenges is likely to be by appreciating social media data as a complement to traditional data sources, and so linking social media data with new or existing survey data [104]. This linkage would provide the ground truth evidence needed to make sense of social media data, as well as ensuring this could be done with the full and informed consent of those taking part.

Data linkage, also called *record linkage*, refers to the process of combining two or more previously unassociated datasets from the same population, and attempting to match the records from each dataset that belong to the same person [105]. Data linkage has become an increasingly popular methodology as populations have amassed and stored more and more data across multiple distinct systems, and is particularly relevant across administrative and health data sources since it allows for powerful understanding of outcomes that could not be measured in other ways [106]. One of the aims of mental health data science, especially in the UK, is to leverage on data linkage improvements in order to make use of large-scale administrative data [8]. There have been many recent innovations in this area. For instance, Pearson and colleagues [107] linked data about moth-

ers whose children had been subject to care proceedings with their NHS mental health records and found that two-thirds of these women had received mental health care, with the majority of these being seen under secondary or tertiary services which indicates severe mental health concerns. Other examples are large-scale linkage projects which house multiple datasets, like the Secure Anonymised Information Linkage (SAIL) Databank which enables anonymised access to linked data across primary care services, education, the fire service and more [108].

Whilst data linkage is an incredibly powerful technique for uncovering new patterns and knowledge, it is not always straightforward to achieve, with the effectiveness of linkage programmes being highly dependent on the common fields available across the datasets being linked, leading to varying margins of error [105, 106]. There are three main methodologies for data linkage which depend on these common fields; these are *deterministic*, *probabilistic* and *machine learning based approaches* [105]. Deterministic linkage applies when there is a single unique identifier that is recorded in both datasets, probabilistic and machine learning methods are applied when such an identifier does not exist and instead records are attempted to be linked with combinations of other fields such as names and dates of birth [105]. One of the benefits of linking social media data is that it depends on a straightforward deterministic method of linkage by asking for a unique username or identifier on the social media platform, or for the user to directly authorise access through an OAuth flow. On the whole we would expect this method to generate minimal error, with the main potential for mistakes to come from misspelled usernames, intentionally false usernames or in instances of a one-to-many account ownership so that a single person might have multiple social media accounts on the same platform which are not all captured. By linking social media data to high quality ground truth data we start to be able to realise the potential for social media data as a detailed longitudinal record of human behaviour, and thus to develop effective digital phenotypes from it.

Of course, there are studies that have conducted online surveys for the purpose of collecting ground truth data for social media mental health inference [41], but relatively few that use existing surveys as an opportunistic means to request data linkage with social media in representative samples. Mneimneh and colleagues [109] evaluated the consent rate to requesting data linkage at the end of existing mental health surveys in the US, Belgium and Saudi Arabi. Their evaluation found that between 20-36% of their survey respondents were Twitter users, and of those that were, 24%, 27%

and 45% of people in the US, Belgium and Saudi Arabia consented to their Twitter data being linked. Notably, the Saudi Arabian survey was the only one that was face-to-face, which may have had an impact on the high consent rate. The team also found that the demographics and health features of those that consented was largely similar to those who did not consent, but that those who reported more sensitive information in the surveys were more likely to consent as were more frequent Twitter users. These findings were similar to two other previous studies that attempted to link Twitter and survey data [110, 111], had similar consent rates as well as Twitter users, and also found that consent rates were higher when consent was requested in a face-to-face context.

The studies conducted in this area so far show promising potential for Twitter data to be linked to existing survey data, and that while even though consent remains relatively low, it does not appear to be biased by demographic characteristics [109, 111]. By linking data in an observational study design, rather than one specifically intending to predict mental health, we also have the opportunity to explore many other variables that may not typically be collected. A limitation of these studies however is that they have been specific to data collected in panel samples at particular time points; this means that the individuals sampled and with linked data may not be followed up again in the future, or have previous data available at other survey time points. Cohort studies are an alternative model of longitudinal, population representative data collection that would allow for longitudinal data comparison, as well as being able to make use of typically expensive or time-intensive variables that have been collected over many years such as neuroimaging, genetic or health record data. These would be challenging financially and practically to collect all at one time point for an individual study. I will go on to discuss the potential for cohort studies to deliver particular benefit if they can be used successfully to link and store digital phenotypes.

1.6 The case for cohort studies

In Sections 1.2 and 1.3 I have laid out the exciting potential for social media to complement current innovations in mental health early-intervention, monitoring and treatment, as well as outlining the practical difficulties in obtaining data from robust and well-characterised samples so that we can conduct reproducible and high-quality science. In Section 1.4 I have discussed the additional challenge of using social media in an ethical way with appropriate consent from those whose data is being collected, and in Section 1.5 I have summarised the emerging practice of linking social

media and health survey data to overcome some of these difficulties, but acknowledge the challenge of collecting rich longitudinal information. In this section, I make the case that cohort studies are a promising vehicle for addressing these limitations, and introduce the Avon Longitudinal Study of Parents and Children which is the focus of this thesis.

1.6.1 High quality ground truth data

Cohort studies, also known as *Longitudinal Panel Studies* (LPS), are observational research studies that prospectively follow the same group of people, a cohort, over time. Cohort studies are a particularly important resource due to their ability to track changes within individuals over time, and the opportunity to observe the aetiology of diseases and their progression using data measured before and after a disease has developed, and in those who may or may not have exposures of interest [112]. These studies may start from the time that the participants are born, known as *birth cohort studies*, and may be representative of a local or national population, known as *population-based cohort studies*. Cohort studies usually run for long periods of time, and collect data over many years about their participants, resulting in abundant datasets that characterise their cohort in significant detail. In Section 3 I discussed the limitation of single-survey point collection of ground truth; Russ et al. [10] hypothesise that effective screening tools for mental health “would need to use longitudinal clinical assessments and social context, alongside physiological, genetic and imaging data where available” to make the best possible guess of whether intervention is needed and likely to be useful to the patient in a clinical context. These kinds of data would be extremely challenging to collect in any context outside of a cohort, given that they require substantial resources and participant time. In a cohort however we can make use of the fact that these resource-intensive data sources already exist, and in many cases exist longitudinally, and so are ready to be linked and used alongside new data sources [113]. Not only can we use the core cohort data, but we can also make use of the cohort as a central hub for multiple data linkage projects, which could allow us to link data from administrative or health datasets with digital phenotypes via the study. The availability of this data gives us an excellent basis for high-quality ground truth data for research, with which we can accurately test algorithms in population representative samples.

Reciprocally, linked digital data generated by our day-to-day activity can supplement traditional research methods, and particularly the rich data already available in cohort studies, to maximise

the utility of the data already collected [113, 114]. Attrition and missing waves are a considerable issue in long term studies, with it being highly unlikely that any individual in a cohort successfully completes every single data collection exercise. By using digital phenotypes that involve passive data collection without any effort on the part of participants we may be able to supplement the data they provide in other ways, and fill periods of missing time with valuable information that they generate through other means than surveys [113].

1.6.2 Ethical data collection and sharing

While there are many ethical concerns about using social media data, as discussed in Section 3, linked data in cohorts can help researchers to address these. Specifically on the themes of consent, disclosure, security and archiving as laid out by Sloan et al. [116], cohorts can ensure that there is appropriate consent in place for the linkage of this data, that disclosure is managed by the cohort team with an emphasis on protecting the privacy of participants, and that only anonymised and relevant data is released to researchers. Archiving can also be managed with specialised software built for the use of cohorts (see Epicosm [117]), and led by dedicated data management teams that are already established within the cohorts, and experienced in managing and handling sensitive, disclosive data. The long running nature of cohort studies and their ongoing dialogue with participants does also mean that consent can be handled more flexibly than in other contexts, with participants able to choose to opt-in or opt-out over time, so that the data provided to researchers will reflect this change in their wishes for their data to be included.

The existing infrastructure for cohort studies and their data management teams may also enable the facilitation of wider data harmonisation and sharing practices among existing consortia. For instance, the Cohort and Longitudinal Studies Enhancement Resources (CLOSER) consortium that currently facilitates data harmonisation, linkage and sharing across 19 UK longitudinal studies, including ALSPAC³.

Whilst cohorts are in theory a promising place to conduct data linkage, there are still outstanding questions on how to do this sensitively with respect to the cohort participants. Cohort participants, due to the long-term nature of the studies, have established relationships with their studies and sometimes with the study staff. They continue to allow the study to collect their data because they

³<https://www.closer.ac.uk/explore-the-studies/>

trust that it will be used for the good of others, and that it will be kept safe. As a result, it is very important to ensure that data linkage projects in cohorts are undertaken with the general support of the cohort, and that participant data is used in a way that they find acceptable [113]. If we can develop a framework that allows this to happen, then we may be able to start drawing on this potential.

1.6.3 The Avon Longitudinal Study of Parents and Children

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a prospective, population based, multi-generational cohort study that began with the recruitment of pregnant women resident in Avon, UK [118–120]. At the time of writing, the index cohort of participants are approximately thirty years old. The health and development of the index children and their parents have been followed since the initial pregnancies. Within ALSPAC the original pregnant women and their partners are referred to as **Generation Zero (G0)** and the index children as **Generation One (G1)**. There are also now **Generation Two (G2)** participants enrolled in ALSPAC, who are the offspring of the G1 participants.

Specifically, ALSPAC recruited 14,541 G0 pregnant women who were resident in Avon, UK with expected dates of delivery from 1st April 1991 to 31st December 1992. Of these initial pregnancies there were a total of 14,676 fetuses, resulting in 14,062 G1 live births and 13,988 children who were alive at 1 year of age. Additional offspring that were eligible to enrol in the study have been welcomed through major recruitment drives at the ages of 7 and 18 years, and through opportunistic contacts since the age of 7. A total of 913 additional G1 participants have been enrolled in the study since the age of 7 years with 195 of these joining since the age of 18. This additional enrolment provides a baseline sample of 14,901 G1 participants who were alive at 1 year of age. On top of the G1 cohort, the G0 mothers and their partners are also participants in the study, and have been followed longitudinally since the time that they were recruited.

The types of data available in ALSPAC are vast, and can be explored through the variable search tool available online.⁴ These data range from genetic profiling to placenta samples to annual surveys of behaviour and life events. ALSPAC also has an extensive data linkage programme, with a history of working closely with participants to establish limits of acceptability and to build trust

⁴<http://variables.alspac.bris.ac.uk/>

in the cohort's approach to sharing their data [121]. Given the range of data available, and the variation in the exact sample that participated in each data collection event I have described the precise characteristics of the sample used in each chapter individually.

ALSPAC is an ideal cohort with which to conduct novel social media data linkage for a variety of reasons. First, the G1 cohort at the age of thirty represent a generation of children who grew up in the early stages of the internet, and who have seen the development of social media throughout their young adulthood. Their age group comprise the highest proportion, 38.5%, of Twitter users world-wide,⁵ and so they are a viable cohort to approach for social media, and specifically Twitter, data linkage. Second, this linkage project is an opportunity for the cohort to align with strategic priorities for population cohorts from the Medical Research Council [113] and the Wellcome Trust [122], who are keen for new types of data linkage to be used in order to maximise the value of existing cohort data. Third, the ALSPAC management and data teams are experienced at working with ALSPAC participants and researchers to enable innovative data linkage strategies, which have also included the exploration of transactional data linkage [121, 123].

1.7 Summary

In Sections 1.2 to 1.5 of this introduction I have described some of the latest innovations in the field of mental health data science, but have also illustrated that there are many outstanding areas which require better evidence to inform more robust inferences about the links between social media and mental health. I have posited that cohort studies are an efficient means of addressing these limitations, given the vast amounts of individual level data available for linkage that are also available longitudinally. This being said, there are certain challenges related to linking data in a cohort that require investigation, such as whether enough people use the platforms in question, whether this type of linkage is acceptable, and whether the population of people who are willing to consent to data linkage are representative of the population we are interested in studying. This thesis is concerned with testing these questions, and also then exploring whether social media that is linked in a cohort study can actually help us to achieve the aim of better characterising training samples, and conducting mental health data science that gives us the ability to make more robust

⁵<https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/>

conclusions about the role social media can play in mental health inference.

**Part A: Establishing the use of digital
phenotypes in ALSPAC**

Chapter 2

The mental health and well-being profile of social media users

This chapter is adapted from Di Cara NH, Winstone L, Sloan LS, Davis OSP, & Harworth CMA (2021). The mental health and well-being profile of young adults using social media. Under review at *npj Mental Health Research*.

Using the definitions provided by the CRediT¹ framework for academic attribution: I was responsible for data curation, formal analysis, investigation, methodology, visualisation, and writing (original draft and reviewing and editing). OD and CH were responsible for funding acquisition, conceptualisation, methodology, investigation, supervision and writing (reviewing and editing). LW and LS contributed to revisions and the methodology.

¹CRediT framework: <https://casrai.org/credit/>

Abstract

Background The relationship between mental health and social media has received significant research and policy attention. However, there is little data about who social media users are which limits understanding of confounding factors between mental health and social media.

Method Here we profile users of Facebook, Twitter, Instagram, Snapchat and YouTube from the Avon Longitudinal Study of Parents and Children population cohort (N=4,083). We provide estimates of demographics and mental health and well-being outcomes by platform.

Results We find that users of different platforms and frequencies are not homogeneous. User groups differ primarily by sex and YouTube users are the most likely to have poorer mental health outcomes. Instagram and Snapchat users tend to have higher well-being than the other social media sites considered. Relationships between use-frequency and well-being differ depending on the specific well-being construct measured.

Conclusion The reproducibility of future research may be improved by stratifying by sex and being specific about the well-being constructs used.

Aims

One of the limitations of conducting analyses with internet-based data is that the samples being used often lack data on their demographic characteristics, and characteristics related to the outcome of interest, like associated mental health outcomes. This chapter aims to provide a thorough description of the demographics of social media users in ALSPAC, and also to describe how rates of mental health outcomes differ across social media platforms. This is helpful for contextualising analyses in future chapters, as well as more generally for understanding the typical prevalences of mental health outcomes on each platform.

2.1 Introduction

The trails of data left online by our digital footprints are increasingly being used to measure and understand our health and well-being. Data sourced from social media platforms has been of particular interest given their potential to be used as a form of ‘natural’ observational data about anything from our voting intentions to symptoms of disease. There is not a single, widely agreed definition of the term ‘social media’ [124], but for the purposes of this study we understand it to be a broad category of internet-based platforms that allow for the exchange of user generated content by ‘users’ of that platform [42]. Both the huge volumes of data available on such platforms, and their increasing uptake across the population [125] have led to two main fields of interest in the intersections of social media and mental health. These are the prediction of mental health and well-being from our online data [41] and, somewhat reciprocally, the influence of social media on our mental health, particularly in the case of children and young people [72, 126]. These fields both ask fundamental questions about the mental health and well-being of social media users, to either understand the ways our mental health influences our social media behaviour, or how our social media behaviours influence our mental health.

Across both contexts a wide range of psychological outcomes have been studied, including predicting suicide at a population-level [127] and individually [128], mapping the influences of social media platforms on disordered eating [129] and self-harm [130], understanding the impacts of cyberbullying through social media platforms [131, 132], and even ethnographic research into online support networks [133]. As highlighted in a recent review which considered research on the relationship between social media use and well-being in adolescents [70], there has tended to be an inherent assumption that social media is the cause of harm when examining the effect of social media on our health. However, recent investigations such as those by Orben and Przybylski [134, 135] and Appel and colleagues [136] illustrate that the role of social media in causing harm may be overestimated. It seems likely that there is some reciprocal relationship between mental health and social media, that requires longitudinal research studies to begin to understand the complexity, coupled with large representative samples to explore the heterogeneity [137, 138]. A developing research area is seeking to answer this question using high time-resolution methodologies such as ecological momentary assessment (EMA), which may provide the level of detail needed to make progress in understanding the nuances of causal effects [139]. Further, there is increasing attention

on the role of within-person effects that see impact change between contexts, as well as individual differences [142]. Meanwhile, attention has also been drawn to the comparative lack of investigation into the potential benefits of social media, such as access to peer support and the ability to readily connect with friends and family, or into the psychological well-being of social media users as opposed to focusing on pathology. Similarly, most psychological prediction tasks using social media focus on predicting illness rather than wellness [41, 143].

Regardless of the direction of interest in the relationship between social media and psychological outcomes, researchers face common challenges, with one of the primary issues being a lack of high-quality information on the characteristics of the whole population of social media users [144]. Valuable demographic information on social media users in the United States is regularly produced by the Pew Research Centre [145], but often researchers rely on algorithmic means to make predictions about the demographics of the groups they study online if they are not recruiting a participant sample whose demographics are known and can be recorded [41, 144, 146]. What we do know about social media users is that they are not homogeneous. The demographic features of populations using them vary across platforms and do not tend to be consistent with the characteristics of the general population [78, 145–147]. This work on the demographic context has been important in understanding the samples that can be drawn from social media platforms, but there remains a lack of information about other characteristics of social media users that are relevant to study outcomes, including mental health and well-being. Consequently, attempts to compare user well-being and mental health between platforms may be unknowingly confounded by differences in the mental health profile of each individual platform. Mellon and Prosser [147] investigated this form of selection bias with respect to differences in political opinion between Facebook and Twitter, and noted the potential for study outcomes to be biased when the outcome variable of interest is associated with the probability of being included in the sample [148]. This also has implications for our assessment of mental health and well-being classification algorithms. For instance, if using Twitter data to classify depression in a random sample of users how many of these users should we expect to be depressed? Should we expect to find more depressed users on Facebook or Instagram? This bench-marking would allow the research community, who frequently face the challenge of establishing reliable ground truth in social media research, to contextualise the sensitivity and specificity of developed models [41, 144].

This study aimed to address the gap in the availability of high quality descriptive data about social media users by describing social media use in a representative UK population cohort study, the Avon Longitudinal Study of Parents and Children (ALSPAC) [118]. I aimed to profile the users of the social media platforms Facebook, Instagram, Twitter, Snapchat and YouTube by considering a range of mental health and well-being measures that are regularly studied, with the objective of better characterising social media users against variables of interest to researchers. These measures included disordered eating, self-harm, suicidal thoughts, and depression as well positive well-being outcomes which are sometimes neglected in the context of social media research [70, 135, 142] like subjective happiness, mental well-being and fulfilment of basic psychological needs. In answering my research questions I also sought to illustrate how cross-sectional data from a representative population cohort *can* provide meaningful contextual information that informs the way we interpret past and future research about social media users and their mental health. Unlike other studies using cross-sectional data [70] I had no intention of exploring causal questions, but aimed to address unanswered questions of who social media users are, and whether selection bias across platforms may have the potential to unintentionally bias outcome statistics about mental health and well-being.

Specifically, my research questions were:

- 1) Are there demographic differences in patterns of social media use (e.g. frequency)?
- 2) Are there demographic differences in the user groups of different social media platforms?
- 3) Are there differences in the mental health and well-being of those using social media sites at different frequencies?
- 4) Are there differences in the mental health and well-being of user groups of different social media platforms?

2.2 Methods

2.2.1 Sample Description

The sample for this study is drawn from the Avon Longitudinal Study of Parents and Children (ALSPAC) [118–120]. Pregnant women resident in Avon, UK with expected dates of delivery from 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled was 14,541. Of these initial pregnancies, 13,988 children were alive at 1 year of age. When the oldest children were approximately 7 years of age an additional 913 children were enrolled. The total sample size for ALSPAC of children alive at one year of age is 14,901. However, since this time there has been a reduction in the sample due to withdrawals, deaths of those in the cohort and also people simply being lost to follow up. As such the exact number of participants invited to each data collection activity changes with time. Please note that the ALSPAC study website contains details of all the data that is available through a data dictionary and variable search tool (<http://www.bristol.ac.uk/alspac/researchers/our-data/>). Study data were collected and managed using REDCap electronic data capture tools hosted at the University of Bristol [149].

The analysis presented in this study is based on a sub-sample of 4,083 participants who responded to a self-report questionnaire at a mean age of 24 years old in 2016/17. The survey was sent to 9,211 currently enrolled and contactable participants, of whom 4,345 (47%) returned it. To maintain a consistent sample throughout the following analyses I considered the 4,083 observations with complete cases for questions related to self harm, suicidal thoughts, disordered eating, and social media use, and without the respondents who said that they ‘didn’t know’ whether they had a social media account ($n < 5$); no respondents stated that they did not have a social media account. As well as the survey at age 24, I considered the responses by those in the main sample to a survey one year previously, at age 23, which collected the well-being measures and the Moods and Feelings Questionnaire, matched to their social media use responses at age 24. This resulted in a sub-sample of 2,862 participants who had responded to both surveys. Table 2.1 gives a comparison of the demographic breakdowns across these samples.

Table 2.1: The number of participants in each of the two samples used in this study, subset by demographic characteristics.

Demographic	ALSPAC Cohort N = 14,901	Main Sample N = 4,083	Sub-Sample N = 2,991
Sex			
Female	49%	66%	69%
Male	51%	34%	31%
Missing (N)	23		
Ethnicity			
Ethnic Minority Groups	5.0%	3.7%	3.5%
White	95%	96%	97%
Missing (N)	2,829	461	332
A Levels			
No	23%	19%	19%
Yes	77%	81%	81%
Missing (N)	10,801	1,384	786
Parental Employment Class			
Non Manual	68%	76%	77%
Manual	32%	24%	23%
Missing (N)	3,406	566	407

Note:

Parental employment class was collected pre-birth of G1 cohort

2.2.2 Measures

This study considered the participants' responses to a range of mental health and well-being measures, as well as demographic data. A brief overview of each of the measures used is given below.

Demographics

Throughout this paper, I use *Male* and *Female* to refer to the participant's assigned sex at birth. Participant ethnicity was reported by their parent/s, and is available in the data as *White*, *Ethnic Minority Group*, or *Unknown*, where Ethnic Minority Group was only available as one group rather than broken down into specific ethnicities. There were two variables relevant to socio-economic status. The first was whether the participant had achieved an A Level or equivalent qualification by age 20, the second was their parents' occupation. Parental occupation was measured using

the Registrar General's Social Class schema [150], and was collected prior to the birth of the index cohort; I took the higher occupational class of the participant's parents where available and grouped the overall schema of six categories into those in *manual work*, and those in *non-manual work*.

Social Media Use

Social media use was measured using three questions. These were: (1) *Do you have a social media profile or account on any sites or apps?* with possible responses of 'Yes', 'No' or 'Don't know'; (2) Given a list of social media sites, *Do you have a page or profile on these sites or apps, and how often do you use them?*, where the social media sites were listed and response options were 'Daily', 'Weekly', 'Monthly', 'Less Than Monthly' or 'Never'; (3) *How often do you visit any social media sites or apps, using any device?* with response options being 'More than 10 times per day', '2 to 10 times per day', 'Once per day' or 'Less than once per day'. Here, the definition of 'social media sites' in questions (1) and (3) was left to the participant to interpret, whereas in (2) a specific list was provided. In the following analyses I have summed responses for the use frequencies per platform from question (2) so that 'Weekly', 'Monthly' and 'Less than monthly' are combined to represent 'Less than daily'.

Mental Health

Depressive symptoms were measured using the short Mood and Feelings Questionnaire (MFQ) [151], a 13-item scale that has been validated for measuring depressive symptoms in adolescents [152] and in young adulthood [153]. Scores range from 0 to 26, with a higher score indicating more severe depressive symptoms [152]. Here I applied a cut-off score of 12 or above as indicating depression [153].

Suicidal thoughts were assessed with the question *Have you ever thought of killing yourself, even if you would not really do it?* with those who indicated that they had 'within the past year' being included. Similarly, intentional self-harm was assessed by asking if participants had *hurt [themselves] on purpose in any way* and I included those who said this had happened at least once within the last year.

Disordered eating was a composite variable that included participants who indicated that they

had been told by a healthcare professional that they had an eating disorder (anorexia nervosa, bulimia nervosa, binge eating disorder or another unspecified eating disorder). Participants were also included if they indicated they had engaged in any of the following behaviours at least once a month over the past year with the intention of losing weight or avoiding weight gain: fasting, throwing up, taking laxatives or medication. This classification of disordered eating followed a similar methodology to that used by Micali and colleagues [154].

Well-being

Well-being was measured using seven questionnaires. The **Warwick Edinburgh Mental Well-being Scale (WEMWBS)** is a fourteen-item questionnaire that has been validated for measuring general well-being in the general population [156, 157], as well as in young people [158, 159]. There are five response categories for each question, and the total score is between 14 and 70. All items in the WEMWBS are positively worded, and it is focused on measuring positive mental health.

The **Satisfaction with Life Scale** [160, 161] is five-item questionnaire designed to measure global cognitive judgements of satisfaction with one's life. Each question uses a seven-point Likert-type measure and the total score is between 5 and 35. The **Subjective Happiness Scale** [162] is a four-item questionnaire based on seven-point Likert-type questions, with the overall score being a mean of the four questions, lying in the range of 1 to 7.

The **Gratitude Questionnaire (GQ-6)** is a six-item measure that uses a seven-point Likert-type scale to assess individual differences in proneness to experiencing gratitude in daily life [163]. Each score is summed to a total between 6 and 42. The **Life Orientation Test (LOT-R)** is a measure of dispositional optimism that has ten items asked on a 5-point Likert-type scale [164]. The overall score is in the range of 0 to 20.

The **Meaning in Life** questionnaire has 10 items designed to measure two dimensions of meaning in life: (1) Presence of Meaning (how much respondents feel their lives have meaning), and (2) Search for Meaning (how much respondents strive to find meaning and understanding in their lives) [165]. Respondents answered each item on a seven-point Likert-type scale, with the two sub-scales scored in total between 5 and 35.

The psychological constructs of autonomy, competence and relatedness associated with self-determination theory were measured using the **Basic Psychological Needs in General (BPN)** questionnaire [166]. This questionnaire has 21 seven-point Likert-style questions with the final score for each of the three sub-domains being the mean of the responses for that sub-domain. As such each of autonomy, competence and relatedness were scored overall from 1 to 7.

All of the well-being measures listed were scored in a positive direction, where higher scores indicate higher alignment with the construct being measured.

2.2.3 Ethics

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. The full list of ethical approval references for ALSPAC can be found on their website (<https://www.bristol.ac.uk/alspac/researchers/research-ethics/>).

2.2.4 Data and code

The datasets analysed in this chapter are not publicly available as the informed consent obtained from ALSPAC participants does not allow data to be made freely available through any third party maintained public repository. However, data used for this chapter can be made available on request to the ALSPAC Executive, with reference to project number B3227. The ALSPAC data management plan describes in detail the policy regarding data sharing, which is through a system of managed open access. Full instructions for applying for data access can be found here: <http://www.bristol.ac.uk/alspac/researchers/access/>. The ALSPAC study website contains details of all the data that are available (<http://www.bristol.ac.uk/alspac/researchers/our-data/>).

The code used to produce the results in this study can be found at <https://osf.io/rkxm6/>.

The descriptive statistics were calculated using the R programming language (v4.0.1) [167] in

RStudio (v1.3), primarily using the `tidyverse` (v1.3.0) package [168] for data manipulation and `ggplot2` (v3.3.1) [169] for visualisation.

2.3 Results

2.3.1 Demographics

I first consider the demographics of social media users across different frequencies of use, and across the five social media platforms: Facebook, Twitter, Instagram, Snapchat and YouTube. These are both taken from the main sample, as described in the Methods. Table 2.2 presents the frequency that participants reported using any social media sites each day, based on sex, ethnicity, education, and their parents' occupational group. Table 2.3 gives the percentage of participants from each demographic group who reported being a user of each platform with any use frequency. The breakdown of every demographic by frequency of use on each platform is provided in full in

Table 2.2: The percentage of each demographic group by their self-reported frequency of using any social media each day.

Characteristic	% of Group Using Social Media At Each Frequency			p-value
	> 10 times a day N = 1576 (39%)	2-10 times a day N = 2144 (53%)	Once a day or less N = 356 (8.7%)	
Sex				<0.001
Female	40	53	7.1	
Male	35	53	12	
Ethnicity				0.4
Ethnic Minority Groups	35	58	6.8	
White	39	52	9.0	
Unknown	41	52	7.2	
A Levels				0.3
No	38	52	9.9	
Yes	38	54	7.9	
Unknown	39	51	9.6	
Parental Employment Class				0.5
Non Manual	38	53	9.1	
Manual	41	51	8.0	
Unknown	39	52	8.3	

Note:

p-value calculated using Pearson's Chi-squared test

Supplementary Table A.1. Figure 2.1 illustrates this breakdown for sex, which is the demographic by which all the following results are stratified due to the imbalance in the sample and the results in Table 2.2 and Table 2.3. Social media use and mental health and well-being outcomes are also

Table 2.3: The percentage of each demographic group who indicated that they had an account on each of the social media platforms considered.

Characteristic	% of Group Using Each Platform				
	Facebook N = 3977 (97%)	Twitter N = 2294 (56%)	Instagram N = 2803 (69%)	Snapchat N = 2864 (70%)	YouTube N = 2989 (73%)
Sex					
Female	98	56	76	73	68
Male	97	57	54	64	83
Ethnicity					
Ethnic Minority	95	58	68	73	74
Groups					
White	98	57	68	70	73
Unknown	96	52	70	72	73
A Levels					
No	98	51	68	72	70
Yes	98	58	68	70	73
Unknown	97	55	71	70	74
Parental					
Employment Class					
Non Manual	98	57	68	69	73
Manual	98	55	71	72	73
Unknown	96	54	70	72	73

known to vary according to gender [170–172].

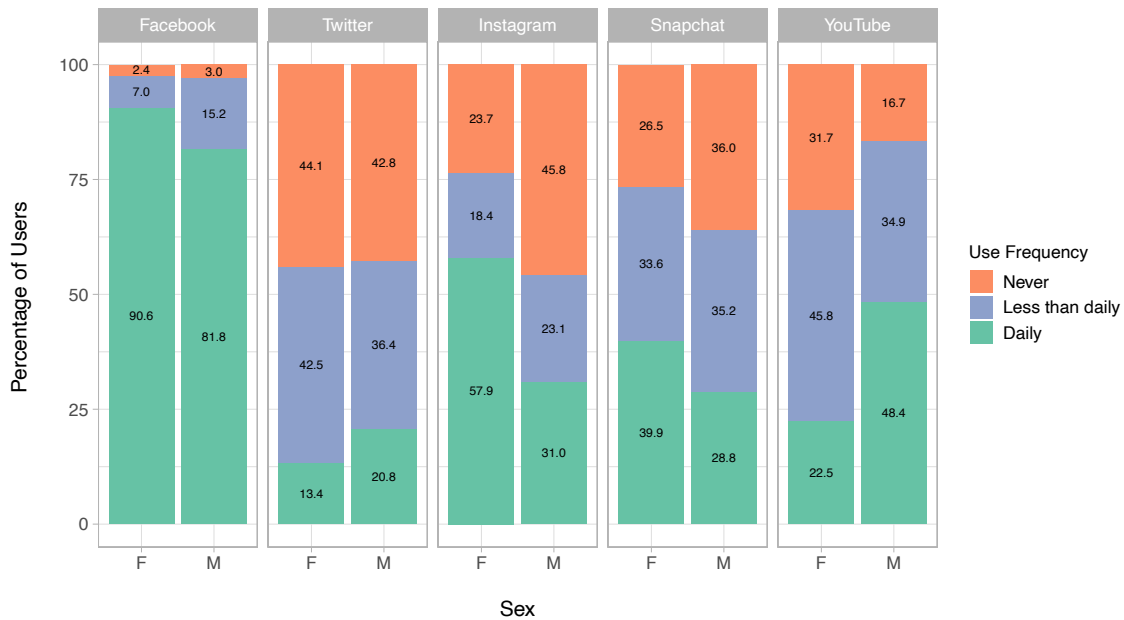


Figure 2.1: Percentage of participants using each of Facebook, Twitter, Instagram, Snapchat and YouTube stratified by the frequency of using that platform, and sex.

2.3.2 Mental Health and Well-being

By frequency of use

First I will consider well-being and indicators of poor mental health across different use frequencies. Figure 2.2 shows how indicators of poor mental health vary across the three frequencies of use, which are more than 10 times a day, 2-10 times a day and once per day or less; no participants reported using no social media at all. These frequencies are contextualised by the prevalence of each outcome in all users of social media. This figure shows that the lowest category of social media use, that is once per day or less, has the highest proportions of disordered eating, self-harm and suicidal thoughts among women. As seen in Table 2.2, only 7.1% of women and 12% of men used social media less than once per day, and so these measurements are subject to wider confidence intervals. Here, depression is defined as being present in those who scored above the cut-off score of 12 in the Short Mood and Feelings Questionnaire (MFQ) [153].

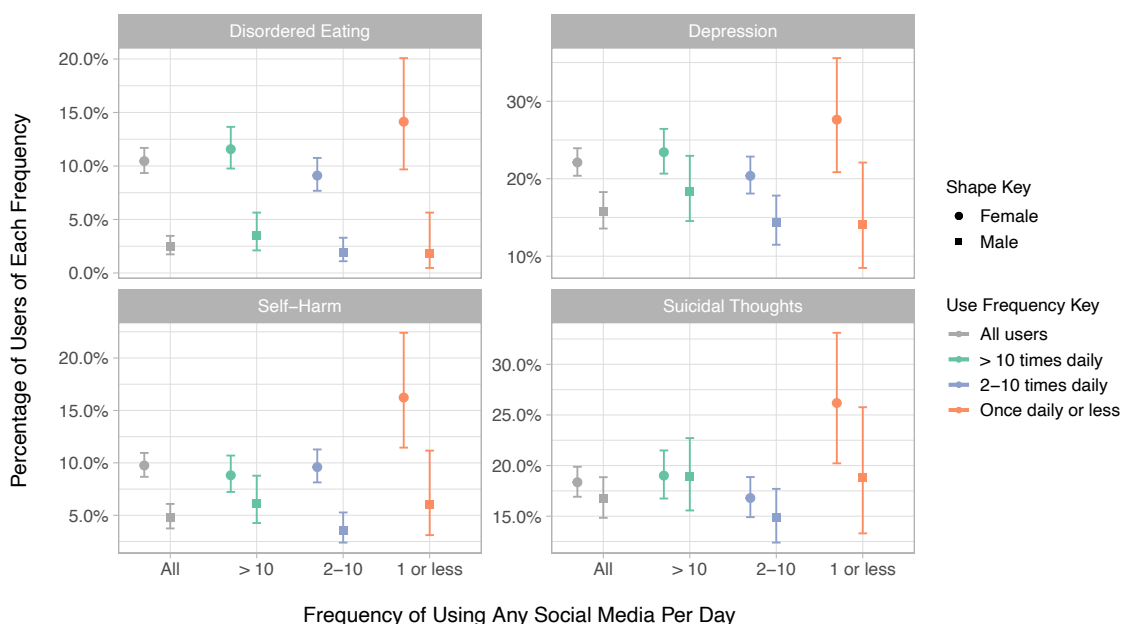


Figure 2.2: Percentage of participants who reported disordered eating, self-harm or suicidal thoughts in the past year, or who met the threshold for depression, differentiated by sex and frequency of any social media use with 95% confidence intervals.

Similarly, each well-being construct is presented in Figure 2.3, and contextualised by the result for all users of social media, regardless of frequency. Separate outcomes are presented for the three sub-scales of the Basic Psychological Needs (BPN) scale and the two sub-scales of the Meaning in Life (MIL) scale. The Life Orientation Test measures optimism, and the Warwick Edinburgh Mental Well-being Scale (WEMWBS) measures overall positive well-being.

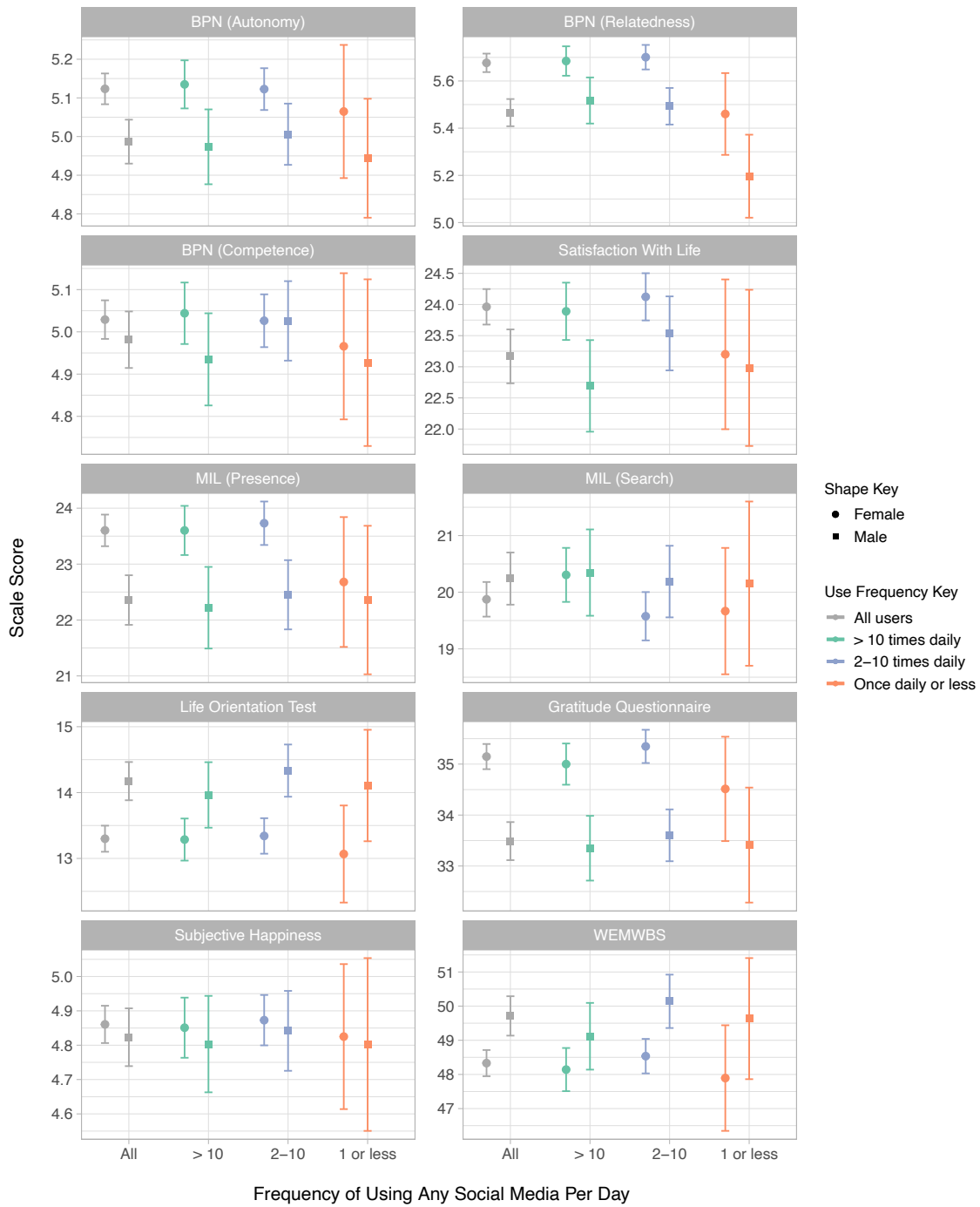


Figure 2.3: Mean scores for seven well-being measures, stratified by sex and overall frequency of using any social media platform, with 95% confidence intervals.

By platform

Here I consider the characteristics of **daily** users of each platform. The relative percentage of daily users against other types of users for each platform can be referred to in Figure 2.1.

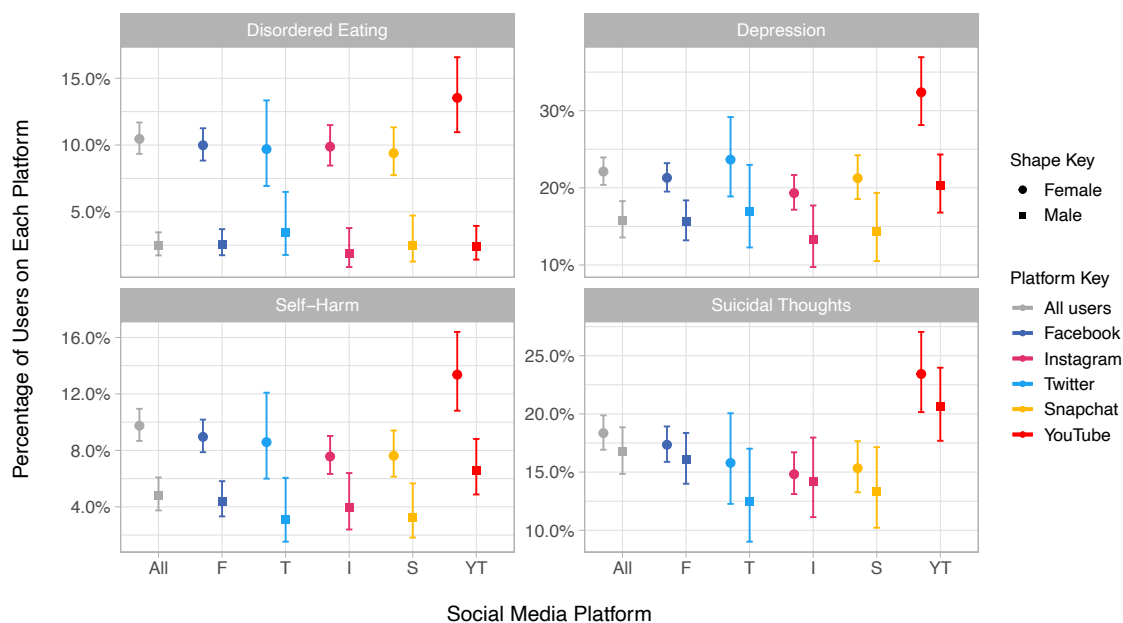


Figure 2.4: Percentage of participants who reported disordered eating, self-harm or suicidal thoughts in the past year, or who met the threshold for depression, differentiated by sex for daily users of each social media platform, with 95% confidence intervals.

Finally Figure 2.5 gives the mean well-being score across each platform for each of the seven well-being measures.

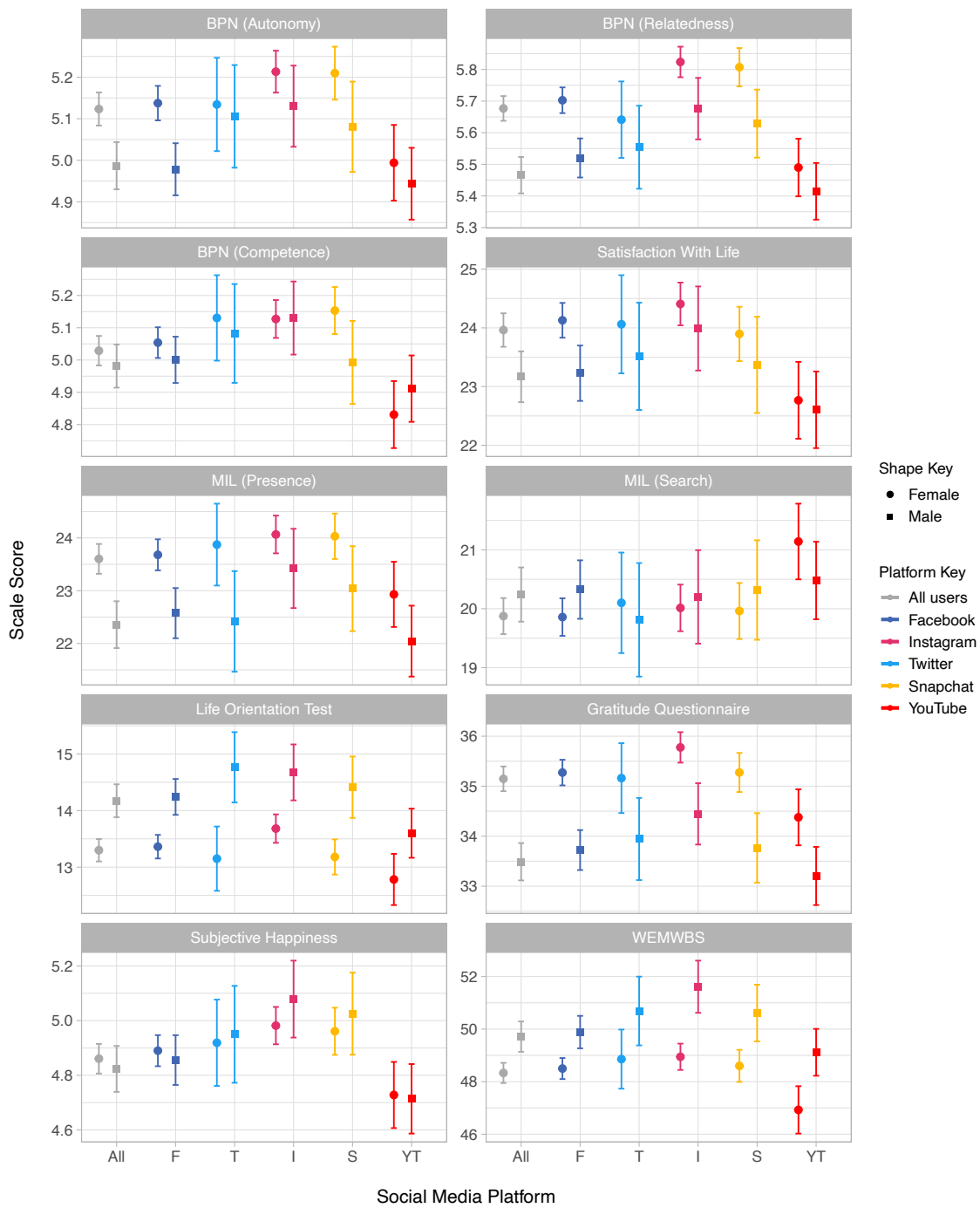


Figure 2.5: Mean scores for seven well-being measures for daily users of each platform, stratified by sex, with 95% confidence intervals.

2.4 Discussion

This study used data from a UK population cohort study to describe the demographics and key mental health and well-being indicators of social media users by their self-reported frequency of using social media and five different platforms used at ages 23 and 24. Overall, I saw that there were differences in demographics and mental states of users across use-patterns and platforms used. In the following sections I detail and discuss the implications of these findings for future research across the themes of demographics, use-frequency and platform used.

In general, just over half of participants reported using social media 2-10 times per day, with more than ten times per day still being common at 39%, and only approximately one in ten participants using social media once per day or less. The results showed that those who rated their social media use at the highest frequency (more than ten times per day) were more likely to be women, more likely to be White and more likely have parents who worked in manual occupations. However, sex was the only demographic that appeared to have a statistical relationship with frequency of use, based on a Chi-squared test. Davies and colleagues [173] saw similar results from a Welsh population survey of social media use that found there was a difference in social media use across genders, but not by measures of deprivation.

Figure 2.1 showed that Facebook is, unsurprisingly, the most popular platform both in being used by 97% of the participants and being the most used platform on a daily basis. Instagram and YouTube showed substantial differences in use patterns across male and female users, with approximately double the percentage of women using Instagram daily as men and, conversely, approximately double the percentage of men using YouTube daily as women. Snapchat also saw higher proportions of daily and overall female users, though this difference between sexes was not as dramatic as for Instagram and YouTube. These patterns of use generally agree with the demographics of users on these sites reported for 18-29 year old US adults by the Pew Research Center [145], although the sample used here saw slightly more Twitter users than their estimated 38%, and fewer YouTube users than their estimated 91% (see Table 2.3). This difference in YouTube users may be partly explained by the fact that it is the only platform with a substantially higher proportion of men than women using it (68% of women vs 83% of men), and that men were under represented in the sample overall compared to women. This emphasises the importance of stratifying results

by sex.

Previous research into the demographics of UK Twitter users also aligns with my findings that men and those from higher socio-economic backgrounds are more likely to be Twitter users than women [146, 147]. Here, I also saw that those from ethnic minority groups are more likely to be Twitter users than White participants, though this is limited by the fact that I could not further separate out results for people with different ethnicities due to the variables available. Across the sample, Twitter was the only social media platform that had a noticeably higher proportion of both A Level educated participants and parents in non-manual occupations. Snapchat saw the reverse pattern with a higher proportion of participants who did not have A Level qualifications and a higher proportion of participants whose parents worked in manual occupations.

Overall, the sex differences between all male and female users varied across outcomes. For instance, a higher percentage of women experienced depression, disordered eating and self-harm overall, but the gap in the prevalence of suicidal thoughts between men and women was much smaller. This concurs with evidence from the last UK-wide psychiatric morbidity survey, in that ‘common mental health disorders’ are more prevalent in women than men [174]. When it comes to well-being, I saw that women also display higher mean levels of well-being across most measures. Exceptions are the Life Orientation Test, which showed men generally had higher levels of optimism, the Subjective Happiness Scale where scores were roughly equivalent, and the WEMWBS where men’s general well-being was slightly higher. These results, apart from the WEMWBS, are consistent with findings on UK wide well-being at the time of the survey, and that men tend to have higher optimism in general [175, 176]. Previous research into the WEMWBS has not generally found large sex differences, but there is evidence that in younger samples there are differences that may be explained by socio-economic status [156, 157, 177]; I note that higher attrition of men in this sample was likely to lead to a bias towards men who are more socio-economically privileged, which may explain why they had higher well-being.

The patterns of mental health outcomes by use frequency displayed in Figure 2.2 showed some support for the so-called ‘Goldilocks theory’ of social media use that hypothesises a quadratic, rather than linear, stimulus-response relationship between social media use and mental well-being [178]. This would mean that moderate use of social media, rather than very little or excessive use, is best for well-being. However, this pattern did not consistently apply. For instance, there was an

inverse relationship between social media use and percentage of women who self-harm, and in men only the group with the highest level of social media use had more severe depressive symptoms. Here I was using self-reported use-frequency where, as I discuss further in the limitations, an individual's assessment of their own use-frequency may be biased by their relationship with social media and overall mental health [179].

When considering the results by well-being measure in Figure 2.3 we saw that subjective happiness and optimism as measured by the Life Orientation Test both appeared relatively consistent across use categories. Relatedness presented the clearest difference across use categories, with relatedness in women being higher for the two most frequent use frequencies. However, perhaps the most notable outcome was the inconsistency between well-being scales which implies that the choice of scale could affect the interpretation of the impact of well-being on social media use. Research into the relationship between social media use and well-being has been said to suffer from what is known as the 'jingle-jangle' paradox where the term 'well-being' is used as a catch-all for anything from depression rates to life satisfaction [180]. This conflation of different well-being measures leads to comparisons of different psychological constructs which may interact differently with social media use: this is hypothesised as one of the reasons that researchers find conflicting evidence for this relationship [180], which my results support. This also adds to the picture of researcher degrees of freedom in choosing how to measure psychological constructs, which has been shown to have a substantial impact on the outcome of analyses of social media and mental health [134]. Subjective well-being is a complex and multi-faceted psychological concept [182], and these findings illustrate the importance of recognising that different measures of well-being could imply different relationships between social media and 'well-being'.

When considering participant outcomes by daily users of each platform more consistent patterns emerge than for use-frequencies. I saw that, particularly for women, YouTube had the highest proportion of users reporting disordered eating, self-harm, suicidal thoughts and depression, with higher prevalence of depression in female users of YouTube compared to male users (Figure 2.4). Whilst overall mental well-being across platforms, as measured by the WEMWBS in Figure 2.5, shows YouTube as being marginally but not drastically lower than other platforms, other well-being measures illustrated some key differences. For instance, YouTube users had lower life satisfaction, relatedness and, particularly for female users, levels of competence (Figure 2.5). Conversely, daily

users of Instagram, and in some cases Snapchat, appeared to have the highest subjective well-being across most measures, with this being particularly noticeable for relatedness, gratitude and happiness (Figure 2.5). The role of self-determination theory in social media use has previously been explored for Facebook and social media in general [184] with relatedness hypothesised as a key motivating factor for social media use. Previous findings have shown that Instagram and Snapchat are used more for social interaction than Twitter and Facebook [84], and so my results may corroborate the importance of relatedness in the use of particular platforms. Regardless of the specific measure, my results have illustrated that there is variation amongst platforms which further challenges the idea that ‘social media’ or ‘social networking sites’ are a homogeneous group, and reiterates the importance of understanding the context of research about or using social media [84, 147].

At face value these results appear to directly contrast with the outcomes of the *Status of Mind* report published by the Royal Society for Public Health [185], where young people rated YouTube as being the most beneficial site for their well-being and Instagram as the worst, based on health-related outcomes such as their anxiety and depression. My findings that a higher prevalence of YouTube users suffer from poorer mental health and well-being may mean that whilst some platforms are seen as ‘worse’ for young people’s mental health, that does not equate to finding more unwell young people on those platforms. One explanation may be that those experiencing poorer mental health are more likely to use YouTube because they experience more benefits from using it, such as community building and peer support [133], than they do from spending time on sites like Instagram. However, this is certainly an interesting area for further exploration in future quantitative and qualitative research.

Whilst this research draws evidence from a robust and well-documented study and the sample being from a birth cohort means that these results are not confounded by age, there are limitations to the cohort sample that I have used. Firstly, the cohort measures a specific age group so I can only infer information about a single age group at each measurement time point. I suspect that different patterns might be found at different ages, knowing that rates of various mental health conditions such as anxiety, depression and suicidality change over the course of childhood, adolescence and adulthood [186], and since each generation may use social media differently [187]. It is also important to note that the two data collection points used in this study were taken a year apart, and

so not all measures were taken exactly at the same time. This means that although I have primarily considered the data cross-sectionally there is a potential for some longitudinal effects to have influenced the data. ALSPAC has also seen differential attrition over time and so, as seen in Table 2.1, the sample for this study when the index cohort were in their early twenties has fewer men than women, and more participants from privileged socio-economic groups in terms of education and class background [118]. As well as this, typical social media use changes over time and by age [145], and so further assessment of social media use across a variety of population-representative age groups would be the most effective way to understand differences between generations. As discussed in the Methods section, there was also a limitation in that ethnicity was only available as two categories (White or Ethnic Minority Groups) and so it was not possible to look further into differences in social media by users of different ethnicities. Given these limitations of the sample it would be valuable to conduct similar research in other cohorts over the coming years.

Another limitation of this study is a lack of specificity about the nature of social media use that participants are referring to when responding. It is possible that activities related to ‘using’ social media, such as posting content versus passive use, change depending on platform used and that there are individual preferences to account for [84, 170, 188, 189]. For instance, YouTube is distinct from other platforms in this study in that its primary function is passive content consumption as opposed to social networking. Previous research has suggested a reciprocal association between passive social media use and lower subjective well-being [188], whilst using social media for direct communication has been positively associated with perceived friend support [190]. This may better reflect the uses of platforms like Snapchat. As well as the subjective nature of ‘use’, there are also ongoing concerns about using self-reported measures of use-frequency to measure social media behaviours. Emerging evidence is showing that self-reports do not align well with objective measurement due to recall bias and differences in interpreting how to include notifications or fleeting checks of social media [191, 192] with self-reported smartphone pickups underestimating associations with mental health compared to objective measures of use [179]. It might be that different ways of measuring social media use, such as types of use, are more useful when considering associations with mental health and well-being outcomes [170]. It is worth noting that the use-frequency measures used in this study are distinct from screen-time, and equivalent use-frequency across platforms may have different time implications; someone may spend short amounts of time

on Instagram or Snapchat checking notifications, but do so frequently, versus visiting YouTube once in a day but spending several hours watching content. These nuances are challenging to capture, but by reporting on mental health prevalence across the available responses in a cohort study we can add to the growing understanding of how self-reported social media use frequency is related to mental health.

In summary, these results amplify the importance of attending to complexity when measuring and analysing social media use and mental health and well-being. It is important to note that the results do not, and cannot, imply that different types of social media use cause poorer or better health outcomes in young people, but they do provide vital contextual information on user groups that can help us better understand the reasons that previous research has found conflicting results. I have provided estimates of seven well-being measures and the prevalence of four key mental health outcomes (depression, disordered eating, suicidal thoughts and self-harm) across the five platforms Facebook, Twitter, Instagram, Snapchat and YouTube, as well as across three use frequencies. My findings have shown that the demographic and mental health footprint of each platform is different. Primarily users differ by sex, but when it comes to platforms YouTube is particularly likely to have both male and female users with poorer mental health and well-being across a range of indicators, alongside evidence that daily Instagram users have better overall well-being than daily users of other platforms. My findings also indicate that relationships between use-frequency and multiple mental health and well-being outcomes are often non-linear, which supports the importance of considering non-linear dose-response relationships between social media and mental health and well-being in future research. Lastly, I showed that the relationship between use-frequencies and well-being changes depending on the measure of well-being used. This means that we cannot conflate different types of well-being, and doing so will likely result in low replicability.

This research has implications for both those who conduct research on the relationship between social media and mental health, and those who study mental health prediction. We must ensure that we are considering both platform-specific and outcome-specific effects rather than conflating types of social media use, social media sites and well-being as single entities. Future research should also stratify results by sex since it is unlikely that studies with differently balanced samples will replicate. My findings on use-frequencies also suggest that we cannot assume linear relationships between social media use and mental health. The understanding of these methodological issues

would be improved by examining profiles of different user age-groups, as well as examining relationships between these variables longitudinally to understand the potential for reciprocal effects. The differences between platforms should be further considered too, as to how different content types and communication modes on different platforms may affect mental health differently.

Chapter 3

Participant views on social media and its linkage to longitudinal data

This chapter is adapted from Di Cara NH, Boyd A, Tanner AR, Al Baghal T, Calderwood L, Sloan LS, Davis OSP & Haworth CMA (2020). Views on social media and its linkage to longitudinal data from two generations of a UK cohort study. *Wellcome Open Research*¹.

I was responsible for data curation, formal analysis, investigation, methodology, and writing (original draft and reviewing and editing).

OD and CH were responsible for funding acquisition, conceptualisation, methodology, investigation, supervision and writing (reviewing and editing). ART assisted with the focus groups themselves, and AB provided resources for the focus groups, as well as both being involved in reviewing and editing. All of AB, TAB, LC, and LSS were involved in the conceptualisation of the original funding bid, and reviewing and editing of the latter drafts.

¹<https://doi.org/10.12688/wellcomeopenres.15755.2>

Abstract

Background Cohort studies gather huge volumes of information about a range of phenotypes but new sources of information such as social media data are yet to be integrated. Participant's long-term engagement with cohort studies, as well as the potential for their social media data to be linked to other longitudinal data, may give participants a unique perspective on the acceptability of this growing research area.

Method Two focus groups explored participant views towards the acceptability and best practice for the collection of social media data for research purposes. Participants were drawn from the Avon Longitudinal Study of Parents and Children cohort; individuals from the index cohort of young people (N=9) and from the parent generation (N=5) took part in two separate 90-minute focus groups. The discussions were audio recorded and subjected to qualitative analysis.

Results Participants were generally supportive of the collection of social media data to facilitate health and social research. They felt that their trust in the cohort study would encourage them to do so. Concern was expressed about the collection of data from friends or connections who had not consented. In terms of best practice for collecting the data, participants generally preferred the use of anonymous data derived from social media to be shared with researchers.

Conclusion Cohort studies have trusting relationships with their participants; for this relationship to extend to linking their social media data with longitudinal information, procedural safeguards are needed. Participants understand the goals and potential of research integrating social media data into cohort studies, but further research is required on the acquisition of their friends' data. The views gathered from participants provide important guidance for future work seeking to integrate social media in cohort studies.

Aims

The aim of this chapter was to better understand the acceptability of linking longitudinal cohort data to social media data, in order to inform the Twitter data linkage project.

3.1 Introduction

The analysis of data collected from social media is a rich and growing area of current research in a wide variety of fields. Social media data has been used to track the spread of disease [193], predict the results of key elections [194] and gauge public reaction to events [195]. Not only are these data widely accessible but they provide a wealth of rich information on views, feelings and interests. Whilst these data are highly valuable in health and social research there are few reliable data sources that can link social media data to factual information about users' lives, with most applications so far focussed on identifying broader trends using 'big data' methodologies. Without these so-called 'ground truth' (empirical, rather than inferred) data, it is not possible to adequately validate social media sentiment analysis methods, or to infer the relevance of the patterns observed to the general population [147]. At present, longitudinal population studies (LPS) remain an untapped resource in terms of obtaining this empirical information. Conducting social media data linkage in this way also has the potential reciprocal benefit of adding significant value to the data already available in the LPS. Those participating in LPS are already familiar with the process of collection and use of sensitive data, have evidenced commitment to providing their personal data for the advancement of science, with these data being readily accessible to researchers. As highlighted by Wellcome [196], and the Medical Research Council (MRC) [113], a key future direction for such studies is to conduct more data linkage. Data linkage within existing and prospective datasets has the potential to reduce the burden on participants and maximise the benefit of research data collected [196], whilst using new types of data that allow for remote data capture, such as social media, is hypothesised by the MRC as a method that could address a lack of engagement and offer cost-effective modes of data collection [113]. The definition of the term 'social media' is left to be explored and defined by the participants within this study.

Whilst such data sources present exciting possibilities, organisations and those working in the emerging population data science field are conscious of the need to understand public views and expectations around the novel use of such data in research [197, 198], and that a process of public/participant dialogue is needed to ensure new activities do not undermine trust in the study and can be seen to provide public benefits [199]. Within the UK the failure of the care.data program is cited as a reminder that even where data science initiatives are legal and technically feasible they can still fail if they lack the 'social licence' needed for public and key stakeholder support [200].

Existing research in the field of record linkage, outside of social media linkage, has found that there is a general acceptance of this work from the public [201–203], even when conducted without consent if data is appropriately anonymised [203], but that these decisions are ultimately complex and conditional on the situation [199, 201–204]. Therefore, it is essential that any novel data linkage activity, or a novel use of existing data, is informed by exploring participant views towards its acceptability, as well as researchers exploring the participants understanding of the data and how it will be used. In this manner participants can inform studies’ efforts to reach a consensus on the best practices for collecting these potentially sensitive data and sharing these with researchers in a secure and ethical way that protects participant anonymity.

The use of social media data for research has also had its own ethical challenges concerning privacy and informed consent [116, 205], as well as difficulty defining what mediums are included in the definition of social media at all [42]. A systematic review by Golder et al. [206] in 2017 found that social media users and researchers tended to be conflicted about whether informed consent was necessary for data collected from public social media sites and, similarly to data linkage issues, this debate tended to rest on the nature of the content, which source the data came from and how the data would be used [204, 207–209]. Subsequent ethical guidelines developed for the field have placed special consideration on the reporting of social media research to ensure users’ privacy [210, 211], and reflects participants’ views that increased sensitivity and personal identifiability of the subject matter should increase the level of anonymity with which it is reported [121, 205, 207]. Previously, research participants have found that photos are more personal than text data [207, 212], however the level of trust in the study or the researchers conducting it may also influence their decision of whether or not to share [200–202]. There is evidence to suggest that there are a body of users who expect their data to be collected as ‘necessary evil’ of day-to-day social media use; these users tend to see information privacy as the responsibility of the individual rather than the company holding the data [205]. There may also be an age related aspect to participants’ willingness to share their social media data, with younger people more likely to agree [213, 214].

Collecting social media data from LPS therefore appears to be promising, as participants will always have given explicit consent, are likely to have a good awareness of how their data is kept safe, and have trust in the study to use and report their data responsibly. This may mean their agreement to share their data and link it to existing data is more likely. However, it is particularly

important in these studies to maintain the trust that has been built with participants by co-creating an understanding of what is acceptable with regards to their information, especially since LPS participants may have specific concerns about the linking of their social media activities to the large volumes of diverse sets of data already held about them by the study. In addition, the series of high-profile online data scandals, the introduction of the new General Data Protection Regulation (GDPR) and the significant concern about the manipulative political use of social media data in what has been called the “Cambridge Analytica scandal” may have had the potential to influence participants’ views about what they consider to be acceptable in terms of data collection, linkage and reporting on their social media data [215]. As such, this study into participants’ views aims to ensure our knowledge of what is considered acceptable practice for social media data linkage remains current in the evolving landscape of online privacy, and to ensure that we consider the specific views of participants in LPS.

In this study I report on participants’ views on social media data linkage in an on-going birth cohort study, the Avon Longitudinal Study of Parents and Children (ALSPAC), also known as ‘Children of the Nineties’ [118–120]. Focus groups were held separately with participants from the index offspring cohort, who are now in their late-twenties, and with the parent cohort, and included semi-structured discussions on their views on, firstly, how they would define social media and what they use it for, and then their opinions on social media research and data linkage. Due to the ambiguous nature of social media we made it a priority to first understand how participants view it as a medium and how they report interacting with it, prior to trying to interpret their views around their data privacy.

3.2 Methods

3.2.1 Sample and Recruitment

ALSPAC is a trans-generational prospective observational study which recruited pregnant women living in Avon, UK; those with expected dates of delivery between the 1st April 1991 and the 31st December 1992 were invited to take part [118–120]. The initial number of pregnancies enrolled was 14,451 and of these pregnancies 13,988 children were alive at one year of age. This was supplemented when the index children reached approximately age seven where eligible cases who

had not joined originally were invited to the study, resulting in a total of 14,901 children alive at one year of age for which there is data from age seven. Since joining the study both parents and index children have been routinely assessed on a number of environmental and psychological measures, provided biological samples and their genetic data. The wide variety of longitudinal data from both generations has provided valuable opportunities for a breadth of research into health and social outcomes for children and young people, as related to genetic, environmental and social factors. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (<http://www.bristol.ac.uk/alspac/researchers/our-data/>). For the purposes of this study, we recruited two separate participant groups from the available sample to take part; the first contained the index children themselves, and the second group contained participants from the parent cohort. A random sample of participants in the study who lived in the Bristol area were invited to take part, to allow easy access to the study location. Inclusion criteria were that participants had a social media account, and spaces were filled on a first-come-first-served basis. The index child group was made up of nine participants aged 26 to 28, with four males and five females. The parent group was made up of five participants aged 53 to 65, with one male and four females. All participants were reimbursed for their travel expenses and offered a £10 shopping voucher for taking part.

3.2.2 Ethics

Ethical approval for this study was provided by the ALSPAC Ethics and Law Committee and the Local Research Ethics Committee, at the University of Bristol. All participants gave their written consent for participation and audio recordings. Fair processing information describing the study was provided in a postal invitation pack. Participant's consent was obtained on arrival at the focus group.

3.2.3 Data collection

Data were collected using focus groups, defined as “semi-structured discussions of 4–12 people that aim to explore a specific set of issues” [216]. Focus groups were seen as the most appropriate data collection method, as opposed to surveys or one-to-one interviews, due to the ability to resolve and discuss conflicting views and information through group interaction, clarify individual and

shared perspectives, as well as directly explore the relative emphasis on certain topics in order to understand their subjective importance [216, 217].

Two focus groups, one for each generational group, took place consecutively at the ALSPAC offices on the morning of Saturday 22nd September 2018. Each focus group lasted 90 minutes and was led by the Principal Investigators of the study Dr Haworth [CH] and Dr Davis [OD], who were assisted by three members of their research team [AT (PhD), JA (MSc), ND (MSc)] and one member of ALSPAC study staff. None of the facilitators had previous relationships with the participants. Both Principal Investigators have previous experience in conducting focus groups and provided guidance to the assisting members of the research team. There were three female [CH, JA, ND], and two male [OD, AT] facilitators present. The participants were made aware that those present were interested in the potential of social media to improve health and well-being, and were later introduced to previous research into expressions of happiness and anxiety on Twitter during the presentation² given by Dr Davis in the middle of the group. Beyond this they were not made aware of the facilitators' specific research interests, which I am reporting in line with the CORE-Q criteria for qualitative research [216]. Throughout the focus groups the facilitators used a 'funnel' style approach to questioning by starting with general questions about individuals' views on what they believed social media to be and how often they used it and then becoming more specific as topics of interest were narrowed down.

The focus groups were structured into the following three parts.

Part 1: Personal views on social media in the UK today

After basic introductions, we introduced a discussion on what the participants believed social media to be, how they tended to use social media, and what they used them for. Participants were subsequently asked to classify themselves as high, medium or low social media users based on the definitions proposed by NatCen [207] as seen in Table 3.1, and their own understanding of what social media are. The definitions given were used in both focus groups to identify the range of social media use represented by the participant group.

Part 2: Presentation on applications of social media data

The participants were given a presentation (Supplementary Material 1, Extended data [218])

²Slides available in the online supplementary materials (doi: [10.17605/OSF.IO/HYD9G](https://doi.org/10.17605/OSF.IO/HYD9G))

Table 3.1: Definitions of levels of social media use as given by NatCen.

Level of Use	Description
High	Those who use social media several times a day
Medium	Those using social media from twice a week up to once a day
Low	Those who do not use social media or use them once a week or less

on the applications of social media data in health and social research which included examples of population level disease symptom tracking, and sentiment analysis of Twitter data, as well as then introducing the spectrum of identifiability of data, using resources from ‘Understanding Patient Data’ (<https://understandingpatientdata.org.uk/what-does-anonymised-mean>), which clarifies the difference between raw, de-personalised and anonymous data. These three terms were understood in this study as:

Raw: Information in its original format, for instance a status update, with no attempts to remove identifiable information.

De-personalised: Information which has had identifiable features removed, but could feasibly still be re-identified. **Anonymised:** Information that has been processed, for instance into a numeric score or aggregated, so that there is no recognisable association between an individual and the piece of information.

Part 3: Views on using social media data for research

Following the presentation, we provided each participant with a template as presented in Figure 3.1, with ‘blank’ spaces which could be filled with cards labelled as described in Table 3.2 to form a possible research scenario. This exercise could produce up to 108 unique scenarios for discussion which were designed to explore participants’ views around linkage of different types of data, and how they would expect these data to be shared and presented to different types of researchers. To illustrate, an example of a completed template is given in Figure 3.2. Participants chose the cards at random from a pack to minimise the bias in the selection of options, then discussed the scenarios in small groups of 4–5, with a facilitator moderating each group.



Figure 3.1: The template that participants filled with options from Table 2 to provide discussion points.

This template was used by participants, in conjunction with the options from Table 3.2, to discuss

Table 3.2: The options presented to participants to fill each ‘blank’ in the statements in Figure 1.

People	Description	Type of data	Platform
‘Children of the Nineties’ staff	Raw	Friends	Facebook
Researchers	Depersonalised	Network	Instagram
Computers	Anonymised	Likes	Twitter
		Text	
		Images	
		Location	

their views around a wide range of different data access scenarios. This allowed the research team to unpick which types of variation in the scenario might make it more or less acceptable to the participant group. One option from each column was randomly selected by participants to complete an activity which explored their views on possible scenarios in which different types social media data might be accessed.

With my consent, researchers access
anonymised text data from
for important health and

Figure 3.2: An example of a completed template from the situational exercise.

This provides an example of how the templates were used during the focus groups to explore a particular situation in which their data might be accessed. Variation of the item in any one

of the boxes could be changed to explore how this might alter the participants' opinion on the scenario.

3.2.4 Analysis

Immediately following the focus group, a short debrief was held between the members of the research team to collate their experiences and any trends they had noticed. An analysis of themes was then later completed from the audio recordings by myself. The analysis procedure was first, familiarisation with the data through listening to the recordings of the focus groups and transcribing them, then identifying material that was not neutral (neutral material included comments from facilitators, or conversations which were not relevant) and finally, systematically coding the relevant material and organising it into themes. Due to the semi-directed nature of the focus group, and relatively small volume of data, the data were first coded deductively [219] into the existing structure given by Part 1 and Part 3 of the focus groups above, rather than inductively deriving broad themes. Within the sub-theme 'What are social media used for?', participant responses were inductively coded and summarised into themes. As such the results are presented as narrative summaries of participants of the two parts, with any identified sub-themes as relevant.

The software package NVivo 12 was used in the analysis stage in order to code the participants' comments digitally and allow for comments on particular areas to be viewed collectively. Nvivo is primarily used for the analysis of non-numerical unstructured data used in qualitative research. Participants were not invited to check the results following their participation in the focus group, but we reviewed all quotes to ensure that participants could not be identified.

3.3 Results

The following results are a narrative description of the outcome of my analyses, arranged by the two discussion sections of the focus groups, and by each main topic within these sections. As such, the first section describes the participants' general views on social media and how they use it. The second section then summarises their views on the use of social media in research with respect to the different variables presented to them in the template exercise, given in Table 1. Quotes have been provided where relevant to further illustrate the discussion.

3.3.1 Personal views on social media in the UK today

What ‘counts’ as social media, and who is using them?

The groups were first asked to discuss what social media was. This prompted both examples of platforms, and descriptions of what made certain platforms ‘count’ as social media. In both groups there was an agreement that defining an application as a type of ‘social media’ was dependent on whether it was possible to share content and interact with others. For instance, the young people agreed that WhatsApp was a type of social media only due to recent updates that allowed users to share ‘daily stories’ and status updates, rather than solely message specific individuals.

“I would just say anything where anybody could join in a, sort-of, general discussion.”

“Anything where you socialise through media... yeah”

— Parents

Facebook was the most widely used and discussed form of social media in both the young people and parent groups, shortly followed by Instagram and Twitter. This is consistent with Facebook being the most common site on which adults have a social media profile [220]. The parent group shared a mutual agreement that they had all had Facebook for the longest time, and preferred it as it was “easier” and more “familiar” than alternative platforms; some participants in this group stated they had had Facebook for over 10 years. In both groups many of the participants stated they were WhatsApp users too. Proportionately more of the young people described actively engaging with Instagram than the parents, and whilst several people across the groups stated they had Twitter profiles, the majority of those said they looked at what other people posted rather than creating content themselves. A sub-section of the younger group noted that they did not regularly post to Facebook, and agreed between them that Facebook was used less by young people than it previously was; this observation is consistent with reported data on the changing demographics of Facebook users [221].

A number of the young people stated that they use Snapchat, and whilst the parent group were aware of Snapchat none of them were users of this platform.

“Facebook mainly. That’s probably the only one I use... and Whatsapp.”

— Parent

“Facebook seems like a platform for older people now.”

“Exactly, my mum uses Facebook more than I do.”

— Young People

Some of the parent generation identified fitness tracking apps such as FitBit and Strava as their more commonly used social media and noted that setting challenges for other users and participating in group competitions were social elements they enjoyed. Several of the younger generation also used fitness apps.

What are social media used for?

In the parent cohort all five participants identified themselves as ‘high’ social media users. In the younger cohort seven out of the nine were also ‘high’ users, with the remaining two identifying as ‘low’ users. These categories of use followed the definitions given by NatCen [207] in their 2014 report where they had equal numbers of participants in each user group, but the proportion of ‘high’ users was much larger in this study. This may represent sampling bias in the present study, and the inclusion criteria of being a social media user may have dissuaded ‘low’ or ‘medium’ users from taking part. However, there was still a wide range of frequency of use within the ‘high’ user group, from checking Facebook once a day on a laptop to using it on waking and then multiple times a day on a portable device. This may suggest that the definition of ‘high’ use is no longer representative given increases in social media use and prevalence over the past five years [220, 222]. When asked to explain what they ‘do’ on social media participants discussed a wide range of activities, which have been individually set out and summarised below.

Interacting with friends and family The main reported use of social media was to keep in touch with friends and family, especially those who were not easily accessible to meet with face-to-face. This was consistent across both groups.

“I use Facebook for like similar to you just like keeping in touch with family and friends that I wouldn’t necessarily see.”

— Parent

Several participants noted the ability to “keep in touch” with others online but the means of com-

munication was not always clear. Some participants clarified that they would have conversations on a private messaging feature of the platform, such as Facebook Messenger, whereas others described it through interactions on content posted to their profile.

“I don’t speak to [my friends and family] directly but they like the things I post.”

— Young Person

Posting content Several participants gave examples of posting content to social media. These examples included pictures of evenings out, asking questions to their online network, pet updates, and exercise records on apps such as FitBit or Strava. As well as personal content a handful of participants also mentioned sharing articles and news through their profiles.

Competitive activities There were two types of competitive activities described by the participants. One was playing games through social media, and the other was taking part in competitions with others on fitness apps. The discussion around fitness competitions was specific to the parent group.

“So when you say about health and fitness there are some amazing apps that you can socialise with other people. And I think they encourage you, or you compete.”

— Parent on fitness apps

Games were discussed in both age groups, but most of the parents described engaging with online games which operate through social media, usually Facebook, compared to a minority of the young people. The games included Scrabble, Candy Crush and FarmVille, with some using these as a way of engaging with friends and relatives on a regular basis.

“I actually use Scrabble [on Facebook] to keep an old lady company every night.”

— Parent

Events Using Facebook as a platform for event planning and organisation was raised between the young people, but not among the parent group. The types of events included those within their immediate social network as well as entertainment advertised through Facebook.

“I use Facebook to know what’s going on in terms of events, the biggest thing I would use it for is going to a party. It’s so useful for ‘there’s a band playing’ or ‘your friend is having a party’. I think for me that’s Facebook’s biggest use.”

— Young Person

Passive or observational use Whilst the participants used social media to stay in touch with others, many of the behaviours described involved ‘scrolling’ through Facebook or Twitter as a consumer of other people’s news.

“I have Facebook but I can’t remember the last time I posted anything on there”

— Young Person

" I've got a Twitter account but all I do is look at stuff, I never actually put anything on Twitter"

— Parent

This passive social media use could be reflective of reported increases in social media being used as a vehicle for viewing content such as videos and news, along with decreases in users creating content [223] which may be to avoid negative feedback on their own user-generated content. The participants described using social media to stay up to date with news articles and celebrities as well as news in their social networks. In parallel with this the young people discussed their perceived need to be discerning about where they got their news from, and which sites they regarded as more trustworthy; there was a disagreement between participants about whether Facebook or Twitter was the more trustworthy news-source and views appeared to be based on preference rather than experience or evidence.

“I would have more faith if something was trending on Twitter than a load of random news articles on Facebook.”

— Young Person

In both groups there were references to using social media to find information about or posts from other people. This is colloquially known as ‘Facebook stalking’ [224]. The parent group were open about using this feature to see what their children were doing, and also to research their children’s

friends and romantic partners. The young people discussed the impact of this phenomenon on dating and how image management on social media may result in gaining an inaccurate impression of what a potential date may be like in real life.

“I stalk my children!”

“Yes, I do stalk my children’s boyfriends”

— Parents on ‘Facebook Stalking’

" Dating apps are part of social media. Some people will research the person they’re going to meet."

— Young Person

Opinions on social media in general.

There were a wide variety of opinions about social media platforms in general. Conversations about concerns were far more prevalent in both groups than discussion of the positive aspects of social media, and both groups were critical about the potential implications of social media. In the vast majority of the participants’ conversations the concerns were speculative, in some cases directly observed and there was only one example of someone being directly impacted by cyber-bullying. However, not all participants may have been happy to speak about this in the group setting. In the parent group the main concerns were about the behaviour of others online. One example given was the use of Snapchat by young people to bully others by reaching a large number of people very quickly, and other examples were given around behaviours of others their own age posting derogatory comments on other people’s content. Those present expressed their disdain for this behaviour, and hypothesised that the anonymity of social media allowed others to act in this way.

“Sometimes I look at the comments people in my age range put up, and I think ‘wow, would you say that out loud in a room full of people?’ ”

— Parent

The parent group also expressed concern about the immediacy of social media, and the “pressure” to respond instantly to messages and communication online. There was a feeling that this would

be difficult for younger people to manage, though was not a concern raised by the younger group. The younger groups' concerns were centred around their inherent distrust of their data security on social networking platforms, as well as the way that peoples' lives may be curated on social media with a bias towards positive events. This was also noted by the parent generation as a risk for young peoples' mental health when using social media.

“If you got engaged and put it on Facebook that’s immediately like 500 likes, it’s blown out. [If you put] ‘I’ve had a really rough day and my dog is sick’, maybe 1 like. We’re so busy chasing the happiness that the duality is never there.”

— Young Person

Amongst the younger generation there was a general consensus that for those who were high users of social media their ‘offline’ and ‘online’ worlds were inherently linked, from the events they choose to attend in the ‘offline’ world to the conversations they had with friends. However, there was debate amongst the young people on whether the representations they made of themselves online was a genuine or biased reflection of their true selves. Some openly acknowledged that they preferred their online persona, whereas others did not feel there was any difference between their online and offline selves. One positive that the young people noted was the ability to curate their online appearance for employment sites such as LinkedIn when desired.

Summary

The majority of participants were ‘high’ social media users [207] and used similar social networking sites across both groups, such as Facebook and Twitter, though with more younger people using Instagram and Snapchat. The main reason participants used social media was to keep in touch with friends and family, but this was being performed in different ways including playing games, posting content and looking up other people on social media sites. Participants also described getting news and information from Facebook and Twitter. Both groups expressed concerns about the role social media plays in their lives and the lives of others, with anecdotal examples of risks of social media use to personal well-being, information security and obtaining misleading information. The discussions highlighted that there are myriad types of data possible to link, and that which of these data we choose to link may impact participants’ views on data linkage. In the next section I explore

participant views on specific types of social media data linkage.

3.3.2 Views on using social media data for research

After being given a presentation on the possible uses of social media data, participants took part in a novel exercise where they discussed different possible scenarios for data collection, with the variables as given in Table 2. The participants engaged well in this exercise and distributing the possible options for each variable amongst the group encouraged a rich, collaborative discussion around what would be deemed acceptable. The results of the analysis below are grouped by each potential variable that participants were asked to consider.

Who has access to the data

Participants discussed how they would feel about different people accessing their data, with the options being ‘Researchers’, ‘Children of the Nineties staff’ or ‘Computers’. Participants understood ‘Researchers’ to be those not necessarily affiliated with the study who may apply to use their data for health and social research, with the study staff being the administrative staff who facilitate the study. No participant had any concerns about the data stored about them being accessed by the staff or researchers associated with the cohort study, with several participants voicing their trust in the study and its protection and safe use of their data. As such, they trusted that their data would not be sold on or used for purposes other than health and social research, which in other contexts was a widely expressed concern in both groups, but more prominently in the young people.

“‘Children of the Nineties’ is fine because I know you’re not going to sell it.”

— Young Person

" I would trust emphatically the ‘Children of the Nineties’."

— Parent

Out of the three possible options most discussion was had over the use of computers (i.e. automated harvesting of information) in accessing participant data, with evident confusion over what that meant in practical sense for both participant groups, and easy misinterpretation of what it would be possible for computers to achieve with their data. For instance, most participants needed clarification on exactly how a computer would be able to access their information, and in one in-

stance there was a concern from participants in the parent cohort that a computer could access and manipulate layers of private data whilst mining social media platforms.

“Once it gets into your systems it can get everything else it wants out of it. Once you’ve put something into Google it can find every email you’ve sent.”

— Parent

The facilitators provided clarification that the cohort research staff would not be able to access private user information held by the social media platform, only the data that would be available to anyone accessing Twitter information. Facilitators also clarified the difference between a malicious virus on a computer and authorised means of accessing online data through an application programming interface. This was generally understood, but some participants were still uneasy about the use of computers. Ultimately, when asked at the end of the sessions if they would be happy for computers hosted by the cohort to collect their Twitter text data all participants agreed they would consent to this.

Level of data anonymity

A difference in attitude between the groups characterised the concerns around the level of data anonymity consented to in each scenario. The parent cohort mentioned on several occasions that they would not post anything on social media that they were not happy for the public to see, and that they regarded it as a public platform. This appeared to lend a relaxed view to all those attending the focus group towards sharing their data in any form, once it was clear that the data would not be sold on and that only the information they consented to sharing would be collected. However, some members of the group said that they would feel more comfortable for computers to have only their de-personalised data due to concerns about what computers could theoretically do with their information; this was related to a misunderstanding about what computers could do as described above. The younger cohort appeared more discerning about the level of data they were willing to provide, with several noting that they felt “safer” or “happier” when raw data was not being used.

“I suppose when things become anonymised it all seems a lot more fine. If it’s being reduced to numbers and data points I would be much more likely to give my consent.”

— Young Person

It was necessary to clarify for the participants that their private messaging data would not be collected, and that computers would only collect the data they had agreed to. Once this had been explained all participants agreed they would be willing to provide their personal text and ‘like’ data in any form. Photo data created more discussion and the younger participants generally felt they would require more anonymity with photo and location data. Anonymity of photo data in this sense referred to distilling a photo down to numeric data, such as the hues and colours used in an image, or to a description of the image.

What types of data are acceptable to collect

The type of data collected generated the most discussion in both groups. In principle the vast majority of the participants agreed that they would provide all forms of data (Table 2), apart from friends’ data, given that it would be collected and distributed by a study that they trusted to store and use their data well. Here, ‘friends’ data’ refers to data produced by a friend or connection, rather than data produced by the user about their friends such as a list of connections, which would be considered ‘network data’.

In the younger cohort there was general agreement that text and ‘like’ data could be collected. Photo data created more discussion, with a feeling among some that it was more personal than their text data, and some feeling that if it was anonymised in the same way then it was no different to text data.

“There’s something intricately linked between your privacy and yourself. A picture of yourself is much more private and identifiable than anything I would write down. So definitely I think I would feel uncomfortable.”

— Young Person

" If the photo is just being broken down into a code then it’s the same as text. What’s being taken and the final product is the same thing."

— Young Person

Other reservations about photo data included a distrust in the ability to reliably code the features of

a photo with the technology currently available, as well as concerns about involving friends who are in the photos without their consent.

In the parent cohort, the same concerns about photos were not expressed, with all agreeing that they only post photos that they would be happy for anyone to see. A similar view was demonstrated by some members of the younger focus group too, given an awareness of how their social media data might be seen by others.

“The only photos I put up on Facebook are ones I’m happy for the whole world to see.”

— Parent

“I’m conscious that any job I go to is going to investigate it [social media profile] themselves so I don’t think anything is very private. I’m always very conscious about what I post.”

— Young Person

The younger cohort had some discussion about location data, but ultimately did not express concern since the majority of participants did not regularly ‘check-in’ on social media or report their location. Those that did were not concerned about the study accessing this information about them and did not feel it was any more personal than the other data the study held on them. The area of data collection that caused the most concern was data related to their friends accessed through their account. For many of the participants this did not vary across platforms, even if all the data collected would be public anyway, for instance on Twitter. Both sets of participants felt uncomfortable at the idea of giving consent for others to access these data when their friends had not agreed, as well as concerns in the younger generation about ownership of data and what would happen if those data had been collected by the study but the original content was later deleted by the user.

“I object to it strongly... my friends haven’t agreed to that.”

— Parent

" Twitter is public, but only for the time you leave it up. If you give permission for someone to store and access your data you’re giving them permission to have it for as

long as they want it."

— Young Person on giving consent for collection of friends' data

When a sub-group of the young people were explicitly asked whether they would consent to the positivity or negativity of their friends' posts being anonymised and scored in order to gauge whether this impacted on their own posts they agreed to this, which was a contradiction to the same group previously stating they would not consent to any of their friends' data being accessed. Their agreement was more readily given for Twitter than for Facebook. Some participants actively voiced their dilemma, that they knew these data would be "used for good" but that they felt morally conflicted about actively allowing it. Other participants reasoned that they would probably give their consent, and that the focus group environment was encouraging them to think more deeply about it than they usually would. Collectively the sub-group of young people came to a decision that they would "probably" allow for these data to be collected, on the basis that it would be used for "important health and social research".

"We're always sharing our data with loads of people all the time who are using our data for advertising and selling it on. At least with this we would have given our consent and knew it was for a good cause."

— Young Person

Across both cohort groups there was a consensus that though they had preferred platforms they would ultimately consent to sharing their information from any platform once they understood how the data would be protected and what it would be used for. This agreement was usually given with reference to their trust in the study that their data would be safe.

"Anything I was sharing on my social media, including location, I'd be happy to share with 'Children of the Nineties' "

— Young Person

When considering friends' data, this did not differ across platforms, and participants were mostly consistent that they would not provide consent for this. This was with the exception of the sub-group who expressed more willingness to consent to their friends' data being collected from Twitter, as a public platform, than Facebook. This subgroup ultimately stated that they would consent

to this data being collected from both platforms if it was being used for ‘good’, though their initial reaction was that this would not be acceptable to them.

" I wouldn't be happy if someone consented on my behalf. And that's the same on every platform. It's not my place to consent on someone's behalf."

— Young Person

Data linkage

Data linkage refers to joining together previously unassociated information about an individual in order to build a comprehensive collection of data about them from different sources, for instance attaching health data to educational outcomes. In this study this would involve adding relevant social media data to each individual's cohort study data profile for use by researchers (which can include information from health and other official records). When asked whether this would be acceptable, many of the participants had already assumed that their data would be linked if it were collected as part of the study, and were agreeable to this happening. In fact, all the participants agreed that this would be the best way to get the value from their data and had ideas about which research questions might be answered by doing so. This view was consistent across all participants in both groups.

“I think it's important. Because to get the fullest roundest picture you need to do that anyway don't you.”

— Parent on linking their existing data

Views on suggested research methodology

The final part of the focus group was suggesting a possible research methodology to participants, to see if they would be inclined to agree to it. The methodology was threefold, and in each case was presented as being with participants' consent for use in important health and social research:

- “Computers access my raw text data from Twitter”
- “Children of the Nineties staff access my raw text data from Twitter”
- “Researchers access my anonymised text data from Twitter.”

Participants unanimously endorsed these three methods of data access for the different groups con-

cerned, understanding that study staff could hypothetically access their personal data but that they would not routinely need to do so. Participants reasoned that this was equivalent to the requirement for study staff to access any of their other sensitive data held by the study, such as health evaluations.

Summary

On the whole participants agreed they would be happy to share their text, 'like', location and network data with researchers from the study in any form, though it was deemed easier to accept the further the data were anonymised. Photo data were a slightly more sensitive data type, with the majority agreeing they would share this in its raw format, but a minority considering photo data too identifiable to share unless anonymised. Participants frequently cited their trust in the ALSPAC study, and subsequently their trust in anyone who was given the data, though there were reservations and confusion for a small number of the older generation participants on the role of computers in the data collection process due to misunderstandings about what computers could do. As participants were agreeable to sharing most of their data, they did not have reservations about which platform this was done through; however, the platform had some influence when considering whether to share information about their friends. For instance, some participants felt it would be more acceptable to collect information posted publicly to Twitter than to Facebook private profiles. The majority of participants were not agreeable to allowing collection of their friends' data, and those who did agree only did so when given a clear scenario describing the type of research question that this would be useful in answering and the anonymisation approaches that may be used.

When considering the differences between generations, the groups' levels of technological insight had a varying effect on their opinions and thresholds for agreement. For instance, in the older group the use of computers in the research generated unease due to the perceived likelihood of them distributing malicious viruses or leaking personal data. The older generation agreed they would be comfortable for the study to collect any data, apart from friends' data, because they had nothing to hide. In the younger group the privacy of the actual content was not widely considered, however the issue of information security and desire to protect their private information was more apparent. Along with their knowledge of this area came a sense of inevitability of how often loss of

privacy already happens, and some reference to incidents such as the Cambridge Analytica scandal. The participants' willingness to share their data was then because they were aware that their data were already being used for the benefit of private companies and they would rather it was being used 'for good' as well.

All the participants also showed good insight into how the data could be useful to the cohort study, with several participants offering suggestions of how their social media data could contribute to health and social research and other platforms (such as exercise or dating apps) that the study should consider researching.

3.4 Discussion

The question of how participants in cohort studies feel about sharing their social media data has been largely unexplored, but the majority of the findings from the present study are consistent with those from previous focus groups on users' views of data linkage and social media research [207, 225]. Ultimately all the participants agreed that they would consent, if asked, to the study collecting their social media data in a scenario where computers (managed by study staff) accessed the raw data, and converted this into anonymised numbers which were then distributed to researchers. They would also consent to these data being linked to their existing data in the study. The most acceptable data types to collect were text, 'likes', location and network data, with images being slightly less acceptable to some, and friends' data being particularly contentious, with only a minority agreeing. When discussing the use of friends' data, participants' views changed depending on how the question was phrased. When first asked if they would share their friends' data, participants were firm that they would not. However, some agreed when presented with a specific scenario. The participants noted their own difference in opinions, and despite discussion around this there was no overall resolution of opinions for any given situation.

The findings on the participants' general views on social media showed similar themes to those noted by O'Reilly and colleagues [225] in their focus groups with adolescents, where participants held a view that social media is bad for mental health, that it was a platform for bullying and some reference to the 'addictive' nature of social media. Certainly, most of the discussion about social media was about its negative attributes, with words such as 'dangerous' and 'unsafe' used.

The belief that social media can be detrimental extended to those participants who said that they had not directly experienced negative consequences. However, the participants' acknowledgement of the negative side of social media was held alongside specific examples of the benefits such as keeping in touch with friends or providing company to lonely older people. This illustrates the participants' awareness of the advantages and disadvantages of social media, and represents a considered decision to continue engaging with it.

As well as a generally negative view of the impact of social media, there was distrust of online data security amongst both generational groups. In the older generation this presented as concerns about the use of computers in the research, and in the younger cohort presented as increased awareness of their digital privacy on social media sites and the inevitability of the exploitation of their data, which appeared to make them more discerning than the parent group on what they would agree to share. This could be seen as consistent with the younger group having grown up with technology available to them and having a different awareness of how it operates than their parents do. However, this view of younger generations as 'digital natives' [226] can be misleading; whilst age is associated with someones' likelihood to be immersed in technology, it is not the only relevant factor [227]. As such the generation differences I observed may not solely be attributed to the participants' age. Interestingly, this contradicts findings by Wellcome [212] on the publics' views of general data linkage, where younger people were more likely to agree to share their data, and older generations were less likely. A broader sample of participants would be helpful in order to thoroughly investigate the nuance in the reasoning of both groups, and how it relates to the types of data being shared.

Despite having some privacy concerns about their social media data, situating this type of data against the level of privacy of other data held on them in the cohort study, such as health assessments and genome-wide genetic data, allowed participants to make reasoned and informed judgments about what they would consent to. However, within 'social media data' there were layers of types of data which held different levels of sensitivity to participants and, similarly to reports by both NatCen [207] and Wellcome [212] on users views, photo data was slightly more sensitive than other types of data. NatCen's report found that researcher affiliation had an impact on whether a participant would consent to a scenario, and I saw this influence with the ALSPAC participants who were openly confident in the study and its intentions and told us that this gave them confidence

in sharing data with the study. I hypothesise that a study which is not using an LPS sample may find more resistance to the disclosure of those data considered more sensitive. Similarly, while the NatCen participants had reservations around the efficacy of social media research the participants in the present study displayed accurate insight into how their data may be useful to researchers and why it was important to gain their views, which could be attributed to their long-term participation in the study.

A common theme throughout both focus groups was reference to their trust in the study, and their belief that their data would be used to benefit others, which supports the use of LPS as a valuable source of ground truth data due to participants' existing investment in participating in research and the depth of data already available on the cohort. The differences in concerns between generations suggests a need for informed consent to be obtained in a thoughtful and well explained way which meets the needs of all age groups, particularly those who are not 'digital natives' [116, 226].

The variety of opinion around the use of friends' data which were found in this study warrant further exploration, particularly given the current digital privacy environment, and the apparent lack of concern over the use of 'network' data such as lists of friends. The difference in opinion depending on how the question was phrased may suggest that only specific, controlled scenarios are acceptable to participants and understanding the thresholds of this decision making is important in considering the ethical implications of this type of work. Similarly to other work on non-consensual data linkage [203], the differences in stance may also be a reflection of the complexity of the decision and the ethical dilemma it presents to participants.

There were limitations to this study, particularly around the sample of both ALSPAC and of the focus groups in particular. Although Bristol was at the time of recruitment representative of UK cities, there is estimated to be a shortfall in the recruitment of less-affluent families, and mothers from ethnic minority backgrounds [118], as well as differential attrition over time [228]. With regards to the focus groups specifically, the sample size was relatively small, especially for the parent group, and it may be likely that those who agreed to attend a focus group on social media would be more willing to share their social media data. Similarly, those participants who actively participate in the study by attending focus groups may be more likely to feel positively towards the study. A future direction may be to survey the whole cohort on their views through an annual survey, which would give a quantitative perspective on the issue of acceptability. It is also important

to recognise that a focus group methodology has drawbacks, particularly with regard to the ability to generalise the results, the ability to cover broad topics in a relatively short time-frame, and the understanding that the views of participants are socially constructed within the environment of the focus group itself [229]. These results should be interpreted with an awareness of these limitations.

3.5 Conclusions

The focus groups have provided an insight into the views of cohort study participants on using their social media data in research. All participants agreed they would be happy to share their anonymised social media data with researchers affiliated with ALSPAC for health and social research, apart from data about their friends. Whilst there was a preference for anonymised data, most participants felt that their trust in the study would allow them to share any level of data with researchers, often motivated by the positive intention of the research. It is acknowledged that the sample that chose to attend the focus group was small and may have been biased in their willingness to agree to the hypothetical scenarios.

The engagement and willingness of the participants to discuss social media and its applications in research suggest that LPS could be a valuable source of ground truth data, especially given the opportunity to link their social media data to other measures taken since birth. This would give researchers a valuable opportunity to learn more about who uses social media and start to study the attributes of this population.

Insights from this research can inform studies designing social media data collection strategies, particularly describing which categories of content are seen as more sensitive than others. Feedback from the participants emphasised the importance of clear information for any participants involved in the suggested research, especially with regard to the involvement of computers in accessing their data and safeguards used to protect it.

The participants' views on which of their data they would be happy to share could be revealing if explored further, especially the distinction between accessing network data against accessing friends' data. This work paves the way for future work integrating social media with LPS data, which will be beneficial for both the studies and those conducting social media research.

Chapter 4

Twitter data linkage: features of consenting participants and their data

Abstract

Background Social media are exciting and potentially valuable data sources for health research. However, to ensure that research using these data can be applicable in wider contexts, they must be linked to high quality ground truth data. This chapter describes the process and outcomes of linking Twitter data in the Avon Longitudinal Study of Parents and Children, as well as investigating potential sources of bias in the data collected.

Method There were 26,205 participants invited to link their Twitter data, of which 4,261 had Twitter and 654 were successfully linked after opting in. Their Twitter data are now continually being collected, with their tweet histories also collected up to the maximum number of historical tweets Twitter allowed. For the purposes of this study, I used Twitter data available up to the 31st October 2020, linked to measures of depression, anxiety and well-being taken in April to May 2020. I compared the characteristics of G1 linked participants to the Twitter user group in Chapter 2 (N=2,347) and to the depression, anxiety and well-being measures collected from the whole cohort in April to May 2020 (N=6,827). I also tested whether rates of tweeting were related to the

sex or generation of participants.

Results 15.3% of those who used Twitter in ALSPAC had their Twitter data successfully linked. Of these N=654 participants, 224 were from the G0 generation of ALSPAC and 430 were from the G1 generation. Their characteristics align with the population of Twitter users described in Chapter 2, and largely reflect the general cohort in terms of their levels of depression, anxiety and general well-being. In the most recent year 471 linked participants had tweeted at least once and a quarter of those tweeted less than 6 times. Tweeting frequencies were not found to be statistically associated with sex or generation of participants.

Conclusion The data linkage programme successfully allowed for Twitter data to be directly compared with the mental health and well-being outcomes of participants. This dataset can now be linked to a wide range of outcomes that have been collected in ALSPAC and represents a realistic range of tweeting behaviours with which to train and test future models.

Aims

This chapter gives a summary of which ALSPAC users agreed to link their data, and how they are different from the rest of the cohort as well as the Twitter users identified in Chapter 2. This information provides useful context for Chapter 6.

4.1 Introduction

The development of digital phenotypes from social media as a means of understanding human health and behaviour is a relatively recent development in medicine and epidemiology, and one with exciting potential [16, 41]. There have been multiple advances in this area over the past ten years, where social media have been used to make inferences about population and individual mental health with relative success [10, 41]. However, ease of access to large quantities of internet data does not necessarily mean the data are high quality and there are several practical and ethical challenges to meet in order to be able to make the most of internet data [104, 116].

Concerns about the quality of the data in the field have persisted for many years and have led to concerns about the validity of inferences that can be made from the published literature so far (for a comprehensive review see Chancellor and De Choudhury [41]). For instance, social media data do not tend to include demographic information about the individuals whose data is collected, and so cannot be accurately characterised to understand demographic effects or biases [230]. This is particularly important since the populations of those using social media are self-selecting, and do not tend to reflect the general population in demographics or mental health outcomes, as seen in Chapter 2 and elsewhere in the literature [78]. Similarly, the data do not usually include information about the outcome variable being inferred, unless a medical diagnosis is stated in a tweet itself. Many studies use such self-disclosures of mental health conditions as positive indications of a mental health disorder, however these cannot be verified and online self-disclosure is likely to be influenced by gender and cultural norms which can then confound analyses [230, 231]. In order to use social media to its full potential, we need it to be linked to well characterised and, ideally, longitudinal datasets that can provide the ground truth data needed to label individual characteristics and outcomes [10].

As well as practical considerations about the quality of the data available for research, there are ongoing concerns about the ethical collection of social media data for research, particularly with respect to informed consent, which can rarely be guaranteed in an online study [102, 230]. It is relatively easy to amass data on a large number of individuals, who in theory have consented through the terms of service of each platform to their data being shared, but in practice social media users often do not realise this is the case [97, 99]. Alternatively, their data may be scraped

from the web without consideration of the terms of service of the platform that the data is being taken from, and whether or not they consented to these.

Data linkage in cohort studies has the potential to address both these practical and ethical concerns, since it requires the explicit consent of all participants, and linkage to high quality longitudinal data that already exists about them. As well as benefiting social media researchers, Al Baghal and colleagues [111] have drawn attention to the potential of social media to reciprocally add value to these longitudinal studies, which often suffer from attrition or missing waves of data. It may be possible to use the data we collect from social media to enhance the data that already exists in the studies, and as such bolster the information available in long-term studies, which is currently a strategic priority amongst large cohort study funders [113]. This potential has been demonstrated by linking Twitter data in a handful of studies in the United Kingdom including Understanding Society and the British Attitudes Survey [83] as well as across the world [109]. Twitter tends to be a particularly popular source of data for research, partly due to the ease of accessibility to its data compared with other platforms such as Facebook and Instagram, who do not tend to allow for research access to their data, and partly due to the value of textual data that it contains.

While data linkage is a promising avenue for addressing the quality of digital footprint data there are some important questions about representativeness and data asymmetries that result from these forms of linkage. Data asymmetry occurs when the volume of linked digital footprint data amongst a participant group varies widely across participants, and in comparison to the volume of questionnaire that it has been linked to [83]. For instance, users may have linked their accounts but only have tweeted once, whilst others may produce thousands of tweets. Much like with missing data, we would hope that this asymmetry is distributed randomly, but if it is associated with particular characteristics of the participant group then this could inadvertently cause confounding in later analyses. Representativeness can also be an issue since we rely on participants to opt-in to data linkage programmes, and so may find that those who opt-in are biased towards certain characteristics, or may have particular mental health profiles that mean our data will not be population representative. Given these issues it is clear that whilst data linkage in longitudinal cohorts has many benefits it also has its own limitations which need to be well understood in order to use the data effectively.

This chapter aims to develop this understanding by focussing on the dataset that has been pro-

duced by linking Twitter data in the Avon Longitudinal Study of Parents and Children (ALSPAC), a multi-generational birth cohort study which aims to compile a rich databank containing information on participants' health and social exposures and subsequent outcomes across the life course [118–120]. A large proportion of the ALSPAC young people are regular users of social media, as seen in Chapter 2, and their age group currently makes up the biggest group of Twitter users [232]. The cross-generational sampling demonstrated here in ALSPAC also means that the technical framework used will be widely applicable across other cohorts including those in the Cohort and Longitudinal Studies Enhanced Research (CLOSER) consortium. I will give an overview of the methodology developed for linking the data, and then go on to describe the data that has been collected through the linkage programme so far, with a focus on the quantities of Twitter data generated by the linkage programme, and the demographic and mental-health features of the linked participants compared to the general cohort. Specifically, the research questions explored are (1) what are the consent rates to Twitter data linkage and how are they comparable to other data linkage studies, (2) how do the characteristics of those who linked their Twitter data compare to Twitter users in the cohort and (3) what is the extent of data asymmetry in the linked data, and is it biased by participant characteristics?

4.2 Methods

4.2.1 Cohort description

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study [118–120]. Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled was 14,541 (for these at least one questionnaire has been returned or a “Children in Focus” clinic had been attended by 19/07/99). Of these initial pregnancies, there was a total of 14,676 foetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. The total sample size for analyses using any data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 foetuses. Of these 14,901 were alive at 1 year of age. Since ALSPAC collects data on multiple generations of participants, the generations are referred to from G0 to G2, where the G0 are the parents of the original study children, G1 the index children, and G2 the children of

the G1 participants.

The Twitter data linkage was conducted on a subset of the full adult ALSPAC cohort, that is the G0 mothers and their partners, and the G1 index cohort. The process for this data linkage is described in the following section, with references to the G0 cohort from this point on referring both to the mothers and their partners.

4.2.2 Twitter Data Linkage Programme

Participant Consultation, Ethics and Informed Consent

Prior to the participant consent campaign the data collection and consent processes were designed with input from a variety of stakeholders. First there were interviews with leaders of the CLOSER cohorts to obtain the broadest possible overview of challenges addressed by the framework; these interviews were conducted by researchers and staff within ALSPAC. Two focus groups were then held in September 2018 for the purpose of exploring participant attitudes towards sharing their social media with the cohort. This included asking what they considered to be acceptable anonymisation for sharing their data, and to understand their general views about the use of social media in health and social care research. These focus groups involved individuals from the G1 (N=9) and G0 (N=5) cohorts and the results from these focus groups are described in detail in Chapter 3. We found that participants were generally accepting of the linkage of their social media data with their consent in order to facilitate health and social care research. They felt that their trust in the cohort would make them more likely to take part, and that they would prefer the use of anonymised data derived from their social media text to be shared with researchers rather than raw data. This research informed the development of the data linkage programme, and subsequent guidelines for data sharing with other researchers by the cohort team.

The study and data linkage programme itself was approved by the ALSPAC Ethics and Law Committee.

Invitation and Reminder Strategy

Following the invitation strategy being designed by the ALSPAC team, all adult members of ALSPAC (N=26,205) were contacted to opt in to the data linkage programme via post or email. Of

those who were contacted, 21,944 said they had no Twitter account (8,500 of which were emailed and 13,444 of whom were contacted via post) and 4,261 were contacted who said they did use Twitter (3,662 via email and the remaining 599 via post). 19.6% of participants of ALSPAC who had a Twitter account provided an account name to link. Of those, 654 participants had their data successfully linked to their cohort data on an ongoing basis, which represents a 78.3% success rate in linking those accounts that were provided. This inclusion flow is represented graphically in Figure 4.1.

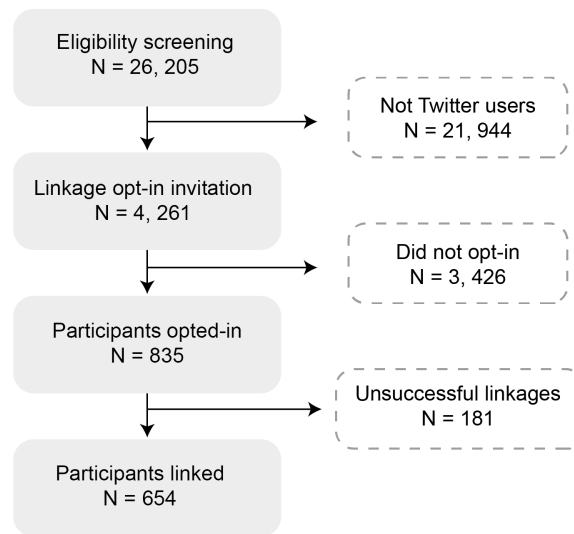


Figure 4.1: The number of participants included at each stage of the process for identifying participants for linkage, as well as the reasons for not being included in the final sample.

Twitter Data Harvesting and Linkage to ALSPAC

For participants who agreed, the data collection and linkage were performed using a software programme named *Epicosm* built for the collection of social media data in cohort data safe-havens [117]. *Epicosm* harvests Twitter data from the consenting cohort participants, and automatically calculates and stores sentiment scores for the tweets. For further details on the technical functioning of *Epicosm* see Tanner et al. [117]. Participant Twitter accounts were linked deterministically to their ALSPAC unique identifiers, with 181 users whose matches were unsuccessful due to errors in the given account names, or accounts being private. Historical tweets span as far back as April 2009 and data collection has been run every three days since to update with the most recent tweets. The Twitter REST API which is used to collect the tweets was limited to collecting approximately 3,200

historical tweets per person. Twitter has since released an updated API which allows approved researchers to collect up to 10 million tweets per month, with no limit on user timelines. However, the data used in this chapter was collected prior to this API becoming available. Data collected about tweets includes: tweet text, tweet type, public metrics such as likes and retweets, and the datetime of the tweet.

Once collected data are stored in a MongoDB database and can be linked within the ALSPAC data safe-haven. The data are then anonymised by the ALSPAC Data Management Team before sharing. Figure 4.2 gives a diagrammatic overview of this flow.

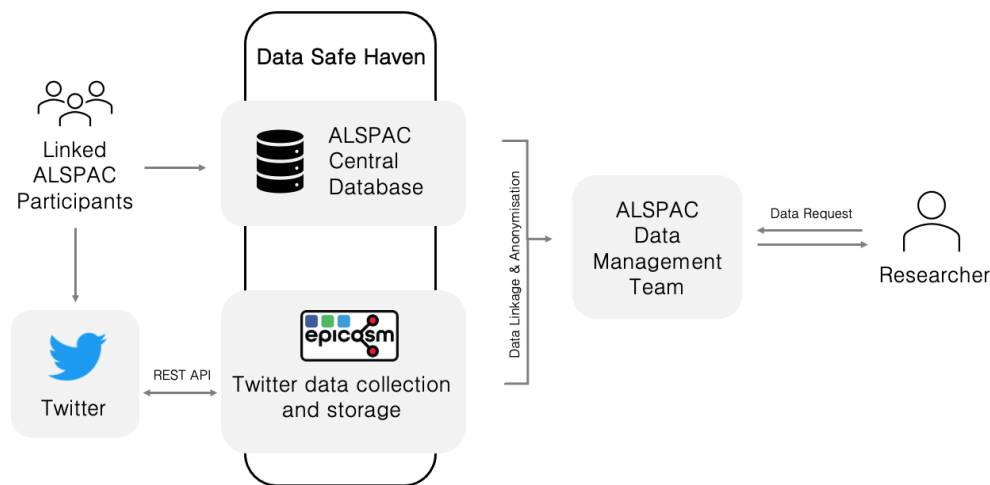


Figure 4.2: A diagram illustrating data flow between the ALSPAC participants, data safe-haven, data management team and researchers.

Data access and pre-processing

In order for tweets to be shared with researchers, the ALSPAC data access requirements stipulated that data needed to be anonymised so that no information that could potentially identify participants is released. As a result, the raw text of tweets are not directly available to researchers which is in line with the wishes of participants (see Chapter 3). Instead, the ALSPAC data management team pre-process textual data inside ALSPAC's data safe-haven and can then release suitably anonymised data linked to the researchers' required outcomes in the cohort. Other forms of dis-

closure control may be used to avoid the re-identification of individuals such as aggregation of timestamps into wider time windows, and removing potentially identifying information in the case of small cell counts. Small counts are five or fewer tweets in any given day or four-hour time-period, which is in line with standard statistical disclosure guidelines [233].

At the time of writing four sentiment analysis algorithms were available for processing tweets with this dataset, with a view to allowing researchers in the future to submit their own pre-processing and analysis algorithms provided they provide sufficient anonymisation for the textual data. This feature is currently under development. The four current algorithms available were the Valance Aware Dictionary for sEntiment Reasoning (VADER) algorithm [234], labMT [235], the Linguistic Inquiry Word Count (LIWC) 2015 [236], and TextBlob sentiment analysis [237]. These algorithms are further discussed in Chapter 6. Textual pre-processing performed by *Epicosm* involved removing urls from the text, and then removal of special characters for LIWC analysis. Special characters were not removed for VADER.

The sentiment algorithms were all applied to each tweet as an individual documents, and so sentiment is measured at the tweet level. Sentiment scores for each tweet were then reported, as well as whether the tweet was a retweet, the date, and the four-hour window of the tweet.

4.2.3 Datasets and measures

In order to make comparisons between the linked sample, Twitter users in the cohort and the whole cohort for this study it was necessary to use several datasets collected by ALSPAC. The following four sections detail Datasets A to D which were each used for a different comparison purpose.

Dataset A: Twitter data

In the present analysis I used data collected from Twitter for the 654 linked participants up to the 31st October 2020. To protect participant anonymity and reduce the risk of disclosure dates with less than five tweets were suppressed prior to being shared (this affected 0.09% of the original tweets). Similarly, four-hour windows on any given day with less than five tweets were also suppressed (this affected 2.6% of the original tweets). The data was collected in 2020 and goes as far back as 2009, but due to restrictions on the Twitter API the full history for all participants may

not have been collected. Basic information about participants such as their sex, age, ethnicity and generation was available with this dataset.

Dataset B: Mental health data

The G0 and G1 cohorts (that is, the parent and index cohort) completed a survey between 9th April and 15th May 2020 that collected data relevant to the COVID-19 pandemic. It was sent to all those for whom there was a valid email address on record. A detailed data note about this resource and the questionnaire design is available [238]. The survey included mental health measures that were of interest for this study, and was chosen to link because of its relatively high response rate, proximity to the tweet harvest date, and because it is one of few points in ALSPAC where both the G0 and G1 participants completed the same mental health measures in the same context at the same time. This made it easier to make full use of the data available from the linked participants from both generations.

Three mental health and well-being measures were taken in this survey. These were depression, measured with the Short Moods and Feelings Questionnaire (MFQ) [239], anxiety with the General Anxiety Disorder-7 (GAD-7) questionnaire [240], and general well-being using the Warwick Edinburgh Mental Well-being Scale (WEMWBS) [156]. These measures are used in this chapter as continuous scale scores.

As well as the mental health data from this survey, some standard attributes like age, sex and the G1's ethnicity and parental socio-economic status were also available. Age was the participant's age in 2021 and sex, ethnicity and parental socio-economic status were all recorded following the birth of the G1 participants.

Dataset C: Linked Twitter data

Dataset C, the linked Twitter data in ALSPAC, were the data from Dataset B for those participants who were linked and who responded, alongside their Twitter data from Dataset A. The overlap between those linked and those who responded was $N=479$. Due to the need to ensure anonymity of these data, the IDs for participants in this dataset were re-anonymised by the ALSPAC data management team, meaning that individuals could not be directly compared between dataset C and any of the other datasets. This means that the individuals in Dataset C are contained within

Datasets A, B and D and so these datasets cannot be assumed to be independent.

Dataset D: Twitter Users in ALSPAC

To understand how the demographic distributions of linked participants compare to all of the known Twitter users in ALSPAC I also compared the information from the linked dataset described above to the data discussed in Chapter 2 regarding the demographics of the N=2,294 Twitter users in ALSPAC and their attributes. These data were collected when the participants were 24 years old, approximately three years before the request for data linkage was sent. For a full description of this data please see Chapter 2.

4.2.4 Analysis

All analyses and data visualisations were conducted in the *R* programming language, version 4.0.3, with RStudio v1.4 [167]. I primarily used the `tidyverse` (v1.3.0) package [168] for data manipulation, `ggplot2` (v3.3.3) [169] for visualisation and `gtsummary` (v1.4.1) [241] and `kable` (v1.3.4) [242] for tabulation.

In the comparisons of results between the whole cohort and linked participants for mental health outcomes, it was not possible to consider only those who had not linked their data because IDs for linked participants go through a secondary anonymisation process. As a result, mental health comparisons are made between the whole cohort (including linked participants), and just the linked participants.

4.3 Results

4.3.1 Features of linked participants

I first want to consider the demographic characteristics of the participants who agreed to link their data in ALSPAC such as their age and sex. I will then also consider their mental health characteristics, and compare this to the rest of the cohort at the same time-point in order to understand how representative their outcomes are.

Demographic characteristics

Sex and age of the linked participants (Dataset A) are presented in Table 4.1, split by the two cohort generations whose data were linked. The linked Twitter sample is made up of approximately one third of the older generation to two thirds of the index generation. Using data from the last time ALSPAC index (G1) participants were asked about their social media use (Dataset D) I can also compare the demographics of the G1 participants to those who filled in that last questionnaire, and to the ALSPAC cohort as a whole (Dataset B). This will tell us whether the demography of those consenting to linkage are similar to the available Twitter sample in ALSPAC. These results are presented in Table 4.2.

Table 4.1: Demographic split of the index cohort with linked Twitter data

Characteristic	G0 (Parents), N = 225	G1 (Index cohort), N = 430
Sex		
Male	39%	36%
Female	61%	64%
Unknown (N)	<5	<5
Age	59 (56, 62)	28 (27, 28)
Unknown (N)	58	70

¹ %; Median (IQR)

Table 4.2: For the index cohort (G1) only: demographic characteristics of the full index cohort, those who said that they had a Twitter account at age 24, and those who agreed to link their Twitter data.

Characteristic	Full G1 Cohort	G1 Twitter Users	G1 Linked Participants
	N = 14,901	N = 2,294	N = 430
Sex			
Female	49%	66%	64%
Male	51%	34%	36%
Unknown (N)	23		<5
Ethnicity			
Ethnic Minority Group	5.0%	3.8%	4.4%
White	95%	96%	96%
Unknown (N)	2,829	241	42

Mental health characteristics

Next, I will consider the depression, well-being and anxiety outcomes of the linked and non-linked participants in April 2020. As noted in Section 4.2.3, it was not possible to obtain the complement of the full cohort dataset and the linked participants, and so the full cohort outcomes (Dataset B) are compared with the linked participant outcomes (Dataset C). The comparisons between these groups is made graphically in Figure 4.3, where the distributions of the continuous scale scores for anxiety, depression and mental well-being are displayed as box plots and density graphs. Given that there are differences in rates of mental health outcomes between men and women (as seen for the G1 cohort in Chapter 2), the equivalent graphs for only female and male participants are given respectively in Supplementary Figures B.1 and B.2. These figures show that in men a similar pattern emerges in that anxiety and depression are slightly higher in the linked cohort, and in men well-being in particular is a little lower. However, in women anxiety and well-being are approximately the same between both the linked and general cohort, but depression is higher in linked participants. Overall, these results suggest that the linked cohort are largely comparable to the overall cohort.

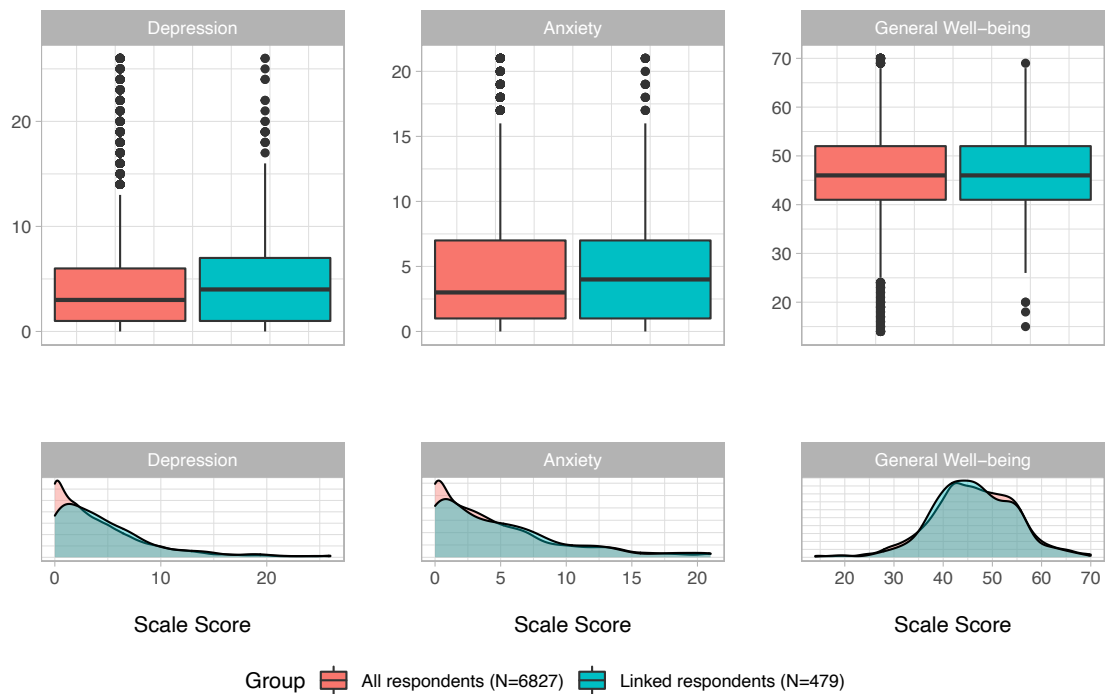


Figure 4.3: A comparison of the distributions of participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.

4.3.2 Features of linked Twitter data

Overall 654 participants were successfully linked. In total, their Twitter data ranges from April 2009 up to 31st October 2020, with a total of 496380 tweets (excluding the 447 tweets with no datetime data). Due to the disclosure control rule of there needing to be five tweets minimum per time window and per day in order for the data to be released, data were suppressed for 447 out of 496,827 tweets in total, and for the four-hour time windows of 12,802 tweets. The rates of datetime information suppression is highly skewed towards earlier years of data collection where either fewer participants were tweeting each day, or less participants' data had been collected that far back. Using the resulting data I now describe the overall volumes of tweets by type, and then the frequencies of tweeting by participants.

Tweet types

Twitter can be used in different ways, with one of the primary choices being whether a tweet is authored by the account holder (a *tweet*) or the re-sharing of an already published tweet (a *retweet*). Figure 4.4 shows the monthly counts of tweets and retweets collected from all of the participants whose data have been linked. There is an increase in tweet volume over time which could be caused by the data collection software being limited to 3,200 tweets per person, so that the full history was only collected for some people, and/or by a slow increase in the popularity of Twitter as a social networking site. There is also a very clear spike in tweet volume at the beginning of the COVID-19 pandemic.

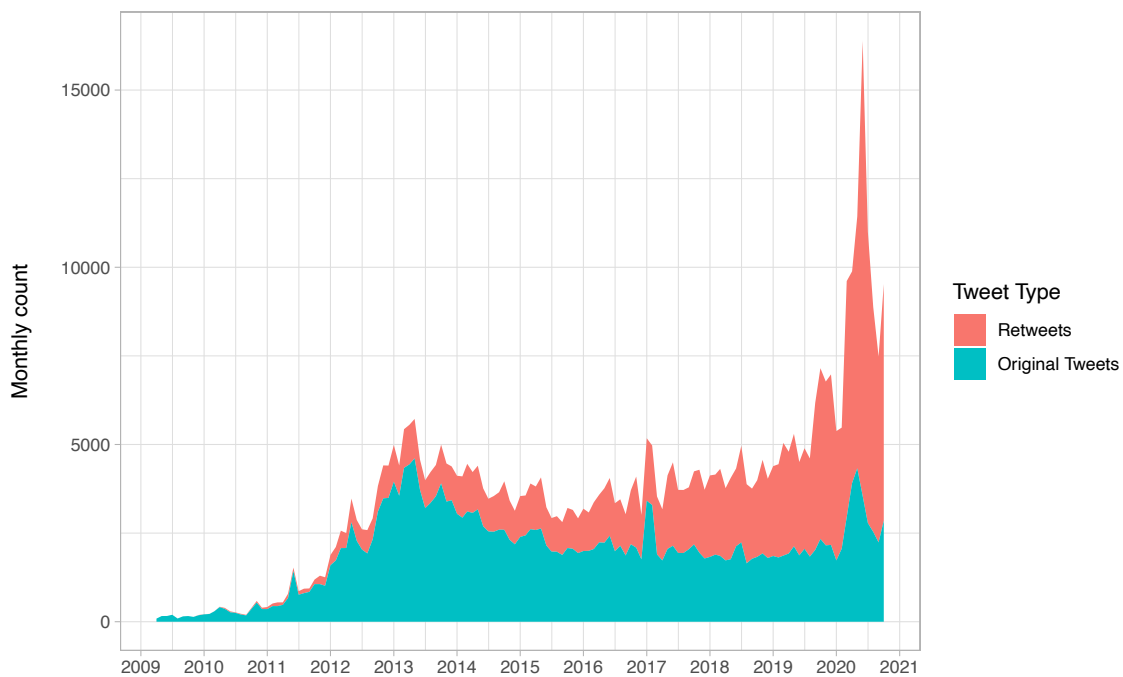


Figure 4.4: Monthly counts of original tweets and retweets collected for all participants. The original tweet and retweet values are layered to fill to the total number of tweets for each month.

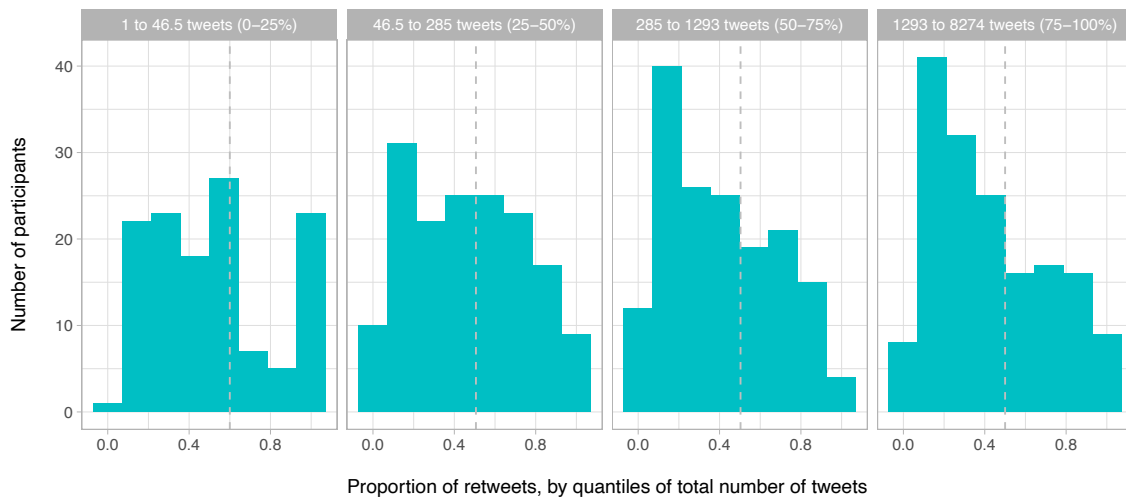


Figure 4.5: Proportions of total tweets that were retweets, split by participants who are in each of the quantiles of total number of tweets. The dashed grey line indicates the median proportion per quartile.

Over the whole sample the percentages of original tweets and retweets were 55.4% and 44.6% respectively. Of course, not all users will be retweeting at the same rate. To investigate the changes in re/tweet proportions Figure 4.5 shows proportions of retweets, split into the four quantiles of overall tweet volume. The overall pattern shows that the distribution of the number of retweets people send becomes more skewed towards fewer retweets as their overall tweet volume gets higher, although the overall median proportion of retweets remains roughly the same.

Frequencies

Since participants tweets can only be collected up to the limits of the Twitter API I will concentrate the description of tweeting frequency to the most recent year of data. This is 31st October 2019 to the 31st October 2020.

Over this most recent year of data 471 participants had used Twitter at least once, with the minimum number of tweets per person being 1, and the maximum being 8274, with an overall mean of 231.6 tweets over the year. Here ‘tweets’ includes both original authored tweets and re-tweets. This gives an average frequency of 4.5 tweets per week. Of those who tweeted in the past year, the first quantile of participants tweeted up to 6 times, the second quantile up to 32 times, the third quantile up to 132 times and the fourth quantile up to 8274 times. A histogram of tweets per person is given

Table 4.3: Weekly tweet frequency for the most recent year of data, split by sex and generation.

Characteristic	Sex			Generation		
	F, N = 277	M, N = 192	p-value	G1, N = 298	G0, N = 172	p-value
Weekly Tweet Frequency			0.14			0.6
Mean (SD)	3.3 (8.5)	6.1 (17.6)		3.9 (12.1)	5.4 (14.6)	
Median (IQR)	0.5 (0.1, 2.2)	0.8 (0.1, 3.4)		0.7 (0.1, 2.6)	0.5 (0.1, 2.4)	
Range	0.0, 75.2	0.0, 159.1		0.0, 159.1	0.0, 109.7	

¹ Wilcoxon rank sum test

in Figure 4.6.

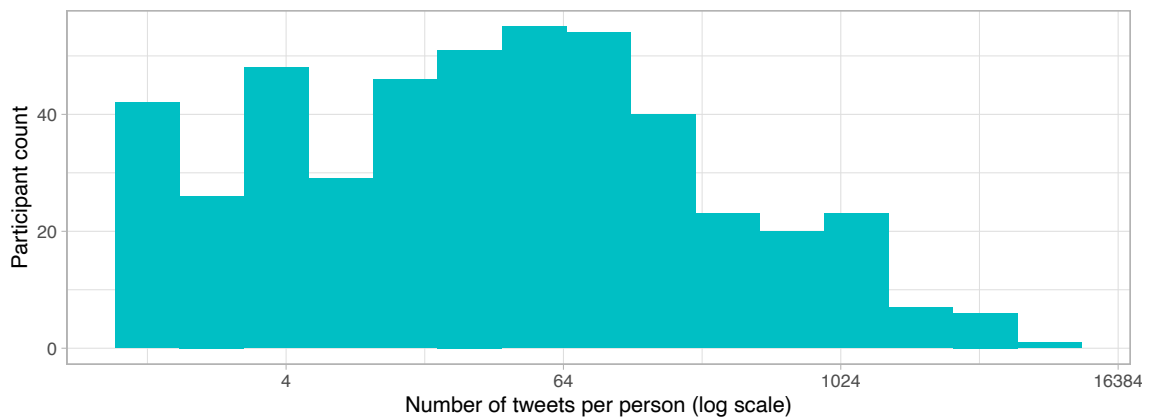


Figure 4.6: Histogram of the number of tweets per person in the most recent year of data (31st Oct 2019 to 31st Oct 2020), with tweets per person transformed with the binary logarithm.

Considering this most recent year of data, I also looked at whether frequencies vary by sex or generation of the participants. Table 4.3 presents a variety of descriptive statistics of the weekly frequencies of tweets by both of these characteristics, as well as the results of the Wilcoxon rank sum test to test the null hypothesis that both groups have the same underlying distribution. The Wilcoxon rank sum test does not assume normality of the data, which is appropriate in this case given the skew observed in Figure 4.6. In terms of daily trends, of the tweets for which the time window was available, 68% of tweets were sent between 8AM and 8PM, 22% sent between 8PM and midnight, and then 9% sent between midnight and 8AM.

4.4 Discussion

In this chapter I have given an overview of the data that have been collected through ALSPAC's Twitter data linkage programme, and used a dataset linked to a recent mental health survey to describe the distribution of different attributes in the dataset and examine potential biases or asymmetries.

4.4.1 Consent and successful linkage rates

Of the ALSPAC population that said they used Twitter (N=4,261), 19.6% agreed to linkage. However, due to 181 unsuccessful matches, 15.3% were ultimately linked. There was a large difference in the number of Twitter users between participants invited by post versus by email, where 30% of the emailed participants used Twitter, but only 4% of the participants invited by post did. The rate of successful linkage is slightly lower than that obtained by Al Baghal et al. [111] in their linkage experiments with panel surveys, which varied between 27% and 37%. However, this may be because the age group for this study included those aged 56 to 62 years old, rather than mainly young people who are more likely to be Twitter users [232], and that it included postal surveys rather than being only web-based.

There is evidence from previous studies that survey mode can have an impact on a participant's likelihood to consent to Twitter data linkage [109, 111], where higher consent rates for data linkage were obtained when participants were invited by an interviewer in person. This evidence is also consistent across other types of data linkage requests in longitudinal studies, and it is thought that this finding is because there are lower levels of understanding of requests made online compared with being able to converse with a study representative [243, 244]. The level of understanding that participants have of a request is then associated with their likelihood to consent [244]. When taken alongside the strong messages from ALSPAC participants in Chapter 3 around the importance of personal privacy and trust when it comes to their social media data, it is reasonable to hypothesise that rapport with an interviewer or clinic staff may have increased the level of trust in the safe use of participant data, and therefore have increased the opt-in rate. This is a step that could be implemented at a later date or considered for future data linkage projects with digital footprints.

4.4.2 Characteristics of the linked participants

Knowing that the linked participants represent 15% of Twitter users in ALSPAC, I then investigated whether there was a bias in the characteristics of those who did choose to opt-in versus those Twitter users who did not. I considered both demographic and mental health characteristics against the ALSPAC cohort, and also demographic characteristics against the group of Twitter users in ALSPAC. Differences between Twitter users in ALSPAC and the general population are discussed in more detail in Chapter 2. When considering representativeness in ALSPAC, it is important to note that the characteristics of the index population (G1) were designed to be population representative, but their parents (G0) were not. Of those who were linked in the G1 cohort, the sex distribution compares relatively well to the overall ALSPAC G1 population, as well as to the Twitter population for this cohort described in ALSPAC in Chapter 2. The linked sample does reflect the general pattern of attrition in ALSPAC, in that women are more likely to remain in the sample over time [120], and so whilst the sex distribution reflects overall ALSPAC patterns, they are not necessarily population representative (Chapter 2). I did also see that for the G0 cohort of parents, there was a higher proportion of men to women than in the G1 cohort.

Overall these patterns concur with previous research that found there was not a statistical association between opting-in to data linkage and demographic features [243, 245]. However, there is evidence from the literature that level of education is associated with opting-in, which is thought to be linked to the participants level of understanding of what they are being asked to do [245].

In terms of mental health and well-being, Figure 4.3 shows that linked participants are fairly well represented in terms of their anxiety, depression and well-being distributions against the full cohort, with just slightly higher proportions of participants with higher anxiety and depression in the linked sample than the cohort overall, but very similar rates of general well-being. This follows what we might expect from Chapter 2, where we saw that Twitter users did exhibit slightly higher rates of depression than the cohort overall, but had very similar rates of general well-being. As such, the differences seen in Figure 4.3 between the distributions for all participants and linked participants are likely to be attributable to linked participants being Twitter users, rather than to opting-in to the data linkage programme. This is further supported by evidence from Mneimneh et al. [109] who found that the presence of mood or anxiety disorders or suicidal ideation had no impact on opting-in to Twitter data linkage across three studies.

4.4.3 Characteristics of the linked data

Collecting Twitter data from a population or general survey sample means that we should not expect all participants to produce the same volumes of data, unlike in a study where we could select a sample from the Twitter API based on their tweet frequency. In this study, of those with linked data two thirds had tweeted in the past year. Tweet frequencies of those linked show that a quarter of this sample have sent 6 tweets or fewer in the past year of data, with the overall distribution of tweeting frequencies highly skewed towards fewer tweets per person. This, as seen in Table 4.3, can impact summary statistics like the average number of tweets per person. When considering the average tweet frequency by generation the older generation (G0) appear to tweet more frequently if using the mean, but the younger generation (G1) tweet more frequently when using the median. This is likely due to a small number of participants tweeting at very high rates as seen from the distribution in Figure 4.6 which skews the mean.

Whilst the data is asymmetrical as expected [83], I have also found that variations in tweet frequency are not statistically associated with the sex or generation of the participants (Table 4.3). That said, women do appear to post less than men in terms of summary statistics, and previous studies have found that women post less often too [83]. Other research has also found that those with higher level qualifications post more regularly, and those without post less [83], which is something that could be tested with the linked dataset at a later date. Other potentially influential patterns include that retweets are more frequent in those who tweet less often as seen in Figure 4.5.

Overall the presence of data asymmetry is expected, and indeed is the purpose of linking data in a population representative cohort. By having a broad range of tweeting patterns we can attempt to replicate a more realistic experimental setting for training algorithms using Twitter data, as opposed to only training with an ideal dataset that does not transfer effectively to ‘real world’ use. There are suggestions that there may be minor demographic differences in tweet frequency, which should be considered in the use of the dataset, especially since those who are less well represented are those who are already more likely to be harmed by other forms of structural bias.

4.4.4 Strengths and limitations

There are some limitations to this investigation. Firstly, due to the dual anonymisation process of participant IDs the linked Twitter data cannot be compared directly to other research datasets from the cohort. This limited my ability to statistically test differences between samples since the samples cannot meet assumptions of independence. The data collection was also limited by the Twitter API itself, which did not guarantee that every Twitter user's full history would be collected. Following on from this, the most recent (and therefore most likely to be complete if the full history was not collected) year of data includes the beginning of the COVID-19 pandemic which, as can be seen in Figure 4.4, aligns with a huge increase in tweet volume that is likely to have biased the tweet frequency figures calculated. Lastly, since only the anonymised sentiment data was available for analysis it was not possible to conduct any detail analysis of patterns in the textual data collected.

In terms of strengths, the data linkage project described presents a number of advancements to the collection and sharing of research data from digital footprints that may allow for new advances in this space. Knowing who does and does not use Twitter, and who consented to data linkage, means that we can accurately understand bias generated from differential opt-in patterns and Twitter use. Researchers can now request for Twitter data to be linked to any of the huge number of measures available in ALSPAC over all time, which is a significant improvement in the availability of ground truth information available for model training. Up until now much of the field of digital footprint research has relied on single studies either inferring data labels from Twitter information or attempting to gather ground truth labels directly from participants at a single time point (this will be discussed further in Chapter 5). Secondly, the data being linked in a population cohort gives researchers an opportunity to understand how algorithms might behave on data that looks more like what might be found naturally on Twitter. That is, a wide range of tweeting patterns, a mix of demographics and also two different age ranges to test with. Lastly, the data being managed by the cohorts data management team allows for digital footprint data to be managed ethically and securely, which is one of the primary concerns about the use and sharing of such data, especially from social media [116].

4.5 Conclusion

The Twitter data linkage programme that has been described and presented in this chapter allows for Twitter data from a group of participants to be studied alongside their longitudinal data collected by the ALSPAC cohort study. Whilst there are some asymmetries in the data, these are in line with what would be expected based on the literature, and reflect a realistic complement of Twitter users. The knowledge of the characteristics of the Twitter data and the participants it belongs to can be used in future studies to train and test new or existing models for classifying health conditions on Twitter.

Part B: Mental health data science using Twitter

Chapter 5

A review of methodologies for monitoring mental health on Twitter

Abstract

Background The use of social media data in predicting mental health outcomes has the potential to allow for continuous monitoring of mental health and well-being, and to provide timely information that can supplement traditional clinical assessments. However, it is crucial that the methodologies used to create models for this purpose are of high quality from both a mental health and machine learning perspective. Twitter has been a popular choice of social media due to the accessibility of its data, but access to big datasets is not a guarantee of robust results. Here I review the current methodologies used in the literature for predicting mental health outcomes from Twitter data, with a focus on the quality of underlying mental health data and machine learning methods used.

Method A systematic search was used across six databases with keywords relating to mental health disorders, algorithms, and social media. 2,759 records were screened, from which 165 papers were analysed. Information about methodologies for data acquisition, pre-processing, model creation and validation were collected, as well as replicability and ethical considerations.

Results The 165 papers reviewed used 120 primary datasets. There were an additional 8 datasets identified that were not described in enough detail to include, and 10 papers did not describe

their datasets at all. Of these 120 datasets, only 16 had access to ground truth data (i.e. known characteristics) about the mental health disorders of social media users. The other 104 datasets collected data by searching key words or phrases, which may not be representative of patterns of Twitter use for those with mental health disorders. The annotation of mental health disorders for classification labels was variable and 68 out of 120 datasets had no ground truth or clinical input on this annotation. Despite being a common mental health disorder, anxiety received little attention.

Conclusions The sharing of high-quality ground truth datasets is crucial for the development of trustworthy algorithms which have clinical and research utility. Further collaboration across disciplines and contexts is encouraged to better understand what types of predictions will be useful in supporting management and identification of mental health disorders. In communicating their studies researchers should be explicit in exactly what their models predict, use appropriate evaluative metrics, and ensure their reported methodologies allow for accurate interpretation of their results.

Aims

This chapter aims to understand where the current limitations of the literature around mental health inference from social media lie, and identify areas which could be targeted to improve the quality of future research.

5.1 Introduction

The detection of signals of mental health through heterogeneous data, known as *digital phenotypes* [15], is a rapidly evolving field of research that requires interdisciplinary expertise in both the behavioural psychology of mental health, and the computational modelling of associated behaviours using digital footprint data [246]. Social media has been a popular platform for accessing data to investigate digital phenotypes [247, 248], and has provided a promising opportunity to model individual and interpersonal behaviours to further understand typically private topics such as hate speech [249] and political ideation [250], as well as mental health. Whilst there are a range of possible social media platforms which could be used for analysis, Twitter (<https://twitter.com>) has been a popular choice for research due to its public-facing design and readily available application programming interface (API) which enables easy access to data for research [57, 251].

Currently mental illness is one of the leading causes of the overall global disease burden [252], with depression estimated to be one of the most prevalent diseases worldwide [253]. The implications of mental ill health are profound on both a micro and macro scale, from personal relationships to the global economic burden [254, 255]. As a result there has been increasing interest in the potential of data-driven methods to provide a new approach to early detection and prevention of mental health disorders [10, 256–259], particularly for young people [257], which could serve to promote access to mental health care, and improve opportunities for self or clinical monitoring. The use of data created through day-to-day technology use could even contribute to clinical assessments by health care professionals, who typically use questionnaire-style diagnostic tools which can be biased by a patient’s retrospective recall [260] and so cannot always provide an accurate overview of a patient’s well-being for weeks, or months, at a time. Additional benefits of using social media data are the ability to collect data on populations of people with less common mental health disorders such as schizophrenia or PTSD, which is generally not possible outside of a clinical environment.

5.1.1 Themes from previous reviews

There have been a series of reviews on the topic of mental health inference from social media, all of which have focussed on a range of social media platforms. The key reviews identified were by Wongkoblap et al. in 2017 [57], Guntuku et al. in 2017 [143], Chancellor and De Choudhury in 2020 [261] and Kim et al. [40] in 2021. Despite the potential for digital footprint data to drive

advances in the monitoring and detection of mental health outcomes previous research and reviews in the field have raised significant concerns about the current literature. These concerns center around the validity of ground truth mental health data, methodological clarity and the ethics of the research and its proposed applications.

Firstly, there have been concerns about the quality of data used to train models for mental health inference due to poor construct validity in the generation of data labels [41, 57, 262]. For machine learning to be effective, the labels that a supervised-learning algorithm should be ‘learning’ from (i.e. the ground truth) should represent the same construct that the researcher intends for the model to predict in the future; construct validity refers to this equivalence between the label and the construct intending to be predicted. Systematic reviews by both Wongkoblap et al. [57] and Chancellor and De Choudhury [41] found that using self-reports and affiliations were a very common method for constructing datasets. This means that studies use datasets for training that are constructed and labelled based on self-reports of mental health disorders in tweets (for instance a user tweeting “I have depression”) or based on affiliation with accounts about a specific disorder (such as following an account that tweets about experiences of PTSD) [41, 57]. Research by Ernala and colleagues [262] showed that whilst positive cases identified through self-report and affiliations led to fairly good performance for schizophrenia prediction when validated on the same dataset, they performed poorly when validated against a separate dataset whose diagnoses had been assigned by clinicians. The poor performance of models using assumed ground truth information when tested on clinically validated ground truth suggests that the construct validity of using self-report and affiliations as ground truth is likely to be unsatisfactory for transferring models to a real-world setting. Chancellor and De Choudhury [41] found that only 17 of the 75 studies they included used methods to obtain ground truth that had validity external to the training dataset, such as from participants themselves, news reports about their deaths, or their medical records.

As well as concerns about the data being used to train models in the literature, previous reviews [40] have also identified a lack of transparency and clarity in the methodologies used to produce models. It is common for researchers not to declare important details such as the features included in their models [41], and also uncommon for researchers to include data availability statements [57, 94]. The review by Chancellor and De Choudhury [41] found that only 42% of the 75 papers included reported on all five of what they considered to be minimum reporting criteria, which were:

the number of samples or data points, the number of variables/features, algorithm or regression chosen, at least one validation method and their explicit fit or performance metrics. Overall, the lack of clarity and transparency makes it difficult to assess how research has been conducted, and therefore to compare results between papers and determine the quality of research methods [40].

Aligned to concerns about the sourcing of ground truth data, another issue that has been raised is the characterization of mental health in general, recognising that the mathematical modelling of a psychological construct requires making assumptions about the way it can be captured as data [263]. Chancellor et al. [261] conducted a discourse analysis of the ways that researchers wrote about the people behind the data being used in mental health inference from social media, and found it was often unclear whether the research was considering people or individual tweets. Notwithstanding that there is a significant assumption in using a single tweet as being indicative of depression, this also makes it challenging to understand both the analysis and the results of the proposed models since what is being predicted, tweet or individual outcome, is not reported. Additionally, representing mental health outcomes as a binary also implies certain assumptions about the way the researcher has chosen to model mental health outcomes, which does not allow for a range of symptom severity or allow for the possibility of co-morbidity which is generally high amongst common mental health disorders like anxiety and depression [261, 264].

Lastly, all previous reviews have highlighted ethics as an ongoing concern. The ethical concerns generally refer to the privacy of the individuals whose data is often being used without their knowledge or consent, the sharing of datasets that contain inferred information about those individuals (e.g. a suspected mental health disorder) and the implications of sharing models that could publicly infer information about individuals who had no association with the original study. Outside of the research itself, there are outstanding questions regarding the ethics of using the proposed systems in practice, such as the impact of misclassification on patients [143]. It is worth noting that these ethical concerns are also an ongoing discussion in the critical algorithm literature [102, 265].

5.1.2 The purpose of the present study

The most recent systematic review that covered all papers published on the topic of predicting mental health from social media sites was the review by Chancellor and De Choudhury [41] in 2020. They proposed a list of modelling decisions and outcomes that should be reported in all studies to improve methodological clarity in response to their findings of insufficient methods reporting across 57% of the 75 included studies. This review included literature up to 2018 and considered research on a range of 12 social media sites.

Since this review took place there have been 4 years of new literature to account for. In this time there has been a significant trend in the sciences, especially psychology, towards open science and the improved sharing of data and methodological decisions fuelled by the so-called replication crisis [266, 267]. Ethical concerns have also received greater attention in the past few years, especially in fields using social media data, in the wake of the Cambridge Analytica scandal. The scandal, which broke in 2018, revealed that millions of people's Facebook data were used to analyse and infer their personal characteristics for political advertising. Given these wider cultural changes, the time since previous reviews and also the opportunity for recommendations from previous reviews in 2017 [57, 143] to have been incorporated into new research, I intend to provide an updated review in the area of mental health inference from social media. Specifically this review focusses on the social networking site Twitter, since the updated nature of the review includes the time period where research access to the Facebook and Instagram APIs, two of the most popular social media sites, were removed to provider tighter controls on user data. No such controls were implemented on Twitter.

I conducted a review of the existing literature on prediction of mental health disorders and mental well-being from Twitter by implementing a systematic search to find papers published between 2013 and December 2021. My aims were similar to those posed in previous reviews [41, 57, 143], in that they focus on methodological processes rather than necessarily the results of the research. I set out to evaluate:

- the machine learning methodologies used, such as the ways that pre-processing, feature selection, modelling and validation were conducted,
- the datasets that were being used by each paper, such as how the datasets were collected and

- how mental health outcomes were labelled in these datasets to achieve construct validity,
- the replicability of each paper,
 - whether or not each paper discussed any ethical considerations.

Uniquely this review aimed to include well-being constructs as well as mental health disorders, and also aimed to understand methods to construct datasets as separate to the methods to model mental health, which allowed for analysis of the prevalence of dataset re-use and which datasets are particularly popular.

As is crucial in interdisciplinary work, I first wish to establish some shared understanding with the reader on the use of terminology through this paper [268]. Here, I take ‘prediction’ to be an algorithmic decision to assign an unseen piece of data to a category (e.g. depressed or not depressed), without meaning prediction of the future [269]. I also make distinctions between *mental health*, and a *mental health disorder* with the term *mental health disorder* reserved for references to a medical condition, and is separate from, but related to, general mental health and well-being [270, 271]. *Mental health outcomes* refers to both mental health disorders and specific well-being constructs (for instance, general well-being, happiness, life satisfaction or self-esteem).

5.2 Methods

5.2.1 Search Methodology

On the 7th May 2019 and with two update searches on 26th October 2019 and 6th December 2021, a search of six electronic databases was conducted, (Web of Science, Scopus, PubMed, and Ovid MEDLINE, PsychInfo and PsychArticles), as well as a Google Scholar Search. The search was for peer-reviewed articles or papers that contained terms related to mental health disorders and well-being, machine learning and Twitter in their title or abstract (see Supplementary Material in Section C.1 for the full list of search terms). Each search was refined for the requirements of the database. Results were required to have been published in 2006 or later, to avoid unrelated publications from before Twitter was created.

Several key review papers in the field of mental health prediction from social media were identified prior to the systematic review [41, 57, 143, 272] and 16 other review papers in related fields were

identified through the systematic review process [40, 273–287]. Secondary citations from all of these reviews were included in the screening phase if they had not already been identified through the database search. Lastly, a small number of papers were identified through recommendations from colleagues and referencing software.

5.2.2 Screening Methodology

Rayyan software [288] was used to identify and remove duplicates from the results and was used to review the titles and abstracts to screen papers for a full-text review. At this stage, papers which appeared to be irrelevant, for instance relating to personal social networks as opposed to online social networks, or having no relevance to mental health, were removed.

A full text review was then conducted on the remaining 651 papers. At this stage inclusion criteria were as follows:

- 1) The paper considered data from Twitter in order to build the algorithm. Despite being similar to Twitter, Weibo was excluded due to some differences in data types available and the nature of use.
- 2) The paper was not considering a specific group of people, such as veterans or new mothers.
- 3) The paper considered a mental health disorder or specific well-being construct, rather than a less specific concept such as stress. This was based on the paper's title and what it stated it was predicting.
- 4) The paper was training a model for the purposes of inference, rather than solely analysis of features.

This full text review left 165 papers which met the criteria for inclusion in the analysis.

5.2.3 Data Collection

Literature searching, screening and analysis were completed by myself. For each included study I recorded the details of the mental health outcome studied, machine learning algorithms used, features and model input, useful validation and evaluation strategies used to assess models, and the reported results. For each primary dataset identified, meaning those where data was collected by the research team and not reused from an existing study, I also recorded the method of data

collection, the key characteristics of the dataset, how data was annotated and any quality control processes used. A complete record of identified and reviewed papers is included in the online Supplementary Material (doi: [10.17605/OSF.IO/HYD9G](https://doi.org/10.17605/OSF.IO/HYD9G)).

5.3 Results

Figure 5.1 illustrates the number of papers included at each stage of the screening process.

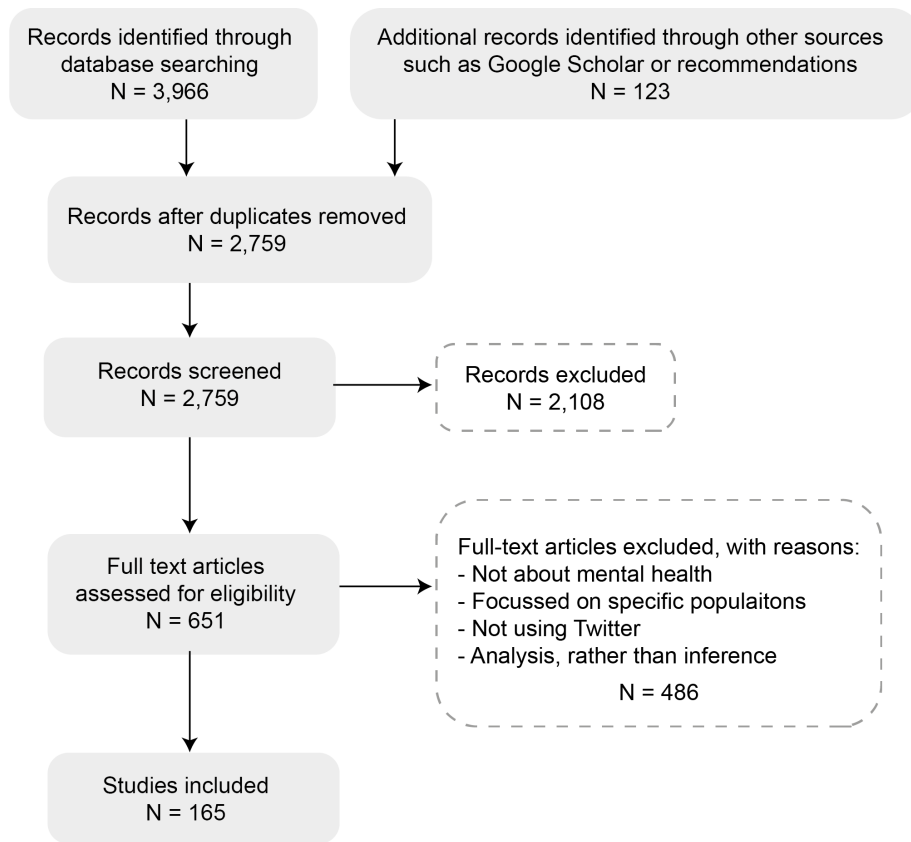


Figure 5.1: PRISMA flow diagram of inclusion and exclusion figures for the literature search

Table 5.1 gives how many of the papers included were published in each year, and shows that 45% of papers identified on this topic were published in 2019 onwards, which is after the date that previous reviews have included.

Table 5.1: The number of papers included in the review that were published each year.

Year	N Papers
2013	3
2014	6
2015	11
2016	7
2017	16
2018	13
2019	34
2020	36
2021	39

5.3.1 Mental health outcomes predicted

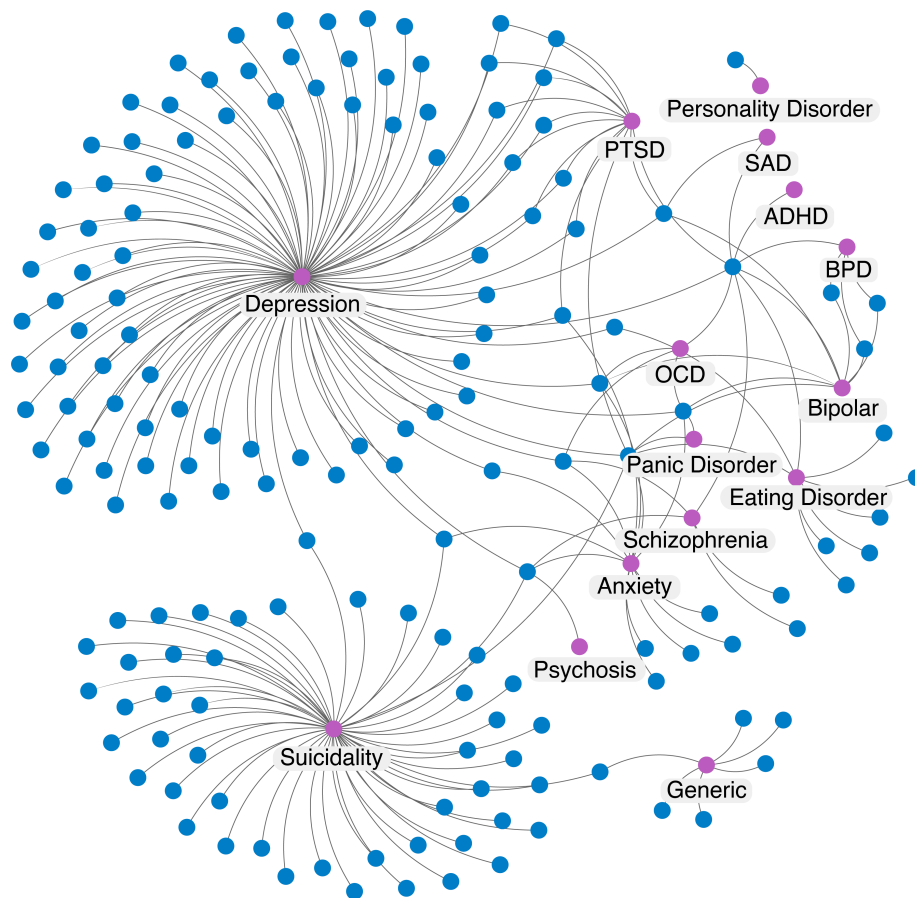


Figure 5.2: Network digram showing which mental health disorder (pink) each paper (blue) attempted to infer.¹Depression and suicidality have been the most popular, with most papers attempting to predict a single outcome.

Figure 5.2 outlines the network of mental health disorders the included papers covered. It illustrates that depression was the most common target, and was predicted in 93/165 papers (56%), followed by suicidality (31%), PTSD (8%) and anxiety (8%). It was most common for studies to approach this problem as a single-class prediction, though 26 of the 165 papers considered more than one mental health disorder.

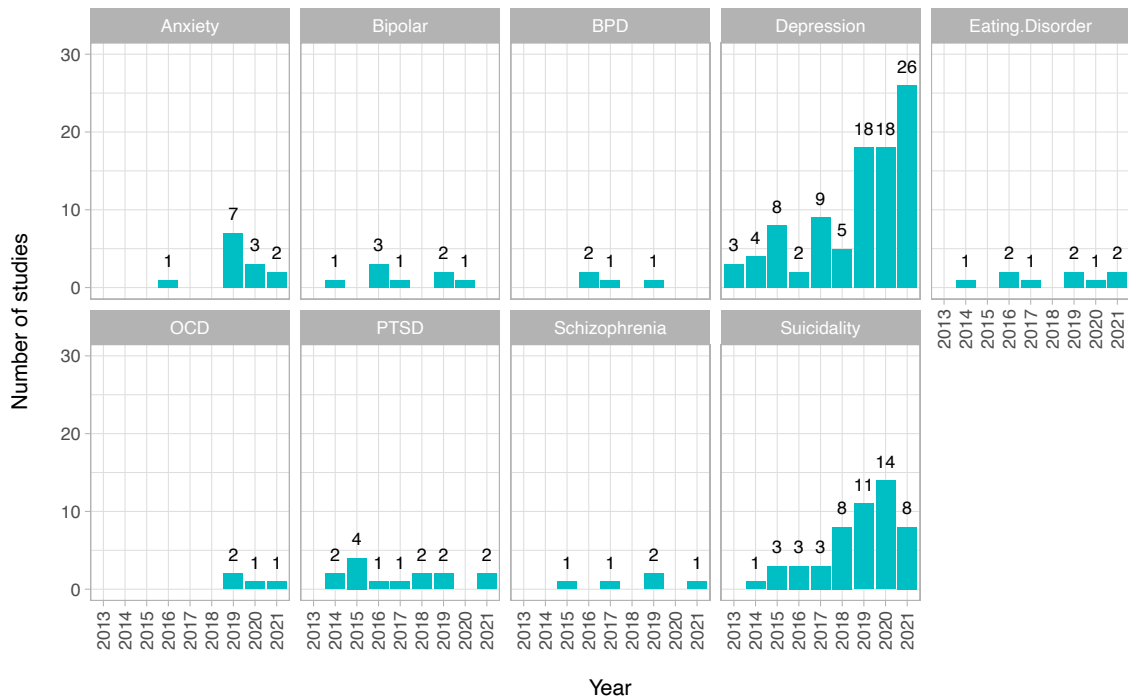


Figure 5.3: The number of studies considering each mental health disorder published each year. To be presented in this figure the mental health disorder needed to be included in more than two studies.²

Figure 5.3 shows there has been an increase since 2019 in the number of studies being published on this topic, but they are dominated by studies about depression and, to some extent, suicidality. Analysis of other disorders has remained fairly static over time. Whilst there is a tendency overall to focus on mental health disorders, there was one study that included prediction of happiness and self-esteem [289].

¹ Acronyms: Attention Deficit Hyperactivity Disorder (ADHD), Borderline Personality Disorder (BPD), Obsessive Compulsive Disorder (OCD), Post Traumatic Stress Disorder (PTSD), Seasonal Affective Disorder (SAD)

² Acronyms: Borderline Personality Disorder (BPD), Obsessive Compulsive Disorder (OCD), Post Traumatic Stress Disorder (PTSD)

5.3.2 Datasets

One of the aims of this review was to analyse the unique datasets that were being used for prediction of mental health outcomes across the included studies. Overall, I identified 128 unique datasets as coming from 155 papers included in this review which I will refer to as *primary* datasets; 10 papers did not provide a description of the dataset to understand the dataset being used. Of these 128 datasets, 8 were not described in enough detail for it to be possible to generate any detail for analysis. This was usually due to links to dataset sources being invalid, or links to online datasets which were not actually described in the text. This left 120 unique datasets that contained enough detail to be analysed.

All the studies identified for this review used an annotated dataset to train prediction models. *Annotation* refers to the process by which each observation or data point that will be used to train the model is given an outcome that the model is trained to predict. In this case, the annotations are expected to be a mental health outcome.

Different studies take different approaches to the process of collecting and annotating their datasets, and in this section I give an overview of these processes for the 120 datasets that were adequately described. Then, since some studies used primary datasets that were developed and shared by others, I also give a brief description of which datasets were those that were most commonly re-used. The full table of results with the data extracted from each paper and dataset are available in the online Supplementary Material (doi: [10.17605/OSF.IO/HYD9G](https://doi.org/10.17605/OSF.IO/HYD9G)).

Descriptions of data collection

To understand approaches to data collection I recorded whether the description of the dataset specified the number of tweets included in the final dataset, how many individual users were in the dataset, the time period over which Twitter data were collected, the API or tool used to access the Twitter data and the search query or strategy used to collect the data. These were chosen because they represent basic descriptive information that is important for interpreting the results of the studies, and also represent reasons that some studies may find differing results. For example, using data from different time periods, different APIs and different search queries to access data would all result in different samples, and these may then yield different predictions when addressing the

same core question.

Out of the descriptions of the 120 datasets included I found that 68/120 (57%) datasets had the number of users in the dataset, 96/120 (80%) how many tweets were in the dataset, 66/120 (55%) the time period over which the data were collected from Twitter, 84/120 (70%) which API or tool was used to access the Twitter data and 108/120 (90%) the search strategy they used to query the API. The smallest described dataset was Coello-Guilarte et al. [290] with 200 annotated tweets, the largest being Shen et al. [291] with over 300 million tweets from users they determined to be depressed, and 10 billion control tweets.

Annotating mental health outcomes

Next, I recorded information about how the data were annotated with mental health labels. This included the method used to attribute labels to the tweets or users, and whether there was any secondary quality control conducted by human annotators if an automated method was used. Additionally I evaluated the range of methods that were used to develop control samples of tweets or users who do not display the mental health outcome that is being predicted.

I originally intended to also record whether annotations were being made at the tweet or user level, but unfortunately it was not common for studies to specify which of these approaches they were taking, and so not possible to summarise the frequencies seen in the papers reviewed.

Table 5.2: Overview of the different methods used to annotate datasets with ground truth labels.

Ground truth type	Description	Count	Quality Control	Example
Validated				
Self report	Completion of a standardised measure, or disclosure of affected time periods by the individual	12		User scored >30 in Centre for Epidemiological Studies Depression Scale (CES-D)
Secondary report	News reports of death by suicide, or data donation by family following death	4		CES-D score used as a continuous variable. Name reported in the media was searched on Twitter for a user account.
Data driven				
Affiliation	The account either follows or interacts with a system or other accounts known to be associated with the mental health disorder being considered.	2		Accounts that had retweeted tweets from a list of accounts about depression were annotated as being depressed
Keywords	A certain number/combination of keywords used to search the Twitter API, believed to indicate the presence of the mental health disorder.	52	Expert 21 Non-expert 18 None 13	User used the string “depress” more than five times in 2 weeks, and their timeline was reviewed by a clinical psychologist to confirm the assessment was reasonable (expert). User used ‘depression’ at least once in a tweet (none).
Self disclosure	A phrase such as ‘I have been diagnosed with X’ is used to search the Twitter API, and used to indicate the presence of the mental health disorder.	29	Expert 2 Non-expert 14 None 13	String ‘I have been diagnosed with depression’ was used without checking the context (none). String ‘I have been diagnosed with depression’ was used following verification by a clinical psychologist (expert) or a computer science researcher (non-expert).
Sentiment label	Some threshold is decided based on a sentiment polarity score that maps it to a mental health outcome.	2		Sentiment score below -1 meant the user was annotated as depressed.
Other				
Random sample	A random sample of tweets are taken from the Streaming API or based on some other criteria, such as a particular language being used, and screened for inclusion	5		Tweets in a particular language were accessed from the streaming API and annotated as suicidal if the researcher thought it indicated suicidality.
Unknown	Not enough information provided to understand the method for generating ground truth labels.	14		

Note:

In the row headings, ‘Validated’ refers to data annotations that have not been assumed from the data collected, and have been validated by either the user themselves or an external source. ‘Data-driven’ refers to annotations that are derived from the data collected from social media. In terms of Quality Control, ‘Expert’ annotation was when annotation was done by those who were called experts in the paper, or who were reported as having some educational or practical background in mental health. ‘Non-expert’ annotation was done by anyone not in the ‘Expert’ category, for instance undergraduate students or computer science researchers.

As Table 5.2 illustrates the datasets were annotated in many different ways, but only 16 of the datasets overall were validated by offline ground truth. That is to say that the label was not assumed from the data collected. Even within those studies that did use validated scales for ground truth, they could define the threshold score for presence of a disorder from the same scale differently. For instance, a CES-D score greater than 30 or a score greater than 22 were both used as cut-off scores for the classification of depression in different studies. Due to the variety of methods presented comparisons between studies could be comparing datasets that have very different definitions of the same mental health outcome.

Some studies attempted to increase the accuracy of keyword or self-disclosure based annotation by introducing human annotators to the process. However, a handful of studies using this method reported that annotators found it difficult to decide on the category that tweets should be placed in, especially when they were seen without the context of other tweets from the same user [292, 293]. To overcome this some annotated datasets used more than one annotator in order to assess agreement between annotators, or introduced a third annotator to provide a deciding opinion on conflicting assessments (for example [294, 295]). As might be expected, there was generally a relationship between the size of a dataset and the level of quality control; highly curated data with labels produced by experts and multiple coders tended to be smaller in volume, and those using largely automated methods were able to produce vast datasets with little human input on the target classification labels.

The vast majority of studies defined mental health as a binary or categorical outcome, as opposed to using a continuous scale. This is important since the outcome being predicted indicates a different research question, and ultimately a different purpose. For instance, classification of tweets that are ‘risky’ or ‘not risky’ in terms of suicidal expression, versus a longitudinal view of change in depressive symptoms. This was largely influenced by the approaches to data labelling, where the presence of keywords or self-disclosure do not allow for a measurement of symptom intensity and instead necessitate a binary or categorical approach.

Since most datasets took a categorical approach to mental health, most also collected control users, which are users who were judged not to have the mental health disorder being identified. Approaches to developing a control sample included taking a random sample of tweets from the Streaming API on a particular day, searching for a word or phrase (like “the” or “today is my

birthday”) in the Search API and using the results as controls, or simply using all those users who were not labelled as positive from the original keyword/phrase search. In some instances studies conducted checks to ensure there were not overlaps between the positive and negative samples, but this was not always stated as being the case. In terms of the balance of cases to controls in the datasets, some studies developed datasets in order to intentionally balance cases and controls [296–306], whereas others searched using their chosen criteria and took the “naturally occurring” number of cases from their dataset [251, 291, 293, 302, 307–309].

Dataset re-use

Of the primary datasets identified there were two that were re-used more often than others. The dataset on depression and PTSD which was produced for the Computational Linguistics and Clinical Psychology (CLPsych) Workshop 2015 [302] was used a total of 10 times, and the dataset produced by Shen et al. [291] for depression prediction in 2017 was used the most often at 14 times. The other most frequently reused datasets were those produced by Burnap et al. [310] in 2017 for suicidality (4 uses), Jamil et al. [311] in 2017 for depression (3 uses) and Vioules et al. [312] in 2018 for suicidality (3 uses). Another dataset used in four studies, despite not being produced for mental health prediction, was the ‘sentiment140’ dataset. This is a Kaggle³ competition dataset where tweets are labelled with their sentiment polarity.

Finally, the remaining datasets were created by authors for their own use, and occasionally re-used by the same authors over two studies. In most cases datasets were created specifically for the task the study was focussed on, and included datasets of tweets in other languages like Spanish [290], Bengali [313], Japanese [289] and Arabic [314].

5.3.3 Modelling workflows

After identifying the dataset to use for training, there are typically a series of stages to go through in order to develop and assess a predictive model. First the researcher must prepare the dataset for use (known as pre-processing), select the features that will be used in the model (known as feature selection), choose and apply an algorithm to create a model from, and then finally validate the model to assess how well it performs on unseen data.

³Kaggle is a website where individuals and teams can participate online in data science challenges.

Not all studies reported their methodologies along each of these four key stages. In summary I found that 121/165 (73%) studies described at least some of their pre-processing steps, 138/165 (84%) described the features or feature selection process, 160/165 (97%) described the algorithm/s used and 135/165 (82%) gave some description of their model validation process. Figure 5.4 illustrates that there has not been much change in reporting standards since 2020, and in fact the areas of algorithm choice and feature selection have been reported in fewer papers more recently. In the following sub-sections I report on those studies which did include this information by summarising the methodologies that were used across the literature in each stage.



Figure 5.4: The proportion of studies that reported each of the stages of modelling that I considered, split by those published before 2020 (N=90) and those published in 2020 or later (N=75).

Pre-processing

In natural language processing (NLP), the computational interpretation of written text, it is typical to pre-process or *clean* textual data to prepare it for feature generation and selection. These steps tend to focus on making the text less noisy by removing data that is unlikely to be useful in the predictive task such as stripping non-alphanumeric characters, removing stop words (common or filler words), lemmatising the text (transforming words to their root), or tokenisation (splitting sentences or documents into separate tokens delimited by spaces).

However, for data taken from social media some pre-processing stages may be adapted to reflect the inherent meaning that, for instance, non-alphanumeric characters and stop words contribute to the text. These characteristics of text may also be expected by some sentiment analysis algorithms

such as VADER [234]. Another consideration around internet language is the inclusion of emoji in text. Emoji can have meaning in natural language [315], and so their inclusion is likely to be relevant in NLP-type tasks.

A minority of the studies who described their pre-processing stages regarded Twitter’s native language of interaction such as hashtags and @-mentions as parts of natural language and retained this information in the tokenisation stage by replacing @-mentions with an @ symbol, or URLs with the word “URL” (e.g. [298, 316, 317]). Other studies chose to tokenise the text in a more traditional manner, by removing all non-alphanumeric information (e.g. [311, 318–324]). Papers which included emoji as tokens usually did so by replacing the emoji by the word “emoji”; (e.g. [298, 302, 325]) or by a unique code for each emoji (e.g. [291, 316, 323]). Others removed emoji all together from the text (e.g [328]). Variation in these pre-processing strategies means that there are differences in the type of information taken forward to the feature selection and modelling stages.

Some pre-processing decisions that may have impacted the effectiveness of the subsequent model training processes were rarely described. For instance, several studies have replicated a finding that personal pronouns are a useful feature in the prediction of depression (e.g. [319, 329]). However, personal and other pronouns may be included in stop word dictionaries (for instance the popular NLTK [330] stop word list), and so automatically removed from the training data before any feature selection or model fitting has taken place. Additionally, many of the studies used keyword or keyphrase search terms to label ‘positive’ cases for mental health disorders, but it was not made clear whether the terms used to label the data were removed from the training dataset. For example, if the term “depress” used five times identified a user as being depressed, and this term was present five or more times in the training data of every person who had been labelled as depressed at the modelling stage, then the model may learn that “depress” is reliable signal for depression.

Features

To apply a machine learning algorithm to a dataset, a series of features (also known as variables) have to be constructed. Most papers used some combination of each of the feature types, as described in Table 5.3. Overall, textual interpretation and textual features were the most popular. 72 papers used at least one form of textual interpretation, such as word embeddings, and 125 used at

Table 5.3: Overview of feature categories, the number of studies that used at least one feature from each category and a description of the types of features it contains.

Feature type	Number of studies	Description
Text Interpretation	72	Features interpreting the meaning of the text, usually through sentiment dictionaries.
Demographics	15	Known or algorithmically inferred demographic information.
Connectivity	35	Features relating to the user's social network such as the number of followers or @-mentions.
Sharing (When)	25	Features relating to time, such as time between tweets, tweet frequency or times of day.
Sharing (What)	26	Features relating to the type of content being shared, such as URL links or retweets.
Textual features and structure	125	Structural features of the text such as TF-IDF scores, bag of words, word embeddings and language models.
Keywords	39	Counts or distributions of keyword lists like medication names.
Parts of Speech	33	Labelling parts of speech or grammatical features.
Images	12	Use of image data like profile pictures, or shared images.

Note:

TF-IDF refers to term frequency–inverse document frequency, a statistic that reflects word importance across a group of documents.

Table 5.4: The number of studies using each type of algorithm for at least one model.

Algorithm	Number of studies
SVM	83
Tree-based	67
Naïve Bayes	61
Regression-based	52
Deep Learning	40
Other	54
Unknown	2

least one type of textual structure which tended to be either n-grams or term frequencies⁴. It is worth noting that datasets built in languages other than English were often required to derive their own pre-processing and feature selection tools such as sentiment dictionaries or stop word lists due to there not being existing software and tools readily available in their language.

Algorithms

Whilst different studies chose different approaches to modelling the data, the majority used well-recognised algorithms such as Support Vector Machines (SVM), Naïve Bayes, or tree-based algorithms. Table 5.4 illustrates that SVM appears to be the most popular algorithm. However it was not always the primary model, and often provided a baseline measure against more complex approaches such as deep learning, or as part of an ensemble learning approach. Within regression logistic regression tended to be used, which reflects the categorical nature of most of the datasets. Deep learning approaches, for instance convolutional neural networks, have been relatively popular over time, but certainly do not form the majority. Included in the “Other” category are bespoke algorithms written for this problem [306, 331] as well as less popular out-of-the-box options. Examples of these are a hidden markov models [63], a Martingale framework [312], or complex decision lists [291, 332].

Whilst all but one study did describe the machine learning algorithm they used to produce their final model, very few of the studies went into any detail on their hyper-parameter tuning processes

⁴Word embeddings are numeric representations of textual data where words that are ‘closer’ together in their numeric representation are more similar in their meaning. N-grams are groups of n words that appear sequentially in a longer piece of text. E.g. the tri-grams (n=3) of “This is a sentence” would be (“This”, “is”, “a”) and (“is”, “a”, “sentence”).

which refers to the adjustments made to the values that control the model's learning process. It was also not common for studies to justify their choice of algorithm, though the choices were not inappropriate.

Validation

Understanding the effectiveness of a machine learning model allows us to evaluate how well the algorithm might generalise to unseen data. Most often ten- or five-fold cross-validation was used, as well as the area under the Receiver Operator Characteristic curve.

Two issues relevant to model validation were rarely discussed or acknowledged in the papers. Given that some of the datasets were designed to include a small number of controls to high numbers of cases, some standard metrics, particularly accuracy, are likely to over-represent how effective the algorithm is [333]. Secondly, studies rarely clarified how they stratified their data for training, testing and validation. This has implications for assessing the potential for data leakage to create bias in the model's effectiveness and has been shown to be problematic in other applications of machine learning to digital epidemiology [334], as well as specifically creating bias in cross-validation assessment of machine learning for mental health [335].

5.3.4 Ethics

The consideration of ethical approval was assessed for a subset of 100 of the included papers, since ethical approval was only included in the rubric for reporting on studies in this review from December 2021. However, this still represents all studies published in 2020 and later, from which point I had anticipated that ethical considerations should be more prevalent given the cultural impacts discussed in Section 5.1.2, as well as previous reviews suggesting this was an area of concern.

Overall, I found that 85/100 did not discuss any ethical issues as part of their studies. 11/100 studies did discuss ethics thoroughly, and/or were granted ethical approval for their studies. 4/100 studies made reference to ethics as not being applicable to the study.

Whilst some studies simply did not include consideration of ethics, there were a handful that directly contravened ethical guidance published by both the Association of Internet Researchers [95] and the British Psychological Society [94] regarding the use of internet data for research. This

was generally by publishing tweets verbatim, sometimes along with the mental health annotation, and/or by publishing usernames in the paper. Additionally some studies developed web-apps that allowed for a user timeline or tweet to be input and a prediction displayed about whether that user was experiencing the mental health disorder under consideration, though it was not always clear whether these web-apps were still operational.

5.3.5 Replicability

Finally, I assessed the replicability of each paper in terms of the quality of the detail provided. For 44/165 (27%) studies I assessed that there was enough detail for the study to be replicated. For 53/165 (32%) studies it is possible they could be partly replicated but some assumptions about methodological processes (typically the pre-processing stages) would need to be made. However for 68/165 (41%) studies there was not enough detail provided to attempt replication of the study due to key information being missing such as the data annotation process, the algorithm used or the feature construction. In some cases it was clear that publishing formats and word limits had left limited room for description, but authors did illustrate use of external repositories on GitHub and the Open Science Framework to host more detailed methodologies or code which provides a straightforward solution to this issue.

Only 6/165 studies either provided the scripts used to analyse the data or offered to make them available on request. Alternatively, a handful of authors provided pseudo-code for all stages of the model building process as part of the article. Overall, this was an unexpectedly low rate of code sharing given the recent emphasis in both computer science and psychology on greater methodological transparency. Whilst some may not share code for ethical reasons there are alternatives such as offering to make it available on reasonable request, which were not widely used.

5.4 Discussion

5.4.1 Principal Results

This review set out to understand the current scope, direction and trends in the prediction of mental health outcomes from Twitter data. 165 papers published between 2013 and 2021 were included in the review. I saw that the number of papers published in this area has increased year on year

since 2013, and that 45% of the included studies were published in just the two year span of 2020 to 2021. I sought to assess the quality of the published research from both a machine learning and mental health perspective, and to make recommendations that can begin to enable the creation of meaningful outputs that support aims of mental health care provision and support. In the following sections I summarise the principal results and contextualise them against previous work along the themes of methodological clarity, the availability of ground truth characteristics and lastly looking towards developments that would support practical applications of these algorithms in the future.

These discussions lead to a series of recommendations for studies that aim to predict mental health outcomes from social media.

Methodological clarity

Every paper in this review used algorithmic methods for making predictions, with a wide range of novel and exciting possibilities for future development. However, the description of machine learning workflows given was often poor and a lack of clarity was a consistent theme in the results. In 11% (18/165) of studies there was not an adequate description of the datasets to understand the data being used, and in 27% of studies there was no description of model pre-processing. The proportion of studies reporting these details has not increased over time.

As well as missing out on the author's reasoning, poor reporting on modelling methods also reduced replicability, with only 27% of studies assessed as being replicable with the information provided. Despite recommendations to improve the description of methodologies in place since 2017 [57], and the increasing recognition of open science practices [336], I was surprised to find that only 6/165 papers made their code available either open source or on request, where only providing code on request would be a reasonable means of mitigating ethical concerns.

The lack of clarity often started with a poor description of the purpose of the prediction task being attempted, which has an impact on all subsequent modelling decisions and the assessment of their suitability [261, 337]. It also prevents the comparison of results between papers, due to it often being impossible to tell if the same or a different predictive task is being compared.

Availability of ground truth characteristics

I found that the processes for determining what constituted a mental health disorder, and hence the labelling of training data, relied on circular reasoning in 104 out of the 120 primary datasets (87%). This reasoning assumes that those who self-report mental health disorders online, or who use certain combinations of keywords, are truly experiencing the specified outcome. It also means that groups of users who were collected for ‘control’ groups were unlikely to be true controls, given the relatively high prevalence of mental health disorders in the general population [41, 261]. Similarly, keyword-based approaches were also used to derive ground truth for mental health outcome annotations in 43% of the datasets reviewed, with keywords being highly likely to be based on the language of a particular geographical area or age group, and also prone to misspellings when focussing on clinically related keywords [338]. Attempts to work with clinicians to develop a list of keywords for depression detection have also found low levels of agreement between clinicians [339], which suggests that keyword based detection may not be a robust means of detecting genuinely depressed users. This lack of reliable, verified ground truth data about mental health outcomes is a fundamental threat to the quality of models for mental health inference. It also aligns with concerns being raised in other fields that large online datasets can not replace the need for high quality data [248, 338, 340].

Without validated ground truth in the majority of studies, there was not information available to characterise the dataset by key demographics such as age, gender or cultural background. We know that expressions of mental health disorders are cultural, and variable across demographic groups [231, 341], and that those using social media do not represent the general population [146, 147] (also see Chapter 2). Lacking this information means that it is not possible to assess the impact of demographic features on model performance, and so bias may be going unnoticed. Research by Aguirre et al. in 2021 [342] reinforces this, after the finding that the CLPsych dataset (used in ten papers in this review) was not representative of the population demographics of depressed people, and that a classifier produced using this dataset performed most poorly for people of colour.

When models are created with datasets whose ground truth cannot be verified, the importance of validating models on alternative datasets increases [262]. Shared datasets such as the CLPsych Task 2015 [302] and Shen et al. [291] have contributed to numerous studies by providing a dataset available to researchers [343] as well as providing data to develop novel approaches with (though

as discussed by Aguirre et al. [342] these datasets are unlikely to be population representative). Sharing high-quality ground truth datasets would be a beneficial next step for future developments [262]. Due to the sensitivity of these data we would need to think carefully about how data sharing could be managed ethically [98]. Further possibilities lie in the use of data safe havens for controlling sensitive data access, as has been used by workshop tasks such as CLPsych in the past few years, and in the use of synthetic data [291] which is a developing opportunity that allows a dataset with statistical properties similar to the original data without releasing the sensitive data itself. The work of collating available datasets has been started by Harrigian et al. [344] through the development of an open source list of datasets for predicting mental health from social media, many of which are only available on request to comply with ethical guidelines. However, as described in Section 5.3.2, data sharing is impeded by researchers sometimes not even describing the dataset they are using, or providing broken/out of date links to data repositories.

Towards practical applications

This review of the mental health outcomes covered by the 165 papers included showed that there is a significant focus on depression and suicidality, but that anxiety receives much less attention, along with serious mental health disorders like PTSD, schizophrenia and psychosis. Whilst well-being was included in the review keywords, only one study that considered well-being outcomes was identified, which predicted happiness and self-esteem measured using validated scales [289]. More specific keywords relating to different types of well-being may have yielded more results in this area. Though the majority of the focus of the datasets reviewed was on dichotomous outcomes, a future alternative is a greater focus on symptoms of disorders [345]. This has been suggested as a solution to detecting commonly co-morbid illnesses which have many connected symptoms [346], an issue that has arisen in multi-class prediction of mental health outcomes [347]. The majority of the papers reviewed have effectively attempted to classify someone as having a mental health disorder or not, but perhaps social media may have more to offer in the tracking of online behaviours that are strong proxies for specific symptoms of mental health disorders. This is perhaps best illustrated by suicidality, which is a complex concept that has been effectively modelled using machine learning [348].

Another area of development which would benefit from further investigation is using the time-

based features of Twitter data. Considering that one of the main benefits of using social media data for monitoring is the high-resolution time-series information it provides, it was surprising that only 15% of studies used any time-based features in their models, and only one study used ground truth data that was measured at more than one time-point [312]. By considering Twitter data as a time-series we could approach tasks like identifying optimal points for intervention, using methods such as change-point detection, or simply monitoring well-being with time. Having multiple instances of ground truth data for the same individual would also allow us to assess how model performance changes over time, since model drift is a particularly important concern in online settings where language and platform features continuously adapt, potentially resulting in the degradation of a trained model over time [349]. Clinicians have so far expressed interest in using social media to measure overall symptom changes between time points, rather than as a diagnostic tool [350], and so this is an area of work that requires more attention if social media data is to provide a practical use in the future. Since the literature searches conducted for this review took place longitudinal and time-series work has been published more frequently in 2022 [67, 351].

Throughout the literature there appears to be a consensus that more meaningful and deliberate engagement with medical professionals and patients is needed to establish a direction for future research, and explorations into Public Patient Involvement (PPI) and co-production may be effective ways of achieving these aims. Crucially, we do not yet have a broad evidence base about how patients might want to use this technology or what they would not want it to be used for as part of their care [205]. It is clear that for the work so far to develop into a technology with real-world utility further consultation on useful clinical applications and the ethical dilemmas presented by them will be needed [204, 352, 353], but this is still work to be done.

5.4.2 Recommendations

On the basis of this review I have two sets of recommendations. The first is for researchers in this field, building on the recommendations made by Chancellor and De Choudhury [41], which aim to increase the quality, replicability and transparency of mental health inference on social media:

- 1) Explicitly state the prediction task being attempted. This should include whether the outcome predictions are at the user level or the tweet level, and what the intended use of the

resulting model is.

- 2) Specifically state the mental health outcome the model will be attempting to classify and how this outcome has been defined for the purpose of labelling the training data.
- 3) Explicitly state assumptions made about the mental health outcome as part of the modelling approach taken. For instance, what type of variable the outcome has been modelled as (continuous, binary etc), or what time frame it is assumed it will be detectable within.
- 4) When creating new datasets, ensure that they are thoroughly described. I particularly recommend the use of *Data Sheets for Datasets* [354] for thorough dataset reporting, which can be included as supplementary material hosted by an online repository that provides a permanent digital object identifier (DOI) such as the Open Science Framework or a pre-print server.
- 5) Explain pre-processing steps in enough detail that they can be thoroughly understood and replicated. Particular attention should be paid to whether stop word lists used and the train/test/split stratification to ensure they are appropriate for the prediction task being conducted.
- 6) Where possible, conduct error analysis in order to explain how and why data have been misclassified.
- 7) Include a Code and Data Availability Statement, and ensure that any crucial links to materials use a DOI.
- 8) Include an Ethics Statement that describes whether or not the study has received ethical approval, and the ethical considerations that researchers should be aware of when reading, replicating or applying the research. The *Ethics Sheet* for this type of research developed by Mohammad [355] is particularly recommended.

My second set of recommendations are broader, community level aims that focus on developing ways of working that will enable these new technologies to achieve positive outcomes:

- 1) Work towards an understanding of the needs of the public and patient populations who will be the subjects of the models being developed, and ensure that research is advancing in line with their needs.
- 2) Find and agree a means by which high-quality ground truth data and trained models can be shared securely and ethically between research groups, with the purpose of improving the validation of models for predicting mental health on social media.

-
- 3) Maximise the benefits of what social media can add to our understanding of mental health, as opposed to replacing the role of mental health professionals. In particular, the time-series nature of social media has been under-explored so far.

5.4.3 Limitations

Whilst best efforts were made to include all relevant papers in this review, there is always a possibility that relevant studies were missed in the systematic search process. Similarly, the search was conducted using English language search terms, and non-English studies were not reviewed. Previous research from Kim et al. [40] showed that several studies in this area have been published by teams in China, Spain and India, which may not have been included.

This review does not go into detail about the outcomes of the studies identified, such as their results, which models appeared to be most successful or which features have been especially relevant throughout the varying approaches. These are investigations that could yield useful directions for improving future models and refining the process of feature selection.

5.5 Conclusion

In this review I have shown that there is a wealth of research being conducted and published on predicting mental health outcomes from Twitter, but at present the quality of study datasets and dataset descriptions is frequently poor, and the large majority of studies do not provide enough information about their analyses to understand or attempt to replicate them. For this technology to move towards being used for the benefit of the populations they are intended for, the research community needs more sources of high-quality ground truth data with clinically valid labels, that can be shared ethically for benchmarking and model training. A strong partnership between researchers, clinicians, patients and the general public is also needed to ensure that the prediction tasks being developed are those that will be both ethically viable, as well as clinically useful. Given the sensitivity of this research area, researchers have an ethical responsibility to ensure the transparency of machine learning methods, in terms of the data used, the algorithms employed and precise evaluation and reporting of a model's effectiveness.

If we can achieve our aim of using digital data to effectively model mental health then there is

the potential for huge advancements in our understanding, monitoring and management of mental health conditions in the future.

Chapter 6

Modelling mental health using linked Twitter data

Abstract

Background The digital data we create by interacting with online platforms is a novel source of temporal information about ourselves. Sentiment analysis of our social media data has been of particular interest for understanding fluctuations in our mental health and well-being over time. However, it is challenging to obtain data linked to longitudinal information about participants' mental health and so most studies have been conducted at single time points. In this chapter Twitter data from ALSPAC was used to assess the effectiveness of sentiment, and patterns of life, to model of depression, anxiety and general well-being.

Method Ground truth data using the MFQ, GAD-7 and WEMWBS was collected at two time points in April to May 2020 and May to July 2020 for G1 and G0 participants. Multiple regression models were used to model pattern of life and sentiment variables from LIWC, LabMT and VADER against each mental health outcome at the first survey time point. Models were tested using different lengths of training data and different decay functions for older tweet data. Finally, the most successful models were validated at the second survey time point.

Results Two weeks of Twitter data was generally sufficient to predict all three mental health out-

comes. Models trained at the first time point perform relatively well at the second, but error is slightly biased by the generation of participants and, for anxiety and depression, their sex. Whilst models accounted for a modest amount of variance in the outcomes (between 10 and 13%) we also see that prediction intervals are wide.

Conclusion Sentiment and pattern of life features are likely to be most suitable for broad population-level trends rather than individual-level inference. Using models specific to attributes which mental health is known to vary by, like gender, may improve performance of population trend modelling using Twitter. Additionally, whether individual aggregation is necessary for accurate population inference is a question that should be explored further.

Aims

To use the linked Twitter data for modelling mental health outcomes in ALSPAC, and in doing so see if this novel data is practically useful for this purpose, and if it allows us explore new questions about the relationship between mental health and Twitter.

6.1 Introduction

Predicting mental health from our digital data is an area of research that has seen increased interest year-on-year since 2013 [320]. The potential benefits of achieving new digital phenotypes for mental health could be huge, and a means of supporting progress in the provision of, and access to, mental health care. The future applications of these novel technologies have been considered at both the individual and population levels. For instance, they could be a means of providing already overstretched services with a way of ensuring patients receive appropriate support if flagged between long check-up times. Detecting individual poor mental health could also be used to direct early interventions which may prevent more complex issues arising at a later stage at a greater cost for all involved through ‘now-casting’ technologies [356, 357]. Alternatively, monitoring population mental health could be used at a strategic level to put in place adequate services to meet anticipated demand [358]. Here *populations* may refer to specific sub-populations like student groups [19, 359] and emergency workers [360], or to geographic populations [361]. Applications of this type have been of particular interest since the onset of the COVID-19 pandemic, where using internet data to provide high frequency information became very popular, particularly in the form of data visualisation dashboards [362]. Whilst many of these population-level overviews were related to COVID-19 symptomatology and case prediction [363], data relating to mood and mental health were also generated, with the aim to provide timely updates on community psychological well-being and resilience [364–367]. Twitter, as a public and easily available data source for digital footprint data, is often a popular source of data for these types of applications. However, as seen in Chapter 5, the majority of studies that have previously used Twitter data have not had access to ground truth data about the mental health and well-being outcomes that they were attempting to predict, which presents a challenge to the robustness of their results.

In this Chapter I will explore the way that features derived from Twitter data, particularly sentiment-related features, are associated with the mental health outcomes of depression, anxiety and general well-being when working with data from a known and well-characterised population sample. This will include how effective models derived using these features are under a variety of different temporal conditions, such as altering the distance into the past they consider data and testing how well they predict forward into the future. To begin with, I will give an overview of the use of sentiment in mental health inference in general, as well existing evidence on the role of

temporality in mental health inference and understanding.

Given the focus on individual monitoring in the review of existing research in Chapter 5 I will primarily focus here on covering the literature of population-level approaches to mental health monitoring.

6.1.1 Sentiment in population mental health inference

In the literature on population-level mental health inference approaches have covered a variety of outcomes including social anxiety [368], suicidality [369, 370], depression [371–373], well-being [361, 374, 375] and happiness [235, 376]. Interestingly, of 17 articles on population-level outcomes that were found during the systematic search for Chapter 5 five were focussed on positive mental health outcomes like well-being and happiness, whereas only one study did so out of the 165 included in the main review. A dedicated review article on the topic of using big data to study subjective well-being in 2017 found 33 articles [377]. This suggests that it is more common for positive mental health to be studied in the context of populations than individuals.

The modelling approaches used for population-level outcomes tend to be based on sentiment features. For instance, the *Hedonometer* created by Dodds et al. in 2011 to derive temporal patterns of happiness used a crowd-sourced sentiment lexicon for happiness in English language, where words were rated based on how ‘happy’ they were (this lexicon is known as LabMT) [235]. A similar lexicon approach for happiness has been applied in Arabic [376], and other studies have used categories of the Linguistic Word Inquiry Count (LIWC) [236] word lists of *anger*, *anxiety*, *death*, *risk*, and *sadness* to assess psychological distress [378] or *anxiety*, *anger*, *sadness* and *positive emotions* [365]. Other sentiment methods have included using overall sentiment valence measures such as TextBlob [237] and VADER [234] as a proxy for overall mental well-being [374, 379], and the NRC Emotion Lexicon [380] which contains the emotions of anger, anticipation, fear, surprise, sadness, joy, disgust, and trust [367]. As well as sentiment, patterns of life at a population level have also been found to be useful indicators of mental health and well-being of populations [53]. *Patterns of life* originated as term in surveillance and usually refers to using geo-location data to build an understanding of daily routines. However it is used in digital phenotyping to refer to non-language features of data, such as times of day that users are active, the frequency of their interactions with others, or what types of content they are producing [53, 321].

Similarly to findings in the variety of methods for sourcing ground truth in Chapter 5, models for population monitoring are validated in a variety of ways. A popular method is to use figures reported in publicly available national or regional surveys against tweets sourced from those geographic areas [361, 369, 371, 381]. Some studies have manually annotated tweets to train models [368], whereas others took a different approach and used pre-trained models from existing research for classifying mental health outcomes in individuals and applied them to population data [366]. There were also studies which did not attempt to validate their findings against external data, and whose aims were primarily to visualise and describe changes in sentiment features over time or geographic areas [53, 367, 368], often implied as a proxy for overall mental well-being.

For those studies that did attempt validation, the results illustrated that associations between measures such as happiness and sentiment were not always positive, as we might expect. Jaidka et al. [361] tested the difference in inferring subjective well-being using both machine learning and dictionary-based sentiment measures, and found that life satisfaction and happiness outcomes were negatively associated with the LIWC measure of Positive Emotion. Similarly, Gibbons et al. [381] found that population-level self-rated mental health was negatively associated with the LabMT average happiness measure developed by Dodds et al. [235], and the overall sentiment measure produced by the VADER algorithm [234]. Since both VADER and LabMT are coded in a positive direction this implies that as the mental health outcome improves the sentiment score becomes more negative. These contrary patterns do not tend to arise in the individual-level literature, which is usually more focussed on classification than regression (see Chapter 5). On an individual level previous work exploring the relationship between sentiment and depression by Coppersmith et al. [321] found that the LIWC categories of pronouns, swear words, anger, negative emotion and anxiety were significantly different for those with depression, though swear words, anger and negative emotion were highly correlated with each other given that they shared many core words. This research also found that the importance of sentiment versus pattern of life features differed depending on the outcome disorder, having compared depression, Post Traumatic Stress Disorder (PTSD) and Seasonal Affective Disorder (SAD) [321].

More recent sentiment related work has posited that differences between users with and without depression are better explained by the presence of short phrases that represent cognitive distortions, which aligns with Cognitive Behavioural Therapy theory [48]. However, these features are

not standard phrases in current sentiment dictionaries. Other recent research has explored the connectivity between LIWC sentiment categories using network analysis, as opposed to solely their given values [382]. This research found that whilst the LIWC categories were associated with depression symptom severity as expected (i.e. that negative sentiment is positively correlated with depression symptom severity, and positive emotion is negatively correlated) depressive episodes were best explained by increased network connectivity between the nine categories [382].

In summary, sentiment lexicons are a popular approach for representing outcomes in both the population-level and individual-level prediction of mental health. However, there are inconsistencies in how sentiment is associated with mental health, and a wide variety of approaches to modelling different mental health disorders making it challenging to compare results from different studies and across different mental health outcomes. More sophisticated methods, such as language models and word embeddings, are also much more popular features in the recent literature than sentiment in general [383, 384] as the technology to produce them, or adapt existing trained models such as BERT [385], has become more accessible.

6.1.2 Temporality in mental health inference

As well as the features used in mental health modelling, there is a question of how those features should be constructed, and how much influence their construction has on the outcome of interest. Mental health is rarely a consistent, life-long state and instead has a temporal nature which may last different lengths of time for different mental health disorders or facets of well-being [386]. For instance, an episode of depression is required to have lasted for more than two weeks to be formally recognised as a disorder [7] and in research asking participants to record self-identified episodes of depression over the last year, Kelley et al. [382] found that the mean length was approximately 80 days. Other outcomes such as suicidality may vary significantly, with research suggesting that suicidal intent can precede a suicide attempt by as little as 10 minutes [387, 388]. General well-being on the other hand is not a clinically diagnosed disorder, and instead is a positive outcome thought to represent optimal psychological functioning. Given this, there is not a single agreed definition of well-being as a whole, but in general different types of well-being are thought to be relatively stable over time [389]. Despite these variations in the experiential length of mental health disorders and well-being most clinical self-report surveys ask respondents to answer on the basis

of the past two to four weeks.

Whether or not ground truth collected using diagnostic screening questionnaires actually represents the past two weeks of the respondents' emotions is unclear. Bias in memory, such as a bias towards negative events has long been thought to be an important component of mental health disorders like anxiety and depression [390, 391]. Additionally the 'forgetting curve', first proposed by Ebbinghaus in 1885 [392] and recently reproduced [393], suggests that memories are gradually lost over time which may mean that more recent events are more pertinent to self-assessments of mental well-being. The differential temporal signatures of different mental health conditions, and the cognitive memory bias presented by the so-called 'negativity bias' and the 'forgetting curve' may all have an impact on the relationship between an individual's recorded experiences through their tweets over the period a self-assessment asks them to consider, and the symptoms that they then report. Tchalac et al. tested this potential impact in their prediction of depression in individuals using the PHQ-9, a self-completed depression questionnaire, as ground truth, and found that two weeks of tweet data did give the best results [394]. Similarly in predicting depression from tweets de Jesús Titla-Tlatelpa and colleagues also found shorter periods of time generated less error [395]. In the detection of suicidal ideation Sawhney and colleagues found that up to three months of data was useful, but that after this the benefit of additional data saturated [396]. Research modelling suicidality has also tested the impact of a temporal weighting function, that effectively introduces a decay in the influence of a tweet the further it is from the measurement time point, and found that an exponential decay performed better than no weighting at all [397]. These results suggest that there might be different lengths of tweet history relevant to different disorders, and that weightings that represent 'forgetting' could improve the accuracy of inference using Twitter data.

Another question that has been raised in Chapter 5 is how models perform when tested at future time points. This is particularly relevant to research where the task is to monitor mental health for change, as opposed to classification of single tweets for risk levels. For instance, if the task we are attempting to model is a tweet-level system that flags individual tweets that appear concerning for further follow up then a single time-point may be enough. This being said, it still does not tell us how a model might translate to a different time of year, or even a different year altogether. Research has shown [235, 398] that there are strong time-based effects to the emotions we express on Twitter. For instance, emotions as measured through sentiment algorithms change throughout the day fol-

lowing circadian rhythms, and also weekly patterns (known as circaseptan rhythms) [398]. Models that intend to monitor change with time may need to take these rhythms into account, as well as understand how sensitive predictions are to unexpected events that have the potential to disrupt the population's baseline levels of mental health and well-being at a local, national or international level [376]. For instance natural disasters, inter/national conflict, or a pandemic.

6.1.3 The present study

In this chapter I will explore a series of research questions that investigate the efficacy of using Twitter sentiment and patterns of life to model three mental health outcomes: depression, anxiety and general well-being. Specifically, the research questions are:

- 1) How well do patterns of life and standard sentiment codings of positive and negative emotion predict depression, anxiety and general well-being?
- 2) Is prediction of depression, anxiety and general well-being improved by using a larger number of linguistic categories derived from Twitter data?
- 3) What is the effect of changing the window and weightings of Twitter data on prediction performance?
- 4) How do predictions perform over time?

I will use the linked Twitter data described in Chaer 4, collected within the ALSPAC cohort study to explore these questions, which allows for high quality ground truth data about the participants and their mental health to be modelled using data at two time points.

Mental health outcomes will be modelled as continuous variables, represented by the score of the questionnaire used to measure them. This approach allows us to assess changes in mental health as improving or worsening, rather than movement between a states of illness and 'well-ness'. Whilst a continuous scale cannot possibly capture the complexity of the experience of a mental health disorder, it goes some way towards acknowledging most mental illnesses are not binary experiences and symptoms can vary significantly between individuals.

6.2 Methods

6.2.1 Sample

This study uses data from the Avon Longitudinal Study of Parents and Children (ALSPAC) [118–120]. Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The initial number of pregnancies enrolled is 14,541 (for these at least one questionnaire has been returned or a “Children in Focus” clinic had been attended by 19/07/99). Of these initial pregnancies, there was a total of 14,676 fetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. When the oldest children were approximately 7 years of age, an attempt was made to bolster the initial sample with eligible cases who had failed to join the study originally. As a result, when considering variables collected from the age of seven onwards (and potentially abstracted from obstetric notes) there are data available for more than the 14,541 pregnancies mentioned above. The number of new pregnancies not in the initial sample. The total sample size for analyses using any data collected after the age of seven is therefore 15,454 pregnancies, resulting in 15,589 fetuses. Of these 14,901 were alive at 1 year of age.

The linked Twitter data used in this chapter were collected from any adult ALSPAC participants who had a Twitter account and consented to collection of their Twitter data. More detailed information on the collection of these data is given in Chapter 4. There were 654 linked participants in total. The ground truth data used in this study were from questionnaires sent to ALSPAC participants at the beginning of the COVID-19 pandemic. These surveys were sent to all adults in ALSPAC, meaning that both the index children (Generation 1) and their parents (Generation 0) were surveyed. Responses for the first COVID-19 survey were collected between the 3rd April 2020 and the 14th May 2020. Responses for the second COVID-19 survey were collected between the 26th May 2020 and 3rd July 2020. These surveys will be referred to as Survey 1 and Survey 2 throughout this chapter. Figure 6.1 shows the two data collection periods, relative to the number of tweets collected during these periods.

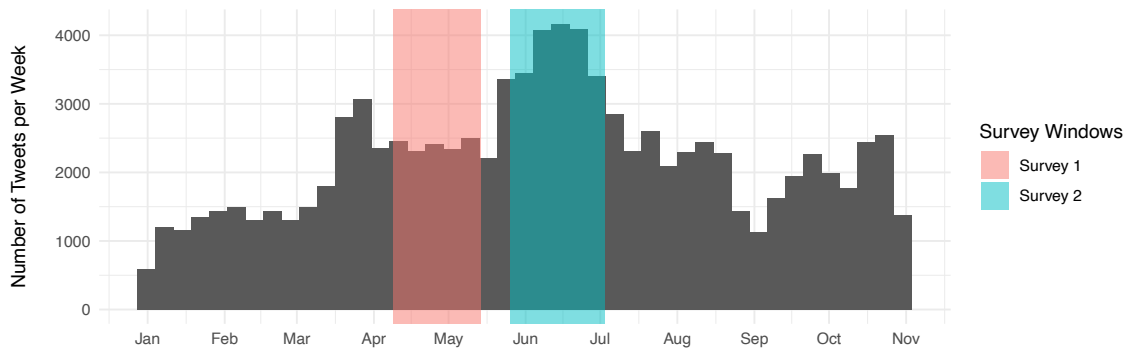


Figure 6.1: The windows of data collection for Survey 1 and Survey 2 where the psychological outcomes were measured.

For the purposes of modelling tweets I chose to initially focus on the two week window for each participant that led up to the date that they completed the survey, which concurs with the time period participants were asked to evaluate in the mental health measures they completed. Participants were required to have tweeted at least twice in those two weeks in order to be included. In total this resulted in 151 participants with linked Twitter data that completed Survey 1, and 136 participants with linked data that completed Survey 2. Since different members of the ALSPAC sample tend to complete each survey, and they may be tweeting at different times, the participants who completed the two surveys do not perfectly overlap. 92 are included in both samples, with a total of 195 participants across both, as illustrated in Figure 6.2.

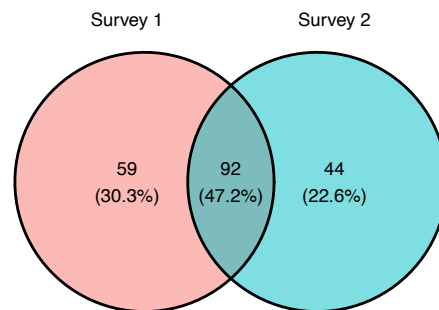


Figure 6.2: The number of linked participants with two tweets in the two weeks leading up to the survey date for Surveys 1 and 2, and how many participants are unique to each survey or overlap across both.

6.2.2 Measures

I will briefly describe the measures used in this study, which includes both the mental health measures collected and the sentiment analysis algorithms used to generate the feature variables.

Psychological Outcomes

Depressive symptoms were measured using the short Mood and Feelings Questionnaire (MFQ) [151], a 13-item scale that has been validated for measuring depressive symptoms in adolescents [152] and in young adulthood [153]. Scores range from 0 to 26, with a higher score indicating more severe depressive symptoms [152].

Symptoms of anxiety were measured using the GAD-7, designed to be a brief measure for generalized anxiety disorder [240] which has been validated in a primary care population. It has seven items, with an overall maximum score of 21. Higher scores indicate more severe symptoms of anxiety.

General well-being was measured using the Warwick Edinburgh Mental Well-being Scale (WEMWBS), which is a fourteen-item questionnaire that has been validated for measuring general

well-being in the general population [156, 399], as well as in young people [158, 159]. There are five response categories for each question, and the total score is between 14 and 70. All items in the WEMWBS are positively worded, and it is focused on measuring positive mental health. Uniquely of the three measures used in this study, higher scores indicate more positive outcomes in the WEMWBS, and it was also designed so that the distribution of scores would be roughly normal, as opposed to the diagnostic questionnaires which tend to be skewed towards lower (and thus not clinically relevant) scores.

For all three measures respondents were asked to consider how they have felt for the past two weeks.

Derived Twitter variables

Pre-processing A detailed description of how the Twitter data in general was linked and derived using the *Epicosm* software [117] is included in Chapter 4. Each algorithm described below was applied to each individual tweet, with the variables being provided for analysis at tweet level. No textual pre-processing was conducted by *Epicosm* for VADER or LabMT. For the LIWC each tweet was tokenised, with a word considered to be a string of alphanumeric characters or underscores delimited from other words by spaces.

VADER Sentiment The Valence Aware Dictionary and sEntiment Reasoner (VADER) [234] is a sentiment analysis tool developed with the specific intention of being sensitive to short social media length text. It aims to convey the orientation of sentiment (positive versus negative) as well as the strength of the intensity of the sentiment. This is achieved through a rule based system that takes consideration of punctuation use and capitalization in the text to infer intensity.

VADER returns an overall figure for each of positive, neutral and negative sentiment. These represent the proportion of the text that falls into each category, without any adjustment for word-order sensitivity, negation or punctuation amplification (which are used to account for intensity). These three scores are all between 0 and 1, and add up to 1. The sentiment intensity adjustments are then made for the *compound* score, which is a weighted composite score of the positive, neutral and negative scores, normalised between -1 and +1 where higher scores indicate more positive sentiment overall [234].

LabMT Sentiment LabMT (Language Assessment by Mechanical Turk) is a sentiment dictionary developed by Dodds and Danforth in 2011 [235], which they designed to measure happiness in text based on human ratings of individual words. It was developed by collecting ratings of 10,222 words on a nine-point scale of happy to sad from Amazon’s Mechanical Turk crowdsourcing platform, with the original word list drawn from Twitter, Google Books, music lyrics and the New York Times [235]. Each word received 50 ratings, and each rater assessed 100 randomly chosen words. The development of this sentiment dictionary was influenced by the ANEW lexicon [400] which is also a popular measure of sentiment from written text and used a similar approach, though for only 1,034 words with a group of student raters.

The overall score for each tweet is the average happiness of the words in the tweet.

Linguistic Inquiry and Word Count (LIWC) The Linguistic Inquiry and Word Count (LIWC) has been in development since 1993 by Pennebaker and colleagues, with most recent version of the tool made available in 2015 [236]. This is version that I use and refer to here. The LIWC is a dictionary based approach that has a series of words in each of its 73 categories. Since the most recent version in 2015 it can also accommodate short phrases and text-based emoticons, such as ‘:-)’ which would be included under ‘Positive Emotion’. The development process for the LIWC was extensive and involved several rounds of rating, consensus and discussion between a group of expert ‘judges’, which started originally by drawing words from common emotion rating scales like the Positive and Negative Affect Scale (PANAS) [401].

The categories for the LIWC generally have a hierarchical structure, which at the highest level are split into *Linguistic Dimensions*, *Other Grammar* and *Psychological Processes*. Within these categories are then further categories, and in the case of *Linguistic Dimensions* and *Psychological Processes*, sub-categories. For instance, *Affective Processes*, a sub-category of *Psychological Processes*, is further split into *Positive Emotion*, *Negative Emotion*, *Anxiety*, *Anger* and *Sadness*, but *Affective Processes* is also a category in its own right that is the combination of all words in its sub-category. In some cases like *Social Processes*, the overarching category also contains words that do not fit into any of its sub-categories. Additionally, many categories and sub-categories share words, so for instance the word “cried” is part of five categories [236]. As a result of this structure LIWC categories are not independent and are often correlated with one another. The overall score

for each LIWC category that is derived for a piece of text is the proportion of words in the piece of text that came from each category. As such this score can range between 0 and 1.

Whilst the LIWC is generally described in the literature as being a sentiment analysis tool, it actually incorporates several different types of dictionary based analysis, including parts of speech in its *Linguistic Dimensions* like tagging of different types of pronouns, verbs and adjectives, as well as a lexicon based approach to topic modelling through recording categories included in the text. Whilst these approaches are not as sophisticated as tools directly developed for the purposes of parts of speech tagging and topic modelling, they do provide some insights which go beyond a typical sentiment analysis tool.

Patterns of life Patterns of life refer to data about people's routine use of Twitter. This can include many different features but in this study refers to the number of tweets someone has sent, the mean and variance of the time interval they sent their tweets within, the proportion of their tweets which were re-tweets and the proportion of their tweets which were sent between midnight and 4am.

Mean time intervals were calculated by numbering each 4-hour time window that tweets were sent within over the day as 1 to 6, with 1 referring to midnight-4am and 6 being 8pm-midnight. The mean and variance of this variable were taken for the aggregated dataset. Pattern of life variables were not weighted for the analyses described in Section 6.2.3 since they are calculated over the whole time window of tweets, rather than on a tweet-level.

6.2.3 Analysis

The analyses of these data and the investigation of the research questions outlined above will be split into four sections.

How well do patterns of life and codings of emotion predict mental health?

To answer this question I conducted individual regression analyses of each of sub-categories for *Affective Processes* in the LIWC, *VADER positive*, *negative* and *compound* and the *LabMT* score as dependent variables against each of the continuous scales of depression, anxiety and general well-being as the independent variables. Every analysis included the sex and generation of the

participants as dependent variables. For each of the sentiment variables the mean and variance was taken for each individual over the two weeks leading up to Survey 1, and Survey 1 was used as the ground truth data source. To be included in the analysis participants must have tweeted at least twice in the two week period. The same approach was repeated for the patterns of life variables. Both depression and anxiety were transformed using the natural log on their raw values plus one due to inherent skewness in these clinical scales that leads to a long right hand tail. General well-being measured with the WEMWBS follows a roughly normal distribution and so transformation was not necessary.

Two weeks of data up to the survey period was chosen because it reflected the time period that each of the MFQ, GAD-7 and WEMWBS measures (used to measure depression, anxiety and general well-being respectively) asked participants to consider when responding.

It has also been seen to be an effective time period in other research using similar measures as ground truth [394]. Survey 1 was chosen as the ground truth data point because it had the higher number of participants, and because there was evidence in the data (see Figure 6.1) and in the literature [402, 403] that the events surrounding the national protests in the Black Lives Matter movement at the beginning on June 2020 had a significant impact on Twitter data sourced during that period, which overlaps with Survey 2.

Is prediction improved by using a larger number of linguistic categories?

Here I considered whether predicting the mental health outcomes using a wider range of variables than just the standard codings of positive and negative emotion was beneficial. In order to do this I used all variables already included in the first stage of analysis, as well as all of the other LIWC categories available, and repeated the regression analyses described in Section 6.2.3 for each of these available variables. Continuous variables were again aggregated by individual using their mean and variance. There were a total of 160 variables tested. Once again I considered these over the two weeks previous to the completion of Survey 1 for each participant, for only those participants with two tweets in that two week period.

Based on this analysis I then considered the variables which were associated with each outcome with $p < 0.05$. The decision not to use the bonferroni adjusted p-value was made ad-hoc given that for two of the outcomes this would have left no associated variables. Additionally, the bonferroni

threshold is likely to be conservative for groups of variables where there are likely to be high correlations between variables. These Twitter features were then used as the dependent variables in a multiple regression model which predicted each of depression, anxiety and general well-being. Some variables were excluded from the model due to multi-collinearity which is discussed further in the Results.

For each model five-fold cross validation repeated ten times was used to obtain a more robust estimate of model error.

What is the effect of changing the window and weightings of data?

Next I considered how the performance of the predictive models changed when the training data was weighted according to a decay function, and when the window of data input to the model was lengthened. The same features and modelling methodology as described in Section 6.2.3 were used.

The decay functions tested for weighting the data were:

$$\frac{1}{2} + \frac{1}{2 + (16 \frac{day}{max(day)})^3} \quad (6.1)$$

$$(1 - \frac{day}{max(day)}) + \frac{1}{max(day)} \quad (6.2)$$

$$\frac{1}{\sqrt{day}} \quad (6.3)$$

$$\frac{1}{day} \quad (6.4)$$

as well as no weighting function. These were chosen to represent a range of options across the potential feature space of weightings, as visualised in Figure 6.3.

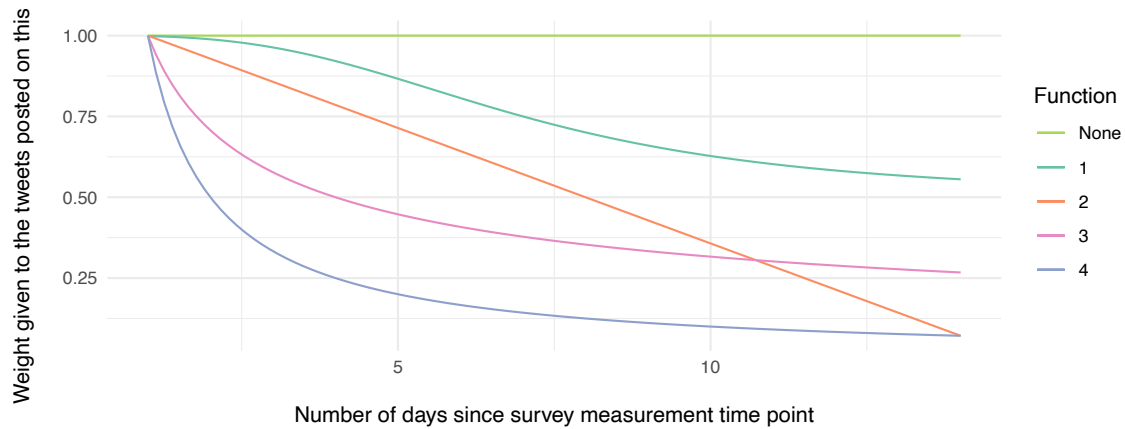


Figure 6.3: The five options of weighting functions

Windows of training data were tested by extending the input data to include 2, 4, 6, 8, and 12 weeks since the date participants completed Survey 1. Each combination of window and weighting options was tested on the same 151 participants against their mental health outcomes at Survey 1. In this experiment the mean result following ten repeats of five-fold cross validation was used for each combination of weighting scheme and window length. On each repeat individuals were randomly allocated into a fold, with each individuals data never contained in both the test and train data.

How do the predictions perform over time?

Based on the best models found from varying the window and weighting on Survey 1 data, I analysed the model error for each outcome, as measured by the Root Mean Squared Error (RMSE), when predicting values at Survey 2. This essentially treated the Survey 2 data as a validation dataset for the models trained on data from Survey 1. Error was calculated for all predicted results, and then also for the sub-groups of sex and generation. For these sub-groups some individuals from the training set would have also been in the validation set because they responded to both surveys (see Figure 6.2), and these individuals were retained due to the impact on sample sizes if they were removed. However, to test the impact of including the same people in both samples, RMSE was also calculated for those who were within the original training sample, and those who were not. Welch's t-test was used to test for differences between the sub-groups.

Following this analysis I applied the predictions made by the models to fortnightly data between

the 1st January 2020 and 31st October 2020 and visualised the predictions made for each outcome over time. Each two week period is considered as a discrete section of the data (that is, none of the data from each period overlapped) and the predictions here were made for any participant who had tweeted in that time, not restricted to any tweet frequency, nor to the original sample of respondents. Prediction error was calculated for these predictions.

6.2.4 Software and code

The Twitter data were collected and pre-processed using Epicosm [117]. ALSPAC survey data were collected and managed using REDCap electronic data capture tools hosted at the University of Bristol [149]. REDCap (Research Electronic Data Capture) is a secure, web-based software platform designed to support data capture for research studies.

Statistical analysis of the data were performed using the R language (v4.0.4) in RStudio (v1.4). Data tidying was primarily conducted using the `tidyverse` suite of packages [168], with data visualization using `ggplot2` (v3.3) [169] and tabulation using `gtsummary` (v1.5) [241] and `kableExtra` (v1.3). Modelling was conducted using the `caret` package (v6.0) [404].

6.2.5 Ethics

As discussed in Chapter 4, the data collection process for the linked Twitter data in ALSPAC was conducted with ethical approval and informed consent from the participants. Secondary data analysis projects using ALSPAC data are not required to go through ethical approval, although the projects themselves must be approved by the ALSPAC Executive Committee. However, internal ethical approval processes do not necessarily capture the potential ethical and societal risks of data science projects, and are primarily concerned with protecting the participants who are taking part in a study.

Since mental health prediction is an area of research that has potentially far-reaching consequences I have included in Appendix D.1 an analysis of this project using the Data Hazards framework. Data Hazards are labels that can identify potential risks generated by the pursuit of a research project, or generation of research data (see Zelenka and Di Cara 2022 [405] for a detailed explanation). The labels chosen for this project were informed by two reflective peer feedback groups in 2021, where other researchers and data scientists considered the potential consequences of this

project and discussed the reasoning for the inclusion or exclusion of each Data Hazard.

6.3 Data Description

To give a thorough overview of the data being modelled, this section will provide a variety of descriptive information about the sample, the Twitter data and its temporal patterns.

6.3.1 Mental health outcomes across the samples

The first descriptions will be of the psychological outcomes themselves. Table 6.1 gives a description of each outcome at each of the survey time-points, and between the linked sample and the samples included in this study. We can see that the samples used in this study have slightly poorer mental health outcomes across all three measures compared with the whole linked sample. We also know from Chapter 4 that the linked sample also has marginally poorer mental health outcomes than the main ALSPAC sample. Summaries of differences across both Survey time points by sex, and generation are available in the Appendix in Tables D.2 and D.3 respectively. When using the generally accepted cut off scores of 12 or more for the MFQ (measuring depression) [153] and ten or more for GAD-7 (measuring anxiety) [240] then there would be 46 people in the sample with depression and 80 with anxiety at Survey 1. At Survey 2 this would be 50 and 83 for depression and anxiety respectively.

It is common for depression and anxiety to co-occur, and for general well-being to be related to both depression and anxiety. As such I would expect for all of these combinations of variables to be somewhat correlated. Spearman's r for each pair are 0.77 between depression and anxiety, -0.63 between depression and general well-being, and -0.61 between anxiety and general well-being. In summary, all of these pairs are moderately correlated, with all correlations rejecting the null hypothesis that there was no relationship.

6.3.2 Twitter features

Sentiment features tend to be fairly highly correlated since many of them are attempting to measure roughly the same concepts. Figure 6.4 shows the intercorrelations between the features that have been identified for this study as representing codings of emotion or affect. The VADER scores are highly intercorrelated as would be expected since their scores are proportional to one another. VADER neutral is also negatively associated with LIWC affect which appears to be best related to

Table 6.1: Summary of key features of the linked Twitter sample against the sample who tweeted at least twice and completed Survey 1 (N=151), and those who tweeted at least twice and completed Survey 2 (N=136)

Characteristic	Linked Sample	Survey 1 Sample	Survey 2 Sample
	N = 654	N = 151	N = 136
Sex			
Female	412 (63%)	92 (61%)	87 (64%)
Male	241 (37%)	59 (39%)	48 (36%)
Unknown	1		1
Generation			
G1	430 (66%)	96 (64%)	88 (65%)
G0	224 (34%)	55 (36%)	48 (35%)
Anxiety (Survey 1)			
Unknown	193	4	17
Depression (Survey 1)			
Unknown	204	9	18
Well-being (Survey 1)			
Unknown	193	7	18
Anxiety (Survey 2)			
Unknown	199	30	6
Depression (Survey 2)			
Unknown	198	27	4
Well-being (Survey 2)			
Unknown	199	32	8

¹ n (%); Median (IQR)

variables that describe positive affect.

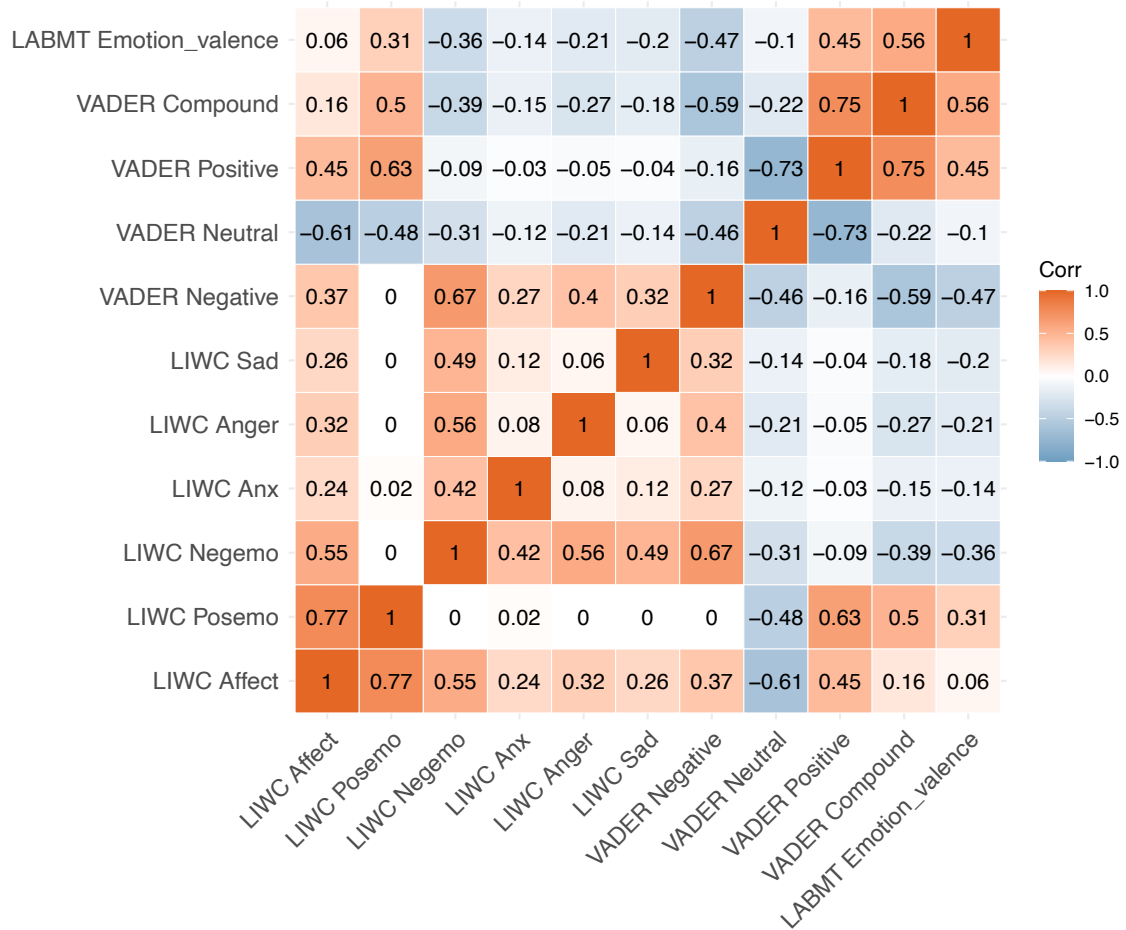


Figure 6.4: A correlation plot of the relationships between sentiment variables that relate to codings of emotion or positive and negative sentiment, without aggregation by individual. Correlations are calculated using Spearman’s Rank, with colour representing the correlation coefficient.

Another important feature of Twitter data is the difference in tweeting frequencies between individuals. Figure 6.5 is a histogram of these frequencies in the two weeks leading up to Survey 1, and shows that there is a huge amount of variability in the volume of tweets produced.

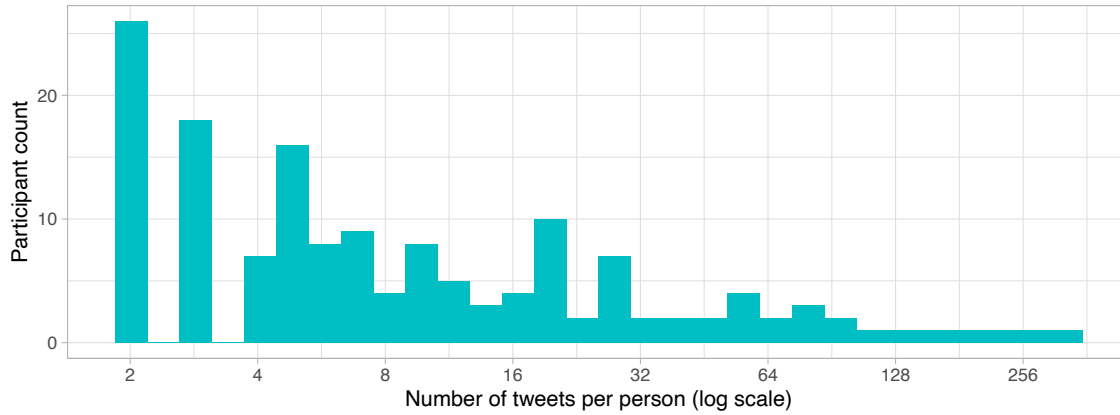


Figure 6.5: Histogram of tweet frequency over the two weeks leading up to Survey 1, for those who have tweeted at least twice. Tweets per person are transformed with the binary logarithm.

6.3.3 Temporal patterns

Next, I will consider how sentiment has changed from January to October 2020 and split this by three different features of the data. These are the sex of participants in Figure 6.6, the generation of the participants in Figure 6.7 and then whether the tweet was a retweet or not in Figure 6.8. VADER was chosen for these summaries as previous research has suggested that it outperforms other sentiment algorithms as an overall summary of sentiment for social media length text [52].

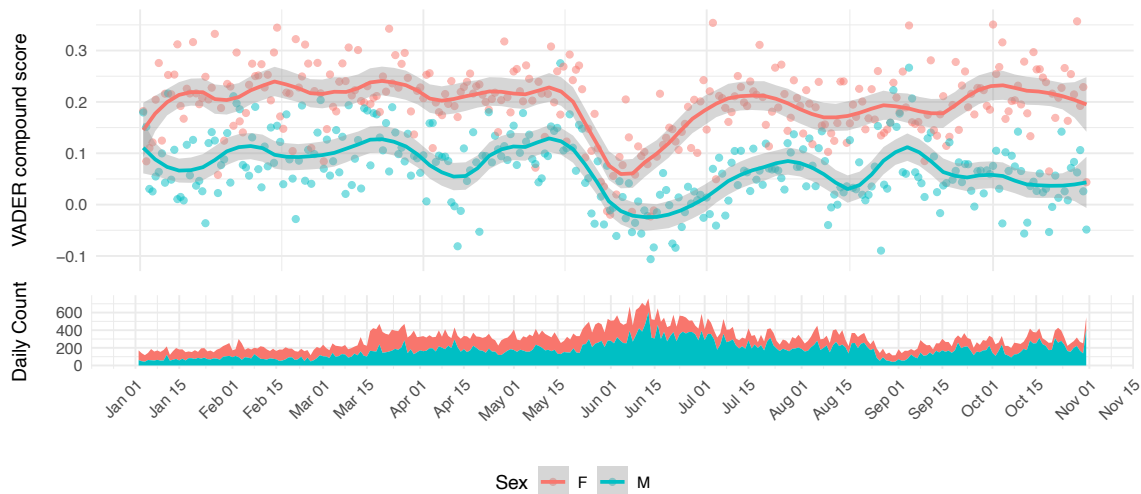


Figure 6.6: In the top figure, VADER compound is split by sex and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of each sex group differentiated by color and stacked on top of each other.

In general the sentiment of women’s tweets is higher than men’s across time. This difference appears to be between 0.05-0.1 units of VADER compound sentiment over time, which is fairly large considering that VADER compound is measured between -1 and 1 overall.

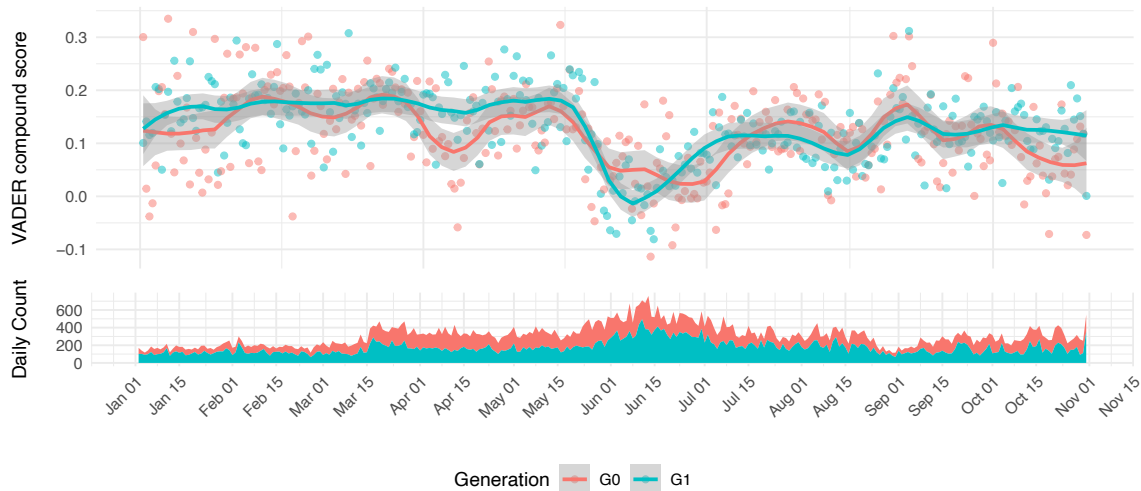


Figure 6.7: In the top figure, VADER compound is split by generation and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of each generational group differentiated by color and stacked on top of each other.

Whilst the difference between the G0 and G1 cohorts does not appear to be large, differences in their patterns of sentiment relate more to timing. G1 participants saw a steeper and slightly deeper dip in sentiment at the beginning of the COVID-19 pandemic, as well as a large increase in tweet frequency that is not as pronounced in the G0 participants. The dip in sentiment for the G0 participants is more gradual and the negative peak appears about 3 weeks later than the G1s, though it is not as negative.

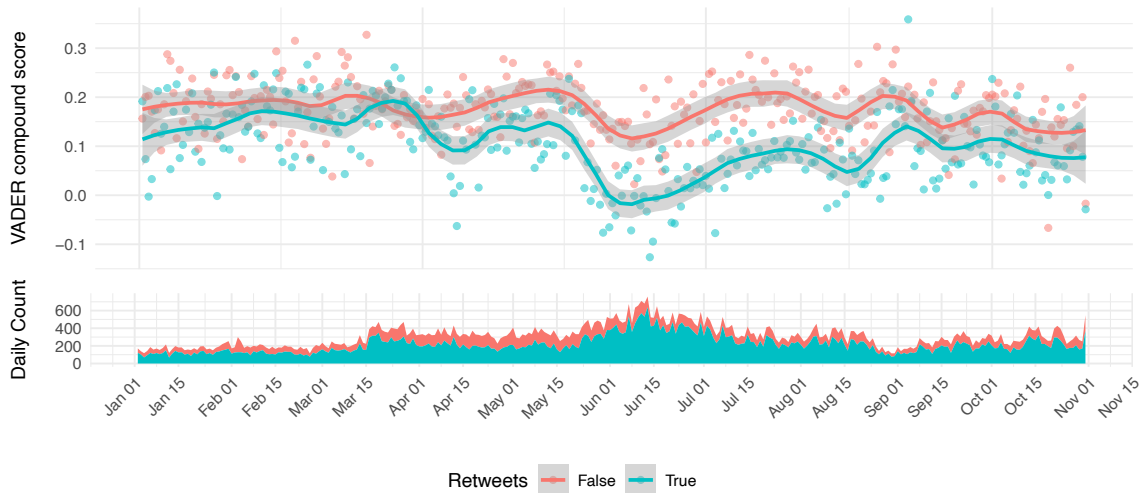


Figure 6.8: In the top figure, VADER compound is split by whether or not a tweet is a retweet and plotted between 1st January 2020 and 31st October 2020. In the lower figure, the total number of daily tweets is plotted, with the contribution of retweets and original tweets differentiated by color and stacked on top of each other.

Lastly we can consider the proportion of retweets versus original tweets over time. In Figure 6.8 we can see that original tweets are generally more positive than retweets, apart from a short period of time in mid-late March 2020. Across all of the timeseries plots of VADER compound we can see that there was a noticeable dip in overall sentiment in early June 2020. This time period corresponds with several important events including the phased re-opening of schools in England following the coronavirus lockdown, and the Black Lives Matter protests across the UK.

Another temporal perspective on overall sentiment scores is throughout the day. Twitter data provided on the ALSPAC participants is mapped to four hour windows, and so in Figure 6.9 we can see the average sentiment at each four hour window in each month throughout the year. Overall daily sentiment starts low just after midnight (though still within VADER’s ‘neutral’ window of

± 0.05), and is highest during the middle of the day, before dropping again in the evening. The only notable exception is for December where the 04:00-7:59 time point is much higher than all the others, which is likely to be due to people posting ‘Happy Christmas’ messages on Christmas morning. There are no obvious patterns in the differences between summer and winter.

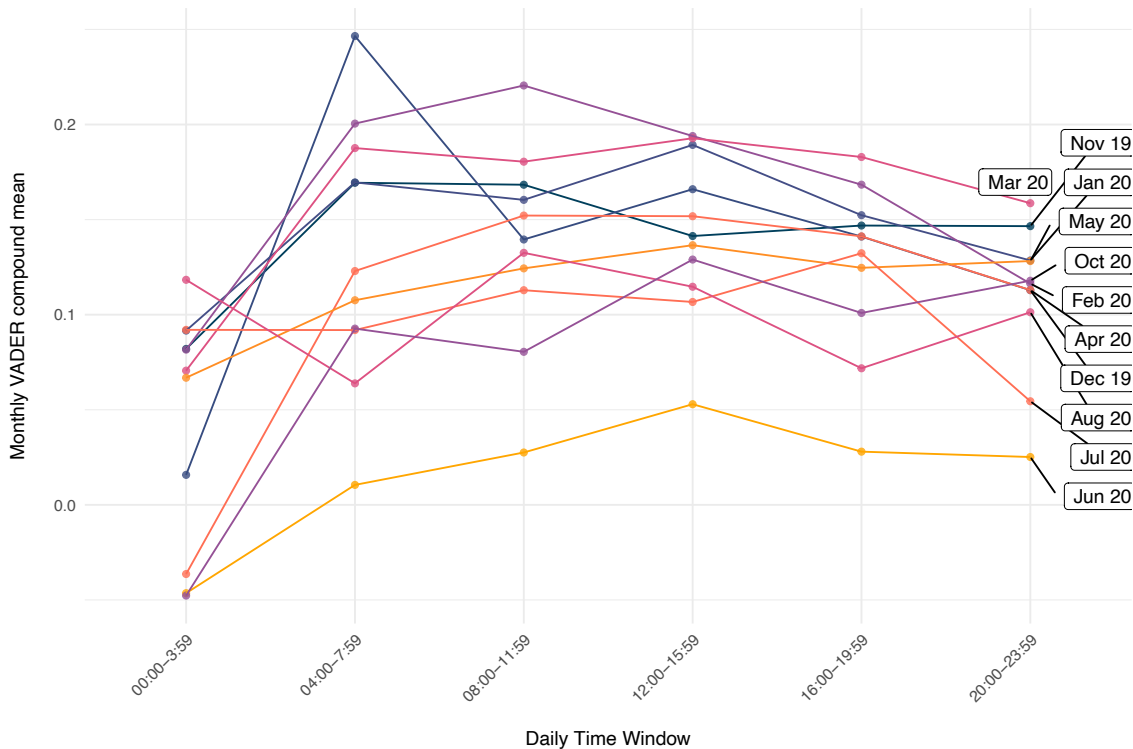


Figure 6.9: VADER compound across the day over 12 months. Warmer months are coloured towards yellow, and colder months towards blue.

Having considered sentiment as an overall outcome we can also look at the categories of the LIWC and how these change over time. As noted in Section 6.2.2 LIWC categories extend beyond sentiment to include areas like parts of speech and topics. Figure 6.10 is a visual summary of how each of the top ten occurring LIWC categories change over time, based on their mean value across all tweets each month. The LIWC category ‘function’ is omitted because it is consistently much higher than all of the other values at approximately 0.28.

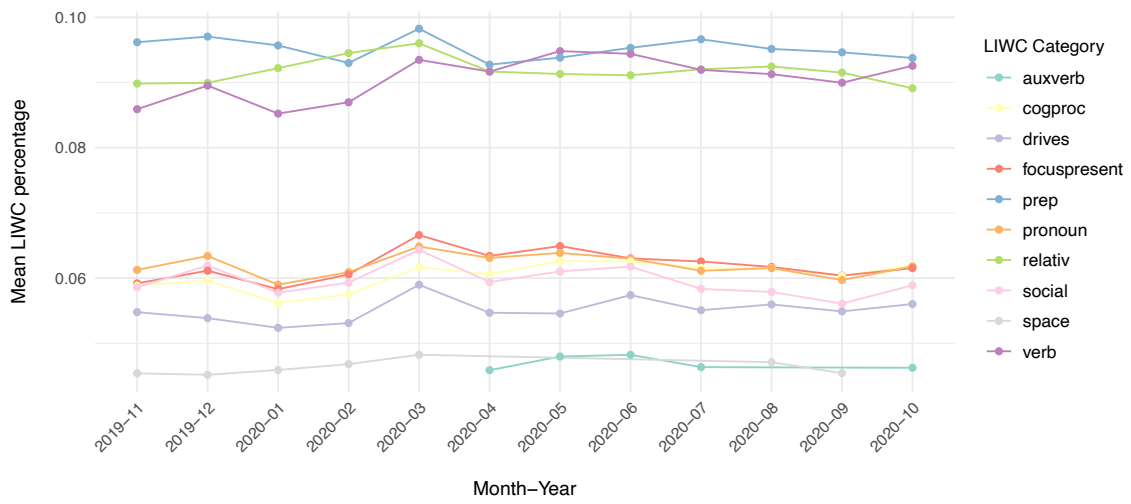


Figure 6.10: The ten most common LIWC values over 12 months (without 'function' which has a mean of approximately 0.28)

6.4 Results

6.4.1 How well do patterns of life and codings of emotion predict mental health?

This first research question considers the overall positive and negative codings generated by the *LabMT* happiness scale and *VADER*, as well as the *Affective Process* dimensions of the LIWC and patterns of life (PoL). Each of these outcomes was tested against depression, anxiety and general well-being using a linear model, with depression and anxiety transformed using the natural log. Models were also adjusted for sex and generation. Figure 6.11 presents the magnitude of each outcome as the percentage of the variance explained after sex and generation were accounted for. This metric was chosen to maximise the opportunity for comparison between different types of features and between different mental health outcomes.

The clearest association is between anxiety and the mean of *VADER positive*. This is also the only association with a p-value beneath the bonferroni adjusted threshold. In fact, anxiety is generally best explained by more of the features shown than depression or general well-being are. The associations of all outcomes and variables are in the directions that would be expected, that is that increasing negative sentiment variables are associated with an increase in poorer mental health and the opposite for positive variables.

The equivalent results are available in Supplementary Table D.4 for the data not aggregated by individual, and illustrates that when using the disaggregated data some variables showed the opposite direction of effect, although their p-values are above 0.05. For example, *LIWC Positive Emotion* was positively associated with increased depressive symptoms, as was *VADER Positive*.

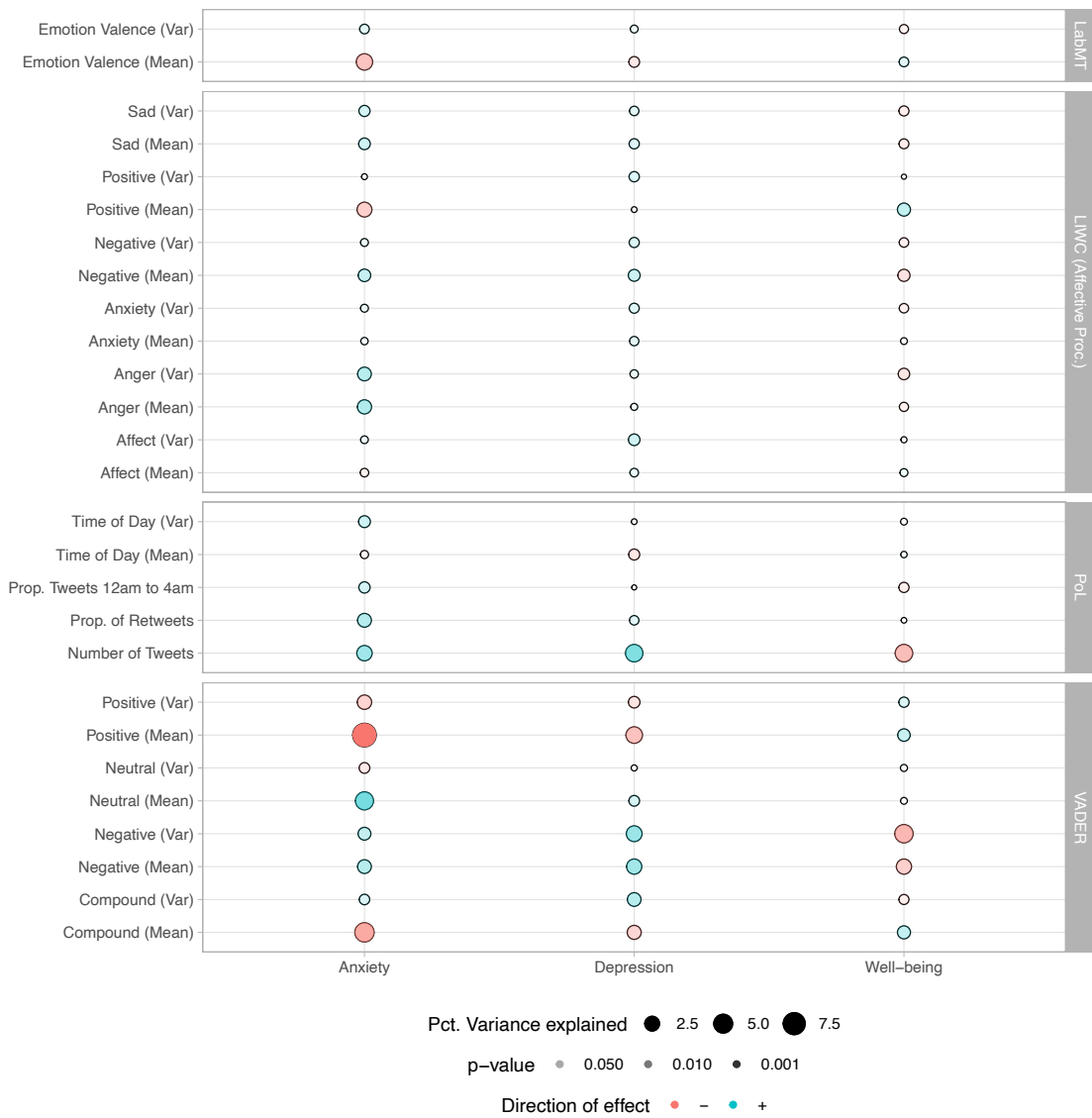


Figure 6.11: The results of regression models of each sentiment variable against depression, anxiety and general well-being, adjusted for sex and generation. The percentage variance explained after sex and generation have been accounted for is given by the size of each point, the sign of the coefficient is given by the colour, and the p-value is represented by the transparency of the point.

6.4.2 Is prediction improved by using a larger number of linguistic categories?

Now I consider the results of only looking at the most successful sentiment features for each of the mental health outcomes. I took variables with a p-value of less than 0.05 for their association with each mental health outcome, with Table 6.2 displaying these sorted by the percentage variance they explained in their individual models (after sex and generation are accounted for). The pairwise correlations between each of the sets of variables associated with depression, anxiety and general well-being are illustrated by correlation plots in the Appendix, in Figures D.1, D.2 and D.3 respectively. Each associated set of variables were then used as dependent variables to model depression, anxiety and general well-being using a linear multiple regression model for each. The results for each of the outcomes is given in the following sub-sections.

Note that depression and anxiety were both log transformed using the natural logarithm. This decision was made to better meet the assumptions of the linear regression, most specifically the normality of the residuals. Number of tweets was also min-max scaled to the range of zero to one for better interpretability against the other coefficients.

Table 6.2: This table displays the variance accounted for by each sentiment variable with $p < 0.05$ when regressed against depression, anxiety or well-being, and adjusting for sex and generation. The 'Effect' column gives the direction of the coefficient. p-values in orange are < 0.001 and those in teal are < 0.01 .

Variable	Effect	Variance Explained	p-value
Anxiety			
VADER Positive (Mean)	-	8.19	0.000
VADER Compound (Mean)	-	4.81	0.005
VADER Neutral (Mean)	+	4.04	0.010
LIWC Leisure (Mean)	-	3.91	0.011
LabMT (Mean)	-	3.06	0.025
LIWC Body (Mean)	+	2.96	0.028
LIWC Health (Var)	-	2.86	0.030
LIWC Quant (Var)	+	2.77	0.033
LIWC Health (Mean)	-	2.70	0.035
Number of Tweets	+	2.53	0.042
LIWC Friend (Mean)	-	2.42	0.047
Depression			
LIWC Verb (Mean)	+	4.45	0.006
LIWC Verb (Var)	+	3.79	0.012
Number of Tweets	+	3.65	0.014
VADER Positive (Mean)	-	3.16	0.022
LIWC Leisure (Mean)	-	3.09	0.024
LIWC Quant (Var)	+	3.03	0.025
LIWC They (Var)	-	2.76	0.033
VADER Negative (Var)	+	2.75	0.033
LIWC Female (Mean)	-	2.68	0.035
LIWC Female (Var)	-	2.49	0.043
VADER Negative (Mean)	+	2.47	0.044
LIWC Focuspast (Mean)	+	2.47	0.044
LIWC Body (Mean)	+	2.44	0.045
Well-being			
LIWC Verb (Var)	-	4.76	0.007
LIWC Verb (Mean)	-	4.56	0.008
LIWC Money (Var)	-	4.35	0.010
VADER Negative (Var)	-	4.22	0.011
LIWC Home (Var)	-	3.88	0.015
Number of Tweets	-	3.71	0.017
LIWC Relativ (Var)	-	3.61	0.019
LIWC Focuspast (Mean)	-	3.58	0.019
LIWC Health (Mean)	+	3.02	0.032
LIWC Time (Var)	-	3.01	0.032
LIWC Sexual (Mean)	-	2.88	0.036
LIWC Shehe (Var)	-	2.76	0.040

Depression

In the depression linear model, *LIWC female* mean and variance had high variance inflation factors and so the variance was removed from the model, since the mean accounted for more variance overall. As such the theoretical model for depression was:

$$\begin{aligned} \log(\text{Depression_Survey1} + 1) = & \alpha + \beta_1(\text{liwc_verb.mean}) + \\ & \beta_2(\text{liwc_verb.var}) + \beta_3(\text{n.tweets01}) + \\ & \beta_4(\text{vader_positive.mean}) + \beta_5(\text{liwc_leisure.mean}) + \\ & \beta_6(\text{liwc_quant.var}) + \beta_7(\text{liwc_they.var}) + \\ & \beta_8(\text{vader_negative.var}) + \beta_9(\text{liwc_female.mean}) + \\ & \beta_{10}(\text{vader_negative.mean}) + \beta_{11}(\text{liwc_focuspast.mean}) + \\ & \beta_{12}(\text{liwc_body.mean}) + \epsilon \end{aligned} \tag{6.5}$$

Table 6.3 gives the results for this model. It shows that less variance in the use of *They* pronouns was associated with increased depressive symptoms, and more use of words relating to *Body* was also positively associated with increased symptoms.

Repeated 5-fold cross-validation (10 repeats) was also used to obtain a more robust estimate of the model's error, which found that the mean R^2 was 0.096 (SD = 0.09), and the mean RMSE was 0.857 (SD = 0.094).

Table 6.3: Summary of the linear model of the 11 chosen sentiment variables and number of tweets against depression measured at Survey 1. The lower half of the table gives summary information for the model overall.

Characteristic	Beta	95% CI	p-value
LIWC Verb (Mean)	4.1	-0.94, 9.1	0.11
LIWC Verb (Var)	14	-41, 69	0.6
Number of Tweets	1.6	-1.9, 5.1	0.4
VADER Positive (Mean)	-0.73	-2.5, 1.1	0.4
LIWC Leisure (Mean)	-0.38	-11, 11	>0.9
LIWC Quant (Var)	219	-84, 523	0.2
LIWC They (Var)	-268	-510, -27	0.030
VADER Negative (Var)	15	-2.9, 33	0.10
LIWC Female (Mean)	-5.2	-34, 24	0.7
VADER Negative (Mean)	-2.9	-7.8, 1.9	0.2
LIWC Focuspast (Mean)	3.7	-6.9, 14	0.5
LIWC Body (Mean)	31	2.6, 59	0.032
R ²	0.204		
Adjusted R ²	0.130		
p-value	0.002		
No. Obs.	142		

¹ CI = Confidence Interval

Anxiety

For the anxiety linear model *VADER Neutral* (mean) and *VADER Compound* (mean) were removed in favour of keeping *VADER Positive* due to high multicollinearity and the fact that *VADER positive* accounted for the highest variance of the three.

As such the theoretical model for anxiety was:

$$\begin{aligned} \log(\text{Anxiety_Survey1} + 1) = & \alpha + \beta_1(\text{vader_positive.mean}) + \\ & \beta_2(\text{liwc_leisure.mean}) + \beta_3(\text{labmt_emotion_valence.mean}) + \\ & \beta_4(\text{liwc_body.mean}) + \beta_5(\text{liwc_health.var}) + \\ & \beta_6(\text{liwc_quant.var}) + \beta_7(\text{liwc_health.mean}) + \\ & \beta_8(\text{n.tweets01}) + \beta_9(\text{liwc_friend.mean}) + \\ & \epsilon \end{aligned} \tag{6.6}$$

The results of the model in Table 6.4 show that anxiety was associated with increased use of words relating to the *Body* category, and that this was the only variable in the model with a p-value < 0.05. Fewer words in the categories of *Leisure*, *Health*, and *Friend* were also associated with anxiety, as well as less variance in the mention of *Health* and more variation in the *Quant* category. *Quant* includes words that are quantifiers such as ‘lots’ or ‘little’.

Repeated 5-fold cross-validation (10 repeats) found that the mean R^2 was 0.127 (SD = 0.084), and the mean RMSE was 0.791 (SD = 0.058).

Table 6.4: Summary of the linear model of the 8 chosen sentiment variables and number of tweets against anxiety measured at Survey 1. The lower half of the table gives summary information for the model overall.

Characteristic	Beta	95% CI	p-value
VADER Positive (Mean)	-1.7	-3.6, 0.16	0.073
LIWC Leisure (Mean)	-1.7	-12, 8.4	0.7
LabMT (Mean)	-0.05	-0.19, 0.08	0.4
LIWC Body (Mean)	44	17, 70	0.001
LIWC Health (Var)	-101	-409, 208	0.5
LIWC Quant (Var)	187	-60, 434	0.14
LIWC Health (Mean)	-15	-37, 7.0	0.2
Number of Tweets	0.51	-2.4, 3.4	0.7
LIWC Friend (Mean)	-5.4	-21, 11	0.5
R ²	0.200		
Adjusted R ²	0.147		
p-value	<0.001		
No. Obs.	147		

¹ CI = Confidence Interval

General Well-being

Lastly the same approach was used for estimating general well-being. Here, no variables were removed from the model due to multicollinearity. The theoretical model for general well-being was:

$$\begin{aligned}
 \text{Wellbeing_Survey1} = & \alpha + \beta_1(\text{liwc_verb. var}) + \\
 & \beta_2(\text{liwc_verb. mean}) + \beta_3(\text{liwc_money. var}) + \\
 & \beta_4(\text{vader_negative. var}) + \beta_5(\text{liwc_home. var}) + \\
 & \beta_6(\text{n. tweets01}) + \beta_7(\text{liwc_relativ. var}) + \hspace{10em} (6.7) \\
 & \beta_8(\text{liwc_focuspast. mean}) + \beta_9(\text{liwc_health. mean}) + \\
 & \beta_{10}(\text{liwc_time. var}) + \beta_{11}(\text{liwc_sexual. mean}) + \\
 & \beta_{12}(\text{liwc_shehe. var}) + \epsilon
 \end{aligned}$$

In the results of the general well-being model in Table 6.5 increased variance in the discussion of *Money* related terms was associated with a decrease in well-being, as was increased variance in the discussion of LIWC *Relativity* terms. Relativity refers to words describing how one thing relates to another and contains terms such as ‘down’ or ‘earlier’. The mean of *Health* related words was associated with increased well-being, and *Sexual* related words was associated with decreased well-being.

Repeated 5-fold cross-validation (10 repeats) was again used to get a more robust estimate of the model’s error, which found that the mean R^2 was 0.14 (SD = 0.101), and the mean RMSE was 7.923 (SD = 1.16).

Table 6.5: Summary of the linear model of the 11 chosen sentiment variables and number of tweets against general well-being measured at Survey 1. The lower half of the table gives summary information for the model overall.

Characteristic	Beta	95% CI	p-value
LIWC Verb (Var)	-292	-787, 204	0.2
LIWC Verb (Mean)	-33	-76, 10	0.13
LIWC Money (Var)	-3,135	-5,647, -623	0.015
VADER Negative (Var)	-96	-215, 23	0.11
LIWC Home (Var)	-1,368	-3,405, 669	0.2
Number of Tweets	-19	-46, 7.9	0.2
LIWC Relativ (Var)	-195	-420, 30	0.089
LIWC Focuspast (Mean)	-17	-115, 80	0.7
LIWC Health (Mean)	139	5.9, 273	0.041
LIWC Time (Var)	-174	-616, 268	0.4
LIWC Sexual (Mean)	-277	-516, -37	0.024
LIWC Shehe (Var)	-1,902	-4,653, 849	0.2
R ²	0.287		
Adjusted R ²	0.222		
p-value	<0.001		
No. Obs.	144		

¹ CI = Confidence Interval

6.4.3 What is the effect of changing the window and weightings of data?

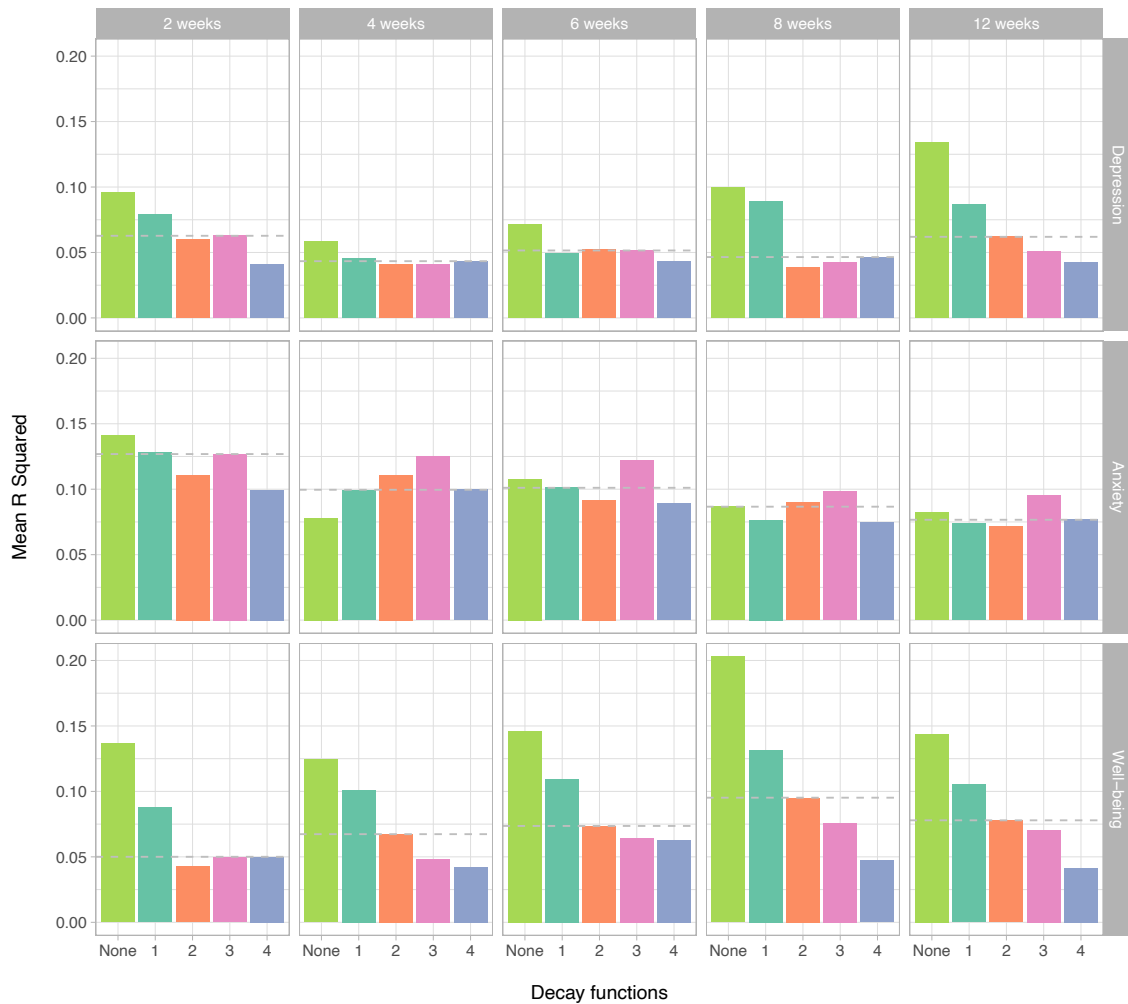


Figure 6.12: The mean R Squared value obtained from 10 x 5-fold cross validation plotted for the outcomes of depression, anxiety and well-being. The input data contains an increasing number of weeks for each graph from left to right, and within each graph the y-axis represents the different weighting functions used on that number of weeks of data.

Here we tested the impact of changing the window of Twitter data by two weekly intervals between two and twelve weeks. For each week I also tested the impact of weighting the continuous variables by decay functions (6.1) to (6.4), as well as using no decay at all. The models were run using the 151 participants from the Survey 1 time point, with Survey 1 as the source of mental health ground truth for depression, anxiety and general well-being. Each model's R^2 value is the mean of ten repetitions of 5-fold cross validation.

Overall, not weighting the data at all appears to be the most effective use of the data (represented by the ‘None’ option in bright green in Figure 6.12). However, we can see that anxiety is generally less effective as the time window increases, with longer windows preferring weightings that emphasise the contribution of the data closest to the measurement time point. Focussing on the unweighted results, depression starts reasonably high, reduces but then starts to climb again at 4 weeks and peaks at 12 weeks. General well-being tends to increase up to 8 weeks, and then reduces again. It is worth noting that the 4 to 12 week period for this data would have included the start of the COVID-19 pandemic.

6.4.4 How do predictions perform over time?

Given Figure 6.12 the next step was to use the models to see how well these predict outcomes at a future time point. For comparability, I continue to use the models trained on two weeks of unweighted data.

Analysis of residual error at Survey 2

By predicting the values of the second survey it is possible to assess how the models perform at a future time point, as well as whether the future predictions are biased by any particular characteristics of the participants. These results are given in Table 6.6 As can be seen in Figure D.4 in the appendix the residual error is roughly normally distributed, as so there does not appear to be any systematic under or over estimation. The R^2 values for depression, anxiety and well-being at Survey 2 were 0.002, 0.002 and 0.0009 respectively.

To investigate whether any particular characteristics were associated with prediction bias I tested the difference in Root Mean Square Error (RMSE) between the groups of sex, generation and also whether or not the individual being predicted was in the original sample. Ethnicity was not explored because it was only available for the G1 cohort which left only one participant in the ‘Ethnic Minority Group’ group. The gap between completion of Survey 1 and Survey 2 is around a month on average.

Table 6.6: Table containing the Root Mean Squared Error (RMSE) for each sub-group of those predicted at the second survey time point (Survey 2), using the model trained on data from Survey 1. The sub-groups include sex, generation and whether or not the individual being predicted was part of the original sample at Survey 1. P-values were calculated on the original errors using Welch's Test.

Predictions	Depression	Anxiety	General Well-being
Original Model (Mean, SD)	0.857, 0.094	0.791, 0.058	7.923, 1.16
All Survey 2	0.875	1.006	8.991
Sex			
Female	0.88	0.939	8.427
Male	0.842	1.128	9.655
p-value	0.015	0.003	0.117
Generation			
G0	0.784	1.004	8.135
G1	0.919	1.007	9.408
p-value	0.001	0.008	0.024
In training sample			
Yes	0.806	0.913	8.438
No	0.998	1.166	9.96
p-value	0.052	0.283	0.077

Predictions over time

Lastly, I used the trained model for each outcome to make predictions over time, and see how effective population models may be for this purpose. To do this I took the models trained on two weeks of unweighted data at Survey 1 and applied them to two weekly intervals of data from the 1st January 2020 to the 31st October 2020. The results of the predictions were averaged at each time point, and included any individuals who had tweeted in each two week window. This was not restricted to a minimum number of tweets, or to those in the original samples for Survey 1 and 2. The results of the mean predicted value at each fortnight is given in Figure 6.13 for each of depression, anxiety and general well-being. The graph is annotated with important national events in England across time. The same graph, but with prediction intervals included, is given in the Appendix in Figure D.5. It shows that the prediction intervals are very wide, and that confidence in trends is likely to be low.

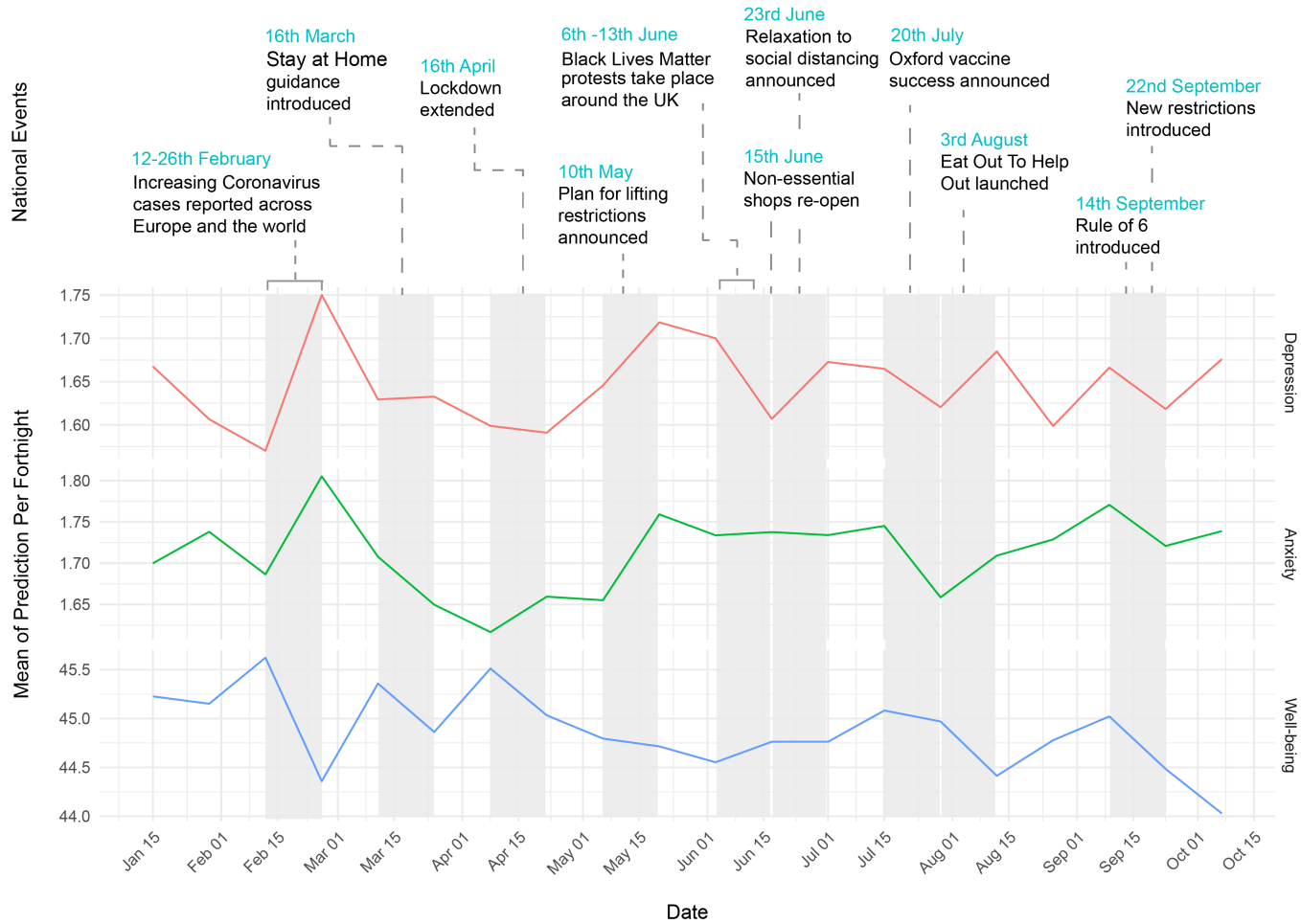


Figure 6.13: Each line illustrates the mean predicted values for depression, anxiety and general well-being as predicted by the linear models for each outcome. Predictions were made on data aggregated by individual over each two week period between the 1st January 2020 and 31st October 2020. The figure is annotated with key national events over the same time-period, particularly those regarding the COVID-19 pandemic, with grey horizontal bars indicating the two-week prediction period that each event happened within.

6.5 Discussion

This study used linked Twitter from two generations of a population cohort study to analyse how effective sentiment and pattern of life features were for inferring continuous measures of depression, anxiety and general well-being.

By using cohort data this study makes a unique contribution to the literature by examining sentiment features against known ground truth data at two different time points. Using this data also allowed for the analysis of potential demographic sources of bias in trained models, and to assess how well they performed on both within and outside sample predictions at a future time point.

Here I will discuss the results against each of the four original research questions, before discussing the overall implications of the study along with its limitations and suggested future directions.

6.5.1 How well do patterns of life and codings of emotion predict mental health?

Overall most standard sentiment codings of sentiment or affective outcomes in the LIWC accounted for between 2-8% of the overall variance in each mental health outcome. More variables accounted for variance in anxiety than depression or general well-being, with the association between *VADER Positive* and anxiety being the only association with a p-value that was under the bonferroni threshold. Previous work predicting stress from Facebook data did also find an important negative association between positive sentiment and stress [366]. Here, pattern of life variables were relatively well associated with the outcome [321], with the number of tweets in particular accounting for the highest amount of variance out of the five variables tested across all three outcomes. There has been mixed success of pattern of life variables in previous research into depression [321]. The variables related to the time of day of tweets were not as successful in this study as previous research suggested they might have been, with depression particularly thought to be associated with tweeting late at night [319, 406, 407]. However, this effect may have been lost due to the summarisation of time-windows in the ALSPAC data, which leaves only the groups 8pm to midnight, midnight to 4am and 4am to 8am. As a result I restricted the night window

to midnight to 4am (similarly to [321]), but previous research using the timings of 9pm to 6am found this variable to be a useful predictor [319]. Additionally, because there are likely to be fewer people tweeting in anti-social hours it is more likely that the detailed timing of these tweets was removed under ALSPAC statistical disclosure policy due to fewer than five tweets being made on that date and time. Future users of ALSPAC twitter data may wish to derive a custom binary variable that represents tweets within antisocial hours of 9pm to 6am, and using 3-hour windows for the ALSPAC data may still give appropriate anonymity but provide more detail for researchers.

Previous studies had seen that at a population level life satisfaction and happiness were both negatively associated with *LIWC Positive Emotion* [361] and that population self-rated mental health was negatively associated with *LabMT* and *VADER Compound* [381], but these results were not replicated in this research. I found that each of the standard sentiment codings were associated with each mental health outcome in the direction that would be expected. That is, that as depression and anxiety symptoms increase, the mean values of *LabMT*, *VADER Compound* and *LIWC Positive Emotion* decline. The reverse direction of effect was found for general well-being, which is to be expected. There is potential that since an increasing number of tweets was a predictor of poorer scores in all three mental health outcomes, over-representation of tweets from individuals who are more likely to have poorer mental health could skew the outcome data at a population-level. This could be why we did not find the same direction of effect as other population level studies [361, 381]. Whilst the effects did not have p-values less than 0.05, we did see that when disaggregated data is used (Table D.4) increased depression was positively correlated with increases in *LIWC Positive Emotion* and *VADER Positive*. Improved general well-being was also associated with increased *LIWC Anxiety* words. Previous work by Jaidka et al. [361] explored the reason for these contrary correlations found in their research, and saw they could largely be attributed to some positive words such as ‘lol’, ‘love’ and ‘good’ being highly used and also coded as ‘positive’. Similar patterns were true for high frequency negative words. When these words were removed from the *LIWC* dictionary the contrary patterns were no longer present. It is possible that grouping or aggregation by individuals could also be a useful method for addressing this feature in population mental health inference in the future, and it would be useful to test this with other population-level datasets.

6.5.2 Is prediction improved by using a larger number of linguistic categories?

When including all the additional categories of the LIWC, which include parts of speech lists and general purpose topics, each outcome had between eleven and thirteen associated features where $p < 0.05$. If the bonferroni adjusted p-value had been used then we would only have retained a single variable, which was *VADER Positive* for predicting anxiety. Taking the unadjusted threshold, many of the variables that are not related to affective processes were included as top associations with each outcome. There was also a combination of mean and variance of different categories, which suggests that in some instances the changes in discussion of a topic are more important than the mean number of times the topic is mentioned. This may be related to theory that increased variance may be an early warning sign of fluctuations in someone's mental health [408].

In a systematic review of studies predicting depression from Twitter that used a validated scale, *LIWC Past Focus*, negative emotions, anger words, and fewer words per Tweet were all found to be associated with depression [40]. *LIWC Past Focus* is thought to be associated with rumination, which is a common symptom of depression [409]. Our analysis did find an association between increased *Past Focus* and depression, and also highlighted the categories of *Leisure*, *Quant*, *They* and *Body*, of which both *They* and *Body* had p-values less than 0.05 in the final model. *Body* could reasonably be expected to be associated with higher depression symptoms given the strong link between depression and somatic symptoms [410]. Meanwhile the increased variance of the use of *They* pronouns was associated with decreased depressive symptoms. This may function in a similar way to the recurrent finding that increased use of I pronouns are associated with depression [319, 320, 411], in that increased variance in the used of *They* represents a healthy balance of discussing both the self and others. Interestingly, when the Twitter data were disaggregated, *I* pronouns were one of the LIWC categories associated with anxiety and depression.

Given that anxiety is under-researched in this area generally (see Chapter 5) there is little precedent for which categories would be expected to be associated with it. We saw, similarly to depression, increased anxiety was associated with increased use of words relating to the *Body*, and less words associated with *Health*. Without being able to investigate the text of tweets it is not possible to tell whether the importance of these words in anxiety was related to the training time point being at the start of the COVID-19 pandemic, but the presence of *Body* could also be explained by increased

somatic symptoms, which are a feature of anxiety as well as depression [412]. Fewer words in the categories of *Leisure* and *Friend* may be a reflection of the impact of anxiety on social functioning [413], with more anxious people less likely to discuss friends or leisure activities. Previous work on predicting anxiety from Reddit posts had found that first-person pronouns and anxiety related bigrams were important features in the prediction of anxiety [414], but the training dataset was built on anxiety related search-terms. The LIWC *Anxiety* category was not associated with anxiety at all in this study, which reinforces the value of using datasets that are not derived using the outcome of interest to predict that outcome.

Lastly, general well-being was the most successful model of all three, with an adjusted R^2 value of 0.22. Many of the LIWC categories included in the general well-being model, such as *Money*, *Home*, and *Health*, all relate to aspects of ones personal circumstances, which are factors in the measurement of subjective well-being [415]. The increase of *Sexual* related words being associated with decreased well-being is an interesting outcome, which may be explained by the high volume of swear words that are included in the *Sexual* category. Swear words have been found to be associated with higher depression in previous studies, which may explain the link [320, 321].

The fact that some categories overlapped between each of the depression, anxiety and general well-being models is encouraging given that all of these outcomes are fairly highly correlated. It is possible that the overlapping categories are linked to common causes of changes in each of the outcomes, for instance a difference in *Past Focus* between depression and well-being, or a difference in the use of *Health* related terms between anxiety and well-being. The relatively high overlap of categories of depression and anxiety (*Body*, *Quant* and *Leisure*) again may then be expected due to the high co-morbidity of these two disorders [416]. Having this overlap does suggest a limitation to the specificity of the models for each outcome, which could be tested by assessing the predictive performance of each model on the other two outcomes. A similar experiment by Kelley et al. did find that specificity was poor between correlated outcomes [417].

6.5.3 What is the effect of changing the window and weightings of data?

Unlike other studies that had success in modelling their data with a decay weighting we found that increasing the window past two weeks or introducing decay weightings did not improve the fit of the model in general [396, 397]. There were some instances, for example 8 weeks of data for well-

being and no decay, where the adjusted R^2 value was higher, but being able to use fewer weeks of data may have more practical use than achieving a better fit. It is also important to note that the period of data included in the extended window would have included Twitter data from the very beginning of the COVID-19 pandemic, and so this may have impacted the patterns found.

Since the two weeks of un-weighted data were used to select the variables for inclusion in each model we might find different results if the feature selection process was completed independently for each set of data. This could be achieved by repeating the original variable selection process used in this study for each model, or by using automated variable selection methods like the elastic-net for each new weight and time-window specification [418].

In summary, it is possible to achieve some gains in model fit by increasing the window of data, particularly for general well-being and depression, but that these improvements are generally not large and are likely to be outweighed by the benefit of using a shorter time period. Whilst there were slight differences in the patterns for anxiety, depression and general well-being we do not see big differences between each mental health outcome to suggest that each is relevant for different periods of time. However, replication of this analysis on data from outside the COVID-19 period would provide stronger evidence for these conclusions.

6.5.4 How do the predictions perform over time?

All suggested uses for Twitter data in the real-time monitoring of mental health require a model trained on data from a particular period of time to apply to future data. Uniquely in this study we had the benefit of a second ground truth time point in order to be able to test how each model performed at a future point. Whilst all the models generally had higher RMSE at the future time point, for depression and general well-being the error was within one standard deviation of the error variance estimated through cross validation in the original model. The error for anxiety was much higher, within four standard deviations of the originally predicted error. Testing the differences between all the groups revealed that predictions of depression in women, and anxiety in men were likely to be made with higher error. This could be explained by the fact that women show higher variability in their depression scores than men do, and that at Survey 2 men showed slightly higher variability in their anxiety scores than women did (Table D.2). Generation also created a distinction in prediction accuracy across all three mental health outcomes, with more error in the predictions

of the G1 cohort than the G0 cohort. Although there were more G1 participants in the training data than G0, G1 participants overall had poorer outcomes with wider variability in their outcomes than the G0 participants (see Table D.3), which again may be the reason for higher error in their predictions.

We also used the models to assess predicted anxiety, depression and well-being across the pandemic period. Encouragingly each of the outcomes predicted appeared to be related in terms of increasing poor mental health or positive mental health, but also behaved independently in terms of the steepness of the change. These predictions were generally coherent with events that occurred over the pandemic period, though at times the reasons for steep inclines or declines were not immediately obvious without access to the textual data that made up the tweets. For instance, predicted symptoms of depression and anxiety both increased between the end of August/early September, but general well-being also increased at this time. These predictions were made further away from the original training time-point and it is possible that the model was over-fitted on text that was highly related to the COVID-19 pandemic, and so as topics of discussion changed more error was introduced into the model.

Despite the apparent utility of these models for tracking changes over time, their prediction error (as seen in Figure D.5) was very wide, and is likely to mean that only general trends at a population-level can be reliably inferred from Twitter data. This is consistent with previous research that found whilst there were associations between Twitter data and mental health outcomes these associations were weak [417]. Modelling of individual-level mental health outcomes is likely to require accounting for different individual baselines [419]. Similarly, population models may be improved if they were trained separately on significant groups such as gender and differing age-brackets, which when aggregated can mask important differences in the impacts of events on different groups [395, 420, 421].

6.5.5 Strengths and limitations

There are several strengths to this study which give it a unique perspective on the challenge of inferring mental health from Twitter. First, the quality of the ground truth data is a significant strength of this study, especially considering the generally poor quality of data in the field as a whole (see Chapter 5). This study has the benefit of using a novel linked Twitter dataset that has

been derived from a well-known and characterised population cohort. As such, I have been able to analyse and model Twitter data against known characteristics of the Twitter users included in the study, and this has also enabled three different mental health outcomes to be explored within the same methodology. Second, having two ground truth time points has allowed me to test the model at a different time-point from the one at which the model was trained, as well as testing the prediction for those who were or were not in the original sample.

This being said, the study is not without limitations. One of the major limitations is the timing of the ground truth time points, which coincide with the beginning of the COVID-19 pandemic. This timing may well effect the generalisability of the findings of this study, and also have created perturbations in the data that increased the likelihood for model error. A second limitation of this study is that the sample was more limited than others in this area, with under 200 participants. These participants were roughly equal to the main linked sample in terms of their distribution of sex and generation, and in their mental health outcome scores. Knowing from Chapter 5 that men are more likely to be higher Twitter users than women, it is also likely that women were over-represented in this sample and therefore the models derived may not generalise as well to Twitter data as a whole.

The last main limitation of this study is that it was not possible to access the raw textual data from participants' tweets. This prevented a more systematic exploration, for instance using word-shift graphs as illustrated in previous research [235, 329, 376], to gain a more robust understanding of the changes in the underlying topics of discussion over time. This also would have allowed us to understand if particular words were influencing the strength of the chosen sentiment categories, such as whether or not *LIWC Sexual* was important because of the large number of swear words it covers. Since we could not interact with the raw textual data we were also unable to test whether differences in pre-processing steps made differences to the model outcomes, though this would be an interesting avenue for future research.

6.5.6 Future directions

Based on this study there are many interesting future directions for research. One of the main ways in which these findings could be further explored is by considering more future ground truth time points. The ALSPAC COVID-19 data was collected four times within the first year of the

pandemic, and by extending the analysis completed here to future time points it would be possible to test whether, and by how much, model error increases as a function of time from the training data point. Similarly, planned improvements to the data collection software will soon allow for every user's entire Twitter history to be collected, rather than just their most recent 3,200 tweets. This will allow future studies to consider ground truth time points further back in time, some of which include a greater variety of mental health outcomes such as the variety of well-being measurements explored in Chapter 2. The use of more historical ground truth and Twitter data would also support questions about longer term interactions between mental health and Twitter.

A further development based on the data considered here would also be to test the benefits of training models for population sub-groups who are likely to have different experiences of mental health outcomes. This is particularly the case for gender, where it would be of benefit to have a more recent variable that describes participant genders, as opposed to their sex assigned at birth. Future advances in individual level prediction are likely to require methods that can account for individual baselines of affect [419].

Whilst this research has begun to touch on the impact of individual characteristics on model bias we are still far from understanding how algorithms to model mental health behave in relation to common issues such as algorithmic bias based on cultural or gendered presentations of mental illness, since very few studies have access to ground truth data that can accurately capture the characteristics of their training sample (see Chapter 5 for a more detailed discussion). On an aligned topic, it would be beneficial to have a clearer understanding of how models behave when making out of sample predictions, with research so far suggesting that performance is fairly poor [335]. This was explored in this study, but with a small sample where the out-of-sample group were drawn from the same original study and so were likely to share many characteristics. Testing models on completely different datasets would be informative for understanding when and how different models are most and least accurate.

Lastly, when comparing the categories chosen for the three mental health outcomes it seemed likely that co-morbidity and correlations between depression, anxiety and general well-being were reflected in the overlaps of the categories best associated with each outcome. If individual symptoms, or thematically similar groups of symptoms [412], are considered rather than just the total outcome measure it is possible that stronger and more specific associations may be found that can

be explained by common causes or common symptoms. The relative success in this study of general well-being prediction also suggests that it is a fruitful area for further exploration in general, and an important population-level outcome.

6.6 Conclusion

This study has explored several questions relating to the use of Twitter data for inferring trends in mental health that have methodological relevance to the modelling of different mental health outcomes. For clarity I will briefly summarise the main findings and their implications for future research:

- Codings of emotion using sentiment dictionaries are generally associated with mental health outcomes, but the relationships are not strong and generally not specific to any mental health outcome. One exception is the strong negative relationship observed between mean *VADER Positive* sentiment and anxiety. Therefore it is not advised to use any single sentiment coding as a proxy for a mental health outcome (Section 6.5.1).
- Similarly, when considering more than one feature at a time model error and prediction bounds are high, so a generic model using sentiment dictionary features is highly unlikely to provide sufficient sensitivity or specificity for individual-level prediction (Section 6.5.2).
- When aggregated by individual, variance in sentiment features can be more strongly associated with a mental health outcome than the mean value of those features (Section 6.5.2).
- General well-being is explained better than depression or anxiety by sentiment and pattern of life features of Twitter data. Depression was the least successful of all three, which suggests that anxiety and general well-being (along with other forms of well-being such as happiness) are good candidates for future research in this area (Section 6.5.2).
- Accounting for the grouped structure of individuals over time within Twitter datasets may be important for the accurate prediction of population-level outcomes using sentiment. This may be due to the highly variability in tweet frequencies between individuals at a population-level and the association between higher tweeting frequencies and poorer mental health outcomes. This should be investigated further (Section 6.5.2).
- In general two weeks of Twitter data is broadly sufficient for modelling, but different mental health outcomes show signs of fitting somewhat better to different lengths of training data

(Section 6.5.3).

Further research into mental health inference on Twitter using data with more ground truth time-points would allow us to test how model error changes as the time from the original trained model increases, and therefore how effective these models are likely to be in the future. Testing models across different datasets would also be informative in terms of the assessment of the portability of models across contexts, and to better understand potential bias. Lastly, future research accounting for co-morbidity between disorders and correlations between mental health disorders and positive well-being is suggested as a means of unpicking associations that are specific to symptom groups rather than individual disorders.

Chapter 7

Discussion

This thesis has taken an interdisciplinary approach to understanding how we can make use of cohort studies to improve research into digital phenotypes for mental health. Across the empirical research chapters, which are summarised in Table 7.1, we have found that linkage is both acceptable and feasible, yielding a novel dataset that has allowed us to test the utility of common methods for inferring mental health from social media. Since each chapter has its own discussion of strengths, limitations and future directions this overall discussion will give a broad overview of the main messages from the findings as a combined body of work. I first give a commentary on the potentials and limitations on three main themes: the use of cohort studies as sources of ground truth for mental health digital phenotypes, how we develop digital phenotypes for the measurement of mental health, and the issues of ethics and acceptability in these areas of research. I will then discuss directions for future research across the field.

Table 7.1: A summary of the main findings and subsequent implications from each empirical chapter of this thesis.

Chapter	Main Findings	Implications
2. The mental health and well-being profile of social media users	<p>We saw that women were more likely to use social media more frequently in general, but that users of different social media platforms have different mental health profiles and are made up of different demographic groups.</p> <p>Different well-being measures are associated differently with the outcomes of social media use frequency and platform of use.</p> <p>In terms of Twitter in particular, both male and female Twitter users have higher rates of depression than the general sample population, but lower rates of suicidal thoughts and self-harm.</p> <p>In the areas of relatedness, satisfaction with life, optimism and gratitude, male users of Twitter have higher well-being than the general sample of men, but female Twitter users have lower well-being than the overall sample of women.</p>	<p>We should not assume that rates of mental health disorders are the same across all social media platforms.</p> <p>Similarly we must pay attention to the different measurements of well-being, rather than assuming one can be a proxy for all of the others.</p> <p>Studies should also ensure that they adjust for sex/gender in their analyses.</p>
3. Participant views on social media and its linkage to longitudinal data	<p>Participants were supportive of the collection and linkage of their data to support health and social care research.</p> <p>Having trust in the cohort was an important factor in this outcome.</p> <p>However, concerns were expressed about the collection and use of friends' data without their knowledge.</p> <p>When given a specific scenario of how this data could be used a sub-group did say they would agree for it to be collected.</p> <p>Participants were most agreeable to their data being collected if it was anonymised before researchers were given access to it.</p>	<p>Trust is a core issue in social media data linkage programmes, with cohort studies particularly well placed to develop these programmes given their existing relationships with participants.</p> <p>The ability to ensure that researchers would not have access to raw textual data was important to participants.</p> <p>Linkage programmes should be aware of the lack of social license for 'blanket' consent to collecting data that belongs to participants' friends.</p> <p>However, providing specific information about the data being collected and how it will be used may give enough information to allow participants to make an informed decision.</p>

<p>4. Twitter data linkage: features of the consenting participants and their data</p>	<p>19.6% of Twitter users in ALSPAC opted in to linking their Twitter data, and 15.3% were ultimately successfully linked.</p> <p>Linked participants represented Twitter users in ALSPAC in terms of their demographics, but had marginally higher rates of depression and anxiety than the general cohort sample. Given the results from Chapter 2 the higher rates of depression could be attributed to linked participants being from the sample of Twitter users, rather than opting in to linkage.</p> <p>There was a very wide range of tweet frequencies in the linked participants, and a quarter of participants only tweeted six times in the past year. Differences in tweeting frequency were not associated with sex or generation of participants.</p>	<p>The populations that agree to social media data linkage are broadly representative of the general population of Twitter users, which is positive for the generalisability of future research.</p> <p>A feature of population representative data from Twitter is that it is highly variable in quantity, and this asymmetry in the data should be accounted for in analysis of population-level data.</p> <p>Future data linkage programs may benefit from using face-to-face opportunities to request consent, which may result in higher consent rates.</p>
<p>5. A review of methodologies for monitoring mental health on Twitter</p>	<p>165 studies had attempted to predict individual mental health outcomes from Twitter data between 2013 and 2021, with 45% of all studies published in 2020 and 2021.</p> <p>Depression and suicidality were the most common outcomes studied, with other disorders being relatively understudied and only one study considered positive well-being.</p> <p>Only 13% of the datasets used had access to information about participant mental health that had not been inferred from their Twitter data.</p> <p>The quality of methodological reporting was generally poor and had not improved with time.</p> <p>In the subset of 100 studies where inclusion of ethics was assessed, 85% did not include any mention of ethical considerations.</p>	<p>Advice from previous reviews on better reporting of methodologies and increased consideration of ethical issues does not appear to have made an impact on the research being published.</p> <p>Without being able to compare the detail of different research questions and approaches it is not feasible to analyse patterns of what works in this area.</p> <p>Since most studies do not have ground truth data about the people whose data they are using there is little understanding of potential bias in the models being created.</p> <p>Similarly, there is minimal understanding of out-of-sample performance of most proposed models.</p> <p>At present the research in this field is still highly exploratory, and requires significant attention to methodological and ethical issues in order to make progress.</p>

<p>6. Modelling mental health using linked Twitter data</p>	<p>Here we found that whilst most positive codings of emotions are associated with improved mental health, and negative codings with worse mental health, these associations are generally not very strong.</p> <p>These associations were contrary to some previous research on population level mental health, which we concluded was due to a lack of individual aggregation causing spurious correlations.</p> <p>We found that positive general well-being was best associated with life-style topics such as *Money*, *Health* and *Home*, rather than positive sentiment.</p> <p>An increasing number of tweets was associated with worse mental health outcomes across depression, anxiety and general well-being.</p> <p>Two weeks of Twitter data was broadly sufficient for inferring mental health outcomes.</p> <p>Models predicted relatively well at a time point a few weeks into the future, and error was mostly seen between men and women, and the two generations.</p>	<p>Linked data from a population sample provides new perspectives on frequently studied associations, and accounts for the natural variation found in the population that is usually not captured by Twitter datasets.</p> <p>Population-level research may benefit from aggregation of individual level tweets since variable data production may generate spurious associations.</p> <p>Both well-being and anxiety are promising areas for future research using Twitter data.</p>
---	---	---

7.1 Commentary on thesis themes

Three central themes of this thesis were the use of cohort studies as sources of ground truth for mental health digital phenotypes, how we can develop digital phenotypes for the measurement of mental health, and the acceptability of these areas of research. Here I will give a brief overview of the conclusions under each theme. This includes an overview of the strengths and weaknesses seen in each area, both in relation to the approach taken in this thesis, and the field in general.

7.1.1 Contributions of cohorts

The core theme of this thesis was the process of developing ground truth data for digital phenotyping using a birth cohort study. Using a cohort study for this purpose is a novel approach to sourcing data for digital phenotyping, and the research presented in this thesis has effectively served a dual

purpose of using the data and evidence obtained from the cohort to conduct new research whilst also testing the effectiveness of linking and using cohort data for this type of research. This integration of new data sources, including social media, into birth cohorts is a core aim of long-term strategies for cohorts by the Medical Research Council [113] and the Wellcome Trust [122]. As a result, an important outcome of this thesis is an assessment of the strengths and limitations of conducting social media data linkage in a birth cohort, which is useful to other cohort studies in the UK and around the world who are attempting to do the same.

To begin with, we saw in Chapter 3 that social media data linkage can be acceptable to cohort participants, provided that conditions for anonymity and boundaries around data collection are met. These conditions are practically feasible, especially with the use of specialised software for this task [117], and are in line with existing guidance on the ethical sharing of social media data [98]. Acceptability among participants is crucial given that a lack of ‘social license’ can cause reputational damage and undermine trust in cohorts which has implications for their long-term success [200]. In Chapter 4 we also saw that those who opted-in to link their Twitter data were generally representative of those using Twitter in the cohort, which supports the technical feasibility for the use of this data for understanding population-level trends. Given the results of Chapter 2 we can also see how a representative population of Twitter users may differ from the general population. Having this linked cohort data means that we can understand samples we are basing our analyses on in much greater detail, which was a key limitation identified in the existing literature in Chapter 5, and elsewhere [144, 344]. In this way population cohorts differ from other research samples in that we always know who is not represented in the sample, not just who is. This is key for understanding where bias might be present in the data, or might be generated from using this data to train predictive models. The strengths of having linked cohort data can be seen in Chapters 2 and 6, where individual-level accurate data about a population means that we can make robust conclusions, and also explain divergences in results seen elsewhere in the literature. This is partly *because of* the asymmetry we see in the rates of data production between participants which, whilst a limitation for some modelling methodologies [83], actually allows us to understand how representative data on Twitter behaves and responds to certain analyses, rather than assuming all users tweet at consistent rates or that tweeting rates are unrelated to outcomes of interest, like mental health status. The strength of cohorts for social media data linkage is also supported by the

availability of multiple time points, which is a key difference between linkage in this sample and other panel survey samples [83, 109].

These explorations into the utility of linking social media data in a birth cohort illustrate many strengths of the method, but like all data collection methodologies there are also limitations to this approach. One of the main limitations is that there is a hard limit on the size of the linked sample. Unlike in solely Twitter-based research where researchers can continue to collect data until they are satisfied that they have enough tweets or users, the linked sample is limited to those who have opted in. This does present a limitation to sample sizes, and as we saw in Chapter 6 due to the asymmetry in the tweeting frequency of populations it is unlikely that the whole linked sample will have tweeted within the period of interest [83]. The other limitations of social media data linkage in birth cohort studies are aligned to the limitations of cohort studies in general. Firstly, due to the long running nature of cohorts, measurement of certain concepts may change over time or be dependent on external funding, and so harmonisation of different types of measurement is a challenge. This might mean, for instance, that depression is measured at different times to other mental health outcomes, and at different time points may be measured with different tools. In ALSPAC this challenge extends to the differences between approaches to surveying the different generations, such that measures for each generational group are likely to be taken at different times and different frequencies. Another general limitation of cohort studies is that they are generally not suitable for studying rare disorders, which may limit the usefulness of using birth cohort data for training models on outcomes such as bipolar disorder, schizophrenia and personality disorder which are relatively popular in the field of mental health inference (see Chapter 5) but relatively uncommon in population samples. However, the population representative nature of the cohort data is likely to mean that it is highly suitable for testing the sensitivity and specificity of trained algorithms, although those suffering from acute mental health disorders are more likely to have been lost to follow up over time.

A final consideration in the use of linked data from cohort studies is the anonymised nature of the data, which is a condition of access desired by participants (Chapter 3) and implemented by the cohort's data management team (Chapter 4). Rather than being a limitation, this should be framed as a methodological consideration for future researchers and cohorts in the careful balance between privacy and trust [422]. By being able to access data that contains more personal and

sensitive information about participants, researchers must be willing to sacrifice some of the ease of access to this data, though the data should still be useful and usable. These sacrifices may mean the data is anonymised, or that it can only be accessed inside a secure research environment. Having had the opportunity to test the experience of working with anonymised data in this thesis I found that anonymisation does limit some detailed understanding of the behaviour of the data (as discussed in Chapter 6), but that this could be mitigated if it was straightforward for researchers to request updated data that allowed them to explore trends ad-hoc. For instance, extracting the most common words from each category in the LIWC for each user would be highly useful, and still does not require access to the raw text of a tweet. Similarly, allowing researchers to submit code to train their own models, such as language models, may be a solution, although such methods are known to have privacy flaws [423]. Potential for privacy leaks from outputs such as language models can also be difficult to assess, especially by those who are not subject matter experts, and this may be a barrier to allowing this type of feature construction from cohort datasets. Other solutions currently being explored include the potential for differential privacy methods, which involve using algorithmic methods to generate useful information about the data without disclosing information about any individual [422].

In summary, we saw encouraging results from the programme of Twitter data linkage in ALSPAC. We suggest that this method of dataset development addresses important limitations of existing datasets and can support important advances in the quality and robustness of results from digital phenotyping studies in the future.

7.1.2 Digital phenotyping for mental health

A second aim of this thesis was to explore the utility of using digital phenotypes to infer mental health. Here I will briefly cover the benefits and limitations of the use of Twitter for mental health inference that I have encountered, as well as the limitations of our measurement of mental health ground truth in general.

The proposed applications of digital mental health inference using social media have been discussed throughout this thesis, and its benefits could be said to come down to *timeliness* and *scale*. As seen in Chapter 6, even with large error margins it *is* possible for Twitter data to provide signals of changes in mental health at a population level much quicker than a traditional survey method

could, which is encouraging. This was using a relatively basic model of multiple regression, but other modelling methodologies that make use of machine learning techniques, as seen in Chapter 5, may be even more successful. There is a trade-off here though, with more opaque techniques potentially being better at making correct inferences but being harder to understand. This trade-off is an area that would particularly benefit from further engagement with the proposed users of these future technologies to understand what level of explainability and transparency would be required for adequate interpretation and trust.

Another of the primary benefits of social media data for mental health inference that I discussed in the introduction to this thesis (Section 1.3.1) is the availability of textual data to use as features for modelling, which sets social media apart from other potential digital phenotypes. We saw in Chapter 5 that textual features are popular, and I have illustrated the use of textual data for sentiment modelling in Chapter 6. However, patterns in word use and linguistics are likely to differ significantly between individuals, groups and settings due to gender, cultural influences, neurodiversity and more [35, 424, 425]. An over-reliance on these methods, sentiment and keyword lists in particular, may result in models that have limited generalisability. This is less likely to be an issue at a population level, particularly for models that are created for specific geographic or demographic populations.

This thesis, and the field in general [41], is largely focussed on the use of the social networking site Twitter, which has its own benefits and drawbacks. From results in Chapter 2 we know that Twitter is one of the social media sites that people interact with least frequently day-to-day. However, it is also one of the sites whose data is easiest to access for researchers. This means that, at the time of writing, Twitter is one of the most feasible platforms with which to conduct population-level health monitoring, since its data is available for this purpose. When it comes to individual-level inference though, data from Twitter is unlikely to be sufficient data source for digital phenotyping in the majority of people (see Chapter 2), and it is likely that more representative and context-aware models would be derived from combinations of data from different platforms. This avenue of work is currently limited by the availability of data from other social media sites, which also restricts the applicability of any research completed using these platforms, even if the data could be accessed for research purposes.

Lastly, this thesis has concentrated on mental health as measured by validated scales, but there

are many different perspectives from which to study and model mental health and its associated behaviours. For instance, there is growing evidence for network or hierarchical perspectives on mental health which move away from traditional models of disorder classification that are set out by diagnostic manuals like the DSM [346, 426], and which may even allow for the incorporation of environmental and genetic effects into our understanding of the interplay between symptoms and behaviours [427]. Outside of disorders we should also consider how we can use digital phenotypes to measure positive mental health. As seen in Chapter 6, and elsewhere in the literature [53, 235], there is certainly evidence that this approach can be successful, which supports the idea that ‘good’ mental health should be more than just the absence of illness, but also the promotion of well-being [428].

7.1.3 Ethics and acceptability

The ethics of digital phenotyping for mental health is an integral and important part of the field. In Chapter 3 I investigated the acceptability of social media data collection and research for cohort study participants, and found that in general participants are supportive of these activities, particularly at the population level. This is consistent with previous findings about users’ views of what mental health inference on Twitter could and should be used for [205]. In Chapter 5 I reported that, despite calls for greater engagement with ethical issues in the field of mental health inference from social media data, the vast majority of studies continue not to include mention of potential ethical issues of this research. It is a concern that by not considering ethics along with our development of these technologies that we build up *ethical debt* [430], a term coined by Fiesler and Garrett for putting off ethical considerations of new technologies until they are deployed, similarly to how *technical debt* refers to the prioritisation of deployment of technology over the quality of the code itself.

Conversations about ethics in this field tend to focus on the important issues of privacy and surveillance [431], but there are a wide variety of other aspects to consider, from the well-being of researchers themselves to technical decisions about performance trade-offs [102]. New ethical frameworks for AI, such as the Data Hazards project discussed in Chapter 6 [405], are efforts to recognise these wider range of ethical issues that can arise from data-based research and encourage researchers to consider them more frequently. Using linked data from cohorts does allow

researchers to address some of these concerns in a way that has not previously been possible with social media data. For instance, informed consent is obtained from all participants, and their data is securely stored and distributed by a trusted, central team [98, 116]. There being no need to manually label data prevents the risk of psychological harm to those who may otherwise be responsible for labelling data [102]. Additionally, we can understand and account for bias in models rather than it being an unknown entity [35, 342]. This presents an exciting development in the ethical quality of social media data available for research, and is an exemplar of how participants and researchers can work together to derive shared boundaries for the safe use of their data [123].

Other issues of ethics relate more to individual decisions made by data scientists themselves, who often make decisions of how to model and represent the social constructs that they are representing, in this case mental health. Barocas and Boyd argued that “[d]ata scientists engage in countless acts of implicit ethical deliberation while trying to make machines learn something useful, valuable, and reliable”, describing how decisions of how to clean data, which algorithm to use and how to weigh interpretability against accuracy are all ethical decisions, though they may not be described as such [432]. Our statistical representation of a social construct, like mental health, is embedded with our assumptions about how it should be understood and modelled [263]. It is in these decisions that the influence of researcher reflexivity, positionality and priorities are likely to become important [433]. This is not to say that such influences are negative, on the contrary, variation in research approaches is how we achieve progress and new ideas, and is an inevitable part of the research process. However, understanding how our assumptions influence the solutions we create, especially when they have the potential to be widely used and concern highly personal human attributes, can help us to develop more inclusive and fairer technologies. For instance, believing that identification of those with a given disorder is the highest priority may result in a focus on reducing false negatives at the cost of increased false positives. Conversely, those who believe that incorrect intervention in someone’s life is more damaging than not intervening when someone has a mental health disorder, may instead choose to prioritise the reduction of false positives.

7.2 Future directions

In the previous sections I have touched on the potentials and limitations of the techniques and methodologies used both in this thesis and in the field as a whole. Here I propose what I believe

will be valuable directions in future research on the basis of these observations.

Firstly, given the encouraging results from the data linkage project described and used in this thesis, future development of social media data linkage programmes in cohorts is recommended. This would be especially powerful if conducted across diverse cohorts in terms of age and geographic location. Future research may then include harmonising data across multiple cohorts to develop safe and controlled opportunities to test models on a variety of population groups in a way that manages data appropriately.

In terms of digital phenotyping research itself, one of the first areas that I believe will be beneficial for future research is not so much a research opportunity as an adaption of how research in this area is currently conducted. This is to be more explicit about the nuance in the question of whether we can infer mental health from social media, by specifying exactly which research question we are trying to answer.

Even subtle changes to the problem specification change the ethical concerns, dataset requirements and modelling methodologies that are relevant to the task at hand. For instance, human rated assessments of individual tweets that are labelled for ‘risk of suicide’ or ‘no risk of suicide’ are likely to be appropriate for a system that intends to flag tweets that should be followed up by a trained human whose role is to offer support. In this case the system is aiming to automate a human screening process. However, this dataset may not be as useful for measuring suicide risk in a large population over time. Similarly, models and datasets used to understand population mental health over time are unlikely to be suitable for measuring individual-level changes in suicidality [434, 435]. These distinct approaches to “inferring mental health with social media data” have unique ethical challenges, require different modelling and validation processes, different considerations of computational efficiency trade-offs, and potentially models with different levels of explainability. By capturing this nuance we can make sure that critique of datasets and approaches are made with respect to the research aims, and also focus comparisons and progress against aligned research questions.

In being more specific about our questions we can also more readily understand how research programmes will benefit from interdisciplinarity, and which disciplines should be involved. For instance, as well as generally benefiting from input from clinical psychology as has been suggested previously [41, 57, 102], population-level monitoring of mental health is likely to benefit from the

involvement of public health professionals and health policy makers as the parties who are to be the end users of these data. Additionally, experts in human-computer interaction (HCI) can support understanding of how these new technologies could be communicated most effectively for decision making [436, 437]. This being said, interdisciplinarity has its challenges too. From different expectations of project timelines and outputs, to different technical vocabularies, it takes time and effort to build a strong interdisciplinary collaborations that allow for useful and meaningful input from all parties. On this basis, we need to ensure that involvement of experts from different fields are not included as ‘check-box’ exercises, but that those with different types of expertise are all given the opportunity to influence this field in a meaningful way. For example, as seen in Chapter 5, it is relatively common to involve mental health experts in data labelling but not in defining the research question, which misses an opportunity to invite critical expertise at an important stage of the project.

Another important next step in the field is working out how to involve potential users of these technologies in their development, particularly those who are from marginalised groups, to better understand how humans and algorithms can work together in the provision of mental health care [438, 439]. An important component of systems that have been found to cause harm and reinforce structural inequalities in the past is that they are employed in settings and institutions that inherently feature power imbalances, like policing, statutory social care, or health insurance [23, 440, 441]. These are settings where it is likely that training data will reflect existing inequalities since it is those inequalities which are often causally related to the outcomes they are predicting. Mental health services are, by design, a setting with significant power imbalances [442, 443], with this power intended to be used to protect those who are most vulnerable, although often experienced as oppressive by those within it [444]. In the UK people who are Black African or Black Caribbean are proportionately more likely to be sectioned than any other ethnic group [445], and there is evidence that people from marginalised groups are more likely to be misdiagnosed with mental health conditions [446]. This not only affects our training data, but also leads to questions of how the social power of an algorithmic system may influence outcomes in settings where mental health care is provided [447]. For instance, women in general are less likely to be considered credible by medical professionals, with this impact compounded for Black women [448, 449]. By involving a diverse group of citizens in the exploration of these technologies we can hope to address risks that

could arise from their deployment and avoid foreseeable future harms as best we can. Doing so is likely to mean that we also need to work on how best to involve lay people in the development of data science projects. This is something that other fields have done with great success such as Public Patient Involvement groups, citizen scientists and peer researchers in health and medical research, or participatory/action research in health and social care [450].

Lastly, we can also use social media data to improve our understanding of mental health. In particular there are interesting avenues in network representations and co-morbidity of mental health disorders, and the temporality of different mental health constructs [386] which social media data may be well placed to inform as an alternative to other resource intensive methods such as ecological momentary assessment [382]. This can also include understanding of the reciprocal interactions between mental health and social media which were discussed in Section 1.3.2 [345]. In doing so, and by continuing to use ground truth data with better specified samples, we could also use social media to develop more robust understandings of the impact of key life events, such as moving away from home, starting families or measuring well-being over the lifespan [451].

7.3 Conclusion

Overall we have seen that social media data linkage in cohort studies is acceptable, feasible and produces novel benefits, particularly in the availability of population representative and longitudinal data as well as the ethical sharing and storage of social media data from participants. There are limitations of these data that align with common limitations of cohort study data in general, but generating new datasets with different limitations to those that exist currently allows us to explore new questions and address different concerns that we have been unable to previously. Twitter is still the most straightforward platform to achieve this outcome with, and restrictions on other popular platforms like Facebook and Instagram present a challenge to digital phenotyping research that can only really be resolved by the corporations who control this data.

Continuing to develop safe, trustworthy and acceptable methods for mental health inference from digital phenotypes can allow us to achieve new understandings of mental health which advance the treatment and prevention of mental illnesses, as well as promote positive mental well-being.

Appendix A

Chapter 2

A.1 Additional sample information

Table A.1: Percentage of the users of each social media site by use-frequency and demographics reported.

Platform	Frequency	Sex		Ethnicity		A Levels		Parental Employment Class	
		Female	Male	Minority Ethnic Groups	White	No A Levels	A Levels	Non- Manual	Manual
Facebook	Daily	90.6	81.8	86.5	87.5	86.7	88.7	87.4	88.8
Facebook	Less	7.0	15.2	9.0	10.2	10.8	8.9	10.1	8.9
Facebook	Never	2.4	3.0	4.5	2.3	2.5	2.3	2.5	2.3
Twitter	Daily	13.4	20.8	9.8	16.5	16.1	15.6	15.9	17.5
Twitter	Less	42.5	36.4	48.5	40.3	35.6	42.6	41.4	37.7
Twitter	Never	44.1	42.8	41.7	43.2	48.3	41.8	42.7	44.9
Instagram	Daily	57.9	31.0	53.8	48.7	43.6	49.3	48.7	49.3
Instagram	Less	18.4	23.2	15.2	19.9	24.2	18.4	19.1	21.7
Instagram	Never	23.7	45.8	31.1	31.4	32.2	32.3	32.2	29.0
Snapchat	Daily	39.9	28.8	44.7	35.9	40.6	33.9	33.9	43.0
Snapchat	Less	33.6	35.2	28.8	34.1	31.3	36.2	35.3	29.4
Snapchat	Never	26.5	36.0	26.5	30.0	28.1	29.9	30.8	27.5
YouTube	Daily	22.5	48.4	34.1	31.6	29.2	30.8	31.1	34.2
YouTube	Less	45.8	34.9	40.9	41.8	41.2	42.8	42.5	39.3
YouTube	Never	31.7	16.7	25.0	26.6	29.6	26.4	26.5	26.5

A.2 Descriptive data on mental health and well-being outcomes

Table A.2: The percentage of the sample who had experienced each of the four categorical mental health outcomes.

Characteristic	Percentages by Sex	
	% Female (CI)	% Male (CI)
Depression	22 (20, 24)	16 (14, 18)
Disordered Eating	10 (9.3, 12)	2.5 (1.7, 3.5)
Suicidal Thoughts	18 (17, 20)	17 (15, 19)
Self-Harm	9.7 (8.7, 11)	4.8 (3.7, 6.1)

Note:

Depression was measured in the sub-sample (N=2,862)

¹ CI = Confidence Interval

Table A.3: Summary statistics for well-being outcomes, all measured in the sub-sample (N=2,862).

Characteristic	Mean (SD) by Sex		Min. Value	Max. Value
	Female	Male		
BPN (Autonomy)	5.12 (0.92)	4.99 (0.89)	1.00	7.00
BPN (Competence)	5.03 (1.06)	4.98 (1.04)	1.00	7.00
BPN (Relatedness)	5.68 (0.91)	5.47 (0.90)	1.25	7.00
Satisfaction With Life	23.97 (6.60)	23.15 (6.72)	5	35
MIL (Presence)	23.60 (6.54)	22.36 (6.89)	5	35
MIL (Search)	19.88 (7.07)	20.23 (7.15)	5	35
Life Orientation Test	13.30 (4.59)	14.16 (4.51)	0.0	24.0
WEMWBS	48.33 (8.87)	49.67 (8.97)	14	70
Gratitude Questionnaire	35.15 (5.72)	33.47 (5.79)	7.0	42.0
Subjective Happiness	4.86 (1.26)	4.82 (1.31)	1.00	7.00

Note:

Basic Psychological Needs (BPN)

Meaning In Life (MIL)

Warwick Edinburgh Mental Well-being Scale (WEMWBS)

Table A.4: Contingency table of suicidality and disordered eating (N=4,083).

Characteristic	Suicidality		
	No	Yes	Total
Disordered Eating			
No	77%	15%	92%
Yes	5.4%	2.3%	7.7%
Total	82%	18%	100%

Table A.5: Contingency table of suicidality and self-harm (N=4,083).

Characteristic	Self-harm		
	No	Yes	Total
Suicidality			
No	79%	2.9%	82%
Yes	13%	5.2%	18%
Total	92%	8.1%	100%

Table A.6: Contingency table of disordered eating and self-harm (N=4,083).

Characteristic	Self-harm		
	No	Yes	Total
Disordered Eating			
No	86%	6.2%	92%
Yes	5.9%	1.8%	7.7%
Total	92%	8.1%	100%

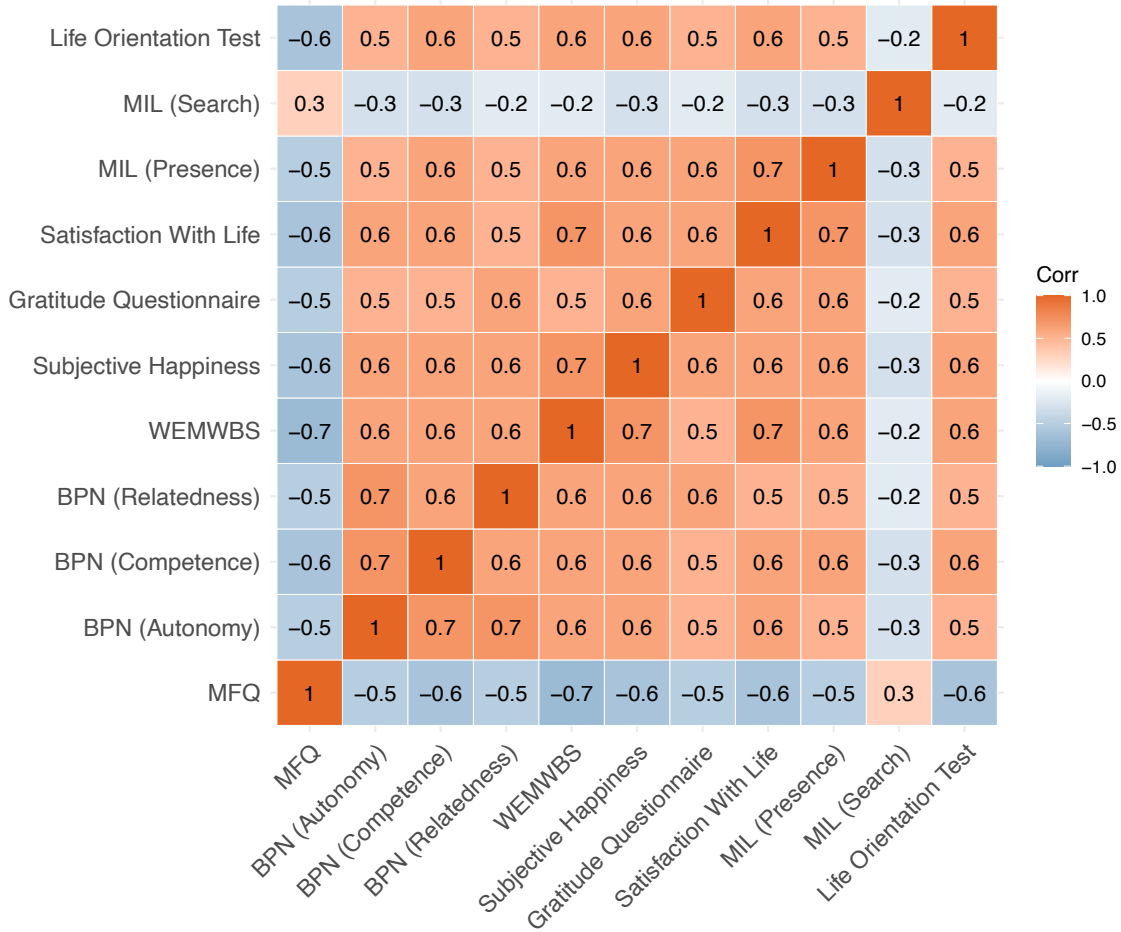


Figure A.1: A correlation matrix for all continuous mental health and well-being variables using Spearman's Rank coefficient (all $p < 0.000$).

A.3 Outcomes by platform for all social media users

In the main text the graphs by platform only include daily users of each. These graphs have an expanded sample to include every participant who said they used each platform with any frequency.

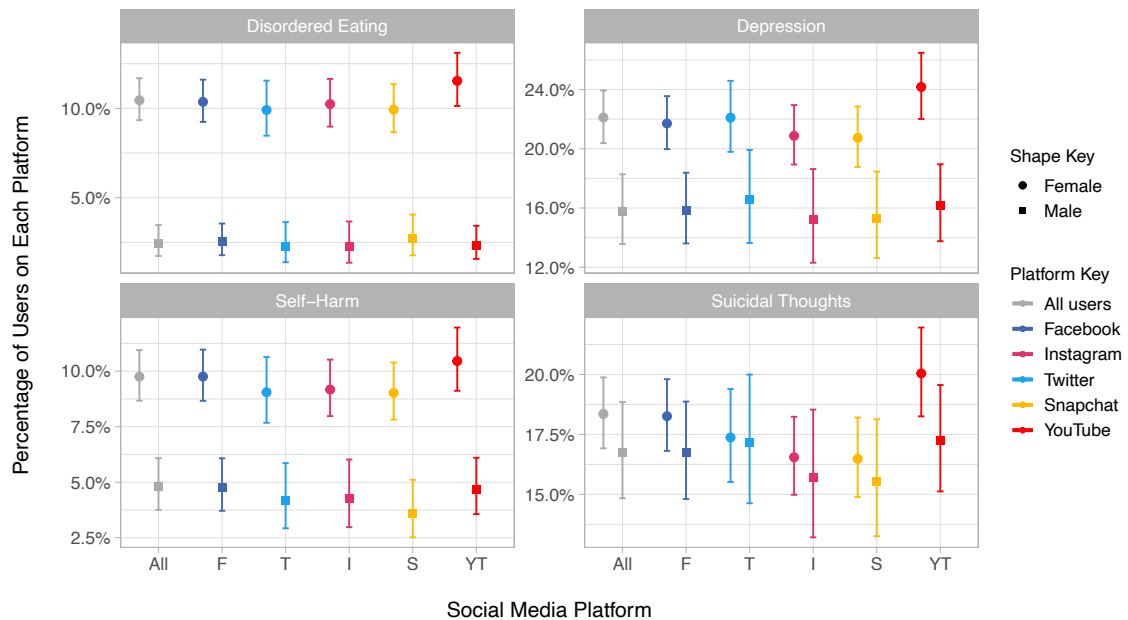


Figure A.2: Percentage of participants who reported disordered eating, self-harm, depression, or suicidal thoughts in the past year, differentiated by sex for all users of each platform, with 95% confidence intervals.

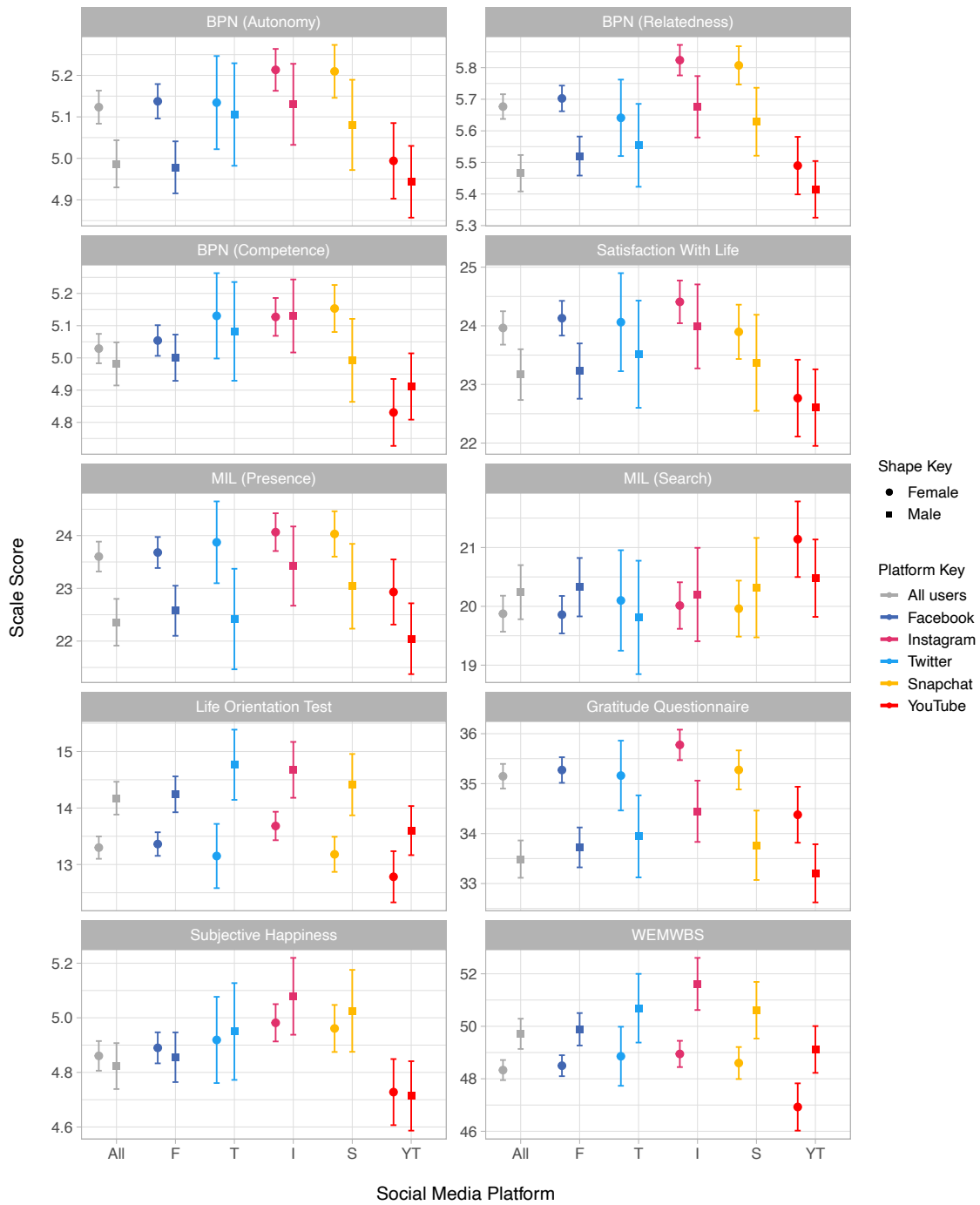


Figure A.3: Mean scores for seven well-being measures for all users of each platform, stratified by sex, with 95% confidence intervals.

Appendix B

Chapter 4

B.1 Mental health sample comparisons by sex

The following two graphs illustrate Figure 4.3 when split by sex, with Figure B.1 showing comparison of Female scores and Figure B.2 showing comparison of Male scores.

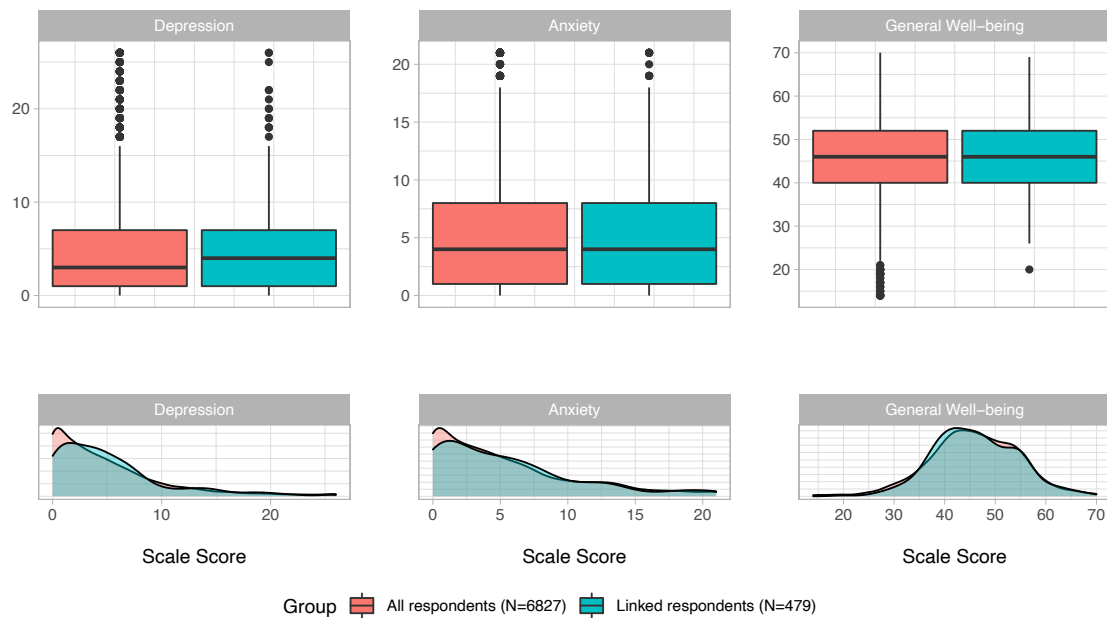


Figure B.1: A comparison of the distributions of female participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.

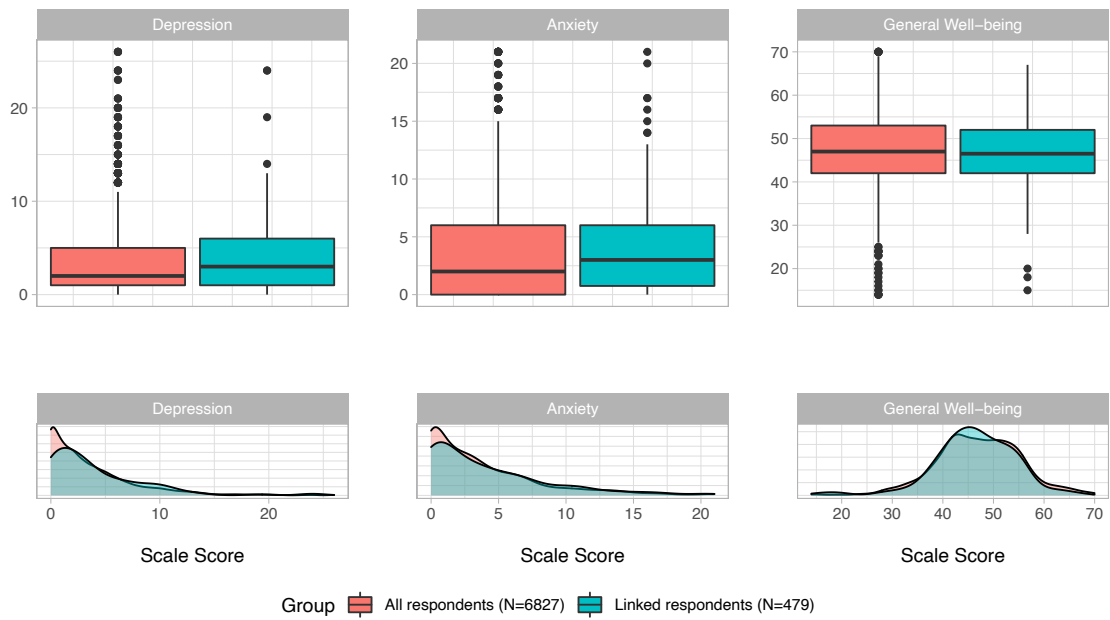


Figure B.2: A comparison of the distributions of male participant scores for anxiety, depression and general well-being between those who agreed to link their Twitter data, and the whole cohort (including linked respondents). The box plot is presenting the median and interquartile ranges.

Appendix C

Chapter 5

C.1 Systematic search key terms

The terms in point (1) are related to classification and machine learning, in point (2) are related to mental health and in point (3) are related to Twitter:

- (1) algorithm OR predict* OR detect* OR understand OR perceiv* OR “machine learning” OR “deep learning” OR “artificial intelligence” OR AI OR interpret OR character* OR classif* OR model* OR analy* OR machine OR recogni* OR sentiment
- (2) depress* OR bipolar* OR wellbeing OR PTSD OR “post traumatic stress disorder” OR suici* OR “mental health” OR mentalhealth OR anxi* OR “personality disorder” OR “eating disorder” OR " ED " OR “disordered eating” OR DSM* OR ICD* OR (mental AND well*) OR (mental AND ill*) OR schizophren*
- (3) twitter OR tweet* OR social media OR social network*

These three sets of terms were searched in each database as (1) AND (2) AND (3). Where the database allowed this was restricted to titles, keywords and abstracts.

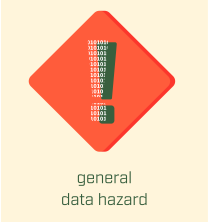

Appendix D

Chapter 6

D.1 Data Hazards Analysis

The Data Hazards project was created and developed jointly by myself and Natalie Zelenka. The Data Hazards labels (<http://datahazards.com/>) are a series of potential consequences of data science projects that present a risk to the ethical integrity of the project. These labels were developed to help researchers acknowledge risks to data science research that may not fit into the remit of research ethics committees, but still require consideration to ensure that data science is done safely. The project encourages reflective thinking by researchers, and provides a series of resources to help make regular reflection on ethical issues a more prominent aspect of applied data science.

Table D.1: An analysis of different potential ethical hazards that might be presented by this project, using the Data Hazard labels.

Hazard	Reasoning	Safety Precautions
 <p data-bbox="298 569 391 611">general data hazard</p> <p data-bbox="240 653 332 737">General data hazard</p>	<p data-bbox="500 600 959 684">Yes. This applies because the project uses data science.</p>	<p data-bbox="987 600 1442 789">I have aimed to be explicit throughout about what types of data science are being used and how they have been implemented so that they can be scrutinised.</p>
 <p data-bbox="298 972 391 1014">reinforces existing bias</p> <p data-bbox="240 1056 467 1140">Reinforces existing biases</p>	<p data-bbox="500 1003 959 1623">Yes. This Hazard applies because the models being generated will learn patterns from the data being input about participant mental health. The ALSPAC sample does not include many people from ethnic minority backgrounds, and so there is less training data available for people who are not White. Other biases may include that the ALSPAC parent sample is most likely to include people who are heterosexual.</p>	<p data-bbox="987 1003 1442 1245">An error analysis was conducted to assess how the predictions may be biased by characteristics like sex and age. However, this was limited by the variables available.</p>



Ranks or classifies people

Yes. This model could be used to rank people based on their predicted mental health outcome score.

In this study I was not interested in individual rankings, though in theory they could be derived using the models described in the study. I also intentionally chose to use continuous measures of mental health to avoid classifying people as having a mental health disorder or not.



High environmental cost

No. The methods used in this study were not reliant on high performance computing or precious materials beyond those usually used in the creation of standard computers and laptops.

NA



Lacks community involvement

Yes and No. The community of ALSPAC participants whose data was linked for this study were consulted. However, individuals who experience poor mental health were not involved in the design or development of the study itself.

The study has focussed on population-level inference instead of individual-level inference. Future research on individual-level inference would benefit from involvement of more stakeholders.



Danger of misuse

Yes. The inference of mental health states from public social media data has potential for misuse by other individuals who may be capable of using models to attempt to infer information about people who were not in the original study.

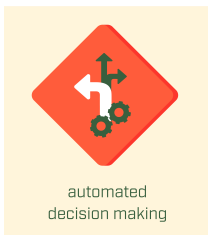
See the reasoning provided in 'Risk to Privacy'.



Difficult to understand

No. The models in this study have intentionally been developed using methods that are transparent and relatively easy to understand and explain. The data used for this study is only available on request, but is accessible to others.

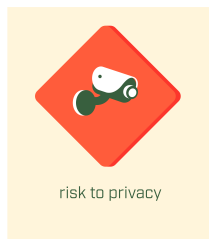
NA



May cause direct harm

No. These models do not have the capacity to cause direct harm to an individual in their current use case.

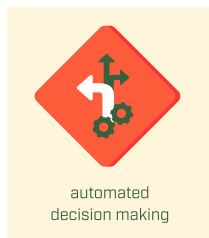
NA



Risk to Privacy

Potentially. For the reasons given in ‘Danger of Misuse’ there could be a risk to privacy if someone attempted to predict individual level mental health using these models. However, there is no risk to privacy for participants in the study, since their data is protected by the central ALSPAC team.

The models used in this study have significant ranges of error for individual level prediction. As such it is highly unlikely that an accurate estimate could be obtained for an individual, especially out of sample. I have made this clear in the presentation of results to ensure that readers are aware of this.



Automates decision making

No. Whilst the tools produced in this study provide information that could be used for decision making they do not actually make any decisions.

NA



Lacks informed consent

No. All participants took part with informed consent.

NA

D.2 Descriptive mental health data by sex and generation

Table D.2: Summary of numbers of tweets and the mental health outcomes between men and women at Survey 1 and Survey 2.

Characteristic	Survey 1		Survey 2	
	F, N = 92	M, N = 59	F, N = 92	M, N = 59
Number of tweets	7 (4, 18)	6 (3, 20)	7 (4, 18)	6 (3, 20)
Depression (MFQ)	5.0 (2.0, 9.0)	3.0 (1.0, 6.0)	6.0 (3.0, 10.0)	4.0 (1.5, 6.0)
Unknown	5	4	15	12
Anxiety (GAD-7)	5.0 (2.2, 9.0)	3.0 (1.0, 6.0)	5.0 (3.0, 8.0)	3.0 (0.0, 6.0)
Unknown	2	2	15	15
Well-being (WEMWBS)	45 (39, 52)	46 (42, 51)	42 (39, 50)	46 (40, 54)
Unknown	4	3	20	12

¹ Median (IQR)

Table D.3: Summary of numbers of tweets and the mental health outcomes between G0 and G1 at Survey 1 and Survey 2.

Characteristic	Survey 1		Survey 2	
	G1, N = 96	G0, N = 55	G1, N = 96	G0, N = 55
Number of tweets	7 (3, 16)	6 (3, 39)	7 (3, 16)	6 (3, 39)
Depression (MFQ)	5.0 (3.0, 10.0)	2.0 (1.0, 6.0)	6.0 (4.0, 10.0)	4.0 (1.5, 5.5)
Unknown	3	6	19	8
Anxiety (GAD-7)	6.0 (2.2, 10.0)	3.0 (1.0, 5.0)	5.0 (3.0, 9.0)	3.0 (0.2, 6.0)
Unknown	2	2	21	9
Well-being (WEMWBS)	44 (39, 49)	47 (42, 55)	41 (38, 46)	47 (43, 53)
Unknown	3	4	22	10

¹ Median (IQR)

D.3 Correlations between model features

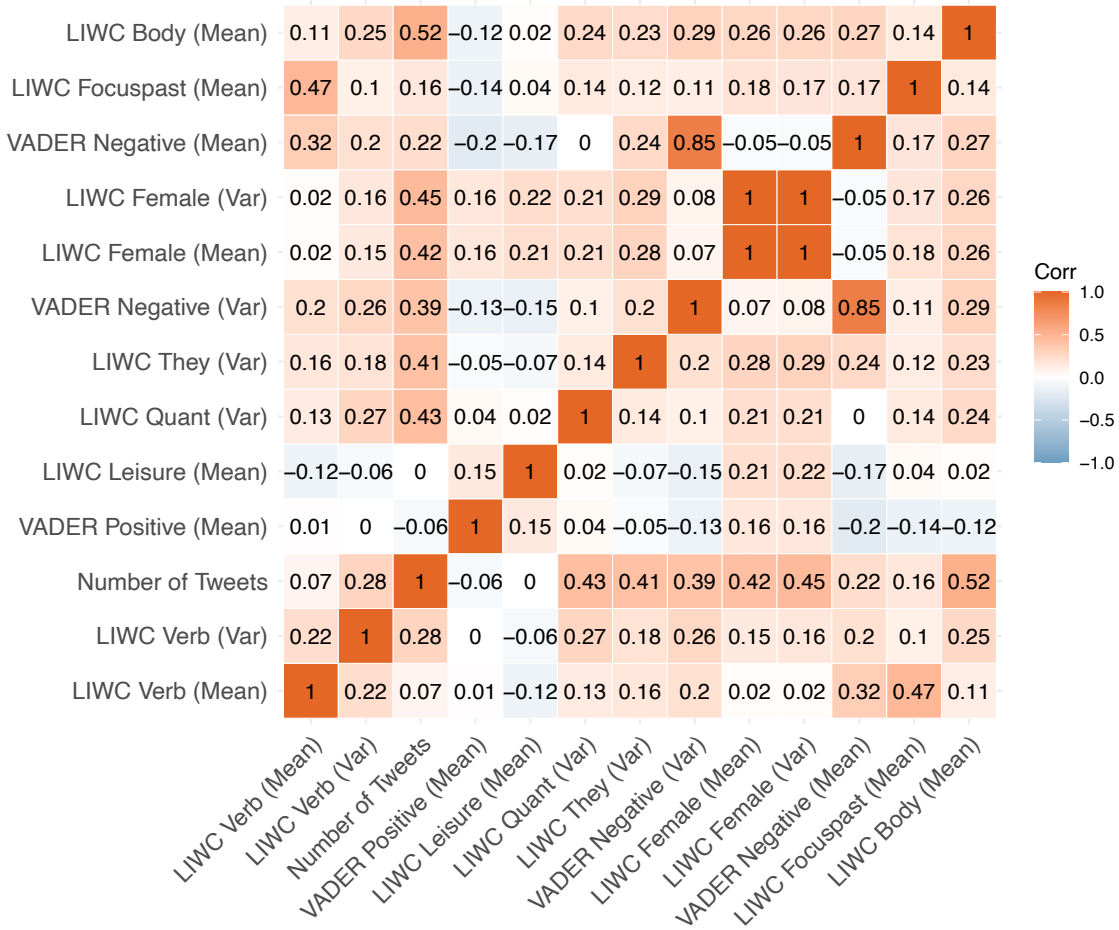


Figure D.1: Correlations between the variables which were best associated with depression

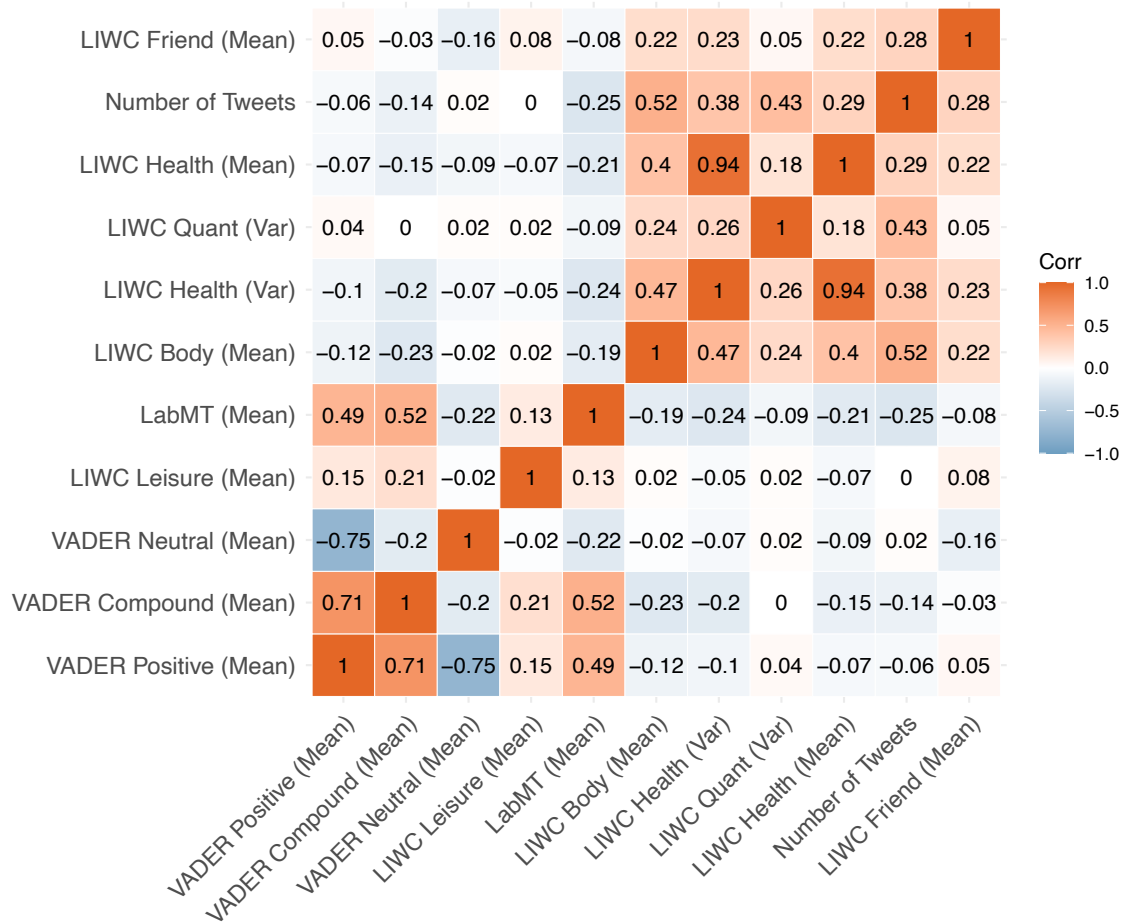


Figure D.2: Correlations between the variables which were best associated with general anxiety

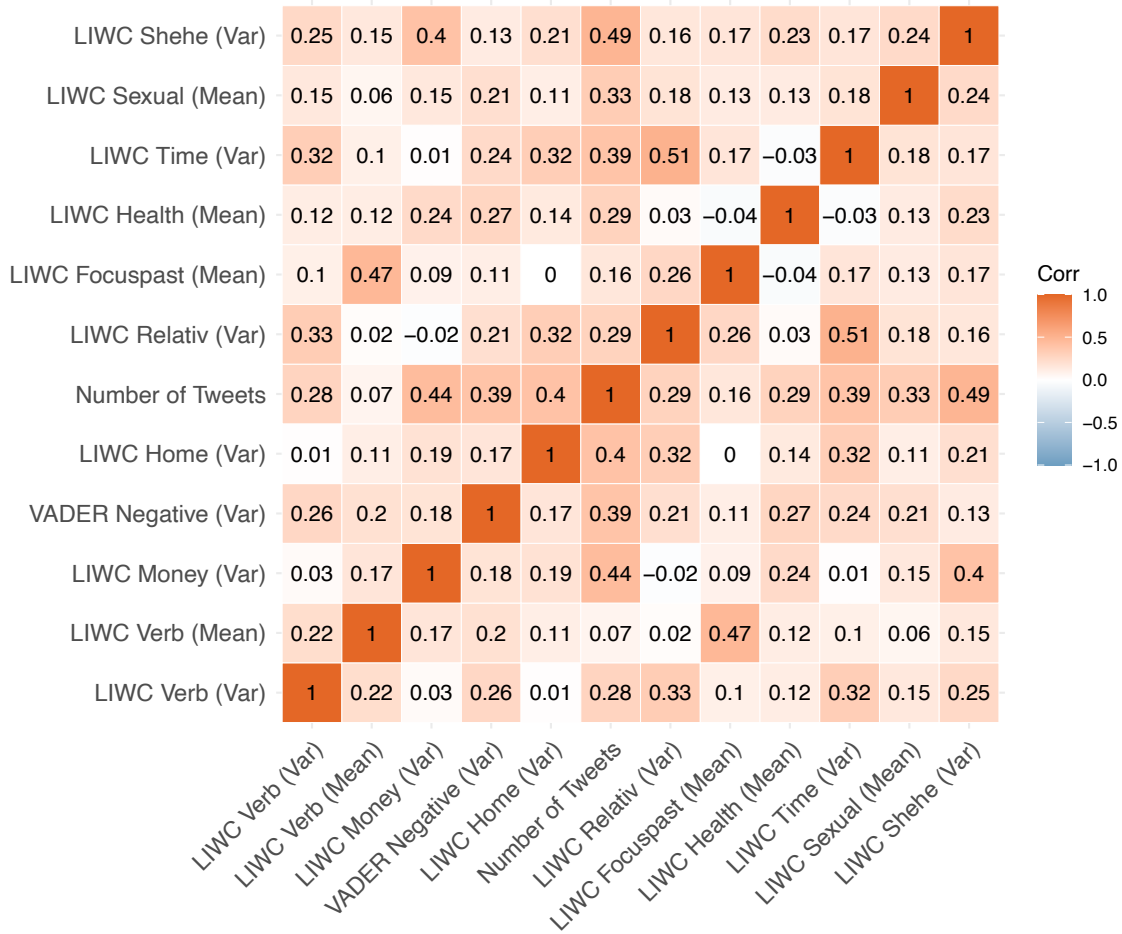


Figure D.3: Correlation between the variables which were best associated with general well-being

D.4 Associations for disaggregated data

Table D.5 gives the results of testing all associations of sentiment variables against each of the mental health outcomes when no aggregation by individual was conducted.

Table D.4: This table displays the variance accounted for by each of the sentiment variables that align with standard codings of emotion, as well as patterns of life, when regressed against depression, anxiety or well-being, and adjusted for sex and generation.

Variable	Direction of Effect	Variance Explained	p-value
Depression			
LIWC Affect	+	0.090	0.077
LIWC Anger	+	0.034	0.280
LIWC Anx	+	0.016	0.461
LIWC Negemo	+	0.043	0.223
LIWC Posemo	+	0.035	0.267
LIWC Sad	+	0.038	0.250
VADER Compound	-	0.041	0.232
VADER Negative	+	0.106	0.055
VADER Neutral	-	0.072	0.115
VADER Positive	+	0.003	0.736
LABMT Emotion_valence	-	0.406	0.000
TIME Num	+	0.161	0.018
Anxiety			
LIWC Affect	+	0.087	0.080
LIWC Anger	+	0.785	0.000
LIWC Anx	+	0.079	0.097
LIWC Negemo	+	0.693	0.000
LIWC Posemo	-	0.101	0.061
LIWC Sad	+	0.001	0.831
VADER Compound	-	1.693	0.000
VADER Negative	+	1.346	0.000
VADER Neutral	-	0.054	0.169
VADER Positive	-	0.399	0.000
LABMT Emotion_valence	-	1.109	0.000
TIME Num	+	0.932	0.000
Well-being			
LIWC Affect	-	0.000	0.947
LIWC Anger	-	0.398	0.000
LIWC Anx	+	0.011	0.542
LIWC Negemo	-	0.331	0.001
LIWC Posemo	+	0.197	0.010
LIWC Sad	-	0.030	0.321
VADER Compound	+	0.966	0.000
VADER Negative	-	0.751	0.000
VADER Neutral	+	0.048	0.206
VADER Positive	+	0.177	0.015
LABMT Emotion_valence	+	0.719	0.000
TIME Num	-	0.482	0.000

Table D.5: This table displays the variance accounted for by each sentiment variable with $p < 0.001$ when regressed against depression, anxiety or well-being, and adjusting for sex and generation with no aggregation by individual.

Variable	Direction of Effect	Variance Explained	p-value
Anxiety			
VADER Compound	-	1.69	0.000
VADER Negative	+	1.35	0.000
LABMT Emotion_valence	-	1.11	0.000
TIME Num	+	0.93	0.000
LIWC Prep	-	0.85	0.000
LIWC Anger	+	0.79	0.000
LIWC Negemo	+	0.69	0.000
LIWC Affiliation	-	0.57	0.000
LIWC Relativ	-	0.56	0.000
LIWC Space	-	0.50	0.000
LIWC Relig	-	0.44	0.000
VADER Positive	-	0.40	0.000
LIWC Differ	+	0.37	0.000
LIWC We	-	0.36	0.000
LIWC I	+	0.32	0.001
LIWC Swear	+	0.31	0.001
Depression			
LIWC Prep	-	0.51	0.000
LIWC Space	-	0.46	0.000
LABMT Emotion_valence	-	0.41	0.000
LIWC Relativ	-	0.38	0.000
LIWC I	+	0.35	0.001
Well-being			
VADER Compound	+	0.97	0.000
VADER Negative	-	0.75	0.000
LABMT Emotion_valence	+	0.72	0.000
LIWC Prep	+	0.57	0.000
TIME Num	-	0.48	0.000
LIWC Anger	-	0.40	0.000
LIWC Health	+	0.39	0.000
LIWC Swear	-	0.38	0.000
LIWC Negemo	-	0.33	0.001
LIWC We	+	0.33	0.001

D.5 Tables of Results Accompanying Figures

Table D.6: Table of the results displayed in Chapter 6, Figure 6.11. The results of regression models of each sentiment variable against depression, anxiety and general well-being, adjusted for sex and generation. The percentage variance explained after sex and generation have been accounted for is given.

Variable	Anxiety		Depression		Well-being	
	Estimate (p-val)	Variance Explained	Estimate (p-val)	Variance Explained	Estimate (p-val)	Variance Explained
LabMT						
Emotion Valence (Mean)	-0.1449 (0.025)	3.065	-0.07591 (0.26)	0.776	0.5824 (0.377)	0.518
Emotion Valence (Var)	0.0113 (0.362)	0.514	0.007388 (0.57)	0.198	-0.09807 (0.439)	0.397
LIWC (Affective Processes)						
Affect (Mean)	-2.344 (0.482)	0.306	2.435 (0.48)	0.306	19.55 (0.568)	0.217
Affect (Var)	19.09 (0.592)	0.178	45.93 (0.195)	1.025	-78.25 (0.823)	0.033
Anger (Mean)	27.98 (0.071)	1.998	6.705 (0.685)	0.101	-124.6 (0.425)	0.422
Anger (Var)	440 (0.087)	1.800	192.8 (0.502)	0.277	-3165 (0.22)	0.994
Anxiety (Mean)	6.482 (0.641)	0.135	12.1 (0.393)	0.448	-49.33 (0.725)	0.083
Anxiety (Var)	126.5 (0.562)	0.208	219.8 (0.322)	0.602	-1837 (0.401)	0.468
Negative (Mean)	8.907 (0.139)	1.347	8.362 (0.175)	1.125	-82.11 (0.176)	1.213
Negative (Var)	33.91 (0.569)	0.201	61.27 (0.314)	0.621	-509 (0.394)	0.483
Positive (Mean)	-7.574 (0.053)	2.291	-0.6802 (0.866)	0.018	61.04 (0.126)	1.541
Positive (Var)	9.469 (0.846)	0.023	49.27 (0.298)	0.663	31.71 (0.946)	0.003
Sad (Mean)	16.22 (0.192)	1.048	13.01 (0.304)	0.647	-115.9 (0.354)	0.571
Sad (Var)	158.3 (0.229)	0.893	117.5 (0.38)	0.472	-1269 (0.337)	0.612
Patterns of Life						
Number of Tweets	0.002851 (0.042)	2.529	0.004064 (0.014)	3.653	-0.03341 (0.017)	3.708
Prop. of Retweets	0.3318 (0.083)	1.844	0.174 (0.377)	0.479	0.2763 (0.888)	0.013
Prop. Tweets 12am to 4am	0.9491 (0.218)	0.937	0.05192 (0.947)	0.003	-7.64 (0.322)	0.649
Time of Day (Mean)	-0.04907 (0.531)	0.243	-0.09738 (0.235)	0.865	0.2169 (0.784)	0.050
Time of Day (Var)	0.06437 (0.179)	1.109	0.008119 (0.869)	0.017	-0.1705 (0.724)	0.083
VADER						
Compound (Mean)	-0.7282 (0.005)	4.809	-0.4735 (0.074)	1.936	4.149 (0.127)	1.535
Compound (Var)	0.4783 (0.28)	0.720	0.7694 (0.087)	1.788	-4.268 (0.338)	0.610
Negative (Mean)	2.581 (0.085)	1.821	3.118 (0.044)	2.468	-30.06 (0.053)	2.467
Negative (Var)	8.78 (0.124)	1.453	12.37 (0.033)	2.745	-156 (0.011)	4.223
Neutral (Mean)	1.986 (0.01)	4.040	0.9087 (0.259)	0.779	-2.783 (0.724)	0.083
Neutral (Var)	-3.478 (0.261)	0.779	-0.7127 (0.817)	0.033	12.01 (0.693)	0.104
Positive (Mean)	-2.989 (0)	8.193	-1.914 (0.022)	3.158	11.92 (0.155)	1.335
Positive (Var)	-4.465 (0.065)	2.081	-3.227 (0.194)	1.032	24.3 (0.321)	0.654

Table D.7: This table represents the results from Chapter 6, Figure 12. The mean R Squared value obtained from 10 x 5-fold cross validation plotted for the outcomes of depression, anxiety and well-being. The input data contains an increasing number of weeks, and difference weighting functions were used on each number of weeks of data.

Window length	Wave	Weight	Depression	Anxiety	Well-being
2 weeks	1	None	0.0955072	0.1408749	0.1365363
4 weeks	1	None	0.0584345	0.0780402	0.1244502
6 weeks	1	None	0.0713726	0.1077627	0.1462798
8 weeks	1	None	0.1000517	0.0866560	0.2031543
12 weeks	1	None	0.1338802	0.0819419	0.1439200
2 weeks	1	1	0.0792149	0.1279158	0.0876104
4 weeks	1	1	0.0456900	0.0987862	0.1011479
6 weeks	1	1	0.0491740	0.1010832	0.1092509
8 weeks	1	1	0.0892288	0.0765593	0.1316419
12 weeks	1	1	0.0866373	0.0737057	0.1052134
2 weeks	1	2	0.0597991	0.1108137	0.0426631
4 weeks	1	2	0.0406545	0.1108217	0.0673303
6 weeks	1	2	0.0526409	0.0917915	0.0736056
8 weeks	1	2	0.0383619	0.0899619	0.0951973
12 weeks	1	2	0.0619529	0.0713363	0.0779059
2 weeks	1	4	0.0409789	0.0993264	0.0500443
4 weeks	1	4	0.0433536	0.0995721	0.0423736
6 weeks	1	4	0.0430505	0.0888625	0.0628472
8 weeks	1	4	0.0464943	0.0745442	0.0473643
12 weeks	1	4	0.0420944	0.0766625	0.0412523
2 weeks	1	3	0.0627423	0.1268601	0.0494644
4 weeks	1	3	0.0408974	0.1248645	0.0481601
6 weeks	1	3	0.0515325	0.1219212	0.0645939
8 weeks	1	3	0.0425438	0.0984817	0.0757927
12 weeks	1	3	0.0509961	0.0952086	0.0705760

D.6 Prediction error

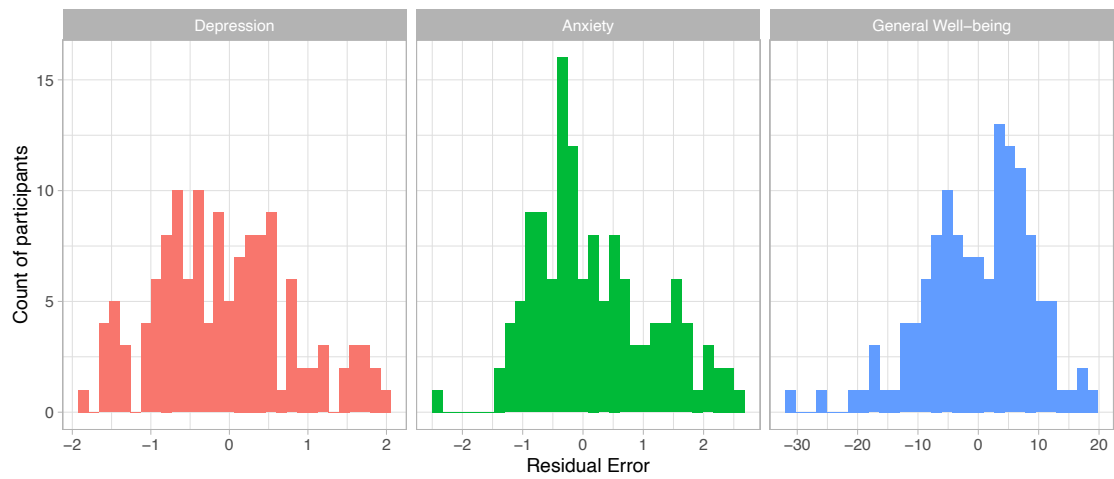


Figure D.4: Histogram of the residual errors from predictions of Survey 2 outcomes, using the linear models trained on Survey 1. Residuals for depression and anxiety were calculated after exponentiating the estimate, which is predicted on a log scale.

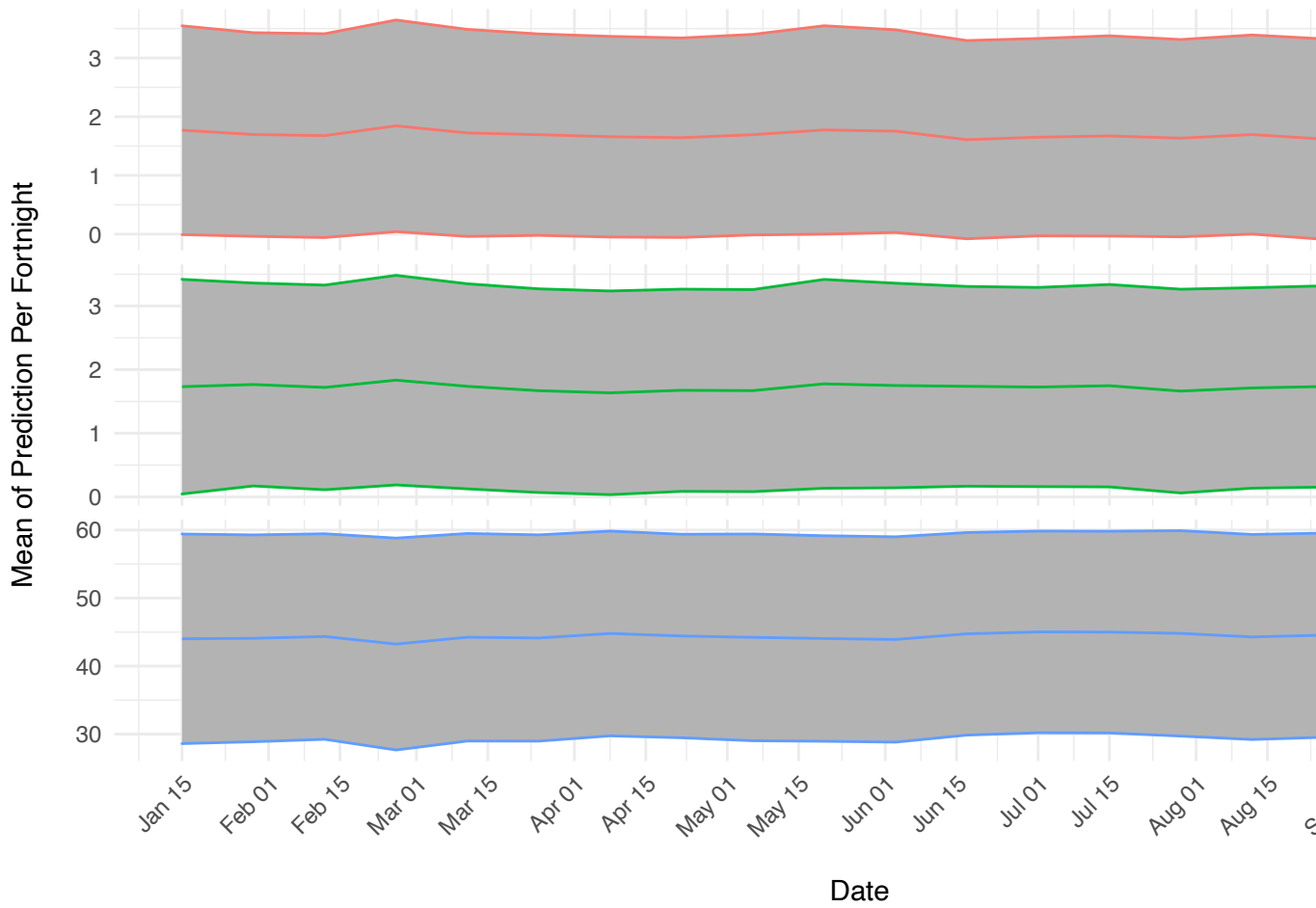


Figure D.5: The graph of predictions made by the models for depression, anxiety and general well-being over the pandemic period using the final models trained on 2 weeks of Twitter data. This version contains the prediction intervals estimated for each fortnightly prediction.

References

- [1] Finlay L, Gough B. *Reflexivity: A practical guide for researchers in health and social sciences*. John Wiley & Sons, 2008.
- [2] Ricker B. Reflexivity, positionality and rigor in the context of big data research. In: *Thinking big data in geography: New regimes, new research*. University of Iowa Press, 2017.
- [3] Dwyer SC, Buckle JL. The space between: On being an insider-outsider in qualitative research. *International Journal of Qualitative Methods* 2009; 8: 54–63.
- [4] Network Global Burden of Disease Collaborative. Global burden of disease study 2019 (GBD 2019) results, <http://ghdx.healthdata.org/gbd-results-to-01> (2019).
- [5] WHO: Mental health and substance use team. *Global health estimates: Suicide worldwide in 2019*. World Health Organisation, <https://www.who.int/publications/item/9789240026643> (2021).
- [6] McManus S, Bebbington PE, Jenkins R, et al. Data resource profile: Adult psychiatric morbidity survey (APMS). *International Journal of Epidemiology* 2020; 49: 361–362e.
- [7] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Arlington, VA.
- [8] McIntosh AM, Stewart R, John A, et al. Data science for mental health: A UK perspective on a global challenge. *The Lancet Psychiatry* 2016; 3: 993–998.

-
- [9] *The data will see you now: Datafication and the boundaries of health.* The Ada Lovelace Institute, <https://www.adalovelaceinstitute.org/report/the-data-will-see-you-now/> (2020).
- [10] Russ TC, Woelbert E, Davis KAS, et al. How data science can advance mental health research. *Nature Human Behaviour* 2019; 3: 24–32.
- [11] Naslund JA, Gonsalves PP, Gruebner O, et al. Digital innovations for global mental health: Opportunities for data science, task sharing, and early intervention. *Current Treatment Options in Psychiatry* 2019; 6: 337–351.
- [12] Torous J, Bucci S, Bell IH, et al. The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 2021; 20: 318–335.
- [13] Su C, Xu Z, Pathak J, et al. Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry* 2020; 10: 1–26.
- [14] Taliáz D, Spinrad A, Barzilay R, et al. Optimizing prediction of response to antidepressant medications using machine learning and integrated genetic, clinical, and demographic data. *Translational Psychiatry* 2021; 11: 1–9.
- [15] Torous J, Kiang MV, Lorme J, et al. New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health* 2016; 3: e16.
- [16] Insel TR. Digital phenotyping: A global tool for psychiatry. *World Psychiatry* 2018; 17: 276.
- [17] US Substance Abuse and Mental Health Services Administration. *Behavioural health workforce report*. Rockville: US Substance Abuse and Mental Health Services Administration, <https://www.mamh.org/assets/files/behavioral-health-workforce-report.pdf> (2020).

- [18] Lwin MO, Lu J, Sheldenkar A, et al. Global sentiments surrounding the COVID-19 pandemic on twitter: Analysis of twitter trends. *JMIR Public Health and Surveillance* 2020; 6: e19447.
- [19] Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: A clinical review. *Evidence-Based Mental Health* 2020; 23: 161–166.
- [20] Honeyman M, Maguire D, Evans H, et al. *Digital technology and health inequalities: A scoping review*. Cardiff: Public Health Wales NHS Trust, 2020.
- [21] Stilgoe J, Owen R, Macnaghten P. Developing a framework for responsible innovation. *Research Policy* 2013; 42: 1568–1580.
- [22] Center for Data Ethics and Innovation. *Review into bias in algorithmic decision-making*. UK Government, <https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making> (2019).
- [23] Clayton V, Sanders M, Schoenwald E, et al. *Machine learning in children’s social care: Does it work?* What Works Centre for Children’s Social Care, https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_technical_report_machine_learning_in_childrens_services_does_it_work_Sep_2020.pdf (2020).
- [24] Davidson BI. The crossroads of digital phenotyping. *General Hospital Psychiatry* 2020; 74: 126–132.
- [25] Areán P, Leshner A, Barch D, et al. *Opportunities and challenges of developing information technologies on behavioral and social science clinical research*. National Institute of Mental Health, <https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/reports/opportunities-and-challenges-of-developing-information-technologies-on-behavioral-and-social-science-clinical-research> (2017).

-
- [26] Foley T, Woollard J. *The digital future of mental healthcare and its workforce*. National Health Service, <https://topol.hee.nhs.uk/wp-content/uploads/HEE-Topol-Review-Mental-health-paper.pdf> (2018).
- [27] Anderson I, Gil S, Gibson C, et al. ‘Just the way you are’: Linking music listening on spotify and personality. *Social Psychological and Personality Science* 2021; 12: 561–572.
- [28] McStay A. *Emotional AI: The rise of empathic media*. Sage, 2018.
- [29] Stark L, Hoey J. The ethics of emotion in artificial intelligence systems. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 782–793.
- [30] HM Government. Equality Act 2010, <https://www.legislation.gov.uk/ukpga/2010/15/contents> (2010).
- [31] Bard J. Developing a legal framework for regulating emotion AI. *Boston University Journal of Science & Technology Law* 2020; 1: 1–46.
- [32] Crawford K. Time to regulate AI that interprets human emotions. *Nature* 2021; 592: 167–167.
- [33] Colvonen PJ, DeYoung PN, Bosompra N-OA, et al. Limiting racial disparities and bias for wearable devices in health science research. *Sleep* 2020; 43: zsa159.
- [34] Bender EM, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.
- [35] Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. *PloS one* 2020; 15: e0240376.
- [36] Müller SR, Chen XL, Peters H, et al. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports* 2021; 11: 1–10.

-
- [37] Kocoń J, Gruza M, Bielaniewicz J, et al. Learning personal human biases and representations for subjective tasks in natural language processing. In: *2021 IEEE international conference on data mining (ICDM)*. IEEE, 2021, pp. 1168–1173.
- [38] Cardoso JR, Pereira LM, Iversen MD, et al. What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics* 2014; 19: 27–30.
- [40] Kim J, Uddin ZA, Lee Y, et al. A systematic review of the validity of screening depression through facebook, twitter, instagram, and snapchat. *Journal of Affective Disorders* 2021; 286: 360–369.
- [40] Kim J, Uddin ZA, Lee Y, et al. A systematic review of the validity of screening depression through facebook, twitter, instagram, and snapchat. *Journal of Affective Disorders* 2021; 286: 360–369.
- [41] Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine* 2020; 3: 1–11.
- [42] Carr CT, Hayes RA. Social media: Defining, developing, and divining. *Atlantic Journal of Communication* 2015; 23: 46–65.
- [43] Darby JK, Simmons N, Berger PA. Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders* 1984; 17: 75–85.
- [44] Bucci W, Freedman N. The language of depression. *Bulletin of the Menninger Clinic* 1981; 45: 334.
- [45] Andreasen NJ, Pfohl B. Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry* 1976; 33: 1361–1367.
- [46] McCulloch G. *Because internet: Understanding the new rules of language*. Riverhead Books, 2020.
- [47] Ligthart A, Catal C, Tekinerdogan B. Systematic reviews in sentiment analysis: A tertiary study. *Artificial Intelligence Review* 2021; 1–57.

-
- [48] Bathina KC, Ten Thij M, Lorenzo-Luaces L, et al. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour* 2021; 5: 458–466.
- [49] Voutilainen A. Part-of-speech tagging. *The Oxford handbook of computational linguistics* 2003; 219–232.
- [50] Zimmermann J, Brockmeyer T, Hunn M, et al. First-person pronoun use in spoken language as a predictor of future depressive symptoms: Preliminary evidence from a clinical sample of depressed patients. *Clinical Psychology & Psychotherapy* 2017; 24: 384–391.
- [51] Van Hee C, Lefever E, Hoste V. Guidelines for annotating irony in social media text, version 2.0. *LT3 Technical Report Series*.
- [52] Ribeiro FN, Araújo M, Gonçalves P, et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 2016; 5: 1–29.
- [53] Frank MR, Mitchell L, Dodds PS, et al. Happiness and the patterns of life: A study of geolocated tweets. *Scientific Reports* 2013; 3: 1–9.
- [54] Nouman M, Khoo SY, Mahmud MP, et al. Recent advances in contactless sensing technologies for mental health monitoring. *IEEE Internet of Things Journal* 2021; 9: 274–297.
- [55] Scott AJ, Webb TL, Martyn-St James M, et al. Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. *Sleep Medicine Reviews* 2021; 101556.
- [56] Perlis ML, Grandner MA, Chakravorty S, et al. Suicide and sleep: Is it a bad thing to be awake when reason sleeps? *Sleep Medicine Reviews* 2016; 29: 101–107.
- [57] Wongkoblaph A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research* 2017; 19: e228.
- [58] Nahum-Shani I, Smith SN, Spring BJ, et al. Just-in-time adaptive interventions (JITAI) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 2018; 52: 446–462.

-
- [59] Chen W, Fan C-Y, Liu Q-X, et al. Passive social network site use and subjective well-being: A moderated mediation model. *Computers in Human Behavior* 2016; 64: 507–514.
- [60] Costello C, Srivastava S, Rejaie R, et al. Predicting mental health from followed accounts on twitter. *Collabra: Psychology* 2021; 7: 18731.
- [61] Kramer AD, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 2014; 111: 8788–8790.
- [62] Manikonda L, De Choudhury M. Modeling and understanding visual attributes of mental health disclosures in social media. In: *Proceedings of the 2017 CHI conference on human factors in computing systems*. 2017, pp. 170–181.
- [63] Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 2017; 6: 1–12.
- [64] Vempala A, Preoțiu-Pietro D. Categorizing and inferring the relationship between the text and image of twitter posts. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2019, pp. 2830–2840.
- [65] Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, et al. What twitter profile and posted images reveal about depression and anxiety. In: *Proceedings of the international AAAI conference on web and social media*. 2019, pp. 236–246.
- [66] Sawhney R, Joshi H, Gandhi S, et al. A time-aware transformer based model for suicide ideation detection on social media. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 2020, pp. 7685–7697.
- [67] Tsakalidis A, Nanni F, Hills A, et al. Identifying moments of change from longitudinal user text. In: *Proceedings of the 60th annual meeting of the association for computational linguistics*. 2022, pp. 4647–4660.

-
- [68] Heffer T, Good M, Daly O, et al. The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: An empirical reply to twenge et al.(2018). *Clinical Psychological Science* 2019; 7: 462–470.
- [69] Faelens L, Hoorelbeke K, Soenens B, et al. Social media use and well-being: A prospective experience-sampling study. *Computers in Human Behavior* 2021; 114: 106510.
- [70] Schønning V, Hjetland GJ, Aarø LE, et al. Social media use and mental health and well-being among adolescents—a scoping review. *Frontiers in Psychology* 2020; 11: 1949.
- [71] Karim F, Oyewande AA, Abdalla LF, et al. Social media use and its connection to mental health: A systematic review. *Cureus* 2020; 12: e8627.
- [72] Dubicka B, Theodosiou L. *Technology use and the mental health of children and young people*. Royal College of Psychiatrists, <https://www.rcpsych.ac.uk/docs/default-source/improving-care/better-mh-policy/college-reports/college-report-cr225.pdf> (2020).
- [73] K Kaye L, Orben A, A Ellis D, et al. The conceptual and methodological mayhem of ‘screen time’. *International Journal of Environmental Research and Public Health* 2020; 17: 3661.
- [74] Ernala SK, Birnbaum ML, Candan KA, et al. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–16.
- [75] Vries LP de, Baselmans BM, Bartels M. Smartphone-based ecological momentary assessment of well-being: A systematic review and recommendations for future studies. *Journal of Happiness Studies* 2021; 22: 2361–2408.
- [76] Sedano-Capdevila A, Porrás-Segovia A, Bello HJ, et al. Use of ecological momentary assessment to study suicidal thoughts and behavior: A systematic review. *Current Psychiatry Reports* 2021; 23: 1–17.

-
- [77] Biernesser C, Bear T, Brent D, et al. Development of an ecological momentary assessment of the impact of social media use among suicidal adolescents. *Archives of Suicide Research* 2021; 1–15.
- [78] Sloan L, Morgan J, Burnap P, et al. Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one* 2015; 10: e0115545.
- [79] Wang Z, Hale S, Adelani DI, et al. Demographic inference and representative population estimates from multilingual social media data. In: *The world wide web conference*. 2019, pp. 2056–2067.
- [80] Morgan-Lopez AA, Kim AE, Chew RF, et al. Predicting age groups of twitter users based on language and metadata features. *PloS one* 2017; 12: e0183537.
- [81] Fosch-Villaronga E, Poulsen A, Søråa RA, et al. Gendering algorithms in social media. *Association for Computing Machinery SIGKDD Explorations Newsletter* 2021; 23: 24–31.
- [82] Ellouze M, Mechti S, Belguith LH. Approach based on ontology and machine learning for identifying causes affecting personality disorder disease on twitter. In: *International conference on knowledge science, engineering and management*. Springer, 2021, pp. 659–669.
- [83] Al Baghal T, Wenz A, Sloan L, et al. Linking twitter and survey data: Asymmetry in quantity and its impact. *EPJ Data Science* 2021; 10: 32.
- [84] Alhabash S, Ma M. A tale of four platforms: Motivations and uses of facebook, twitter, instagram, and snapchat among college students? *Social Media + Society* 2017; 3: 2056305117691544.
- [85] Mancosu M, Vegetti F. What you can scrape and what is right to scrape: A proposal for a tool to collect public facebook data. *Social Media+ Society* 2020; 6: 2056305120940703.

-
- [86] Vittorio A. Facebook move against NYU research stirs scraping policy feud, <https://news.bloomberglaw.com/privacy-and-data-security/facebook-move-against-nyu-research-stirs-scraping-policy-feud> (2021, accessed 3 November 2021).
- [87] Twitter, Inc. Twitter documentation, <https://developer.twitter.com/en/docs>.
- [88] Kim SA. Social media algorithms: Why you see what you see. *Georgetown Law Technology Review* 2017; 2: 147.
- [89] Lazer D, Kennedy R, King G, et al. The parable of google flu: Traps in big data analysis. *Science* 2014; 343: 1203–1205.
- [90] Xie S. New conversation settings, coming to a tweet near you, https://blog.twitter.com/en_us/topics/product/2020/new-conversation-settings-coming-to-a-tweet-near-you (2020).
- [91] Hibbin RA, Samuel G, Derrick GE. From ‘a fair game’ to ‘a form of covert research’: Research ethics committee members’ differing notions of consent and potential risk to participants within social media research. *Journal of Empirical Research on Human Research Ethics* 2018; 13: 149–159.
- [92] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Belmont report: Ethical principles and guidelines for the protection of human subjects of research, report of the national commission for the protection of human subjects of biomedical and behavioral research*. United States Government Printing Office, 1979.
- [93] Nuremberg Military Tribunal. The nuremberg code, 1947. *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law* 1947; 2: 181–182.
- [94] British Psychological Society. *Ethics guidelines for internet-mediated research*. British Psychological Society, <https://www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research> (2021).

-
- [95] Franzke AS, Bechmann A, Zimmer M, et al. Internet research: Ethical guidelines 3.0., <https://aoir.org/reports/ethics3.pdf> (2020).
- [96] Townsend L, Wallace C. Social media research: A guide to ethics. *University of Aberdeen* 2016; 1: 16.
- [97] Proferes N. Information flow solipsism in an exploratory study of beliefs about twitter. *Social Media+ Society* 2017; 3: 2056305117698493.
- [98] Williams ML, Burnap P, Sloan L. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 2017; 51: 1149–1168.
- [99] Fiesler C, Proferes N. 'Participant' perceptions of twitter research ethics. *Social Media+ Society* 2018; 4: 2056305118763366.
- [100] Birhane A. Algorithmic injustice: A relational ethics approach. *Patterns* 2021; 2: 100205.
- [101] Wies B, Landers C, Ienca M. Digital mental health for young people: A scoping review of ethical promises and challenges. *Frontiers in Digital Health* 2021; 91.
- [102] Chancellor S, Birnbaum ML, Caine ED, et al. A taxonomy of ethical tensions in inferring mental health states from social media. In: *Proceedings of the 2019 ACM conference on fairness, accountability, and transparency*. 2019, pp. 79–88.
- [103] Byrd JB, Greene AC, Prasad DV, et al. Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics* 2020; 21: 615–629.
- [104] Stier S, Breuer J, Siegers P, et al. Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review* 2020; 38: 503–516.
- [105] Doidge J, Christen P, Harron K. *Quality assessment in data linkage*. Office for National Statistics, <https://www.gov.uk/government/publications/joined-up-data-in-government-the-future-of-data-linking-methods/quality-assessment-in-data-linkage> (2021).

-
- [106] Bohensky MA, Jolley D, Sundararajan V, et al. Data linkage: A powerful research tool with potential problems. *BMC Health Services Research* 2010; 10: 1–7.
- [107] Pearson RJ, Jewell A, Wijlaars L, et al. Linking data on women in public family law court proceedings concerning their children to mental health service records in south london. *International Journal of Population Data Science* 2021; 6: Online.
- [108] Jones KH, Ford DV, Thompson S, et al. A profile of the SAIL databank on the UK secure research platform. *International Journal of Population Data Science* 2019; 4: Online.
- [109] Mneimneh ZN, McClain C, Bruffaerts R, et al. Evaluating survey consent to social media linkage in three international health surveys. *Research in Social and Administrative Pharmacy* 2021; 17: 1091–1100.
- [110] Henderson M, Jiang K, Johnson M, et al. Measuring twitter use: Validating survey-based measures. *Social Science Computer Review* 2019; 36: 1121–1141.
- [111] Al Baghal T, Sloan L, Jessop C, et al. Linking twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review* 2020; 38: 517–532.
- [112] Webb P, Bain C, Page A. *Essential epidemiology: An introduction for students and health professionals*. Cambridge University Press, 2017.
- [113] Medical Research Council. *Maximising the value of UK population cohorts: MRC Strategic Review of the Largest UK Population Cohort Studies*. Medical Research Council, <https://mrc.ukri.org/publications/browse/maximising-the-value-of-uk-population-cohorts/> (2014).
- [114] Gillan CM, Rutledge RB. Smartphones and the neuroscience of mental health. *Annual Review of Neuroscience*; 44.
- [116] Sloan L, Jessop C, Al Baghal T, et al. Linking survey and twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics* 2019; 15: 63–76.

-
- [116] Sloan L, Jessop C, Al Baghal T, et al. Linking survey and twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics* 2019; 15: 63–76.
- [117] Tanner A, Davis O. Epicosm: Epidemiological cohort online social media (version 1.1), <https://github.com/DynamicGenetics/Epicosm> (2020).
- [118] Boyd A, Golding J, Macleod J, et al. Cohort profile: The “children of the 90s”—the index offspring of the avon longitudinal study of parents and children. *International Journal of Epidemiology* 2013; 42: 111–127.
- [119] Fraser A, Macdonald-Wallis C, Tilling K, et al. Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology* 2013; 42: 97–110.
- [120] Northstone K, Lewcock M, Groom A, et al. The avon longitudinal study of parents and children (ALSPAC): An update on the enrolled sample of index children in 2019. *Wellcome Open Research* 2019; 4: Online.
- [121] Audrey S, Brown L, Campbell R, et al. Young people’s views about consenting to data linkage: findings from the PEARL qualitative study. *BMC Medical Research Methodology* 2016; 16: 1–13.
- [122] Longitudinal Population Studies Working Group. *Longitudinal population studies strategy*. Wellcome Trust, https://wellcome.org/sites/default/files/longitudinal-population-studies-strategy_0.pdf (2017).
- [123] Shiells K, Di Cara N, Skatova A, et al. Participant acceptability of digital footprint data collection strategies: An exemplar approach to participant engagement and involvement in the ALSPAC birth cohort study. *International Journal of Population Data Science* 2022; 5: Online.
- [124] Bayer JB, Trieu P, Ellison NB. Social media elements, ecologies, and effects. *Annual Review of Psychology* 2020; 71: 471–497.

-
- [125] Office for National Statistics. Internet access: Households and individuals, <https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/datasets/internetaccesshouseholdsandindividualsreferencetables> (2020).
- [126] Hollis C, Livingstone S, Sonuga-Barke E. The role of digital technology in children and young people's mental health—a triple-edged sword? *The Journal of Child Psychology and Psychiatry* 2020; 61: 837–841.
- [127] Lee KS, Lee H, Myung W, et al. Advanced daily prediction model for national suicide numbers with social media data. *Psychiatry Investigation* 2018; 15: 344.
- [128] Roy A, Nikolitch K, McGinn R, et al. A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine* 2020; 3: 1–12.
- [129] Santarossa S, Woodruff SJ. # SocialMedia: Exploring the relationship of social networking sites on body image, self-esteem, and eating disorders. *Social Media+ Society* 2017; 3: 2056305117704407.
- [130] Arendt F, Scherr S, Romer D. Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults. *New Media & Society* 2019; 21: 2422–2442.
- [131] Hamm MP, Newton AS, Chisholm A, et al. Prevalence and effect of cyberbullying on children and young people: A scoping review of social media studies. *JAMA Pediatrics* 2015; 169: 770–777.
- [132] Craig W, Boniel-Nissim M, King N, et al. Social media use and cyber-bullying: A cross-national analysis of young people in 42 countries. *Journal of Adolescent Health* 2020; 66: S100–S108.
- [133] Naslund JA, Grande SW, Aschbrenner KA, et al. Naturally occurring peer support through social media: The experiences of individuals with severe mental illness using YouTube. *PLoS one* 2014; 9: e110171.

- [134] Orben A, Przybylski AK. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* 2019; 3: 173–182.
- [135] Orben A. Teenagers, screens and social media: A narrative review of reviews and key studies. *Social Psychiatry and Psychiatric Epidemiology* 2020; 55: 407–414.
- [136] Appel M, Marker C, Gnambs T. Are social media ruining our lives? A review of meta-analytic evidence. *Review of General Psychology* 2020; 24: 60–74.
- [137] Coyne SM, Rogers AA, Zurcher JD, et al. Does time spent using social media impact mental health?: An eight year longitudinal study. *Computers in Human Behavior* 2020; 104: 106–160.
- [138] Primack BA, Shensa A, Sidani JE, et al. Temporal associations between social media use and depression. *American Journal of Preventive Medicine* 2021; 60: 179–188.
- [139] Bennett BL, Whisenhunt BL, Hudson DL, et al. Examining the impact of social media on mood and body dissatisfaction using ecological momentary assessment. *Journal of American College Health* 2020; 68: 502–508.
- [140] Valkenburg P, Beyens I, Pouwels JL, et al. Social media use and adolescents' self-esteem: Heading for a person-specific media effects paradigm. *Journal of Communication* 2021; 71: 56–78.
- [141] Beyens I, Pouwels JL, Driel II van, et al. The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports* 2020; 10: 1–11.
- [142] Weinstein E. The social media see-saw: Positive and negative influences on adolescents' affective well-being. *New Media & Society* 2018; 20: 3597–3623.
- [143] Guntuku SC, Yaden DB, Kern ML, et al. Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences* 2017; 18: 43–49.
- [144] Amir S, Dredze M, Ayers JW. Mental health surveillance over social media with digital cohorts. In: *Proceedings of the sixth workshop on computational linguistics and clinical psychology*. 2019, pp. 114–120.

-
- [145] Pew Research Center. Demographics of social media users and adoption in the united states, <https://www.pewresearch.org/internet/fact-sheet/social-media/> (2021).
- [146] Sloan L. Who tweets in the United Kingdom? Profiling the twitter population using the british social attitudes survey 2015. *Social Media + Society* 2017; 1: Online.
- [147] Mellon J, Prosser C. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics* 2017; 4: Online.
- [148] Heckman JJ. Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society* 1979; 153–161.
- [149] Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics* 2009; 42: 377–381.
- [150] Szreter SR. The genesis of the registrar-general’s social classification of occupations. *British Journal of Sociology* 1984; 522–546.
- [151] Costello EJ, Angold A. Scales to assess child and adolescent depression: Checklists, screens, and nets. *Journal of the American Academy of Child & Adolescent Psychiatry* 1988; 27: 726–737.
- [152] Angold A, Costello EJ, Messer SC, et al. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research* 1995; 5: 237–249.
- [153] Eyre O, Jones RB, Agha SS, et al. Validation of the short mood and feelings questionnaire in young adulthood. *Preprint*. Epub ahead of print 2021. DOI: [10.1101/2021.01.22.21250311](https://doi.org/10.1101/2021.01.22.21250311).

- [154] Micali N, Horton N, Crosby R, et al. Eating disorder behaviours amongst adolescents: Investigating classification, persistence and prospective associations with adverse outcomes using latent class models. *European Child & Adolescent Psychiatry* 2017; 26: 231–240.
- [156] Tennant R, Hiller L, Fishwick R, et al. The warwick-edinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of life Outcomes* 2007; 5: 1–13.
- [156] Tennant R, Hiller L, Fishwick R, et al. The warwick-edinburgh mental well-being scale (WEMWBS): Development and UK validation. *Health and Quality of life Outcomes* 2007; 5: 1–13.
- [157] Ng Fat L, Scholes S, Boniface S, et al. Evaluating and establishing national norms for mental wellbeing using the short warwick–edinburgh mental well-being scale (SWEMWBS): Findings from the health survey for england. *Quality of Life Research* 2017; 26: 1129–1144.
- [158] Ringdal R, Bradley Eilertsen M-E, Bjørnsen HN, et al. Validation of two versions of the warwick-edinburgh mental well-being scale among norwegian adolescents. *Scandinavian Journal of Public Health* 2018; 46: 718–725.
- [159] McKay MT, Andretta JR. Evidence for the psychometric validity, internal consistency and measurement invariance of warwick edinburgh mental well-being scale scores in scottish and irish adolescents. *Psychiatry research* 2017; 255: 382–386.
- [160] Diener E, Emmons RA, Larsen RJ, et al. The satisfaction with life scale. *Journal of Personality Assessment* 1985; 49: 71–75.
- [161] Pavot W, Diener E. The satisfaction with life scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology* 2008; 3: 137–152.
- [162] Lyubomirsky S, Lepper HS. A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research* 1999; 46: 137–155.
- [163] McCullough ME, Emmons RA, Tsang J-A. The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology* 2002; 82: 112.

-
- [164] Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology* 1994; 67: 1063.
- [165] Steger MF, Frazier P, Oishi S, et al. The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology* 2006; 53: 80.
- [166] Deci EL, Ryan RM. The " what" and " why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry* 2000; 11: 227–268.
- [167] R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, <https://www.R-project.org/> (2020).
- [168] Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *Journal of Open Source Software* 2019; 4: 1686.
- [169] Wickham H. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag, 2016.
- [170] Winstone L, Mars B, Haworth CM, et al. Adolescent social media user types and their mental health and well-being: Results from a longitudinal survey of 13–14-year-olds in the united kingdom. *JCPP Advances* 2022; 2: e12071.
- [171] Boyd A, Van de Velde S, Vilagut G, et al. Gender differences in mental disorders and suicidality in europe: Results from a large cross-sectional population-based study. *Journal of Affective Disorders* 2015; 173: 245–254.
- [172] Matud MP, López-Curbelo M, Fortes D. Gender and psychological well-being. *International Journal of Environmental Research and Public Health* 2019; 16: 3531.
- [173] Davies A, Song J, Sharp C. Social media engagement and health. *International Journal of Population Data Science* 2019; 4: Online.

- [174] McManus S, Bebbington PE, Jenkins R, et al. *Mental health and well-being in England: The adult psychiatric morbidity survey 2014*. NHS digital, https://files.digital.nhs.uk/pdf/q/3/mental_health_and_wellbeing_in_england_full_report.pdf (2016).
- [175] Tabor D, Stockley L. *Personal well-being in the UK: October 2016 to september 2017*. Office for National Statistics, <https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/bulletins/measuringnationalwellbeing/october2016toseptember2017> (2018).
- [176] Glaesmer H, Rief W, Martin A, et al. Psychometric properties and population-based norms of the life orientation test revised (LOT-r). *British Journal of Health Psychology* 2012; 17: 432–445.
- [177] Clarke A, Putz R, Friede T, et al. *Warwick-edinburgh mental well-being scale (WEMWBS) acceptability and validation in english and scottish secondary school students (the WAVES project) Glasgow*. NHS Health Scotland, 2010.
- [178] Przybylski AK, Weinstein N. A large-scale test of the goldilocks hypothesis: Quantifying the relations between digital-screen use and the mental well-being of adolescents. *Psychological Science* 2017; 28: 204–215.
- [179] Shaw H, Ellis DA, Geyer K, et al. Quantifying smartphone ‘use’: Choice of measurement impacts relationships between ‘usage’ and health. *Technology, Mind, and Behavior* 2020; 1: Online.
- [180] Kross E, Verduyn P, Sheppes G, et al. Social media and well-being: Pitfalls, progress, and next steps. *Trends in Cognitive Sciences* 2020; 25: 55–66.
- [181] Meier A, Reinecke L. Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review. *Communication Research* 2020; 48: 1182–1209.
- [182] Diener E, Oishi S, Tay L. Advances in subjective well-being research. *Nature Human Behaviour* 2018; 2: 253–260.

-
- [183] Lin J-H. Need for relatedness: A self-determination approach to examining attachment styles, facebook use, and psychological well-being. *Asian Journal of Communication* 2016; 26: 153–173.
- [184] Berezan O, Krishen AS, Agarwal S, et al. The pursuit of virtual happiness: Exploring the social media experience across generations. *Journal of Business Research* 2018; 89: 455–461.
- [185] Royal Society for Public Health. *StatusOfMind: Social Media and Young People's Mental Health and Wellbeing*. Royal Society For Public Health, <https://ed4health.co.uk/wp-content/uploads/2018/12/RSPH-Status-of-Mind-report.pdf> (2018).
- [186] Maughan B, Collishaw S, Stringaris A. Depression in childhood and adolescence. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 2013; 22: 35.
- [187] Di Cara NH, Boyd A, Tanner AR, et al. Views on social media and its linkage to longitudinal data from two generations of a UK cohort study. *Wellcome Open Research* 2020; 5: Online.
- [188] Wang J-L, Gaskin J, Rost DH, et al. The reciprocal relationship between passive social networking site (SNS) usage and users' subjective well-being. *Social Science Computer Review* 2018; 36: 511–522.
- [189] Winstone L, Mars B, Haworth CM, et al. Types of social media use and digital stress in early adolescence. *The Journal of Early Adolescence*. Epub ahead of print 2022. DOI: [10.1177/02724316221105560](https://doi.org/10.1177/02724316221105560).
- [190] Frison E, Eggermont S. Toward an integrated and differential approach to the relationships between loneliness, different types of facebook use, and adolescents' depressed mood. *Communication Research* 2015; 47: 701–728.
- [191] Parry DA, Davidson BI, Sewall CJ, et al. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour* 2021; 5: 1535–1547.

-
- [192] Ernala SK, Burke M, Leavitt A, et al. How well do people report time spent on facebook? An evaluation of established survey questions with recommendations. In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–14.
- [193] Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. *Journal of Biomedical Informatics* 2017; 66: 82–94.
- [194] Burnap P, Gibson R, Sloan L, et al. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies* 2016; 41: 230–233.
- [195] Maynard D, Roberts I, Greenwood MA, et al. A framework for real-time semantic social media analysis. *Journal of Web Semantics* 2017; 44: 75–88.
- [196] Green E, Ritchie F, Webber D, et al. *Enabling Data Linkage to Maximise the Value of Public Health Research Data: full report 2015*. The Wellcome Trust, <https://wellcome.ac.uk/sites/default/files/enabling-data-linkage-to-maximise-value-of-public-health-research-data-phrdf-mar15.pdf> (2015).
- [197] McGrail K, Jones K, Akbari A, et al. A Position Statement on Population Data Science. *International Journal of Population Data Science* 2018; 3: Online.
- [198] Ford E, Boyd A, Bowles JKF, et al. Our data, our society, our health: A vision for inclusive and transparent health data science in the United Kingdom and beyond. *Learning Health Systems* 2019; 3: 1–12.
- [199] Aitken M, Porteous C, Creamer E, et al. Who benefits and how? Public expectations of public benefits from data-intensive health research. *Big Data & Society* 2018; 5: Online.
- [200] Carter P, Laurie GT, Dixon-Woods M. The social licence for research: Why care.data ran into trouble. *Journal of Medical Ethics* 2015; 41: 404–409.

-
- [201] Stockdale J, Cassell J, Ford E. ‘Giving something back’: A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland [version 2; referees: 2 approved]. *Wellcome Open Research* 2019; 3: Online.
- [202] Aitken M, De St Jorre J, Pagliari C, et al. Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies. *BMC Medical Ethics* 2016; 17: 1–24.
- [203] Xafis V. The acceptability of conducting data linkage research without obtaining consent: lay people’s views and justifications. *BMC Medical Ethics* 2015; 16: 1–16.
- [204] Ford E, Curlewis K, Wongkoblap A, et al. Public Opinions on Using Social Media Content to Identify Users With Depression and Target Mental Health Care Advertising: Mixed Methods Survey. *JMIR Mental Health* 2019; 6: e12942.
- [205] Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: A qualitative study. *BMC Medical Ethics* 2016; 17: Online.
- [206] Golder S, Ahmed S, Norman G, et al. Attitudes toward the ethics of research using social media: A systematic review. *Journal of Medical Internet Research* 2017; 19: e195.
- [207] Beninger K, Fry A, Jago N, et al. *Research using social media; users’ views*. NatCen, 2014.
- [208] Moreno MA, Goniou N, Moreno PS, et al. Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior, and social networking* 2013; 16: 708–713.
- [209] Moreno MA, Grant A, Kacvinsky L, et al. Older adolescents’ views regarding participation in Facebook research. *Journal of Adolescent Health* 2012; 51: 439–444.
- [210] Townsend L. Social Media Research and Ethics. Epub ahead of print 2016. DOI: [10.4135/9781526413642](https://doi.org/10.4135/9781526413642).

- [211] Sloan L, Quan-Haase A. *The SAGE Handbook of Social Media Research Methods*. SAGE Publications.
- [212] Millican C, Mansfield C. *Summary report of qualitative research into public attitudes to personal data and linking personal data*. The Wellcome Trust, 2013.
- [213] Al Baghal T, Sloan L, Jessop C, et al. Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. *Social Science Computer Review* 2019; 1–16.
- [214] Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Quality & Safety* 2016; 25: 414–423.
- [215] The Guardian. The cambridge analytica files, <https://www.theguardian.com/news/series/cambridge-analytica-files> (2018).
- [216] Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* 2007; 19: 349–357.
- [217] Harrell MC, Bradley MA. *Data collection methods. Semi-structured interviews and focus groups*. Rand National Defense Research Inst santa monica ca, 2009.
- [218] Di Cara N. Online supplementary material for "views on social media and its linkage to longitudinal data from two generations of a UK cohort study". DOI: [10.17605/OSF.IO/6RX2Z](https://doi.org/10.17605/OSF.IO/6RX2Z).
- [219] Nowell LS, Norris JM, White DE, et al. Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods* 2017; 16: 1609406917733847.
- [220] Ofcom. *Adults' media use and attitudes report*. May, Ofcom.
- [221] Sweney M. Is facebook for old people? Over-55s flock in as the young leave., <https://www.theguardian.com/technology/2018/feb/12/is-facebook-for-old-people-over-55s-flock-in-as-the-young-leave> (2018).

-
- [222] Ofcom. *Adults' media use and attitudes report*. Ofcom, <https://www.ofcom.org.uk/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes> (2013).
- [223] Bayindir N, Kavanagh D. *Social flagship report 2018*. Global Web Index, <https://www.globalwebindex.com/hubfs/Downloads/Social-H2-2018-report.pdf> (2018).
- [224] Graham C, Mathis K. Frappe, stalking and whores: Semantics and social narrative on facebook. In: *Crossing channels, crossing realms: Immersive worlds and transmedia narratives*. Inter-Disciplinary Press, 2013, pp. 135–145.
- [225] O'Reilly M, Dogra N, Whiteman N, et al. Is social media bad for mental health and wellbeing? Exploring the perspectives of adolescents. *Clinical Child Psychology and Psychiatry* 2018; 23: 601–613.
- [226] Prensky M. Digital Natives, Digital Immigrants Part 1. *On the Horizon* 2001; 9: 2–6.
- [227] Helsper EJ, Eynon R. Digital natives: Where is the evidence? *British educational research journal* 2010; 36: 503–520.
- [228] Cornish R, Tilling K, Boyd A, et al. Using Linkage to Electronic Primary Care Records to Evaluate Recruitment and Nonresponse Bias in The Avon Longitudinal Study of Parents and Children. *Epidemiology* 2015; 26: e41.
- [229] Smithson J. Using and analysing focus groups: Limitations and possibilities. *International journal of Social Research Methodology* 2000; 3: 103–119.
- [230] Hunter RF, Gough A, O'Kane N, et al. Ethical issues in social media research for public health. *American Journal of Public Health* 2018; 108: 343–348.
- [231] De Choudhury M, Sharma SS, Logar T, et al. Gender and cross-cultural differences in social media disclosures of mental illness. 2017; 353–369.

-
- [232] Statista Research Department. Twitter – statistics and facts, <https://www.statista.com/topics/737/twitter/> (2021).
- [233] Griffiths E, Greci C, Kotrotsios Y, et al. *Handbook on statistical disclosure control*. UK: Safe Data Access Professionals Group; Safe Data Access Professionals Group, 2019. Epub ahead of print 2019. DOI: [10.6084/m9.figshare.9958520](https://doi.org/10.6084/m9.figshare.9958520).
- [234] Hutto C, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the international AAAI conference on web and social media*. 2014.
- [235] Dodds PS, Harris KD, Kloumann IM, et al. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one* 2011; 6: e26752.
- [236] Pennebaker JW, Boyd RL, Jordan K, et al. *The development and psychometric properties of LIWC2015*. 2015.
- [237] Loria S. TextBlob documentation (v 0.16.0), <https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis> (2020).
- [238] Northstone K, Howarth S, Smith D, et al. The Avon Longitudinal Study of Parents and Children - a resource for COVID-19 research: Questionnaire data capture april-may 2020. *Wellcome Open Research* 2020; 5: Online.
- [239] Messer SC, Angold A, Costello EJ, et al. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents: Factor composition and structure across development. *International Journal of Methods in Psychiatric Research* 1995; 5: 251–262.
- [240] Spitzer RL, Kroenke K, Williams JB, et al. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine* 2006; 166: 1092–1097.
- [241] Sjoberg DD, Whiting K, Curry M, et al. Reproducible summary tables with the gtsummary package. *The R Journal* 2021; 13: 570–580.

-
- [242] Zhu H. *kableExtra: Construct complex table with 'kable' and pipe syntax*, <https://CRAN.R-project.org/package=kableExtra> (2021).
- [243] Jäckle A, Beninger K, Burton J, et al. Understanding data linkage consent in longitudinal surveys. In: *Advances in longitudinal survey methodology*. Wiley Online Library, 2021, pp. 122–150.
- [244] Jäckle A, Burton J, Couper MP, et al. *Understanding and improving data linkage consent in surveys*. Understanding Society at the Institute for Social; Economic Research, https://cadmus.eui.eu/bitstream/handle/1814/73121/WP_202101.pdf?sequence=1 (2021).
- [245] Thornby M, Calderwood L, Kotecha M, et al. *Collecting multiple data linkage consents in a mixed mode survey: Evidence and lessons learnt from next steps*. Centre for Longitudinal Studies, <https://cls.ucl.ac.uk/wp-content/uploads/2017/11/CLS-WP-201713-Collecting-Multiple-Data-Linkage-Consents-in-a-Mixed-Mode-Survey-Evidence-and-Lessons-Learnt-from-Next-Steps.pdf> (2017).
- [246] Correia RB, Wood IB, Bollen J, et al. Mining social media data for biomedical signals and health-related behavior. *Annual Review of Biomedical Data Science* 2020; 3: 433–458.
- [247] Loi M. The digital phenotype: A philosophical and ethical exploration. *Philosophy & Technology* 2019; 32: 155–171.
- [248] Ruths D, Pfeffer J. Social media for large studies of behaviour. *Science* 2014; 346: 1063–1064.
- [249] Williams ML, Burnap P, Javed A, et al. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology* 2019; 60: 93–117.
- [250] Alizadeh M, Weber I, Cioffi-Revilla C, et al. Psychology and morality of political extremists: Evidence from twitter language analysis of alt-right and antifa. *EPJ Data Science* 2019; 8: 17.

-
- [251] Prieto VM, Matos S, Álvarez M, et al. Twitter: A good place to detect health conditions. *PLoS one* 2014; 9: e86191.
- [252] Vos T, Barber RM, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the global burden of disease study 2013. *The Lancet* 2015; 386: 743–800.
- [253] Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: Findings from the global burden of disease study 2010. *The Lancet* 2013; 382: 1575–1586.
- [254] Trautmann S, Rehm J, Wittchen H-U. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? *EMBO Reports* 2016; 17: 1245–1249.
- [255] Fekadu W, Mihiretu A, Craig TK, et al. Multidimensional impact of severe mental illness on family members: Systematic review. *BMJ Open* 2019; 9: e032391.
- [256] Medical Research Council. *Strategy for lifelong mental health research*, <https://www.ukri.org/publications/mrc-strategy-for-lifelong-mental-health-research/> (2017).
- [257] Naslund JA, Gonsalves PP, Gruebner O, et al. Digital innovations for global mental health: Opportunities for data science, task sharing, and early intervention. *Current Treatment Options in Psychiatry* 2019; 1–15.
- [258] Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry connecting with technology. *JAMA Psychiatry* 2016; 73: 3–4.
- [259] World Health Organization. *Comprehensive mental health action plan 2013-2020*, http://www.who.int/mental_health/mhgap/consultation_global_mh_action_plan_ (2013).

-
- [260] Solhan MB, Trull TJ, Jahng S, et al. Clinical assessment of affective instability: Comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment* 2009; 21: 425–436.
- [261] Chancellor S, Baumer EP, De Choudhury M. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM Conference on Human-Computer Interaction* 2019; 3: 1–32.
- [262] Ernala SK, Birnbaum ML, Candan KA, et al. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. *Proceedings of the ACM Conference on Human Factors in Computing Systems* 2019; 1–16.
- [263] Fried EI. What are psychological constructs? On the nature and statistical modelling of emotions, intelligence, personality traits and mental disorders. *Health Psychology Review* 2017; 11: 130–134.
- [264] Hirschfeld RM. The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. *The Primary Care Companion for CNS Disorders* 2001; 3: 24412.
- [265] Conway M, others. Ethical issues in using twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. *Journal of Medical Internet Research* 2014; 16: e3617.
- [266] Camerer CF, Dreber A, Holzmeister F, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* 2018; 2: 637–644.
- [267] Klein RA, Vianello M, Hasselman F, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 2018; 1: 443–490.
- [268] Monteiro M, Keating E. Managing misunderstandings. *Science Communication* 2009; 31: 6–28.

- [269] James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. 7th ed. Springer, 2013.
- [270] Keyes CLM. Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology* 2005; 73: 539–548.
- [271] Slade M. Mental illness and well-being: The central importance of positive psychology and recovery approaches. *BMC Health Services Research* 2010; 10: 26.
- [272] Kim J, Lee D, Park E. Machine learning for mental health in social media: Bibliometric study. *Journal of Medical Internet Research* 2021; 23: e24870.
- [273] Abd Rahman R, Omar K, Mohd Noah SA, et al. A survey on mental health detection in online social network. *International Journal on Advanced Science, Engineering and Information Technology* 2018; 8: 1431.
- [274] Sundarrajan A, Aneesha M. Survey on detection of metal illnesses by analysing twitter data. *International Journal of Engineering and Technology(UAE)* 2018; 7: 37–41.
- [275] Giuntini FT, Cazzolato MT, Reis M de JD dos, et al. A review on recognizing depression in social networks: Challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing* 2020; 11: 4713–4729.
- [276] Edo-Osagie O, De La Iglesia B, Lake I, et al. A scoping review of the use of twitter for public health research. *Computers in Biology and Medicine* 2020; 122: 103770.
- [277] Verma B, Gupta S, Goel L. A survey on sentiment analysis for depression detection. In: *Advances in automation, signal processing, instrumentation, and control*. Springer, 2021, pp. 13–24.
- [278] Bilal U, Khan FH. An analysis of depression detection techniques from online social networks. In: *International conference on intelligent technologies and applications*. Springer, 2019, pp. 296–308.
- [279] Abd Rahman R, Omar K, Noah SAM, et al. Application of machine learning methods in mental health detection: A systematic review. *IEEE Access* 2020; 8: 183952–183964.

-
- [280] Le Glaz A, Haralambous Y, Kim-Dufor D-H, et al. Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research* 2021; 23: e15708.
- [281] Zunic A, Corcoran P, Spasic I, et al. Sentiment analysis in health and well-being: Systematic review. *JMIR Medical Informatics* 2020; 8: e16023.
- [282] Babu NV, Kanaga E. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science* 2022; 3: 1–20.
- [283] Pourmand A, Roberson J, Caggiula A, et al. Social media and suicide: A review of technology-based epidemiology and risk assessment. *Telemedicine and e-Health* 2019; 25: 880–888.
- [284] Castillo-Sánchez G, Marques G, Dorrnzoro E, et al. Suicide risk assessment using machine learning and social networks: A scoping review. *Journal of Medical Systems* 2020; 44: 1–15.
- [285] Beriwal M, Agrawal S. Techniques for suicidal ideation prediction: A qualitative systematic review. In: *2021 international conference on INnovations in intelligent SysTems and applications (INISTA)*. IEEE, 2021, pp. 1–8.
- [286] William D, Suhartono D. Text-based depression detection on social media posts: A systematic literature review. *Procedia Computer Science* 2021; 179: 582–589.
- [287] Skaik R, Inkpen D. Using social media for mental health surveillance: A review. *Association for Computing Machinery Computing Surveys (CSUR)* 2020; 53: 1–31.
- [288] Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews* 2016; 5: Online.
- [289] Mori K, Haruno M. Differential ability of network and natural language information on social media to predict interpersonal and mental health traits. *Journal of Personality* 2021; 89: 228–243.

-
- [290] Coello-Guilarte L, Ortega-Mendoza RM, Villaseñor-Pineda L, et al. Crosslingual depression detection in twitter using bilingual word alignments. In: *International conference of the cross-language evaluation forum for european languages*. Springer, 2019, pp. 49–61.
- [291] Shen G, Jia J, Nie L, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. *IJCAI International Joint Conference on Artificial Intelligence* 2017; 3838–3844.
- [292] O’Dea B, Wan S, Batterham PJ, et al. Detecting suicidality on twitter. *Internet Interventions* 2015; 2: 183–188.
- [293] Sawhney R, Manchanda P, Singh R, et al. A computational approach to feature extraction for identification of suicidal ideation in tweets. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop* 2018; 91–98.
- [294] AlSagri H, Ykhlef M. Quantifying feature importance for detecting depression using random forest. *International Journal of Advanced Computer Science and Applications* 2020; 11: 628–635.
- [295] Yazdavar AH, Mahdavinejad MS, Bajaj G, et al. Multimodal mental health analysis in social media. *Plos one* 2020; 15: e0226248.
- [296] Birnbaum ML, Kiranmai Ernala S, Asra, et al. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research* 2017; 19: e289.
- [297] Coppersmith G, Leary R, Crutchley P, et al. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights* 2018; 10: Online.
- [298] Coppersmith G, Ngo K, Leary R, et al. Exploratory analysis of social media prior to a suicide attempt. In: *Proceedings of the third workshop on computational linguistics and clinical psychology*. 2016, pp. 106–117.

-
- [299] He L, Luo J. ‘What makes a pro eating disorder hashtag’: Using hashtags to identify pro eating disorder tumblr posts and twitter users. *Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016* 2016; 3977–3979.
- [300] Kang K, Yoon C, Kim EY. Identifying depressive users in twitter using multimodal analysis. In: *2016 international conference on big data and smart computing, BigComp 2016*. 2016, pp. 231–238.
- [301] Moulahi B, Azé J, Bringay S. DARE to care: A context-aware framework to track suicidal ideation on social media. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2017; 10570 LNCS: 346–353.
- [302] Resnik P, Armstrong W, Claudino L, et al. The University of Maryland CLPsych 2015 shared task system. 2015; 54–60.
- [303] Tsugawa S, Kikuchi Y, Kishino F, et al. Recognizing depression from twitter activity. In: *CHI '15: Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 3187–3196.
- [304] Waheed T, Aslam M, Awais M. Predicting mental-illness from twitter activity using activity theory based context ontology. *Journal of Medical Imaging and Health Informatics* 2019; 9: 1224–1233.
- [305] Yin Z, Fabbri D, Rosenbloom ST, et al. A scalable framework to detect personal health mentions on twitter. *Journal of Medical Internet Research* 2015; 17: e138.
- [306] Zhou TH, Hu GL, Wang L. Psychological disorder identifying method based on emotion perception over social networks. *International Journal of Environmental Research and Public Health* 2019; 16: 953.
- [307] Braithwaite SR, Giraud-Carrier C, West J, et al. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health* 2016; 3: e21.

-
- [308] Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in twitter. In: *Proceedings of the 8th international conference on weblogs and social media, ICWSM 2014*, pp. 579–582.
- [309] Wang T, Brede M, Ianni A, et al. Detecting and characterizing eating-disorder communities on social media. *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining 2017*; 91–100.
- [310] Burnap P, Colombo G, Amery R, et al. Multi-class machine classification of suicide-related communication on twitter. *Online Social Networks and Media 2017*; 2: 32–44.
- [311] Jamil Z, Inkpen D, Buddhitha P, et al. Monitoring tweets for depression to detect at-risk users. 2017; 32–40.
- [312] Vioules MJ, Moulahi B, Aze J, et al. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development 2018*; 62: 1–12.
- [313] Victor DB, Kawsher J, Labib MS, et al. Machine learning techniques for depression analysis on social media-case study on bengali community. In: *2020 4th international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2020, pp. 1118–1126.
- [314] Alabdulkreem E. Prediction of depressed arab women using their tweets. *Journal of Decision Systems 2021*; 30: 102–117.
- [315] Guibon G, Ochs M, Bellot P. From emojis to sentiment analysis. In: *Workshop affect compagnon artificiel interaction 2016*. 2016.
- [316] Weerasinghe J, Morales K, Greenstadt R. ‘Because... I was told... So much’: Linguistic indicators of mental health status on twitter. *Proceedings on Privacy Enhancing Technologies 2019*; 2019: 152–171.
- [317] Yazdavar AH, Al-Olimat HS, Ebrahimi M, et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017 2017*; 1191–1198.

-
- [318] Burnap P, Colombo W, Scourfield J. Machine classification and analysis of suicide-related communication on twitter. In: *26th association for computing machinery conference on hypertext & social media - HT '15*. 2015, pp. 75–84.
- [319] De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: *Proceedings of the 5th annual ACM web science conference*. 2013, pp. 47–56.
- [320] De Choudhury M, Gamon M, Counts S, et al. Predicting depression via social media. In: *Seventh international AAAI conference on weblogs and social media*. 2013.
- [321] Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in twitter. In: *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. Association for Computational Linguistics, 2015, pp. 51–60.
- [322] Deshpande M, Rao V. Depression detection using emotion artificial intelligence. *Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2017* 2018; 858–862.
- [323] Kumar A, Sharma A, Arora A. Anxious depression prediction in real-time social data. In: *International conference on advances in engineering science management & technology (ICAESMT)*. 2019, pp. 1–7.
- [324] Orabi AH, Buddhitha P, Orabi MH, et al. Deep learning for depression detection of twitter users. In: *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic*. Association for Computational Linguistics, 2018, pp. 88–97.
- [325] Resnik P, Armstrong W, Claudino L, et al. Beyond LDA: Exploring supervised topic modeling for depression-related language in twitter. 2015; 1: 99–107.

-
- [326] Astoveza G, Obias RJP, Palcon RJL, et al. Suicidal behavior detection on twitter using neural network. *IEEE Region 10 Annual International Conference, Proceedings/TENCON 2019*; 2018-October: 657–662.
- [327] Oyong I, Utami E, Luthfi ET. Natural language processing and lexical approach for depression symptoms screening of indonesian twitter user. *Proceedings of 2018 10th International Conference on Information Technology and Electrical Engineering: Smart Technology for Better Society, ICITEE 2018* 2018; 359–364.
- [328] Chiroma F, Liu H, Cocea M. Text classification for suicide related tweets. *Proceedings - International Conference on Machine Learning and Cybernetics* 2018; 2: 587–592.
- [329] Reece AG, Reagan AJ, Lix KL, et al. Forecasting the onset and course of mental illness with twitter data. *Scientific Reports* 2017; 7: 1–11.
- [330] Loper E, Bird S. Nltk: The natural language toolkit. In: *Proceedings of the ACL workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*. 2002, pp. 63--70.
- [331] Saha K, Chan L, De Barbaro K, et al. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the Association for Computing Machinery on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2017; 1: 1–27.
- [332] Pedersen T. Screening twitter users for depression and PTSD with lexical depression lists. In: *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 2015, pp. 46–53.
- [333] Raeder T, Forman G, Chawla NV. Learning from imbalanced data: Evaluation matters. *Intelligent Systems Reference Library* 2012; 23: 315–331.
- [334] Bussola N, Marcolini A, Bruno Kessler F, et al. Not again! Data leakage in digital pathology. In: *Pattern recognition. ICPR international workshops and challenges: Virtual event, january 10–15, 2021, proceedings, part i*. 2021.

-
- [335] Tsakalidis A, Liakata M, Damoulas T, et al. Can we assess mental health through social media and smart devices? Addressing bias in methodology and evaluation. In: *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 2018, pp. 407–423.
- [336] Gewin V. An open mind on open data. *Nature* 2016; 529: 117–119.
- [337] Cho G, Yim J, Choi Y, et al. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation* 2019; 16: 262–269.
- [338] Yin Z, Sulieman LM, Malin BA. A systematic literature review of machine learning in online personal health data. *Journal of the American Medical Informatics Association: JAMIA* 2019; 26: 561–576.
- [339] Leis A, Mayer M-A, Ronzano F, et al. Clinical-based and expert selection of terms related to depression for twitter streaming and language analysis. *Studies in Health Technology and Informatics* 2020; 16: 921–925.
- [340] Schofield P. Big data in mental health research - do the ns justify the means? Using large data-sets of electronic health records for mental health research. *BJPsych bulletin* 2017; 41: 129–132.
- [341] Loveys K, Torrez J, Fine A, et al. Cross-cultural differences in language markers of depression online. Association for Computational Linguistics (ACL), 2018, pp. 78–87.
- [342] Aguirre C, Harrigian K, Dredze M. Gender and racial fairness in depression research using social media. In: *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*. 2021, pp. 2932–2949.
- [343] Conway M, O’Connor D. Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology* 2016; 9: 77–82.
- [344] Harrigian K, Aguirre C, Dredze M. On the state of social media data for mental health research. In: *Proceedings of the seventh workshop on computational linguistics and clinical psychology: Improving access*. Association for Computational Linguistics, pp. 15–24.

-
- [345] Aalbers G, McNally RJ, Heeren A, et al. Social media and depression symptoms: A network perspective. *Journal of Experimental Psychology: General* 2018; 1–9.
- [346] Borsboom D. A network theory of mental disorders. *World Psychiatry* 2017; 16: 5–13.
- [347] Benton A, Mitchell M, Hovy D. Multi-task learning for mental health using social media text. In: *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers*, pp. 152–162.
- [348] Ribeiro JD, Huang X, Fox KR, et al. Predicting imminent suicidal thoughts and nonfatal attempts: The role of complexity. *Clinical Psychological Science* 2019; 7: 941–957.
- [349] Bechini A, Bondielli A, Ducange P, et al. Addressing event-driven concept drift in twitter stream: A stance detection application. *IEEE Access* 2021; 9: 77758–77770.
- [350] Yoo DW, Birnbaum ML, Van Meter AR, et al. Designing a clinician-facing tool for using insights from patients’ social media activity: Iterative co-design approach. *JMIR Mental Health* 2020; 7: e16969.
- [351] Bucur A-M, Jang H, Liza FF. Capturing changes in mood over time in longitudinal data using ensemble methodologies. In: *Proceedings of the eighth workshop on computational linguistics and clinical psychology*. 2022, pp. 205–212.
- [352] Chancellor S, Birnbaum ML, Caine ED, et al. A taxonomy of ethical tensions in inferring mental health states from social media. *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency* 2019; 79–88.
- [353] Young SD, Garrett R. Ethical issues in addressing social media posts about suicidal intentions during an online study among youth: Case study. *Journal of Medical Internet Research* 2018; 5: e33.
- [354] Gebu T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Communications of the Association for Computing Machinery* 2021; 64: 86–92.

-
- [355] Mohammad SM. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics* 2022; 48: 239–278.
- [356] McGorry PD, Ratheesh A, O’Donoghue B. Early intervention—an implementation challenge for 21st century mental health care. *JAMA Psychiatry* 2018; 75: 545–546.
- [357] Tsakalidis A. *Nowcasting user behaviour with social media and smart devices on a longitudinal basis: From macro- to micro-level modelling*. PhD Thesis, University of Warwick, 2022.
- [358] Kolliakou A, Bakolis I, Chandran D, et al. Mental health-related conversations on social media and crisis episodes: A time-series regression analysis. *Scientific Reports* 2020; 10: 1–7.
- [359] Melcher J, Lavoie J, Hays R, et al. Digital phenotyping of student mental health during COVID-19: An observational study of 100 college students. *Journal of American College Health* 2021; 1–13.
- [360] Blair J, Hsu C-Y, Qiu L, et al. Using tweets to assess mental well-being of essential workers during the covid-19 pandemic. In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 2021, pp. 1–6.
- [361] Jaidka K, Giorgi S, Schwartz HA, et al. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences* 2020; 117: 10165–10171.
- [362] Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. *Nature Medicine* 2020; 26: 1183–1192.
- [363] Ivanković D, Barbazza E, Bos V, et al. Features constituting actionable COVID-19 dashboards: Descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. *Journal of Medical Internet Research* 2021; 23: e25682.
- [364] Di Cara NH, Song J, Maggio V, et al. Mapping population vulnerability and community support during COVID-19: A case study from wales. *International Journal of Population Data Science* 2020; 5: Online.

- [365] Pellert M, Lasser J, Metzler H, et al. Dashboard of sentiment in austrian social media during COVID-19. *Frontiers in Big Data* 2020; 32.
- [366] Guntuku SC, Buffone A, Jaidka K, et al. Understanding and measuring psychological stress using social media. In: *Proceedings of the international AAAI conference on web and social media*. 2019, pp. 214–225.
- [367] Xue J, Chen J, Hu R, et al. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of Medical Internet Research* 2020; 22: e20550.
- [368] Lee J, Sohn D, Choi YS. A tool for spatio-temporal analysis of social anxiety with twitter data. In: *Proceedings of the 34th association for computing machinery/SIGAPP symposium on applied computing*. 2019, pp. 2120–2123.
- [369] Choi D, Sumner SA, Holland KM, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA Network Open* 2020; 3: e2030932–e2030932.
- [370] Sinyor M, Williams M, Zaheer R, et al. The association between twitter content and suicide. *Australian & New Zealand Journal of Psychiatry* 2021; 55: 268–276.
- [371] Cohrdes C, Yenikent S, Wu J, et al. Indications of depressive symptoms during the COVID-19 pandemic in germany: Comparison of national survey and twitter data. *JMIR Mental Health* 2021; 8: e27140.
- [372] Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. *International Journal of Environmental Research and Public Health* 2020; 17: 4988.
- [373] Sharma S, Sharma S. Analyzing the depression and suicidal tendencies of people affected by COVID-19's lockdown using sentiment analysis on social networking websites. *Journal of Statistics and Management Systems* 2021; 24: 115–133.

-
- [374] Zhang X, Wang Y, Lyu H, et al. The influence of COVID-19 on the well-being of people: Big data methods for capturing the well-being of working adults and protective factors nationwide. *Frontiers in Psychology* 2021; 12: 2327.
- [375] Tsakalidis A, Liakata M, Damoulas T, et al. Combining heterogeneous user generated data to sense well-being. In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3007–3018.
- [376] Al Shehhi A, Thomas J, Welsch R, et al. Arabia felix 2.0: A cross-linguistic twitter analysis of happiness patterns in the united arab emirates. *Journal of Big Data* 2019; 6: 1–20.
- [377] Luhmann M. Using big data to study subjective well-being. *Current Opinion in Behavioral Sciences* 2017; 18: 28–33.
- [378] Viviani M, Crocamo C, Mazzola M, et al. Assessing vulnerability to psychological distress during the COVID-19 pandemic through the analysis of microblogging content. *Future Generation Computer Systems* 2021; 125: 446–459.
- [379] Priyadarshini I, Mohanty P, Kumar R, et al. A study on the sentiments and psychology of twitter users during COVID-19 lockdown period. *Multimedia Tools and Applications* 2021; 1–23.
- [380] Mohammad SM, Turney PD. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 2013; 29: 436–465.
- [381] Gibbons J, Malouf R, Spitzberg B, et al. Twitter-based measures of neighborhood sentiment as predictors of residential population health. *PloS one* 2019; 14: e0219550.
- [382] Kelley SW, Gillan CM. Using language in social media posts to study the network dynamics of depression longitudinally. *Nature Communications* 2022; 13: 1–11.
- [383] Wang J, Fan Y, Palacios J, et al. Global evidence of expressed sentiment alterations during the COVID-19 pandemic. *Nature Human Behaviour* 2022; 6: 349–358.

- [384] Venugopalan M, Gupta D. An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-Based Systems* 2022; 246: 108668.
- [385] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. Epub ahead of print 2018. DOI: [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [386] Hitchcock PF, Fried EI, Frank MJ. Computational psychiatry needs time and context. *Annual Review of Psychology* 2022; 73: 243–270.
- [387] Deisenhammer EA, Ing C-M, Strauss R, et al. The duration of the suicidal process: How much time is left for intervention between consideration and accomplishment of a suicide attempt? *The Journal of Clinical Psychiatry* 2008; 69: 5230.
- [388] Simon TR, Swann AC, Powell KE, et al. Characteristics of impulsive suicide attempts and attempters. *Suicide and Life-Threatening Behavior* 2001; 32: 49–59.
- [389] Hudson NW, Lucas RE, Donnellan MB. Day-to-day affect is surprisingly stable: A 2-year longitudinal study of well-being. *Social psychological and personality science* 2017; 8: 45–54.
- [390] Mogg K, Mathews A, Weinman J. Memory bias in clinical anxiety. *Journal of Abnormal Psychology* 1987; 96: 94.
- [391] Duyser FA, van Eijndhoven PFP, Bergman MA, et al. Negative memory bias as a transdiagnostic cognitive marker for depression symptom severity. *Journal of Affective Disorders* 2020; 274: 1165–1172.
- [392] Ebbinghaus H. Memory: A contribution to experimental psychology. *Annals of Neurosciences* 1885; 20: 155.
- [393] Murre JM, Dros J. Replication and analysis of ebbinghaus' forgetting curve. *PloS one* 2015; 10: e0120644.

-
- [394] Tlachac M, Rundensteiner E. Screening for depression with retrospectively harvested private versus public text. *IEEE Journal of Biomedical and Health Informatics* 2020; 24: 3326–3332.
- [395] Jesús Titla-Tlatelpa J de, Ortega-Mendoza RM, Montes-y-Gómez M, et al. A profile-based sentiment-aware approach for depression detection in social media. *EPJ Data Science* 2021; 10: 54.
- [396] Sawhney R, Joshi H, Flek L, et al. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In: *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*. 2021, pp. 2415–2428.
- [397] Sinha PP, Mishra R, Sawhney R, et al. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In: *Proceedings of the 28th association for computing machinery international conference on information and knowledge management*. 2019, pp. 941–950.
- [398] Mayor E, Bietti LM. Twitter, time and emotions. *Royal Society Open Science* 2021; 8: 201900.
- [399] Ng Fat L, Mindell J, Boniface S, et al. Evaluating and establishing national norms for the short warwick-edinburgh mental well-being scale (SWEMWBS) using the health survey for england. *Quality of Life Research* 2016; 26: 1129–1144.
- [400] Bradley MM, Lang PJ. *Affective norms for english words (ANEW): Instruction manual and affective ratings*. The Center for Research in Psychophysiology, 1999.
- [401] Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology* 1988; 54: 1063.
- [402] Shugars S, Gitomer A, McCabe S, et al. Pandemics, protests, and publics: Demographic activity and engagement on twitter in 2020. *Journal of Quantitative Description: Digital Media* 2021; 1: Internet.

-
- [403] Patnaude L, Lomakina CV, Patel A, et al. Public emotional response on the black lives matter movement in the summer of 2020 as analyzed through twitter. *International Journal of Marketing Studies* 2021; 13: 1–69.
- [404] Kuhn M. Building predictive models in r using the caret package. *Journal of Statistical Software* 2008; 28: 1–26.
- [405] Zelenka N, Di Cara NH. Data Hazards, <https://github.com/very-good-science/data-hazards> (2022).
- [406] Chen X, Sykora MD, Jackson TW, et al. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In: *Companion proceedings of the the web conference 2018*. 2018, pp. 1653–1660.
- [407] Seabrook EM, Kern ML, Fulcher BD, et al. Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates. *Journal of Medical Internet Research* 2018; 20: e9267.
- [408] Helmich MA, Olthof M, Oldehinkel AJ, et al. Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology* 2021; 41: 51–58.
- [409] Sasso MP, Giovanetti AK, Schied AL, et al. # sad: Twitter content predicts changes in cognitive vulnerability and depressive symptoms. *Cognitive Therapy and Research* 2019; 43: 657–665.
- [410] Trivedi MH. The link between depression and physical symptoms. *Primary care companion to the Journal of Clinical Psychiatry* 2004; 6: 12.
- [411] Holtzman NS, others. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality* 2017; 68: 63–68.
- [412] Bekhuis E, Schoevers R, Van Borkulo C, et al. The network structure of major depressive disorder, generalized anxiety disorder and somatic symptomatology. *Psychological medicine* 2016; 46: 2989–2998.

-
- [413] Barrera TL, Norton PJ. Quality of life impairment in generalized anxiety disorder, social phobia, and panic disorder. *Journal of Anxiety Disorders* 2009; 23: 1086–1090.
- [414] Shen JH, Rudzicz F. Detecting anxiety through reddit. In: *Proceedings of the fourth workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*. 2017, pp. 58–65.
- [415] Linton M-J, Dieppe P, Medina-Lara A. Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open* 2016; 6: e010641.
- [416] Kaiser T, Herzog P, Voderholzer U, et al. Unraveling the comorbidity of depression and anxiety in a large inpatient sample: Network analysis to examine bridge symptoms. *Depression and Anxiety* 2021; 38: 307–317.
- [417] Kelley SW, Mhaonaigh CN, Burke L, et al. Machine learning of language use on twitter reveals weak and non-specific predictions. *npj Digital Medicine* 2022; 5: 1–13.
- [418] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)* 2005; 67: 301–320.
- [419] Pellert M, Schweighofer S, Garcia D. The individual dynamics of affective expression on social media. *EPJ Data Science* 2020; 9: 1.
- [420] Prowse R, Sherratt F, Abizaid A, et al. Coping with the COVID-19 pandemic: Examining gender differences in stress and mental health among university students. *Frontiers in Psychiatry* 2021; 12: 439.
- [421] Rodriguez LM, Litt DM, Stewart SH. Drinking to cope with the pandemic: The unique associations of COVID-19-related perceived threat and psychological distress to drinking behaviors in american men and women. *Addictive Behaviors* 2020; 110: 106532.
- [422] Snoke J, Bowen CM. How statisticians should grapple with privacy in a changing data landscape. *Chance* 2020; 33: 6–13.

- [423] Pan X, Zhang M, Ji S, et al. Privacy risks of general-purpose language models. In: *2020 IEEE symposium on security and privacy (SP)*. IEEE, 2020, pp. 1314–1331.
- [424] Athanasopoulou C, Sakellari E. 'Schizophrenia' on twitter: Content analysis of greek language tweets. In: *International conference on informatics, management and technology in healthcare*. 2016, pp. 271–274.
- [425] Spates K, Ye X, Johnson A. 'I just might kill myself': Suicide expressions on twitter. *Death Studies* 2018; 44: 189–194.
- [426] Conway CC, Forbes MK, Forbush KT, et al. A hierarchical taxonomy of psychopathology can transform mental health research. *Perspectives on Psychological Science* 2019; 14: 419–436.
- [427] Isvoranu A-M, Guloksuz S, Epskamp S, et al. Toward incorporating genetic risk scores into symptom networks of psychosis. *Psychological medicine* 2020; 50: 636–643.
- [428] Slade M. Mental illness and well-being: The central importance of positive psychology and recovery approaches. *BMC Health Services Research* 2010; 10: 1–14.
- [430] Fiesler C, Garrett N. Ethical tech starts with addressing ethical debt, <https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/> (2021).
- [430] Fiesler C, Garrett N. Ethical tech starts with addressing ethical debt, <https://www.wired.com/story/opinion-ethical-tech-starts-with-addressing-ethical-debt/> (2021).
- [431] Akbarialiabad H, Bastani B, Taghrir MH, et al. Threats to global mental health from unregulated digital phenotyping and neuromarketing: Recommendations for COVID-19 era and beyond. *Frontiers in Psychiatry*; 12.
- [432] Barocas S, Boyd D. Engaging the ethics of data science in practice. *Communications of the Association for Computing Machinery* 2017; 60: 23–25.

-
- [433] Haraway D. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*; 14. Epub ahead of print 2020. DOI: [10.2307/3178066](https://doi.org/10.2307/3178066).
- [434] Arah OA. On the relationship between individual and population health. *Medicine, Health Care and Philosophy* 2009; 12: 235–244.
- [435] Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences* 2018; 115: E6106–E6115.
- [436] Raineri P, Molinari F. Innovation in data visualisation for public policy making. In: *The data shake*. Springer, Cham, 2021, pp. 47–59.
- [437] Sanders M, Lawrence J, Gibbons D, et al. *Using data science in policy*. The Behavioural Insights Team, 2017.
- [438] Pos K, Brown L. Becoming an anti-oppressive researcher. In: *Research as resistance: Critical, indigenous and anti-oppressive approaches*. Toronto: Canadian Scholars' Press, 2005.
- [439] Park S, Humphry J. Exclusion by design: Intersections of social, digital and data exclusion. *Information, Communication & Society* 2019; 22: 934–953.
- [440] Eubanks V. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.
- [441] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) *Proceedings of the 1st ACM conference on fairness, accountability and transparency*. New York, NY, USA: PMLR, 2018, pp. 77–91.
- [442] Staniszewska S, Mockford C, Chadburn G, et al. Experiences of in-patient mental health services: Systematic review. *The British Journal of Psychiatry* 2019; 214: 329–338.

-
- [443] Szmukler G. Compulsion and ‘coercion’ in mental health care. *World Psychiatry* 2015; 14: 259.
- [444] Sweeney A, Taggart D. (Mis) understanding trauma-informed approaches in mental health. *Journal of Mental Health* 2018; 27: 383–387.
- [445] Barnett P, Mackay E, Matthews H, et al. Ethnic variations in compulsory detention under the mental health act: A systematic review and meta-analysis of international data. *The Lancet Psychiatry* 2019; 6: 305–317.
- [446] Nakash O, Saguy T. Social identities of clients and therapists during the mental health intake predict diagnostic accuracy. *Social Psychological and Personality Science* 2015; 6: 710–717.
- [447] Beer D. The social power of algorithms. *Information, Communication and Society*; 20.
- [448] Appignanesi L. *Mad, bad and sad: A history of women and the mind doctors from 1800 to the present*. Hachette UK, 2011.
- [449] Werner A, Malterud K. It is hard work behaving as a credible patient: Encounters between women with chronic pain and their doctors. *Social Science & Medicine* 2003; 57: 1409–1419.
- [450] Boote J, Wong R, Booth A. ‘Talking the talk or walking the walk?’ a bibliometric review of the literature on public involvement in health research published between 1995 and 2009. *Health Expectations* 2015; 18: 44–57.
- [451] De Choudhury M, Counts S, Horvitz EJ, et al. Characterizing and predicting postpartum depression from shared facebook data. In: *Proceedings of the 17th association for computing machinery conference on computer supported cooperative work & social computing*. 2014, pp. 626–638.

Abbreviations

ADHD - Attention Deficit Hyperactivity Disorder

AI - Artificial Intelligence

ALSPAC - Avon Longitudinal Study of Parents and Children

API - Application Programming Interface

BPD - Borderline Personality Disorder

BPN - Basic Psychological Needs

DOI - Digital Object Identifier

EMA - Ecological Momentary Assessment

GAD - Generalised Anxiety Disorder

JITAI - Just In Time Adaptive Interventions

LIWC - Linguistic Inquiry Word Count

MFQ - Mood and Feelings Questionnaire

MIL - Meaning In Life

ML - Machine Learning

NHS - National Health Service

OCD - Obsessive Compulsive Disorder

PTSD - Post-traumatic Stress Disorder

RMSE - Root Mean Square Error

SAD - Seasonal Affective Disorder

UK - United Kingdom

US - United States

WEMWBS - Warwick Edinburgh Mental Well-being Scale