# London School of Economics and Political Science: LSE student makes Forbes 30 under 30 list 2022 | India Education | Latest Education News | Global Educational News | Recent Educational News

**COVID-19 Lockdowns and Public Opinion: Did online public opinion impact the stringency of government-imposed lockdowns?**

An Active Learning Approach.

Capstone Project

MSc in Applied Social Data Science

Department of Methodology

London School of Economics and Political Science

Words: 9,946

18 August 2022

# Abstract

This project uses Twitter posts to estimate public opinion on lockdown measures in five Western democracies between April 2020 and April 2021, to determine whether online public opinion on COVID-19 lockdown measures influences the stringency of governments’ lockdown policies. 1.6m tweets related to COVID-19 are collected, and a support vector machine is used to classify tweets as pro or anti-lockdown. Active learning is employed to improve the model. In most specifications, no significant relationship between change in public opinion and change in the stringency of lockdowns is found, controlling for the risk of the virus. A significant relationship between the ratio of pro to anti-lockdown sentiment and stringency is found in the UK, significant at the 1% level, however, the results are not robust.

# TABLE OF CONTENTS

1. **Introduction 1**
2. **Literature Review 3**
   1. Traditional Literature 4
   2. A Machine Learning Approach 6
3. **Research Questions 8**
4. **Data 9**
   1. Explanatory Variable 9
   2. Outcome Variable 10
   3. Case Selection 11
   4. Sampling Strategy 13
5. **Methodology 16**
   1. Classifying Tweets
      1. Speech Labelling 16
      2. Text Pre-Processing 19
      3. Selecting a Model 20
      4. Active Learning 24
   2. Regression Analysis
      1. Operationalization of variables 27
      2. Regression 27
6. **Results 28**
   1. Summary Statistics 28
   2. Main Regression Results 35
   3. Addressing Causality 45
7. **Discussion & Limitations 46**
8. **Conclusion 48**

**Bibliography 49**

**Appendices 54**

# 1 INTRODUCTION

“If public opinion changes and then public policy responds, this is dynamic representation”

- Stimson, MacKuen and Erikson, 1995, p. 543.

The principle that government policy is responsive to the will of its constituents is a central tenet of democratic governance (Wright, Erikson and McIver, 1987). While this principle is often assumed to be true, the endogeneity of public opinion and policy, and the plethora of other factors which influence the two, results in inconsistent empirical evidence on the relationship (Page, 1994).

The COVID-19 pandemic required, first and foremost, a response to a health crisis, but also to a political one. With lockdown measures infringing on people’s freedom of movement, in many instances, support for COVID-19 lockdown policies became a partisan issue (Grossman et al., 2020). National responses varied greatly, not only based on scientific advice but influenced by various political factors, such as cultures of collectivism and trust in government (Xing, Li and Wang, 2021; Chen et al., 2021). In many instances, governments such as those in the UK and Ireland even deviated from the recommendations of their COVID-19 health advisory boards and refused to impose tighter restrictions (House of Commons, 2021, p. 8; Moloney, 2020). This begs the question: what caused governments to deviate from scientific evidence on COVID-19 when they previously followed it, and what role does public opinion play, if any?

This project uses Twitter posts to estimate public opinion on lockdown measures in five Western democracies between April 2020 and April 2021. It aims to capture changes in public opinion over time and investigate the role that public opinion played in determining the stringency of government-imposed lockdowns.

The Oxford COVID-19 Government Response Tracker (OxCGRT) comprises 21 live indicators of government response to the pandemic worldwide (Hale et al., 2021). This includes a “risk of openness” index which “calculates a measure of risk that a country faces from adopting an ‘open’ policy stance” (Hale et al., 2020a, p.1), and a stringency index is calculated based on the lockdown-style policies of each country. Comparing these scores on a national level allows this project to capture the “relative stringency” of lockdown measures. Twitter presents an opportunity to measure public opinion continuously over a short time period, rather than more traditional survey measures, which are infrequent. The TBCOV dataset (Imran, Qazi & Ofli, 2022) provides 1.2 billion COVID-19-related geocoded tweets worldwide. The project samples 1.6m tweets from five English-speaking Western democracies: Australia, Canada, Ireland, New Zealand and the United Kingdom, to formulate estimates of public opinion in these countries over time.

To classify tweets, a supervised machine learning approach is employed, using cross-validation to select a model, and utilising active learning to improve its performance. The project initially creates a training set of 4,000 tweets, which classifies tweets sampled from the TBCOV dataset as neutral, pro-lockdown or anti-lockdown. Using k-fold cross validation and hyperparameter tuning, a Support Vector Machine with a linear kernel and cost of 12 is selected as the best performing classifier. Active learning is then utilised to improve the performance of the model, which increases the F1s of pro and anti-lockdown classes from an estimated 0.26 and 0.25 to 0.5 and 0.34 respectively, labelling almost 4,000 additional tweets during the process. Following active learning, class predictions for tweets are aggregated weekly for each country to derive an estimate of public sentiment toward lockdowns.

This paper implements a cross-sectional time series analysis, the gold standard in the literature (Page, 1994). The proportion of pro and anti-lockdown tweets provide a proxy for public opinion which is regressed on changes to relative stringency, controlling for change in average Western stringency, following the dynamic research design of Page and Shapiro (1983).

Regressing change in pro-lockdown and anti-lockdown sentiment on change in relative stringency, at multiple time lags, yields no significant results. In cases where public opinion saw consistent change, policy moved noncongruently with it, more often than it did not. Based on these model specifications and data used, there is no evidence of a relationship between public opinion and COVID-19 lockdown policy. A final specification of the regression transforms the explanatory variable to a ratio of weekly anti-lockdown tweets to pro-lockdown tweets. While the multi-level model finds no significant relationship with relative stringency, 67% of changes to relative stringency in the UK were congruent with changes in public opinion one week prior. The relationship between change in relative lockdown stringency and change in the ratio of pro-lockdown to anti-lockdown sentiment one week prior is significant at the 1% level. However, as these results are not replicable across other countries, they are not robust. The lack of significant results may be explained by multiple issues with the design of the study. The Support Vector Machine has relatively poor performance, resulting in both inaccurate measures of public opinion, and high variance in estimates from week to week. Further, public opinion and policy are endogenous in this project: the severity of the virus affects both relative stringency and how the public shapes its opinions on policy. Despite a lack of significant findings, this project contributes to the literature on public opinion during COVID-19, providing a coding scheme for the classification of pro and anti-lockdown text and analysing how both sentiments changed over time in five Western democracies.

# 2 LITERATURE REVIEW

While there is consensus in the public policy literature that public opinion influences policy, particularly on salient issues, there is very little consensus on the size of this impact, relative to elites, interest groups, political parties, and mass media, and how generalizable this is across space and time (Burstein, 2003). The endogenous and interactive nature of the relationship between public opinion and policy also means the literature’s results are never without a degree of uncertainty (Page, 1994).

Public opinion is expected to influence policy in broadly two ways. In the longer term, it is expected that elections will change the government's composition to reflect constituents’ preferences, shifting policy as a result. In the shorter term, policymakers will calculate electoral (or other) implications of the current public opinion, and act accordingly, known as “rational anticipation” (Stimson, MacKuen and Erikson, 1995). The public can also be seen as a “thermostat” in the literature. Wlezien (1995) posits that the relationship is more complex and interactive, where the public adjusts its preferences based on policy outputs themselves, sending signals to government for “more” or “less” policy, to which government policy responds.

## 3.1 Traditional Literature

Miller and Stokes’s (1963) study of constituency influence over the US House of Representatives is the first study to test the correlation between survey-based constituency opinion and congressional roll-call behaviour, finding small but significant effects.

A causal modelling approach, presented by Erikson (1978), which posits that Miller and Stokes’s (1963) sample is unrepresentative, makes significant progress in the field by inferring district opinion by simulating constituent preferences on a state level and finds a significantly larger effect than that of Miller and Stokes (1963). Wright, Erikson and McIver (1987) use CBS News and New York Times national opinion surveys to infer state-level ideological identification through causal modelling, concluding that citizen preferences are more important than state social and economic characteristics in determining state policy. However, they note the effect is pronounced on salient issues, and ambiguous for smaller issues, implying that government acts when the public cares enough about an issue to make its position known.

To make claims about causality, rather than correlation, the gold standard in the literature is the cross-sectional time series analysis, rather than a static design (Page, 1994). Page and Shapiro’s (1983) cross-sectional time series study assesses the changes in preferences and policy in the US from 1935 to 1979. By focusing on change in public opinion and policy, the study measures government’s ability to respond to the public will, rather than measuring ideological alignment between constituent and representative. They find considerable congruence between changes in public opinion and policy, particularly on salient issues. The study, however, does not control for world events, interest groups elites, and mass media, which could all produce a spurious relationship (Page, 1994).

Caughey and Warshaw’s (2018) dynamic responsiveness study at a state level is less optimistic. The authors find that although the liberalism of the public predicts future change in policy liberalism, dynamic responsiveness is gradual and a consequence of incremental responsiveness over multiple years. They posit that responsiveness occurs largely through the adaption of incumbents, contrary to the idea that elections are the drivers of change, concluding that their findings are somewhat ambiguous.

The majority of studies take place in the United States, as the US federal state system allows for cross-sectional analysis on a state level, leading to a lack of generalisability in the literature. Hobolt and Klemmemsen (2005) address this, using longitudinal data on responses to the “Most Important Political Issue” in Britain and Denmark, finding public opinion to have a significant impact on policy in this setting. Its most notable contribution to Page and Shapiro’s research design is also regressing policy in time *t-1* on public opinion in time *t*, to examine the possibility of reverse causality.

The studies discussed largely infer opinion from ideological placement of the state public, and not from issue responses directly, limiting the claims of dynamic representation. Hartley and Russell (1992) emulate the classic study of Page and Shapiro (1983), to assess the impact of public opinion on US military spending, controlling for other drivers of policy change. The study directly observes public opinion on military spending, using survey responses related to the topic. While the authors find a significant effect of public opinion on policy, the impact they find is smaller than that of Page and Shapiro (1983). More recently, Lax and Phillips’ (2009) study of gay rights in the US finds a high degree of responsiveness, controlling for interest groups and voters’ ideology, but also finds a significant amount of noncongruence. However, opinion on gay rights is not directly observed, but inferred, by simulating state opinion through national surveys.

The traditional literature on public opinion and policy, discussed above, has made significant steps in disentangling the relationship between public opinion and policy, with a broad consensus on the direction of the relationship, but with divergence on the size of its effect (Burstein, 2003). As Page noted in 1994, the results presented at that time were consistent with the hypothesis that government is responsive, but they couldn’t completely refute claims that government policy is only in line with public opinion some of the time. Jones (1994) argues that responsiveness is only likely on a few issues at any given time. Monroe (1998) finds that policy outcomes are consistent with the majority public opinion in only 55 per cent of cases, arguing that a ‘bias against change’ has become more prominent in US politics, making governments less responsive. The evidence presents a compelling argument that the relationship does exist in some cases, but overall, it lacks generalisability across issues and countries, and the driving force behind this congruence is largely unknown. Therefore, studies in new contexts on the topic will always have value in improving generalisability. COVID-19 presents a particularly interesting scenario, where policy should only be influenced by scientific advice, yet governments consistently deviated from this. The question is: why?

## 3.2 A Machine Learning Approach

Mass use of social media platforms such as Twitter has revolutionised how social scientists can collect data on the public’s perception of policy issues, which of these issues are salient, and how various demographic groups interact on these issues (Murphy et al., 2014). One major limitation of opinion-policy time series analysis is the unevenly spaced data points created by irregularly timed surveys and polls (Page, 1994). Twitter presents an opportunity to measure public sentiment continuously, improving the capabilities of studies, as the influence of public opinion on policy may not follow uniform time lags (Page, 1994).

Much of the internet-based literature has focused on the agenda-setting power of the public on Twitter. Barbera et al. (2019) use a temporal analysis of Twitter messages by the public and legislators during the 113th US Congress, and with unsupervised machine learning, classify tweets into topics. The study finds that the public are more likely to lead, and legislators are more likely to follow in the discussion of public issues, even when controlling for media agenda setting. Gilardi et al. (2021) use dictionary classifiers and a vector autoregression approach on Twitter comments to find that discussions of COVID-19 measures in Switzerland were led by the attentive public and politicians, while parties and newspapers followed.

Social media comments can be used to extract meaningful information to support policymakers in a fast and cheap way, particularly in presenting policy alternatives according to citizens’ opinions for politicians and researchers (Ceron and Negri, 2016). Interviews with pollsters, researchers and journalists by Anstead and O’Loughlin (2015) demonstrate how monitoring social media was valuable for revealing how public opinion formed and shifted during the 2010 UK general election. Multiple studies have also used Twitter to track how public opinions and emotions developed throughout the COVID-19 pandemic (Kwan and Lim, 2020; Lwin et al., 2020; Wang et al., 2020). While public opinion and policy Twitter studies are primarily focused on the agenda-setting power of the public in policy debate, it is feasible that the best place to observe Wlezien’s (1995) interactive public “thermostat”, is online where public opinion on salient issues can be observed. For rational anticipation to occur, all that matters is that politicians have a perception of the most expedient position (Stimson, MacKuen and Erikson, 1995) – Twitter is one of the most effective tools for exposing politicians to their constituents and followers’ opinions.

This study uses the dynamic approach first utilised by Page and Shapiro (1983) to address the traditional question of the field: does public opinion influence policy, and to what extent? However, online data collection presents many advantages, where instead of inferring the public’s opinion on specific issues based on their ideological placement, public opinion on a topic can be directly observed through debates and comments online. COVID-19 provides a unique opportunity to conduct a cross-national rather than a US-based study, with a measure of COVID-19 policy response which is consistent across space and time.

# 3 RESEARCH QUESTIONS

This study analyses tweets related to COVID-19 in five Western democracies between 1st April 2020 and 1st April 2021, to capture changes in public opinion over time, utilising supervised machine learning approaches to classify tweets as pro-lockdown, anti-lockdown, or neutral, based on a coding scheme outlined in Section 5.1.1. This classification of tweets allows the project to find the role that public opinion played in determining the stringency of national lockdown-style policies. Most generally, the project seeks to answer the question:

*Did online public opinion on COVID-19 lockdown measures influence the stringency of governments’ lockdown policies?*

More specifically, this project states 3 hypotheses on the relationship between pro and anti-lockdown sentiment and relative stringency:

H1: When the proportion of anti-lockdown tweets relative to all COVID-19 tweets changes in a country, lockdown measures subsequently change congruently.

Did an increase in the proportion of anti-lockdown tweets result in lockdown measures becoming **less stringent**, when controlling for the risk posed by the virus, and regional trends in COVID-19 response?

H2: When the proportion of pro-lockdown tweets relative to all COVID-19 tweets changes in a country, lockdown measures subsequently changed congruently.

Did an increase in the proportion of pro-lockdown tweets result in lockdown measures becoming **more stringent**, when controlling for the risk posed by the virus, and regional trends in COVID-19 response?

H3: When the ratio of anti-lockdown tweets to pro-lockdown tweets changes in a country, lockdown measures subsequently change congruently.

Did an increase in the ratio of anti-lockdown tweets topro-lockdown tweets result in lockdown measures becoming **less stringent**, when controlling for the risk posed by the virus, and regional trends in COVID-19 response?

# 4 DATA

## 4.1 Explanatory Variable

This project required extensive Twitter data related to COVID-19 and lockdown policy, to gain an accurate representation of public opinion on lockdown measures. Further, to create a measure of national public opinion, Twitter data must be geocoded. While Twitter allows researchers to directly retrieve tweets, using REST API based on predefined parameters, such as keywords, only an estimated 0.85% of tweets are geotagged, with statistically significant differences in the demographics of those who geotag their tweets compared with those who don’t (Sloan and Morgan, 2015). Consequently, this study required a third-party dataset which derived geolocation based on other tweet text and user indicators. The TBCOV dataset (Imran, Qazi & Ofli, 2022) fulfilled this requirement, providing 1.2 billion geocoded tweets worldwide. The authors use a geolocalisation approach which deduces tweet location based on user location, user description, tweet content and time zone. As Twitter does not allow third parties to publish the content of their tweets, external datasets must instead provide Twitter IDs, which are unique 64-bit unsigned integers (Twitter, 2022). These IDs must be “hydrated” by converting back into JSON form by querying Twitter. This project uses the DocNow hydration app for this process (Documenting the Now, 2020).

## 4.2 Outcome Variable

The outcome variable of interest is the stringency of a country’s lockdown policy. However, as COVID policy is largely contingent on the current risk posed by the virus, this study required a measure of lockdown policy which could control for risk. OxCGRT “collects information on policy measures governments have taken to tackle COVID-19” (Hale et al., 2021), comprised of 21 live indicators of government response. OxCGRT presents an overall response index, a containment and health index, a stringency index, economic support, and a risk of openness index.

The stringency index is calculated based on the lockdown-style policies of each country. It is evaluated based on eight indicators: school closures, workplace closures, cancel public events, restrictions on gathering size, public transport closures, stay-at-home requirements, restrictions on internal movement and restrictions on international travel (Hale et al. 2020). The stringency index provides an accurate measure of policies that restrict people’s behaviour, independent of other COVID-19 response policies, such as testing and contact tracing, healthcare investment, vaccination, and income support. This project hypothesises that the stringency index will be the most responsive element of government policy to changes in public opinion, as lockdown-style policies infringe the most upon individuals’ civil liberties (Flood et al., 2020) and ultimately require compliance from the public to be effective (Talic et al., 2021).

The risk of openness index “calculates a measure of risk that a country faces from adopting an ‘open’ policy stance” (Hale et al., 2020a, p.1). It provides a measure of COVID-19 severity which is consistent across countries and time. Other measures of risk such as daily cases, deaths, or hospital admissions are not directly comparable across time due to changing risk assessments based on vaccination and treatment, and across countries due to differing healthcare capacities and treatment capabilities.

The stringency index, scored out of 100, can therefore be offset by the risk of openness index, also out of 100, to create an outcome variable which captures a country’s lockdown stringency, *relative to* the current risk posed by the virus.

## 4.3 Case Selection

Excluding Hobolt and Klemmemsen (2005), the traditional literature on public opinion focuses on state-level changes to public opinion and policy within the US. Hobolt and Klemmemsen’s case study of public opinion in Denmark and the UK is comparative as they compare one highly proportional and one highly majoritarian democracy. The primary aim of this study is not comparison, but to understand a general phenomenon across countries. This project instead employs a collective case study – picking several instrumental (or typical) cases to understand a wider phenomenon, in order to generalise (Stake, 1995). Therefore, I wished to select multiple “Western” Democracies which may be representative of other, non-English speaking, democracies.[[1]](#footnote-1) Australia, Canada, Ireland, New Zealand and the United Kingdom were selected – all sovereign nations, with developed economies and full democratic rights (see Table 1).

This project assesses COVID-19 policy and public opinion on a “sovereign-state” level.[[2]](#footnote-2) While much of the lockdown-style policies vary within these sovereign states,[[3]](#footnote-3) such as countries within the United Kingdom and states within Canada and Australia, consistent measures of COVID-19 risk were only available at the national level. This is an apparent weakness of this research design.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1.** Freedom House Scores for Sampled Countries. | | |  |
| Country | Total Score & Status | Political Rights | Civil Liberties |
| Australia | 95 – Free | 39 | 56 |
| Canada | 98 – Free | 40 | 58 |
| Ireland | 97 – Free | 39 | 58 |
| New Zealand | 99 – Free | 40 | 59 |
| United Kingdom | 93 – Free | 39 | 51 |
| *Source:* *Repucci and Slipowitz (2022) – Freedom House* | | | |

The US could be considered to fit these criteria, but it was ultimately omitted from the study. I believe the US would be an outlier in the study. COVID-19 quickly became a polarising topic within US politics, with evidence of these polarizations being along Democratic-Republican lines within the electorate (Kerr, Panagopulos and van der Linden 2021; Jiang et al.,2020). Given that state COVID-19 policy also differed significantly between Democratic and Republican-controlled states (VanDusky-Allen and Shvetsova, 2021), responsiveness to public opinion on COVID-19 most likely happened at a state level, not a federal one. It was also an outlier concerning freedom scores, with a score of 83 on the Freedom House scale, compared with 93+ for other countries in the sample (Repucci and Slipowitz, 2022). It is a unique case which would fail to generalise to other western democracies.

The five countries selected also create an interesting set, with significantly different COVID-19 strategies. While Australia and New Zealand pursued a zero COVID policy, Canada, UK and Ireland did not.

## 4.4 Sampling Strategy

As the TBCOV dataset was far larger than was computationally feasible to use, a sampling strategy was developed. I wished to maintain the original distribution of the data collected by TBCOV, as it reflects the salience of COVID-19 on Twitter over time, and thus randomly sampled by country and time. The project aimed to have a final sample of between 1-2 million tweets, as it was large enough to have a significant sample for each country, each week, while still being computationally feasible to use active learning on an unlabelled set of this size. As tweet ids must be hydrated, there was a level of uncertainty about what proportion of tweets would successfully hydrate, and therefore the project initially sampled significantly more tweets than the final goal for hydration. Table 2 below summarises the number of tweets sampled and hydrated by country. As 2.6m tweets were successfully hydrated, a sample of 60% of these was used in the final dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 2.** Count of Tweets sampled and hydrated by country. | | | | |
| Country | Sampled Tweets | Successfully Hydrated | Percentage Hydrated | Final Data (60% hydrated) |
| Australia | 1,044,162 | 326,775 | 31.2% | 196,065 |
| Canada | 2,572,321 | 691,897 | 26.9% | 415,138 |
| Ireland | 413,005 | 133,865 | 32.4% | 80,319 |
| New Zealand | 161,737 | 46,423 | 28.7% | 27,854 |
| United Kingdom | 4,565,825 | 1,461,135 | 32.0% | 876,681 |
|  | **8,756,550** | **2,660,095** | **30.4%** | **1,596,057** |

While there is some variation in the proportion of tweets hydrated in each country, they are roughly similar. It is concerning however that less than a third of all tweets were successfully hydrated. While there is no consistent estimate of what percentage of tweets are typically deleted on the Twitter platform, they range from 2.4% (Almuhimedi et al., 2013), to 11% (Bhattacharya and Ganguly, 2016) to 26% (Zhou, Wang and Chen, 2016), and while these studies analyse tweets over a shorter time period than this study (which spans from approximately 26 months to 14 months prior), a 70% deletion rate seems extremely high. This project has very little explanation for this. As it was relatively consistent across country files, it appears to either be an error with the TBCOV dataset, or an unusual phenomenon among COVID-19-related tweets. Next, I conducted a similar test of hydration across time, which is visualised in Figure 1 below.

Chart, bar chart, histogram

Description automatically generated

**Figure 1.** Distribution of sampled and hydrated tweets over time.[[4]](#footnote-4)

It is apparent from Figure 1 that the hydration rate is decreasing over time. In the first week of the sample, 40% of tweets were successfully hydrated, while in the final week, only 16% were. This trend of deletion is occurring in the inverse direction expected, with a greater proportion of newer, rather than older tweets being deleted. There must be a confounding factor here, which is largely unexplained in this project. As mentioned, there may be an unidentified error in the TBCOV dataset, whereby tweet ids were incorrectly input into their TSVs in later months. However, this issue may also be related to the content which was posted during these months. It is possible that the content posted during these later months was more susceptible to deletion. Deleted tweets are more likely to contain negative sentiment and swear words (Bhattacharya and Ganguly, 2016), and to be on regrettable topics, for example, alcohol and illegal drug use, sexual activity, religion and politics, and offensive comments (Almuhimedi et al., 2013; Zhou, Wang and Chen, 2016). An increase in deletion is likely an indication of more negative posting, which likely correlates with both pro and anti-lockdown sentiment. There is little which can be done to rectify this, but it likely indicates that this project will not obtain an accurate measure of public opinion over time, as high levels of deletion will obscure the measures. If this is true, it poses a major threat to the validity of the project’s results.

As an additional sampling test, I wished to ensure there were similar levels of coverage across countries in the dataset. By obtaining data on the estimated number of users per country from Kemp (2022a, 2022b, 2022c, 2022d, 2022e), Table 3 reports on coverage: the number of tweets in the sample proportional to the number of estimated users. In both the sample and hydration, coverage is roughly equal for each country.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 3.** Tweet Coverage by Country | | | |
| Country | Estimated Users | Coverage in Sample | Coverage in Hydration |
| Australia | 3,700,000 | 0.282 | 0.053 |
| Canada | 7,900,000 | 0.326 | 0.053 |
| Ireland | 1,350,000 | 0.306 | 0.060 |
| New Zealand | 552,700 | 0.293 | 0.050 |
| United Kingdom | 18,400,000 | 0.248 | 0.048 |

*Source: Kemp (2022a, 2022b, 2022c, 2022d, 2022e)* *Digital 2022 Reports.*

# 5 METHODOLOGY

A picture containing timeline

Description automatically generated

**Figure 2.** Flowchart of Methodology process.

## 5.1 Classifying Tweets

5.1.1 Speech Labelling

This project opted to employ a supervised machine learning classifier to label data. Supervised models require an amount of labelled text, from which the classifier can learn predictive features. Following the methodologies of Yang et al. (2020), Li et al (2020) and Kabir and Madria (2021), which all sought to build classifiers for COVID-19 tweets, this project undertook manual labelling of COVID-19 tweets. As COVID-19 is a novel topic, and the idea of pro or anti-lockdown sentiment is new, a dictionary-based approach would not have been appropriate.

Yang et al. (2020) provided access to their 10k emotion-labelled COVID-19 tweets. As their classification objective differed, I re-labelled 2,000 of their set to fit my coding scheme, however their labels of “denial”, “anxious”, and “sad” helped greatly with identifying more examples of pro and anti-lockdown sentiment, to increase the prevalence of pro and anti-lockdown observations within my training set. In the relabelled Yang et al. (2020) data, 81 tweets were labelled “Pro-Lockdown”, 159 were labelled “Anti-Lockdown” and 1760 were labelled neutral. I intentionally labelled more “denial” tweets, as preliminary analysis of the TBCOV data showed that the classifier was failing to pick up on Anti-lockdown tweets. Following this, I randomly sampled 2,000 tweets from the 5 countries within TBCOV – and labelled these following the coding scheme outlined below. In the TBCOV sample, 55 were given an NA label, as they were in another language, despite being identified as English in the TBCOV data, 135 were labelled “Pro-lockdown”, 55 were labelled “Anti-lockdown”, while 1755 were labelled “Neutral”. This labelling process indicates a clear class imbalance within the data, which will pose a challenge to the model, giving it significantly less information for two out of three classes.

Figure 3 details the coding scheme for the text classification. “Neutral” was assigned to Class 0, “Pro-lockdown” was assigned to Class 1, and “Anti-lockdown” was assigned to Class 2.

Diagram

Description automatically generated

**Figure 3.** Flowchart for coding scheme of pro-lockdown, anti-lockdown and neutral tweets

There is a great deal of uncertainty in text labelling, making decisions somewhat subjective, particularly as an inter-coder reliability score cannot be reported. Intercoder reliability refers to at least two researchers independently classifying observations, and reporting the consistency between the two, while intracoder reliability refers to the consistency of a researcher’s coding (Given, 2008). As tweets were labelled by a single coder, 1,000 of the TBCOV training set were labelled twice, two weeks apart, and yielded an intracoder reliability score of 0.9805 – 98.05% of all tweets were given the same label in both labelling rounds.

One major source of ambiguity during the labelling process was criticism of government. While a large proportion of tweets criticised governments’ responses to the pandemic, they did not explicitly refer to lockdown measures. As the government response also encapsulated economic support, healthcare, and test-and-trace capacity, alongside lockdown measures, these tweets are coded as neutral, as an individual’s position on lockdown policy cannot be inferred from these tweets. Another source of uncertainty during the process came from discussions surrounding masks. While mandatory mask requirements were part of government policy responses to COVID-19, mask requirements are not included in OxCGRT’s stringency index. Therefore, texts which called for mask mandates or for individuals to comply with mask-wearing were classified as neutral unless they also called for the implementation of, or compliance with, other measures included in the index.

5.1.2 Text Pre-Processing

Before classification, the tweet text must be pre-processed so it is in the appropriate format for analysis – whereby each word can be treated as a variable. As this project is conducted primarily through R, I will be using the *Quanteda* package (Benoit, Watanabe, et al., 2018), to tokenise tweets and create a document-feature-matrix (DFM) for analysis.

Tokenization is the process of splitting text into individual tokens. This project will tokenise words, rather than characters or sentences, as most commonly, words are the most meaningful components of texts (Welbers, Van Atteveldt and Benoit, 2017). Further, tokenisation involves removing tokens which are unhelpful to the model such as stopwords (“on”, “the”, “of”, etc), symbols and numbers. This project removes punctuation, URLs, numbers and stopwords. These tokens are then incorporated into a DFM, a common format for representing a collection of texts in a bag-of-words format (Welbers, Van Atteveldt and Benoit, 2017). Bag-of-words text analysis means only the frequencies of words per text are considered, and the order and position of words are ignored (Welbers, Van Atteveldt and Benoit, 2017).

When creating DFMs, bi-grams (two-word phrases) and tri-grams (three-word phrases) can be included, if they occur in a minimum number of documents. To determine whether including tri-grams would improve this classification model, cross-validation of bi-gram and tri-gram DFMs was performed, using Naïve Bayes and Ridge Regression on the training set of 4,000 tweets. Table 4 indicates that bigrams performed significantly better than trigrams under both models. Therefore, the DFM includes bigrams only.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4.** 10-Fold Cross Validation for Text Pre-processing | | | |
| DFM Type | Model | F1 Class 1 | F1 Class 2 |
| Bigrams | Naïve Bayes | 0.042 | 0.062 |
| Bigrams | Ridge Regression | 0.095 | 0.14 |
| Trigrams | Naïve Bayes | 0.062 | 0.008 |
| Trigrams | Ridge Regression | 0.035 | 0.076 |

5.1.3 Model Selection

The type of data a model is trying to classify, and the objective of classification are important determinants of the appropriate performance metric for comparing machine learning classifiers.

The data in this project is class imbalanced: in the TBCOV training data, 6.75% were labelled as pro-lockdown, and 2.75% as anti-lockdown. A measure such as accuracy, which measures the percentage of observations correctly classified, would not be appropriate for imbalanced data, as an accuracy score of over 90% could be achieved by classifying all tweets as Class 0.

In terms of objective, Classes 1 and 2 are of equal importance in the final regressions, while Class 0, the neutral category, is not included. The performance of models should be assessed on their ability to correctly classify Classes 1 and 2. Therefore, performance measures for individual classes will be used rather than an overall performance metric. Precision measures the proportion of predicted positives that are truly positive, while Recall measures the proportion of true positives that are correctly identified as such. Models will be evaluated based on the F1 scores, the harmonic mean of the Precision and Recall scores. F1 is the most commonly used performance measure for text classification, as aggregating the F1 scores for individual classes accounts for class imbalance (Zhang, Wang and Zhao, 2015).

To assess each model’s performance 2, cross-validation must be performed. Cross-validation is the process of holding out a subset of labelled data during model training and predicting on this held-out subset (James et al., 2009). I have opted to use 10-fold cross-validation, which involves randomly dividing observations into 10 groups of equal size. The first fold is treated as a validation set (held-out set), and the model is fit on the remaining 9 folds, and performance is evaluated on the validation set (James et al., 2009).

10-fold cross-validation is implemented on multiple different machine learning classifiers, recording their Class 1 & 2 F1 scores. For multiple reasons, I did not expect initially strong F1 scores. Firstly, the data is highly imbalanced, lending itself to lower model performance as most machine learning algorithms assume relatively balanced classes (Sun, Wong and Kamel, 2009). Secondly, when imbalanced data is multinomial, many methods for tackling class imbalance are no longer appropriate (Sun, Wong and Kamel, 2009). Further, multiple aspects of the training data were not optimal for strong performance. Tweet labelling was conducted by one coder, presenting the project with two major limitations. Firstly, I was constrained by time, meaning it was only feasible to label approximately 4,000 texts before active learning. As the coding scheme was unique to the project, there were no available training sets prepared by other researchers. Second, having a single coder of all documents creates some error and bias. Assuming there is some coder error, there will be inconsistency across labels, which will reduce the model’s performance on new and existing data.

Models were tested roughly in order of their complexity. Naïve Bayes, Random Forest, SVMs and Logistic Regression are popular choices for multi-class classification (Pranckevičius and Marcinkevičius, 2017). Cross-validation is also performed on LightGBM, a new Gradient Boosting Decision Tree, ideal for high dimensional data, as it achieves state-of-the-art performance in multi-class classification (Ke et al., 2017).

Naïve Bayes models are popular choices for multinomial text classification, as they are simple, efficient, yet well-performing (Frank and Bouckaert, 2006). The Naïve Bayes yielded F1 scores of 0.042 and 0.062 respectively. Regularised regression seeks to constrain, or shrink the coefficient estimates towards zero, significantly reducing their variance (James et al., 2009). λ, the shrinkage penalty, determines how many coefficients will approach zero (James et al., 2009). Using cross-validation λ was selected as 1.075. The model yielded F1s of 0.095 and 0.14. LASSO regression, another form of regularized regression was also tested, which selects a subset of p predictors. λ, the shrinkage penalty must also be selected using cross-validation to determine how many variables converge on 0. Cross-validation selected λ = 0.029, however, this model performed particularly poorly, failing to correctly label any Class 1 or Class 2 texts, with both F1s equal to 0.

Support Vector Machines (SVM) find a hyperplane that separates classes from one another and are considered to achieve state-of-the-art performance on text classification problems (Zhang and Oles, 2001). The SVM, once tuned, was by far the best performing model tested. A randomised hyperparameter search yielded multiple best-performing models, all with linear kernels and a range of C values above 6, with F1s of 0.278 and 0.261 respectively.

Random Forests are a process of repeating simple decision trees n times, but each time taking a random sample of p predictors. However, the model encountered the same problem as LASSO regression, failing to correctly label any Class 1 or Class 2 texts, and simply classifying all texts as Class 0. LightGBM improved on Random Forests when tuned using a random hyperparameter search, however, it only attained F1s of 0.124 and 0.026 respectively.

Cost-sensitive methods were attempted in SVM and Random Forest models due to class imbalance, as was recommended by Sun, Wong and Kamel (2009), but did not yield better results than other models. Finally, I attempted to run an SVM, the best performing model, with some feature selection, which could help improve computational efficiency when working with larger data. Using LASSO regression, I selected all variables which had not converged to zero, where lambda is equal to the value which minimises mean cross-validated error. This model performed significantly worse than the SVM, with F1s of 0.053 and 0.125, therefore I decided to keep all features at the cost of less efficient run time.

|  |  |  |
| --- | --- | --- |
| **Table 5.** 10-fold Cross-Validation results on training data. | | |
| Model | F1 Class 1 | F1 Class 2 |
| Naïve Bayes | 0.042 | 0.062 |
| Ridge Regression | 0.095 | 0.140 |
| LASSO Regression | 0 | 0 |
| SVM (Tuned) | 0.278 | 0.261 |
| Random Forest | 0 | 0 |
| LightGBM (Tuned) | 0.124 | 0.026 |
| SVM LASSO Feature Selection | 0.053 | 0.125 |

With the data sampled and hydrated, the text pre-processed and the best model selected through cross-validation, the project was now prepared for the active learning component.

5.1.4 Active Learning

Active learning is the process of a machine learning algorithm determining which data it learns from, by “querying” the oracle (or human coder) for new labels on the unlabelled data it is most uncertain about (Settles, 2012). This process is repeated to improve model performance. Active learning is typically used in contexts where there is a large amount of unlabelled data, as it is preferable to a “passive” approach to manually label data, as it tends to give better generalization to the model, allowing it to better predict on new data (Cohn, Atlas and Ladner, 1994).

In active learning, when dealing with a large unlabelled dataset, it is difficult to ascertain how much a model is improving over time. It is recommended to impose some stopping criterion – a point in performance where active learning ceases. This can be after a certain number of iterations, or after hitting a performance target. Zhu et al. (2010) propose several stopping criteria, but this project opts for the overall uncertainty method. The method considers the overall uncertainty of all unlabelled examples and assumes that as this uncertainty becomes smaller, the model is improving in classifying the data (Zhu et al., 2010). As active learning will be performed using Support Vector Machines, overall uncertainty will be evaluated on the average distance from the decision boundary. As this is a multi-class classification problem, there are 3 decision boundaries: 0|1, 0|2, 1|2.

I am also including an additional criterion for assessing the performance: a held-out validation set. This allows me to test how the model performs on new data and monitor its improvement over time. After every few iterations of active learning, the model is run on this held-out data to assess the model’s performance. The final performance of the model is evaluated on a held-out validation set of 1,000 tweets.

With each iteration of active learning, the model returns the texts closest to its three decision boundaries for labelling. The process started with returning 10 for each boundary, but it was increased to 15. Due to the large size of the unlabelled data, the active learning model was trained gradually, incorporating more and more data as the model’s performance improved. At each point where new data is added to the process, the validation set is resampled from data the model has not yet seen, and the old set is added to the labelled data.

At each step, the average distance to the decision boundary was reported. As the average distance started at 0.85, I aimed to see a 10% improvement once active learning was completed with all data incorporated – a distance above 0.935. Table 6 reports a summary table from active learning. While active learning was repeated over 70 times, Table 6 simply summarizes the improvements over time. The table includes the size of the unlabelled set from which the model was learning, the number of tweets labelled since the previous test and the size of the training set (labelled size). Three performance metrics are reported on: the F1 scores from 10-fold cross-validation within the training set, the average distance to the decision boundary, and the F1 scores from prediction on the validation set.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 6.** Improving performance of the model through Active Learning. | | | | | | | | |
| Unlabelled Size | Number Labelled | Labelled Size | F1 Class 1 | F1 Class 2 | Average Distance to decision boundary | Val Set Size | Val Set F1: 1 | Val Set F1: 2 |
| 100,000 | Pre | 3,992 | 0.26 | 0.25 | 0.851 | - | - | - |
| 99,700 | 300 | 4,292 | 0.254 | 0.18 | 0.963 | 100 | 0.5 | 0 |
| 299,330 | 400 | 4,692 | 0.301 | 0.208 | 0.889 | 200 | 0.375 | 0.5 |
| 598,958 | 240 | 4,932 | 0.311 | 0.233 | 0.908 | 300 | 0.4 | 0.2 |
| 598,253 | 705 | 5,637 | 0.316 | 0.253 | 0.933 | 300 | 0.5 | 0.25 |
| 597,548 | 705 | 6,342 | 0.298 | 0.289 | 0.946 | 600 | 0.452 | 0.353 |
| 597,098 | 450 | 6,792 | - | - | 0.97 | 600 | 0.438 | 0.375 |
| 996,273 | 180 | 6,972 | - | - | 0.953 | 600 | 0.467 | 0.125 |
| 1,221,328[[5]](#footnote-5) | 645 | 7,617 | 0.364 | 0.344 | 0.952 | 1,000 | 0.464 | 0.125 |
| 1,591,130 | 255 | 7,872 | 0.374 | 0.311 | - | 1,000 | 0.50 | 0.343 |

Once all of the data was included in the model, the average distance to the decision boundary was 0.952, a 12% improvement. However, the F1 for Class 2 was still very poor at 0.125. To address this, I took a smaller subset of 100,000 unlabelled documents, and continued active learning, only labelling tweets on the 0|2 class boundary. After labelling 165 documents, the performance on the validation set is reported in Table 7 below.

**Table 7.** Performance of final model on the validation set.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Reference | | | |  | |  | |  | |  | | |  |  | |
| Prediction | 0 | 1 | 2 | |  | |  | |  | |  |  | Class 0 | | Class 1 | | Class 2 |
| 0 | 929 | 15 | 11 | |  | |  | |  | |  | Precision | 0.973 | | 0.519 | | 0.333 |
| 1 | 13 | 14 | 0 | |  | |  | |  | |  | Recall | 0.9738 | | 0.483 | | 0.343 |
| 2 | 12 | 0 | 6 | |  | |  | |  | |  | F1 | 0.973 | | 0.5 | | 0.343 |

Additional active learning significantly increased the F1 for class 2 from 0.125 to 0.343, while the Class 1 F1 score remained at 0.5. After 71 iterations of active learning, each which labelled between 15 and 45 tweets, there were a total of 8,872 labelled tweets in the training set. 3,992 were labelled in pre-active learning, 3,880 were labelled during active learning, and 1,000 were from the validation set. 1,591,130 unlabelled tweets were then predicted on, resulting in a total of 1,596,057 observations for the final public opinion variables.[[6]](#footnote-6)

## 5.2 Regression Analysis

5.2.1 Operationalization of variables

For the **explanatory variable,** for each week in the analysis, a measure of pro or anti-lockdown sentiment will be taken for country, *i,* which aggregates the proportion of tweets in the TBCOV data coded as either pro or anti-lockdown in week *t.*

For the **outcome variable,** the measure of interest is the lockdown stringency, relative to the current risk posed by the virus, taken from the risk of openness and stringency indices from OxCGRT (Hale et al., 2021).[[7]](#footnote-7)

Countries with a positive score are over-responding to the risk of the virus, and those with a negative score are under-responding.

A control variable, “Average Western Stringency”, is also included. From a preliminary analysis of the data, there are clear regional correlations in relative stringency over the course of the pandemic, where countries moved congruently. This is to be expected due to foreign policy influence on COVID-19 policy, and changes in scientific advice in the global community. “Average Western Stringency” measures the average relative stringency of countries within the UN’s definition of “Western European and other States”, which comprises of 29 countries either in Western Europe or closely allied.[[8]](#footnote-8) This grouping was selected as it included all 5 countries studied. For each country, average western stringency excludes their own score.

5.2.2 Regression

Following the research design proposed by Page and Shapiro (1983), and emulated by Hartley and Russell (1992), Hobolt and Klemmemsen (2005) and Lax and Phillips’ (2009), among others, this paper opts for a change-oriented research design, rather than a static approach. Change in public opinion is regressed on change in relative stringency, to determine whether public opinion and policy are congruent. However, unlike previous authors, who use instances of opinion change as units of analysis, this project will assess the relationship between changes in public opinion and policy on a continuous basis over one year. Due to the country-level groupings in the data, a hierarchical linear model must be employed to account for the shared variance within groups of hierarchically structured data (Woltman et al., 2012).

# 6 RESULTS

## 6.1 Summary Statistics

This section summarises the relationship between tweet classes and relative stringency, prior to determining whether the relationship is causal. Figure 4 presents a visualisation of Class 1 (pro-lockdown) and Class 2 (anti-lockdown) daily tweet counts over time.

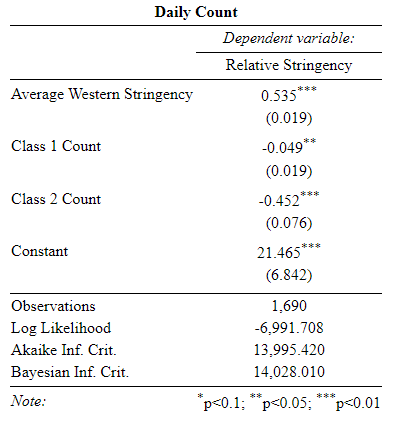
Chart, histogram

Description automatically generated

**Figure 4.** Total daily count of Class 1 and Class 2 tweets from 1 April 2020 to 1 April 2021.

Table 8 regresses the daily count of each class on relative stringency. Controlling for average Western stringency, both Classes 1 and 2 are negatively correlated with relative stringency, significant at the 5% level. These results suggest that, when controlling for average western stringency, discussion of restrictions whether pro or anti-lockdown in sentiment is less prevalent on days when relative stringency is lower. This relationship is far stronger with anti-lockdown sentiment than pro-lockdown sentiment.

**Table 8.** Daily regression of pro-lockdown (class 1) and anti-lockdown (class 2) tweet count on relative stringency. Multi-level model.



However, this relationship may be influenced by the distribution of tweets over time in the TBCOV data. It is apparent from Figure 4 that the count of Class 1 tweets roughly follows the count of total tweets seen in Figure 1. For this reason, Class 1 and Class 2 variables are transformed into proportion of total tweets at time *t* in Figure 5.

**Chart, line chart

Description automatically generated**

**Figure 5.** Total daily proportion of class 1 and class 2 tweets from 1 April 2020 to 1 April 2021.

This transformation has resulted in very high variance in tweet proportion from day to day. While this may be news related, it is also likely related to the lower performance of the models used for prediction, and the class imbalance in the data. To reduce this variance, the proportion of class tweets is aggregated weekly, where each data point marks the average proportion of each class in the previous week. While there is still high variance, the trends over time are now more apparent. Figure 6 displays Class 1 and 2 across time, compared with weekly relative stringency for each county.

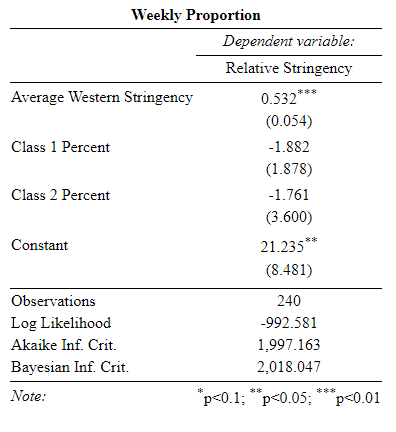
Graphical user interface

Description automatically generated with low confidence

**Figure 6.** The weekly proportion of class 1 and class 2 tweets, and weekly average relative stringency from 1 April 2020 to 1 April 2021.

While each country’s class 1 and class 2 proportions follow different trajectories, pro-lockdown sentiment was higher at the beginning of the pandemic in all countries. It remained high in most countries for several months, but the UK saw a rapid decline in pro-lockdown sentiment from April to July 2020. Pro-lockdown sentiment then hit another peak for Ireland, New Zealand and the United Kingdom during the major third wave beginning in January. Anti-lockdown sentiment does not have such apparent trends. This is likely due to the poor performance of the model, reducing the true variance in anti-lockdown sentiment, however some trends do appear. In Canada, Ireland and the UK, there is a consistent upward trend of anti-lockdown sentiment over time from April to autumn 2020. While it falls in both the UK and Ireland in the face of the third wave leading into the new year, it plateaus in Canada. In New Zealand, inversely, anti-lockdown sentiment seems to gain strength in autumn of 2020. To determine the direction of the relationship between each sentiment and relative stringency, Table 9 regresses Class 1 and 2 on relative stringency, controlling for average Western stringency.

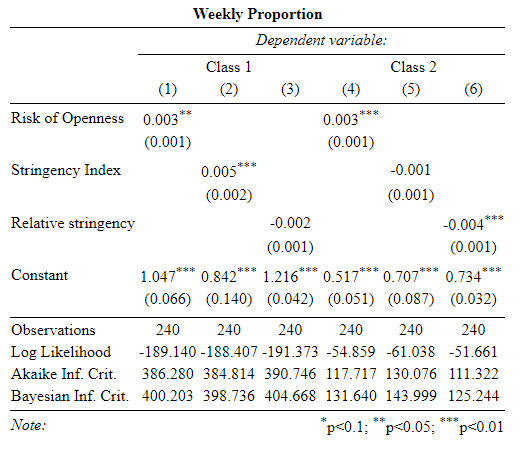
**Table 9.** Weekly regression of pro-lockdown (class 1) and anti-lockdown (class 2) tweet proportion on relative stringency. Multi-level model.



There is no significant relationship between pro-lockdown or anti-lockdown tweet proportion and relative stringency when controlling for average Western stringency. This suggests that the nature of online discussion related to lockdown is not significantly different when relative stringency is higher, compared with when it is lower.

As a final exploratory regression, I individually regress risk of openness, stringency and relative stringency on both Class 1 and 2 weekly proportions to determine the direction of the relationships.

**Table 10.** Weekly regression of risk of openness, stringency index and relative stringency on pro-lockdown (class 1) and anti-lockdown (class 2) tweet proportion. Multi-level model.



Pro-lockdown sentiment has a significant relationship with both risk and stringency, but not relative stringency. The positive correlation between risk and pro-lockdown sentiment is expected, however, the direction of the relationship with stringency is surprising, perhaps suggesting that when lockdown measures are high, pro-lockdown sentiment is also high, defending the measures staying in place, and encouraging others to comply. Anti-lockdown sentiment is significantly positively correlated with risk and negatively correlated with relative stringency. The relationship with relative stringency is to be expected: when stringency is high relative to the virus severity, there is public backlash and criticism, or inversely, as countries’ stringency is low while risk remains high, there is less anti-lockdown criticism of the measures in place as that demographic approve of low relative stringency. The relationship with risk is more surprising. When risk is high, criticism of lockdown measures and doubts about the legitimacy of scientific evidence are high. This phenomenon of COVID-19 denial increasing in the face of higher risk may be related to cognitive dissonance. The Belief-Disconfirmation Paradigm states that dissonance is aroused when people are exposed to information which is inconsistent with their beliefs. If this information does not change their belief, it likely leads to a rejection of that information (Harmon-Jones and Mills, 2019). Greater risk presents anti-lockdown individuals with more information about the threat of the virus, which may result in them rejecting scientific information about the virus altogether, as it is inconsistent with their beliefs that the country should not go into lockdown.

## 6.2 Main Regression Results

Multiple model specifications are included in this results section, with continuous or binary response variables, and continuous or ordinal explanatory variables. Plots of variable distributions can be found in Appendix C. The critical value for this paper is any coefficient found to be significant at the 5% level. As the model performed significantly worse at classifying anti-lockdown over pro-lockdown tweets, some model specifications only include Class 1. All variables are aggregated weekly, therefore, for example, represents the difference in country i, between mean relative stringency this week (t), and mean relative stringency last week (t-1). x represents different time lags in the regression. In this specification, lags are between 0 and 4 weeks.

**Table 11.** Weekly regression of change in pro-lockdown (class 1) and anti-lockdown (class 2) tweet proportion on change in relative stringency at multiple time lags. Multi-level model.

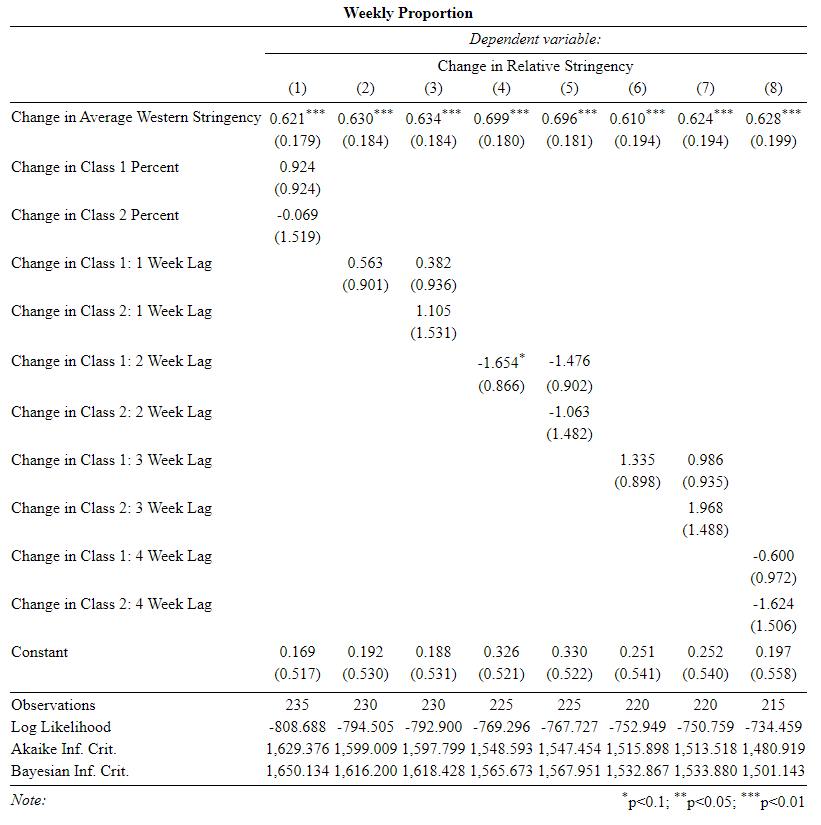
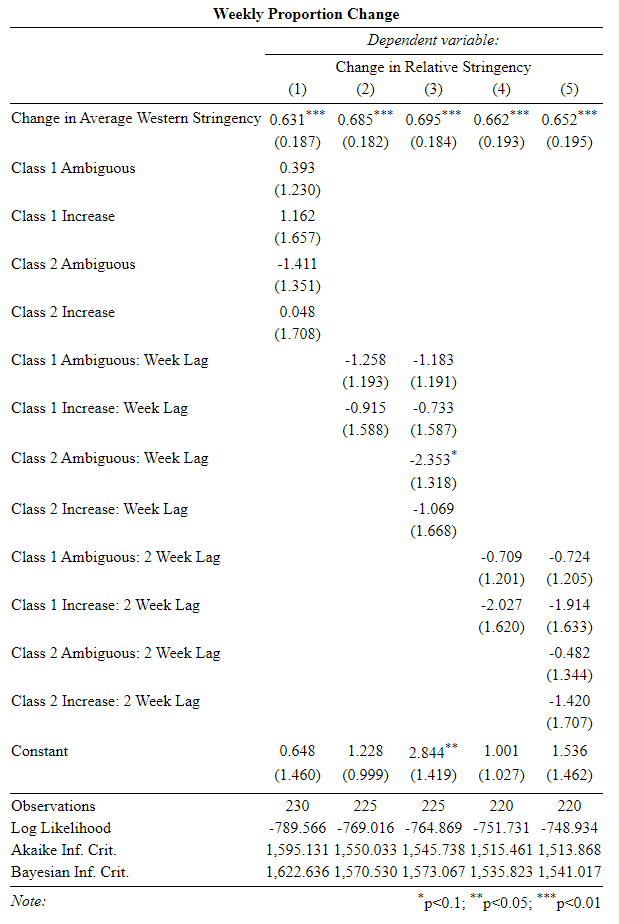


Table 11 presents little evidence to suggest that weekly changes in public opinion had a causal effect on changes to relative stringency, besides the two-week lag in change in class one on change in relative stringency, which is significant and negative at only the 10% level. This negative co-efficient suggests that increasing pro-lockdown sentiment resulted in decreasing relative stringency, contrary to H1, however, this is not significant at the 5% level. While I believe the outcome variable, relative stringency, is better suited to a continuous measure, I have also included a logit model, binarizing changes to policy, in line with the Page and Shapiro’s work. Appendix D contains the regression table, however, in line with Table 11, there is no statistically significant relationship at the 5% level between either pro or anti-lockdown sentiment and relative stringency.

Referring to Figure 8 which displays Class 1 and 2 tweets over time, it is apparent these measures are highly variable. To counteract this weekly variation, the next regression transforms public opinion variables from continuous to ordinal, with 3 response categories. An observation falls into category 1 (decrease) if that class proportion has seen negative change for the past two weeks. An observation falls into category 2 (inconsistent) if, in the past two weeks, it has seen both positive and negative change. An observation is category 3 if that class proportion has seen positive change for the past two weeks.

**Table 12.** Weekly regression of ordinal change in pro-lockdown (class 1) and anti-lockdown (class 2) on change in relative stringency at multiple time lags. Multi-level model.



In line with previous results, there is no significant relationship reported. Tables 13 and 14 report on the change in pro and anti-lockdown sentiment one week prior with change in relative stringency.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 13.** Congruence between pro-lockdown sentiment (one week prior) and relative stringency | | | | | | | | | | | | | | |
|  |  | | | | | | | | | | | | | |
|  | | Total Cases | | | |  | | Cases with clear opinion change | | | |
|  | | % | | | N |  | | % | | | N | |
| Congruent change in opinion & policy | | | 20 | (45) | | |  | | 48 | (45) | | | |
| Noncongruent changes in policy | | | 22 | (49) | | |  | | 52 | (49) | | | |
| Inconsistent change | | | 59 | (131) | | |  | | - | - | | | |
|  | | | 100 | (225) | | |  | | 100 | (94) | | | |
| **Table 14.** Congruence between anti-lockdown sentiment (one week prior) and relative stringency | | | | | | | | | | | | | | |
|  |  | | | | | | | | | | | | | |
|  | | Total Cases | | | |  | | Cases with clear opinion change | | | |
|  | | % | | | N |  | | % | | | N | |
| Congruent change in opinion & policy | | | 17 | (38) | | |  | | 46 | (38) | | | |
| Noncongruent changes in policy | | | 20 | (45) | | |  | | 54 | (45) | | | |
| Inconsistent change | | | 63 | (142) | | |  | | - | - | | | |
|  | | | 100 | (225) | | |  | | 100 | (83) | | | |

In the majority of cases, change in public opinion was inconsistent, increasing one week, decreasing the next, or vice versa. In cases where there was clear change in public opinion, policy moved noncongruently more often than it did not, indicating that based on these model specifications and the data used, this project fails to reject the null hypotheses of both H1 and H2.

The operationalisation of public opinion until this point has been unusual, as it has treated two opposing opinions as independent events. I posit that changes in the strength of pro-lockdown opinion can only be assessed relative to the strength of anti-lockdown opinion. A traditional study, which derives public opinion from survey responses, would measure approval of a policy relative todisapproval (Hartley and Russett, 1992; Lax and Phillips, 2009; Page and Shapiro, 1983), therefore, this project’s measure of public opinion should also be a relative one. For the final specification of the model, I will regress the ratio of anti-lockdown to pro-lockdown tweets at multiple time lags.

**Table 15.** Weekly regression of change in ratio of anti-lockdown (class 2) to pro-lockdown (class 1) tweets on change in relative stringency at multiple time lags. Multi-level model.

A screenshot of a computer

Description automatically generated with low confidence

|  |  |  |
| --- | --- | --- |
| **Table 16.** Congruence: pro-anti lockdown sentiment proportion and relative stringency (one week lag). | | |
|  |  | |
|  |  | |
|  | **Total Cases** | |
|  | % | N |
| Congruent change in opinion & policy | 51 | 120 |
| Noncongruent changes in policy | 49 | 110 |
|  | 100 | (230) |

There is no statistically significant difference at any time lag between the proportion of anti to pro lockdown sentiment and relative stringency. Table 16 shows that noncongruent change in policy occurs almost as frequently as congruent change. Figure 7 displays changes in opinion proportion and relative stringency over time.

**Graphical user interface

Description automatically generated**

**Figure 7.** Weekly proportion of class 1 to class 2 tweets, and weekly average relative stringency from 1 April 2020 to 1 April 2021.

In Canada and Ireland, the proportion of pro to anti-lockdown opinions is quite stable over time, although Ireland sees an increase in anti-lockdown sentiment relative to pro-lockdown sentiment between August and November 2020. The changes to both countries’ relative stringencies appear to be largely independent of changes in public opinion over time. While anti-lockdown sentiment becomes strong relative to pro-lockdown sentiment in New Zealand over time, due to the small sample size of New Zealand opinion, it is difficult to see stable trends, except for the period between July 2020 and October/November, where there is a steady and consistent rise of anti-lockdown sentiment. In Australia, both congruent and noncongruent trends are apparent over time. Between September and November 2020, anti-lockdown sentiment rises relative to pro-lockdown, with relative stringency rising too – moving non-congruently. However, there is also a similar increasing anti-lockdown to pro-lockdown sentiment between April and June, followed by a large decrease in relative stringency during June.

Based on the graph above, there appears to be some evidence of congruence between opinion and policy in the United Kingdom, the country with the largest sample size of tweets in the study. From July to November there is a clear rise in anti-lockdown sentiment to pro-lockdown sentiment, which is met by a congruent decrease in relative stringency, decreasing to its lowest point in October/November. Almost immediately, public opinion does a U turn, and pro-lockdown sentiment increases relative to anti-lockdown sentiment from October to January, and policy follows suit, with relative stringency increasing. From January, there is evidence of non-congruence, whereby anti-lockdown sentiment appears to increase again, while relative stringency increases. Below, Table 17 summarises these trends, broken down at a country level.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 17.** Congruence: pro-anti lockdown sentiment proportion by Country (one week lag) | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |
|  | UK |  | Australia |  | Canada |  | Ireland |  | New Zealand |
|  | N |  | N |  | N |  | N |  | N |
| Congruent change in opinion & policy | 31 |  | 26 |  | 22 |  | 19 |  | 22 |
| Noncongruent changes in policy | 15 |  | 20 |  | 24 |  | 27 |  | 24 |
|  | 46 |  | 46 |  | 46 |  | 46 |  | 46 |
|  |  |  |  |  |  |  |  |  |  |

Table 17 indicates that there is clear congruence between public opinion and policy in the UK. Over a 46-week period, 67% of changes to relative stringency in the UK were congruent with changes in UK public opinion one week prior. Table 18 below regresses changes in class proportion on changes to relative stringency for the UK only.

**Table 18.** United Kingdom. Weekly regression of change in ratio of anti-lockdown (class 2) to pro-lockdown (class 1) tweets on change in relative stringency at multiple time lags. Linear regression.

Table

Description automatically generated

The relationship is significant at the one-week time lag. If the proportion of anti-lockdown tweets to pro-lockdown tweets increases by 100% one-week prior, relative stringency decreases by 12.192 points – pursuing a relatively less stringent lockdown policy. This is significant at the 1% level. Given that results found in the UK are not replicable across the study, I fail to reject the null hypothesis of H3 at the 5% level. However, in the discussion section following, I will discuss some potential reasons why this study may not have found significant results.

## 6.3 Addressing Causality

As proposed in Hobolt and Klemmemsen’s (2005) paper, this project will also regress change in relative stringency in time *t-x* on change in public opinion. Based on Wlezien’s (1995) theory of the public as a thermostat, a significant positive relationship between these variables would be expected – when relative stringency rises, the public reacts with an increase in anti-lockdown sentiment. Table 19 displays the results of regressing different time lags of policy on opinion. There is no significant relationship between current policy and future opinion. While there is no evidence that causality works in the inverse direction from policy to opinion, there should be some evidence that public opinion is reactive to changes in relative stringency. When coupled with the fact that the relationship between public opinion and policy is only significant at one time lag, the results from the UK do not appear to be robust.

**Table 19.** United Kingdom. Weekly regression of change in relative stringency on change in the ratio of anti-lockdown (class 2) to pro-lockdown (class 1) tweets. Linear regression.

Table

Description automatically generated

# 7 DISCUSSION & LIMITATIONS

This section will discuss potential reasons for the lack of significant results in the study, alongside the limitations and weaknesses of the research design.

With the exception of the UK, this project finds no evidence to support its multiple hypotheses relating to public opinion and policy during COVID-19 in five Western democracies. These hypotheses were tested to address the question: *Did online public opinion on COVID-19 lockdown measures influence the stringency of governments’ lockdown policies?* While the results of this project do not provide enough evidence to answer this question definitively, I believe that given some alterations to the design, a future project may provide better evidence.

Firstly, the model used to classify tweets in this project had relatively poor performance, particularly in classifying anti-lockdown tweets. A larger initial training set, with more consistent labelling, cross-checked by multiple coders would likely significantly improve the performance of the model. The poor performance of the model means that significant results may be false or misleading, particularly if the model has significantly different performance across countries or time. Appendix E details how the performance of the model varies across countries and time. Performance appears roughly similar, although the extent of this issue of validity cannot be tested entirely based on the validation-set data.

Beyond the fact that lower performance of the model may lead to inaccurate results, it also led to estimates of public opinion that varied highly from week to week, in particular for countries with a smaller sample of tweets such as Ireland and New Zealand, meaning it was difficult to derive trends over time for public opinion. A future project would ideally both improve the performance of the model and expand the computational capacity of the project, incorporating a larger sample of tweets.

As detailed in the sampling strategy, sentiment estimates may be inaccurate over time, as the proportion of tweets deleted in the sample increased from 60% to 84%, indicating that sentiment later in the sample was more polarised and negative. If this is true, estimates of pro and anti-lockdown sentiment are likely underestimated for the final months of the sample.

Reasons for the lack of significant results go beyond the data and the model’s performance, but into the research design of the project. Public opinion and policy are undoubtedly endogenous in this project. Pro-lockdown sentiment, for example, is correlated with fear of COVID-19, as the coding scheme classifies tweets expressing or invoking fear as “pro-lockdown”. Fear of the virus is most likely correlated with virus severity, which is also a determinant of relative stringency.

Another weakness of the project lies in its definition of the outcome variable. Relative stringency equates a decrease in risk with an increase in stringency of measures. A decrease in relative stringency can indicate that either a country is implementing new policy to increase lockdown measures, or risk is decreasing due to the lockdown measures that are currently in place. While the former is likely to be influenced by public opinion, the latter should not be, yet it is treated as such.

Finally, in line with other research on policy and public opinion, it cannot be said with any certainty that congruence of public opinion and policy is not due to another confounding factor. For example, while there is an upward trend in anti-lockdown opinion over time, alongside a decrease in relative stringency, it is likely that politicians’ preferences followed the same trends as opinion, as pressures from interest or economic groups pushed for opening of the economy, or as lockdown fatigue set in for politicians themselves.

# 8 CONCLUSION

Despite a lack of significant findings, this project contributes to the literature on public opinion during COVID-19, providing a coding scheme for the classification of pro and anti-lockdown text. While numerous studies have worked at classifying COVID-19 Twitter data, in particular emotion labelling (Yang et al., 2020; Li et al., 2020; Kabir and Madria, 2021), the existing literature has not yet attempted to track online public opinion related to governments’ COVID-19 policy. Although the model did not perform as well as anticipated, its’ analysis of trends in COVID-19 public opinion provide a valuable insight into how the public responds to both risk and policy during a pandemic.

# BIBLIOGRAPHY

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N. and Acquisti, A., 2013, February. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 897-908.

Anstead, N. and O'Loughlin, B., 2015. Social media analysis and public opinion: The 2010 UK general election. *Journal of computer-mediated communication*, *20*(2), pp.204-220.

Barberá, P., Casas, A., Nagler, J., Egan, P.J., Bonneau, R., Jost, J.T. and Tucker, J.A., 2019. Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, *113*(4), pp.883-901.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A., 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), p.774.

Bhattacharya, P. and Ganguly, N., 2016, March. Characterizing deleted tweets and their authors. In *Tenth international AAAI conference on web and social media*.

Burstein, P., 2003. The impact of public opinion on public policy: A review and an agenda. Political research quarterly, 56(1), pp.29-40.

Caughey, D. and Warshaw, C., 2018. Policy preferences and policy change: Dynamic responsiveness in the American states, 1936–2014. *American Political Science Review*, *112*(2), pp.249-266.

Ceron, A. and Negri, F., 2016. The “social side” of public policy: Monitoring online public opinion and its mobilization during the policy cycle. *Policy & Internet*, *8*(2), pp.131-147.

Chen, D., Peng, D., Rieger, M.O. and Wang, M., 2021. Institutional and cultural determinants of speed of government responses during COVID-19 pandemic. *Humanities and Social Sciences Communications*, *8*(1), pp.1-9.

Cohn, D., Atlas, L. and Ladner, R., 1994. Improving generalization with active learning. *Machine learning*, *15*(2), pp.201-221.

Documenting the Now. 2020. Hydrator [Computer Software]. Retrieved from https://github.com/docnow/hydrator

Erikson, R.S., 1978. Constituency opinion and congressional behavior: A reexamination of the Miller-Stokes representation data. *American Journal of Political Science*, pp.511-535.

Flood, C.M., MacDonnell, V., Thomas, B. and Wilson, K., 2020. Reconciling civil liberties and public health in the response to COVID-19. *Facets*, *5*(1), pp.887-898.

Frank, E. and Bouckaert, R.R., 2006, September. Naive bayes for text classification with unbalanced classes. In *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, Heidelberg. pp. 503-510.

Gilardi, F., Gessler, T., Kubli, M. and Müller, S., 2021. Social media and policy responses to the COVID‐19 pandemic in Switzerland. *Swiss Political Science Review*, *27*(2), pp.243-256.

Given, L. M. 2008. Inter- and Intracoder Reliability. In: Lisa M. Given Editor, 2008. *The SAGE Encyclopedia of Qualitative Research Methods*, Thousand Oaks, CA: SAGE Publications, Inc. pp. 4-46

Grossman, G., Kim, S., Rexer, J.M. and Thirumurthy, H., 2020. Political partisanship influences behavioral responses to governors’ recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences*, *117*(39), pp.24144-24153.

Hale, T., Angrist, N., Kira, B., Petherick, A., Phillips, T. and Webster, S., 2020. Variation in government responses to COVID-19.

Hale, T., Phillips, T., Petherick, A., Kira, B., Angrist, N., Aymar, K., Webster, S., Majumdar, S., Hallas, L., Tatlow, H. and Cameron-Blake, E., 2020a. Risk of openness index: When do government responses need to be increased or maintained. *Research note, University of Oxford and Blavatnik School of Government, September*.

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S. and Tatlow, H., 2021. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour.* [https://doi.org/10.1038/s41562-021-01079-8](https://www.nature.com/articles/s41562-021-01079-8)

Harmon-Jones, E. and Mills, J., 2019. An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. *American Psychological Association.*

Hartley, T. and Russett, B., 1992. Public opinion and the common defense: who governs military spending in the United States?. *American political science review*, *86*(4), pp.905-915.

Hobolt, S.B. and Klemmemsen, R., 2005. Responsive government? Public opinion and government policy preferences in Britain and Denmark. *Political Studies*, *53*(2), pp.379-402.

House of Commons Science and Technology Committee and the Health and Social Care Committee. 2021, October. Joint Report: Coronavirus: lessons learned to date. https://committees.parliament.uk/committee/81/healthand-social-care-committee/news/157991/coronavirus-lessons-learned-todate-report-published.

Imran, M., Qazi, U. and Ofli, F., 2022. TBCOV: Two Billion Multilingual COVID-19 Tweets with Sentiment, Entity, Geo, and Gender Labels. *Data*, *7*(1).

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2009. ‘Resampling Methods’ in *An introduction to statistical learning,* 2nd edn, New York, Springer, pp. 194-224.

Jiang, J., Chen, E., Yan, S., Lerman, K. and Ferrara, E., 2020. Political polarization drives online conversations about COVID‐19 in the United States. *Human Behavior and Emerging Technologies*, *2*(3), pp.200-211.

Jones, Bryan D. 1994. Reconceiving Decision-Making in Democratic Politics. Chicago: University of Chicago Press

Kabir, M.Y. and Madria, S., 2021. EMOCOV: Machine learning for emotion detection, analysis and visualization using COVID-19 tweets. *Online Social Networks and Media*, *23*, p.100-135.

Kerr, J., Panagopoulos, C. and van der Linden, S., 2021. Political polarization on COVID-19 pandemic response in the United States. *Personality and individual differences*, *179*, p.110892.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kemp, S. 2022a, February. Digital 2022: Australia. *Datareportal.* <https://datareportal.com/reports/digital-2022-australia?rq=australia>

Kemp, S. 2022b, February. Digital 2022: Canada. *Datareportal.* <https://datareportal.com/reports/digital-2022-canada?rq=2022%20canada>

Kemp, S. 2022c, February. Digital 2022: Ireland. *Datareportal.* <https://datareportal.com/reports/digital-2022-ireland>

Kemp, S. 2022d, February. Digital 2022: New Zealand. *Datareportal.* <https://datareportal.com/reports/digital-2022-new-zealand>

Kemp, S. 2022e, February. Digital 2022: The United Kingdom. *Datareportal.* 9<https://datareportal.com/reports/digital-2022-united-kingdom?rq=united%20kingdom>

Kwan, J.S.L. and Lim, K.H., 2020, December. Understanding Public Sentiments, Opinions and Topics about COVID-19 using Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. pp. 623-626.

Lax, J.R. and Phillips, J.H., 2009. Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review*, *103*(3), pp.367-386.

Li, X., Zhou, M., Wu, J., Yuan, A., Wu, F. and Li, J., 2020. Analyzing COVID-19 on online social media: Trends, sentiments and emotions. *arXiv preprint arXiv:2005.14464*.

LSE Candidate Number: 17824. *A Supervised Machine Learning Approach to Analysing Political Speech and Restriction Compliance During COVID-19.*

Lwin, M.O., Lu, J., Sheldenkar, A., Schulz, P.J., Shin, W., Gupta, R. and Yang, Y., 2020. Global sentiments surrounding the COVID-19 pandemic on Twitter: analysis of Twitter trends. *JMIR public health and surveillance*, *6*(2), p.e19447.

Miller, W.E. and Stokes, D.E., 1963. Constituency influence in Congress. American political science review, 57(1), pp.45-56.

Moloney, E. 2020, November. Nine ways Government went against Nphet advice as country set to exit Level 5 Covid-19 restrictions. *Irish Independent.* <https://www.independent.ie/irish-news/nine-ways-government-went-against-nphet-advice-as-country-set-to-exit-level-5-covid-19-restrictions-39801087.html>

Monroe, A.D., 1998. Public opinion and public policy, 1980-1993*. Public Opinion Quarterly*, pp.6-28.

Murphy, J., Link, M.W., Childs, J.H., Tesfaye, C.L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M. and Harwood, P., 2014. Social media in public opinion research: Executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), pp.788-794.

Page, B.I., 1994. Democratic responsiveness? Untangling the links between public opinion and policy. PS: Political Science & Politics, 27(1), pp.25-29.

Page, B.I. and Shapiro, R.Y., 1983. Effects of public opinion on policy. American political science review, 77(1), pp.175-190.

Pranckevičius, T. and Marcinkevičius, V., 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), pp.221-232.

Repucci, S. and Slipowitz, A. 2022. Freedom In The World 2022. The Global Expansion of Authoritarian Rule. *Freedom House.* <https://freedomhouse.org/report/freedom-world/2022/global-expansion-authoritarian-rule>

Settles, B., 2012. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, *6*(1), pp.1-114.

Sloan, L. and Morgan, J., 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, *10*(11), p.e0142209.

Stake, R. E. 1995. The art of case study research. Thousand Oaks, CA: SAGE Publications

Stimson, J.A., MacKuen, M.B. and Erikson, R.S., 1995. Dynamic representation. *American political science review*, *89*(3), pp.543-565.

Sun, Y., Wong, A.K. and Kamel, M.S., 2009. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), pp.687-719.

Talic, S., Shah, S., Wild, H., Gasevic, D., Maharaj, A., Ademi, Z., Li, X., Xu, W., Mesa-Eguiagaray, I., Rostron, J. and Theodoratou, E., 2021. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. *bmj*, *375*.

Twitter Developer Platform. 2022. Twitter IDS. *Documentation.* <https://developer.twitter.com/en/docs/twitter-ids#:~:text=Today%2C%20Twitter%20IDs%20are%20unique,number%2C%20and%20a%20sequence%20number>.

VanDusky-Allen, J. and Shvetsova, O., 2021. How America’s partisan divide over pandemic responses played out in the states. *The Conversation*. <https://theconversation.com/how-americas-partisan-divide-over-pandemic-responses-played-out-in-the-states-157565>

Wang, S., Schraagen, M., Sang, E.T.K. and Dastani, M., 2020. Dutch general public reaction on governmental covid-19 measures and announcements in twitter data. *arXiv preprint arXiv:2006.07283*.

Welbers, K., Van Atteveldt, W. and Benoit, K., 2017. Text analysis in R. *Communication Methods and Measures*, *11*(4), pp.245-265.

Woltman, H., Feldstain, A., MacKay, J.C. and Rocchi, M., 2012. An introduction to hierarchical linear modelling. *Tutorials in quantitative methods for psychology*, *8*(1), pp.52-69.

Wlezien, C., 1995. The public as thermostat: Dynamics of preferences for spending. *American journal of political science*, pp.981-1000.

Wright Jr, G.C., Erikson, R.S. and McIver, J.P., 1987. Public opinion and policy liberalism in the American states. *American Journal of Political Science*, pp.980-1001.

Xing, Y., Li, Y. and Wang, F.K., 2021. How privacy concerns and cultural differences affect public opinion during the COVID-19 pandemic: a case study. *Aslib Journal of Information Management*.

Yang, Q., Alamro, H., Albaradei, S., Salhi, A., Lv, X., Ma, C., Alshehri, M., Jaber, I., Tifratene, F., Wang, W. and Gojobori, T., 2020. Senwave: monitoring the global sentiments under the Covid-19 pandemic. *arXiv preprint arXiv:2006.10842*.

Zhang, D., Wang, J. and Zhao, X. 2015. September. Estimating the uncertainty of average F1 scores. In *Proceedings of the 2015 International conference on the theory of information retrieval*. pp. 317-320.

Zhang, T. and Oles, F.J., 2001. Text categorization based on regularized linear classification methods. *Information retrieval*, *4*(1), pp.5-31.

Zhu, J., Wang, H., Hovy, E. and Ma, M., 2010. Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing (TSLP)*, *6*(3), pp.1-24.

Zhou, L., Wang, W. and Chen, K., 2016, April. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web*, pp. 603-612.

# APPENDICES

## Appendix A

Replication Code for tables, graphs, model selection, active learning, and compiling variables can be found in this Github Repository: <https://github.com/KCCarbery/Capstone>

However, as Twitter does not allow for the redistribution of tweet text, the TBCOV data used will not be published.

Two text files are available:

1. “[Validation\_set\_predictions\_and\_labels.csv](https://github.com/KCCarbery/Capstone/blob/main/Data/Validation_set_predictions_and_labels.csv)” so the reader can better understand the coding scheme. This includes a column “preds”, which indicates the class to which the final model assigned it to see examples of which texts the model correctly and incorrectly classified.
2. “Intracoding.csv” shows the file with two sets of labels for intracoder test.

## Appendix B

Sampled vs Hydrated for each country over time

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

## Appendix C

Distribution plots of variables.

Chart, histogram

Description automatically generated

**Figure C1.** Distribution of Explanatory Variable: Change in Class 1 Proportion

Chart, histogram

Description automatically generated

**Figure C2.** Distribution of Explanatory Variable: Change in Class 2 Proportion

Chart, histogram

Description automatically generated

**Figure C3.** Distribution of Outcome Variable: Change in Relative Stringency

Chart, histogram

Description automatically generated

**Figure C4.** Distribution of Explanatory Variable: Change in Class 2/Class 1 Proportion.

## Appendix D

**Table D.** Weekly regression of change in pro-lockdown (class 1) and anti-lockdown (class 2) tweet proportion on change in relative stringency at multiple time lags. Logit multi-level model.

Text

Description automatically generated with low confidence

## Appendix E

Variation in model performance on the validation set by country and time.

|  |  |  |
| --- | --- | --- |
| **Table E1.** Performance of model across countries. | | |
| Country | Accuracy | N |
| Australia | 0.971 | 137 |
| Canada | 0.960 | 273 |
| Ireland | 0.911 | 56 |
| New Zealand | 0.933 | 15 |
| UK | 0.942 | 519 |
|  | 0.949 | 1,000 |

|  |  |  |
| --- | --- | --- |
| **Table E2.** Performance of model across time. | | |
| Date Range | Accuracy | N |
| 01/04/20 – 13/05/20 | 0.950 | 200 |
| 13/05/20 – 03/08/20 | 0.925 | 200 |
| 04/08/20 – 19/10/20 | 0.945 | 200 |
| 20/10/20 – 06/01/21 | 0.965 | 200 |
| 07/01/21- 03/31/21 | 0.960 | 200 |
|  | 0.949 | 1,000 |

1. As text must be labelled in this project, it is limited to the native language of the coder – English. [↑](#footnote-ref-1)
2. This would be stated as national level, however, while the United Kingdom is often treated as a country, it is a collection of 4 countries, making up a sovereign state. [↑](#footnote-ref-2)
3. Following this clarification, I will now refer to these “sovereign states” as countries for simplicity. [↑](#footnote-ref-3)
4. Appendix B includes individual country graphs over time. [↑](#footnote-ref-4)
5. 1,221,328 was the number of all observations when all data was included, as text duplicates were removed from the sample for active learning. [↑](#footnote-ref-5)
6. Only 4,927 of the 8,872 labelled tweets were included in the sampled data. 2,000 of the training set were from an external dataset, while a further 1,945 of the training set were in the TBCOV data for our countries, they were sampled prior to final sampling for the dataset, and were not selected. The final model did not incorporate the 1,000 labels from the validation set. [↑](#footnote-ref-6)
7. Both variables are a score out of 100, therefore the outcome variable deviates from 0. [↑](#footnote-ref-7)
8. <https://www.un.org/dgacm/en/content/regional-groups> [↑](#footnote-ref-8)