Social Media, Risk Perception, and Social Distancing: Evidence from 15.3 Million Geolocated Tweets

George Tyler

March 3, 2021

Abstract

Does social media predict risk-taking behaviour? I investigate this question in the context of COVID-19 by exploiting a large panel of tweets. Using inferred and explicit geolocation data embedded in the tweets, I study the extent to which public expressions of sentiment such fear, anger, and optimism influence social distancing, as measured by GPS-located smartphone data.

`4430 words in main body, excluding headers and bibliography.`

Contents

1 Introduction

1.1 Overview

The early stages of the COVID-19 pandemic saw an unprecedented shift in behaviour for most citizens of the United States. In a short period of time, a large number changed their habits of working, socialising, and travelling. They did so both as a result of government restrictions in the form of non-pharmaceutical interventions (NPIs) and as a private response to the spread of the pandemic. Economists have taken interest in how citizens formed these behaviour changes, and the role that beliefs and risk attitudes played in determining the response to public policy. A new way to measure belief formation and public sentiment is with social media, an increasingly common platform for expression of opinion. It is plausible that those who express more risk-averse sentiment towards COVID online will be inclined to respond in a stricter fashion to social distancing and other public health regulations. In this dissertation, I study the impact of local expressions of risk attitude on public behaviour in the early months of the pandemic. Specifically, I study whether a measure of risk-averse sentiment on Twitter is linked to increased social distancing behaviour at the county/week level.

This dissertation contributes to two strands of the recent economics literature on the COVID-19 pandemic. First, it investigates the relationship between partisanship and risk preference. Previous papers posit that political preference influences social distancing through risk preference; I contend that *local* risk preference is a separate factor to political partisanship, and has an independent impact on social distancing. More broadly, the paper investigates the relationship between risk preference and economic behaviour, and presents a novel example of economic inference from social media using text analysis.

A key vector for expressing sentiment is social media, with Twitter and Facebook's suite of products[1] being the most widely-adopted, each platform having over 80 million monthly active users in the US. A survey by the Pew Research Foundation indicates that 22% of US adults use Twitter, with 42% of these using it on a daily basis (Perrin & Anderson, 2019). On Twitter, users can share their own text, with the option to link to a website; alternatively, they can 'retweet' another user's text or link. Users can also use 'hashtags' in their tweet, which connects their tweet to a particular topic. If the user has allowed it, Twitter also records the location of the tweet; and it is also possible for the user to set their location on their profile. In this way, it is possible to create a panel of geographically-located tweets about a particular topic.

I exploit GeoCov19 (Qazi et al., 2020), a dataset of 524 million geolocated tweets, to measure the local public sentiment on COVID in the US. The tweets cover the period from 1st February to 1st May, the period I focus on. The particular subset of the data I use contains 33.36 million tweets in total; a small subset are exactly geolocated (the user has provided a GPS location), while most are inferred from the location tab in the user's profile. The tweets were collected using the Twitter Streaming API, querying for tweets containing any of a list of 800 COVID-related keywords. I also use anonymous smartphone location data, collected by the company SafeGraph, as a measure of the extent of social distancing in an area. I present two measures of social distancing at county level: first, the median minutes spent at home during 8am-6pm; second, the proportion of measured devices that stayed at home all day (SafeGraph, Inc., 2020). Demographic controls are also acquired and presented from the American Community Survey and the 2010 US census.

I use dictionary-based text analysis to assess the level of risk sentiment in a tweet. More sophisticated methods of text analysis like latent factor modelling are discussed in the Methods section. In the absence of a lexicon of risk preference, the NRC Emotion Lexicon (Mohammad & Turney, 2013) is used. This is a widely-used mapping of English words to eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Starting from a set of tweets that mention COVID, I assign tweets containing fear-associated words to a risk-averse sentiment. The base unit of analysis is the county-week; as such, I measure the proportion of tweets that contain fearful language in each county and week.

[1] Facebook, Facebook Messenger, Instagram, and WhatsApp

It is plausible that social media is a valid measure for risk appetite. The intuition is that the textual content of a social media post broadly reflects the poster's current opinion of a topic: for example, in response to the first confirmed US COVID death on February 26th, a user might express fearful sentiment, or a neutral sentiment. This opinion of the topic, particularly their level of fear, maps to a user's broader expectations about the course of the pandemic: while other emotions like joy, anticipation, and trust may rely on the context of the discussion, expressions of fear are plausibly consistent in mapping to risk-averse sentiment. When restrictions are implemented, users who initially formed pessimistic expectations may be more inclined to adhere more to them than a user who formed optimistic or netural expectations. The key aspect to the data is that the Twitter conversations provide a real-time insight into local sentiment as NPIs are implemented; this sentiment changes both in response to current, local experience *and* broader partisan interpretations of current events.

The primary econometric specification is a panel model with county and week fixed effects;

$$Y_{it} = \alpha + \beta r_{it} + \mu c_{it} + \tau_i + \delta_t + X_{it}\gamma + \epsilon_{it}$$

where Y_{it} is a vector of social distancing metrics, βr_{it} is the risk perception measure, (i.e. the proportion of total tweets containing fearful language), μc_{it} the number of COVID cases, $\tau_i + \delta_t$ county and week-level fixed effects, and $X_{it}\gamma$ demographic controls.

This research contributes to the recent economics literature seeking to explain the disparities in social distancing in the early stages of the pandemic in the US. In particular, partisanship has been shown to be a significant factor on the practice of social distancing: Allcott et al. (2020), Barrios and Hochberg (2020), and Painter and Qiu (2020) show that areas with more Republicans engaged in less social distancing, are associated with lower perceptions of risk of the pandemic, and exhibited less remote transactions. Ananyev et al. (2020) and Simonov et al. (2020) also measure the causal effect of the right-wing Fox News network on social distancing during the pandemic. This paper builds on Barrios and Hochberg (2020) in particular, which shows that online risk perception is predicted by Trump voter share: by measuring risk perception with a high-frequency geolocated dataset, my approach controls for political alignment and assesses the effect of risk perceptions on their own. In essence, I measure expressions of sentiment regarding COVID risk, and given this data I ask whether local risk sentiment predicts social distancing behaviour beyond political affilitation. Second, this research relates to the recent economics literature around heterogeneous-agent epidemiological models, which endogenise individual behaviour – including social distancing – into the effective reproductive number $R(t)$. These recent models, such as Acemoglu et al. (2020), Brotherhood et al. (2020), and Eichenbaum et al. (2020), assume that preferences over risk are predictive of social distancing behaviour; this paper looks to empirically confirm this key assumption.

This dissertation also contributes to the rapidly-expanding field of text analysis in economics, and presents an example of how the rich sentiment data encoded in social media communication can inform insights into public behaviour. This topic is particularly mature in finance – where sentiment data from public company documents, news media, and social media have been shown to predict stock market reactions (Bollen et al., 2011) – and monetary economics, where central bank statements, coded according to their attitude to inflation, predict fluctuations in Treasury securities (Gentzkow et al., 2019; Lucca & Trebbi, 2009). On the topic of empirical economics, this paper takes a similar approach – by using online data to predict local sentiment – as Stephens-Davidowitz (2014), which uses Google search data to proxy an area's racial animus, and uses this to estimate the Obama vote share. I use geolocated Twitter sentiment to proxy the local attitude to COVID in a given week, and test to see if this predicts social distancing practice.

The argument of the dissertation rests on the following assumptions: first, that social media data is a valid proxy for local risk appetite, and that fear-associated language in COVID-related tweets is an effective estimator of the risk appetite encoded in the tweet. It is also important to note a possible selection effect in the dataset: tweets about COVID may attract a greater level of fear-related language and not reflect an individual's true opinion about social distancing and other preventative measures. I address these assumptions and drawbacks and discuss methods to alleviate them in the Results section.

2 Literature Review

2.1 Text analysis in Economics

Context and applications of Text Analysis in economics; technical section describing the various methods used in the literature is in the methods section.

2.2 Usage of 'Digital Trace' and geolocation datasets in economics

General; description of previous uses of geolocation data

2.3 COVID-19: Empirical estimates of drivers of social distancing behaviour

Summary of empirical research on COVID

2.4 COVID-19: Economic Models

Discussion of different augmentations to SIR models; incorporation of endogenised social distancing / individual choice. Particular focus on risk-attitude heterogeneity.

3 Data

3.1 GeoCoV19 and geolocation inference of Twitter datasets

The primary dataset of tweets I present is a subset of GeoCoV19, a project which collected tweets relating to COVID-19 between February 1 and May 1 2020. The Twitter Streaming API provides a live filter function for a number of keywords and hashtags, and returns all tweets matching any of the search terms. The terms were chosen to cover a broad base of COVID-related talk, including searches for symptoms (e.g.'breathing difficulties'), behaviour (e.g. '#masks4all'), and popular hashtags (e.g. '#IStayHome, #FlattenTheCurve') in addition to central discussion topics like 'coronavirus'. The full list of search terms used is in a separate appendix. The dataset was collected with multilingual use in mind, but a majority of the tweets were based in the US; tracking with the fact that the US makes up most of Twitter's user base. In total, 524 million tweets were collected during the time period; during the first three weeks of February this number was lower, reflecting lower general interest, and increasing to around 6.4 million per day during March and April. We filter for tweets that are in English and are geotagged to originate from the United States. This primary dataset is supplemented with exact-geolocation Tweets from two other similar datasets, Banda et al. (2021) and Lamsal (2020). The collection method for both datasets is broadly similar to GeoCoV19, using the Streaming API and a set of COVID-related keywords. Of the final dataset, around 150,000 have exact geolocation embedded in the tweet. This is due

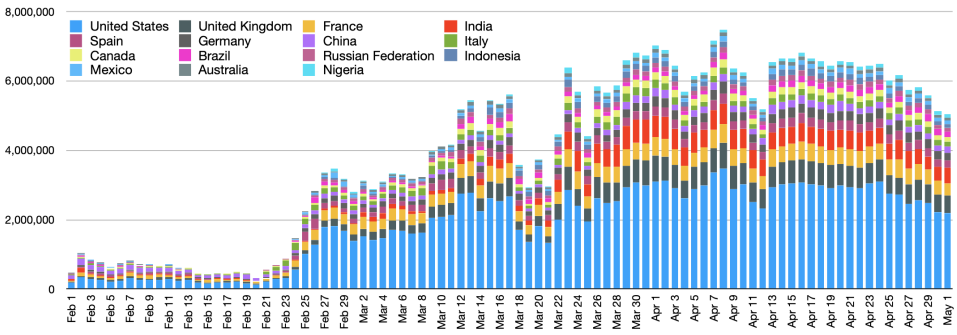
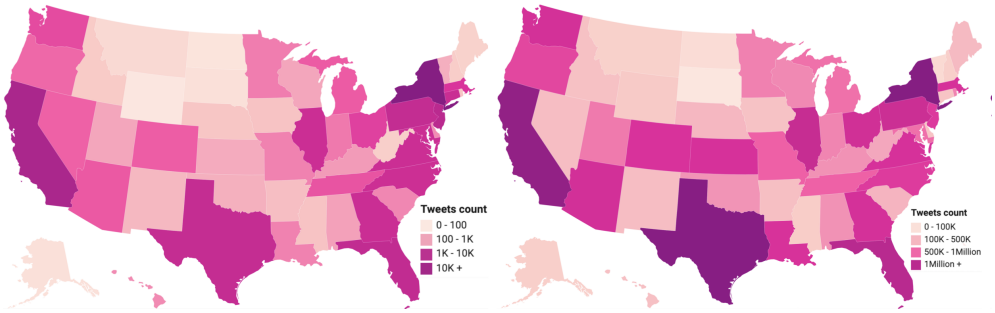


Figure 1: Daily distribution by country of GeoCoV19 tweets, Feb 1st to May 1st, 2020 (Qazi et al., 2020)

to the fact that geo-tagging is an explicit option that needs to be set for each tweet, involving activating location data on the mobile app. Another option for geo-tagging is to select a place from a search box; yielding a 'place' in the metadata. Both of these types of metadata involve accurate locations, but make up a small proportion of the total geolocated tweets. A third method of geolocating tweets is used to identify most of the locations: when activating a Twitter account the user is strongly encouraged to set their location in their profile. Although it is a free field (generated Place suggestions are included, but are optional), most users set this to their current location; this metadata is included with every tweet. The maintainers of the dataset then employ a toponym extraction approach to elicit the location of the location field. The text of the user location field is first cleaned of non-text characters and symbols. Candidates are then created from the remaining unigrams (single words) and bigrams (pairs of adjacent words), ensuring that two-word place names like 'Los Angeles' are included. Groups of three or more words are not considered. Each remaining candidate is filtered against a list of stopwords (see section 4), and against the 'World Cities Database'[2], an index of 3.1 million worldwide place names, covering 141,989 locations in the US. The remaining candidates are sent as one query to Nominatim, the OpenStreetMap search engine, yielding a best-attempt geolocation: the procedure works best when state and place name is given. Cross-checking the procedure with GPS-geolocated tweets, the dataset shows good coverage and accuracy across US counties, and so makes a panel approach viable. The maintainers of the dataset also presented locations derived from the text of the tweet itself using the same procedure, but we filter these tweets out of the final dataset due to low accuracy. A drawback to this gazeteer approach is that, since users can set their profile location freely, users from other countries or states could masquerade as Americans in particular locations, or the classification process could mis-classify a foreign (particularly English) place as a US location due to sharing a name – for example, York County, Maine. In order to account for this we remove counties from consideration that have a share of total tweets significantly higher than their population share of the US: Earth, Texas, is the most prominent example of this. Other than misclassification, deliberate setting of the profile location to a US location is possible. This may be a concern for large, well-known cities like Los Angeles and New York, but it is less likely that a given user will set their location to a less well-known American county. Finally, users may move county and not change their location.



(a) USA: Organic geotagged tweets (b) USA: User distribution

Figure 2: Geographic distribution of GeoCoV19 tweets and users (Qazi et al., 2020)

Twitter's terms of service restrict the large-scale sharing of Tweet datasets; hence public-facing tweet datasets can only be made available 'dehydrated', with only the universal identifiers (a long number that maps to the tweet in Twitter's database) given, instead of the tweet text and assorted metadata. In order to use the dataset to analyse tweet content, researchers wishing to use the dataset must apply and be accepted for a Developer account. This gives access to a password called an API key, which is used to query the Twitter API with the identifiers to 'rehydrate' and gain access to the full tweet text and metadata. Since the tweets are delivered from the servers at download time, this procedure entails that a proportion of Tweets – those identified as containing misinformation or those sent by users whose accounts have since been suspended or deleted – will be unavailable on request from the

[2]Available at `https://www.kaggle.com/max-mind/world-cities-database?select=worldcitiespop.csv`

service when the dataset is 'rehydrated'. This presents a selection problem for topics like misinformation, where Twitter enacts a stringent and continuous policy. A prominent concern for my data might be the suspension of Donald Trump's prolific Twitter account; his tweets were a touchstone for right-wing US online conversations. This does not present an issue to my data, however, because tweets interacting with tweets from a deleted account are still available from the API. The deletion rate ultimately encountered during the rehydration process was around 20%, reflecting the normal removal of machine-generated content.

3.2 SafeGraph: Geolocated smartphone data for measuring social distancing

A second central dataset is provided by SafeGraph, a company which usually collects data on commercial footfall, but made their dataset available for academic research in light of the pandemic. Their dataset consists of over 45 million anonymised smartphone GPS pings located to an accuracy of a 150m square location (SafeGraph, Inc., 2020). Since the data was collected in 2019 and 2020, year-on-year change can also be presented. The data is presented at district level, but for the purposes of the analysis I aggregated this to a county-level daily metric. For each device on each day, the most common night-time location is determined from the previous 6 weeks. This location is then used to determine the number of devices in a county which leave their home, the length and time of day they leave and return, the district they travel to, and the points of interest they visit. From this rich data, I derive two primary variables: the median minutes spent at home during 8am-6pm and the proportion of measured devices that stayed at home all day. It should be noted that the differential anonymisation means that the exact sum of devices is inaccurate; although there is a possibility to bias the proportions I calculate, this only arises in areas which are sparsely populated. These areas usually do not yield enough geo-located tweets to be included in the final analysis, so the problem is largely circumvented.

3.3 Data collection process

Technical details of process

3.4 Descriptive statistics

4 Text analysis: an overview

Section summarises Gentzkow et al, Text As Data (JEL, 2019) In the literature review I summarised the current uses of Text Analysis in the economic literature. In this section, I present a detailed look at the inference problem for text analysis of English documents, and the various types of approaches possible to tackle it. The section introduces concepts that are used in the later Methods section, and also serves as an introduction to the field of text analysis for economics.

4.1 The inference problem

For economists, the fundamental concept is that relevant information exists within a corpus of text. This information can be represented as a low-dimensional variable, relevant to a model of economic decisionmaking: for example, interest rate expectations or risk aversion. However, this low-dimensional variable is expressed in text, a noisy, extremely high-dimensional format: a list of documents which are n words long, drawn from a vocabulary of size p, has a dimension of p^n. Given this computational constraint, the central challenge is to isolate the latent variable from the high-dimensional noise in a robust manner. Gentzkow et al. (2019) present a useful notation for a two-step process of this extraction: first, the raw text $\mathcal{D}$ is mapped to an array of tokens, $\mathbf{C}$. These tokens – usually in the form of words, phrases, or sentences – are the fundamental units of analysis. Second, the token array is mapped to the outcome array $\hat{\mathbf{V}}$, which is an estimate of the latent variable $\mathbf{V}$. $\hat{\mathbf{V}}$ is then used in the final analysis. In these two mappings, there are two challenges: first, to include only tokens relevant to the variable in $\mathbf{C}$; second, to accurately estimate $\mathbf{V}$ using $\mathbf{C}$.

4.2 Mapping $\mathcal{D}$ to C: text pre-processing

In Gentzkow et al. (2019)'s notation, the vector $\mathcal{D}$ is the corpus of text to be analysed, consisting of documents $\mathcal{D}_i$): a central bank announcement, for example. $\mathbf{C}$ is a numerical matrix, where the columns correspond to tokens, and rows correspond to documents.[3] $\mathbf{C}_{ij}$ is a scalar representing the *term frequency*: how many times each token appears in a document. In other words, $\mathbf{C}$ is a frequency matrix, with each member of the matrix representing the counts in a document of a particular token. Each column is an element of the vocabulary – the set of unique tokens in the whole corpus – and so the number of columns equals the size of the vocabulary. Representing raw text in this fashion is a central part of the *information extraction* problem (Manning & Schütze, 1999, p. 529). Ultimately, we wish to discard all tokens which do not convey information relevant to the latent variable, and quantify the amount of information each token conveys.

The first decision is to define the token. The simplest method is to assign tokens to single words, or 'unigrams'. This method, known as the 'bag-of-words', reduces the dimensionality of the document by the maximum amount by ignoring any dependence between the tokens. However, it simplifies language to a large extent, as (by definition) word order, grammar, ambiguous terms ("hard", "line"), and modifiers ("not") are lost in the mapping. These problems can be moderated by instead considering n consecutive words to be a token: "in the beginning" maps to ("in.the", "the.beginning". Since the words overlap, the size of $\mathbf{C}$ increases exponentially, requiring a corresponding increase in computational power. $\mathbf{C}$ also becomes more sparse, requiring more observations. However, even considering pairs of words yields a significant improvement in extracting meaning (Cheng et al., 2006), as we expand the remit to two-word phrases. As computational power has increased, it is possible to efficiently include a wider context than 2- or 3-grams by using the word embeddings technique.

There are several steps that can reduce the size of the vocabulary – the number of columns of $\mathbf{C}$ – without a great degree of information loss. When the bag of words method is used, $\mathbf{C}$ is a word frequency matrix. It will therefore reflect the structure of English in that each document will have a large frequency of structural words like "from", "the", "could". These are called stop-words, and do not convey meaning; the first step of a text analysis is to remove these words by filtering on a list. There are standard lists, but it is important to take the domain of the corpus into account when filtering stop-words; for example, Twitter has domain-specific stop-words "@", "#", but filtering punctuation risks removing emoticons (":)"). A closely related step is converting all words to lowercase, on the general assumption that word meaning does not depend on whether it starts a sentence (Denny & Spirling, 2018). However, in the context of social media, this may be problematic: emotion and intensity of sentiment is often expressed by various forms of capitalisation. Internet-targeted sentiment analysers usually attempt to take this into account. Words are also expressed in English in different forms ("laughing, laughs"), so grouping words by their stem ("laugh") is also done (Porter, 1980). These filtering steps are very effective in reducing the dimensionality of $\mathbf{C}$, but the problems with word ambiguity and word order cannot be avoided.

Maximising information by stop-word removal is often extended by applying term frequency weighting to the matrix. The linguistic principle is that the amount of information a word conveys about a sentence is the inverse of its frequency in the wider text. Stop-words occur most frequently in language, and so convey the least amount of information about the content of a document; conversely, tokens with a lower frequency will convey more information about the content of a document (for example, place names or precise terminology like 'monopsony'). The most amount of information will be given by terms that appear in a small number of the total set of documents – they have a low document frequency – but are repeated in the document being considered – a high term frequency. Dividing these two frequencies gives the "term frequency - inverse document frequency" metric, which represents the amount of information about the document a given token conveys (Manning et al., 2008, p. 100).

For economists, a vitally important concern in these pre-processing steps is reproducibility. The decisions made during this initial process can significantly affect the final result: the necessity of intensive data transformation combined with the inherent variability of the process leads nearly inevitably to the charge that an analysis could lead to a different result if different preprocessing steps had been taken. This contingency of the analysis upon the data preparation is called the "forking paths" problem (Gelman & Loken, 2014), and with the observational data common in economics not much can be done to alleviate it: replication and pre-registration are often unviable. Aside from fully specifying the details of the data pre-processing, Denny and Spirling (2018) propose calculating the distance between alternative document

[3] It is sometimes called the 'Document Term Matrix (DTM)'

term matrices as an ad-hoc metric.

4.3 Methods of estimating $\hat{\mathbf{V}}$

Once the document term matrix has been created, the final step is to estimate the variable. There are a number of methods; the primary factors in deciding the estimation procedure are the computational power available and the theorised characteristics of the latent variable. A central component, particularly for economists, in choosing the technique is the direction of causality. This can flow either way in a text analysis: the information in the text encoded in $\mathbf{C}$ can cause the variable of interest $\hat{\mathbf{V}}$, but in other cases the text is a function of the latent variable. In other words, approaches might model either $p(\mathbf{v}_i|\mathbf{c}_i)$, or $p(\mathbf{c}_i|\mathbf{v}_i)$; the discussion so far has centred on the latter in the form of latent variable modelling, where an unobserved outcome variable *generates* the text; but text analysis can also work by mapping text frequencies to observed outcomes. To illustrate this, contrast Jegadeesh and Wu (2013), where positive words in company report filings (C cause higher stock market returns V, with Bandiera et al. (2020), where the latent variable 'CEO behaviour' V generates text in their diaries (C. Approaches incorporating $p(\mathbf{v}_i|\mathbf{c}_i)$ as the structural form usually take the form of text regressions; models of $p(\mathbf{c}_i|\mathbf{v}_i)$ are broadly termed generative models. Dictionary-based methods are the simplest, and do not directly involve statistical analysis; we begin with an overview of these.

4.3.1 Dictionary-based methods

The first, and simplest, way to perform inference on a latent variable is to use a simple dictionary matching technique. This is the most well-established method used in the economics and finance literature, largely because it is computationally simple and easy to implement. In a nutshell, the method is to use a known function $f(\cdot)$ to stipulate $\hat{\mathbf{v}}_i = f(\mathbf{c}_i)$. No information is inferred from the language corpus; rather, known attributes of $\mathbf{c}_i$, established by linguistic and domain-specific research, are leveraged. This is the method that I take, which is explained in detail in section 5.1. Beyond computational ease, the an advantage of this technique is that it avoids 'model stacking', which is the case if $\hat{\mathbf{V}}$ is subsequently used in the primary analysis. This approach is particularly useful if the variable of interest has no prior observed outcomes (which excludes supervised learning approaches, for example). However, it accordingly must rely heavily on well-established and often domain-specific information about the mapping $f(\cdot)$. When using a dictionary-based approach, the researcher must justify the relevance of this mapping to her application, as general dictionaries may not transfer well to particular contexts.

4.3.2 Text regressions: regularised OLS under high dimensionality

Text regressions are a popular method of inference in cases where we wish to estimate an outcome based on the content of a text corpus. In a text regression, we predict $\mathbf{v}_i$ from $\mathbf{c}_i$ in the normal way, by using ordinary least squares (Gentzkow et al., 2019, p. 541). Assuming a linear form, we approximate the conditional expectiation with

$$E[\mathbf{V}_i|\mathbf{C}_i = c] = \beta^T c = \sum_{k=1}^{K} \beta_k c_k$$

, and the corresponding estimator is

$$\hat{\beta}^{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} (V_i - \mathbf{C}_i^T \beta)^2 = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T V_i$$

. $\mathbf{C}^T\mathbf{C}$ is an $N \times K$ matrix. As K approaches N, each covariate adds the same amount of variance (σ^2/N) to the estimator, regardless of the true coefficient of the covariate. So as K increases, the precision of the estimator decreases (Davidson & MacKinnon, 2004, p. 101). If we have extremely high dimensionality and $K > N$, as is likely the case when each covariate β_k corresponds to a token in a large text vocabulary, we cannot invert the matrix $\mathbf{C}^T\mathbf{C}$ and obtain the OLS estimate. The general method of addressing this is to introduce a penalty term for each additional covariate; we minimise

$$\hat{\beta}^{POLS} = \underset{\beta}{\operatorname{argmin}} \sum_{n=1}^{N} (V_i - \mathbf{C}_i^T \beta)^2 + \lambda(||\beta||_q)^{\frac{1}{q}}$$

, where $||\beta||_q$ is the L^q distance function[4] (Athey & Imbens, 2019, p. 696). Choices of q correspond to different methods: the LASSO method has $q = 1$, for example (Tibshirani, 1996). The effect of this penalisation is to shrink β_k towards zero with increasing λ, thereby maximising the precision of the estimator. The process for the econometrician is then to choose the value of λ, the 'tuning parameter', according to some measure of model fit such as the Akaike or Schwartz Information Criterion.

4.3.3 Latent Dirichlet Allocation and machine learning methods

Finally, machine learning methods have seen an expansion in the academic economics literature; while I have focussed on more traditional methods, they are rare in industry. Industrial applications of text analysis are usually characterised by access to an extremely large, high-frequency corpus: search data is a salient example. In this case, large language models are developed, often using unsupervised machine learning and related techniques. However, some smaller-scale machine learning methods are feasible and can be applied to text analysis for economic research.

5 Methods

5.1 Text analysis for Twitter data: approach taken

5.1.1 NRC Emotion Lexicon

5.1.2 `szuyhet` R package and workflow

5.1.3 The VADER package

5.2 Econometric Approach

5.2.1 Inference with Panel Fixed Effects

5.2.2 Specifications and hypotheses

6 Results

6.1 Robustness checks

7 Discussion and conclusion

References

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2020, July). *Testing, Voluntary Social Distancing and the Spread of an Infection* (w27483). National Bureau of Economic Research. Cambridge, MA. Retrieved January 20, 2021, from http://www.nber.org/papers/w27483.pdf. (Cit. on p. 3)

Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020, April). *Polarization and Public Health: Partisan Differences in Social Distancing during the Coronavirus Pandemic* (Working Paper No. 26946). National Bureau of Economic Research. Cambridge, MA. (Cit. on p. 3).

Ananyev, M., Poyker, M., & Tian, Y. (2020). The safest time to fly: Pandemic response in the era of Fox News. *Covid Economics*, (49) (cit. on p. 3).

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725 (cit. on p. 9).

[4] $L^q = \sum_{k=1}^{K} |\beta_k|^q$. To illustrate, L^2 is therefore the familiar Euclidean metric and L^1 is the Manhattan (taxicab) metric.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. (2021, January 3). *A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration*. Zenodo. Retrieved January 6, 2021, from https://zenodo.org/record/4414856. (Cit. on p. 4)

Bandiera, O., Prat, A., Hansen, S., & Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, *128*(4), 1325–1369 (cit. on p. 8).

Barrios, J., & Hochberg, Y. (2020, April). *Risk Perception Through the Lens of Politics in the Time of the COVID-19 Pandemic* (Working Paper No. 27008). National Bureau of Economic Research. Cambridge, MA. Retrieved January 8, 2021, from https://www.nber.org/system/files/working_papers/w27008/w27008.pdf. (Cit. on p. 3)

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. Retrieved January 13, 2021, from https://linkinghub.elsevier.com/retrieve/pii/S187775031100007X (cit. on p. 3)

Brotherhood, L., Kircher, P., Santos, C., & Tertilt, M. (2020). *An Economic Model of the COVID-19 Epidemic: The Importance of Testing and Age-Specific Policies* (Discussion Paper No. 13265). IZA Institute of Labor Economics. Bonn. (Cit. on p. 3).

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International journal of corpus linguistics*, *11*(4), 411–433 (cit. on p. 7).

Davidson, R., & MacKinnon, J. G. (2004). *Econometric theory and methods*. Oxford University Press. (Cit. on p. 8).

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189 (cit. on p. 7).

Eichenbaum, M., Rebelo, S., & Trabandt, M. (2020, March). *The Macroeconomics of Epidemics* (w26882). National Bureau of Economic Research. Cambridge, MA. Retrieved January 20, 2021, from http://www.nber.org/papers/w26882.pdf. (Cit. on p. 3)

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*(6), 460–466 (cit. on p. 7).

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. Retrieved January 13, 2021, from https://pubs.aeaweb.org/doi/10.1257/jel.20181020 (cit. on pp. 3, 6–8)

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of financial economics*, *110*(3), 712–729 (cit. on p. 8).

Lamsal, R. (2020). Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*. Retrieved January 13, 2021, from http://link.springer.com/10.1007/s10489-020-02029-z (cit. on p. 4)

Lucca, D. O., & Trebbi, F. (2009, September 17). *Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements* (w15367). National Bureau of Economic Research. Retrieved January 13, 2021, from https://www.nber.org/papers/w15367. (Cit. on p. 3)

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Retrieved February 15, 2021, from http://ebooks.cambridge.org/ref/id/CBO9780511809071. (Cit. on p. 7)

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press. (Cit. on p. 7).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, *29*(3), 436–465. Retrieved January 13, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x (cit. on p. 2)
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00460.x

Painter, M., & Qiu, T. (2020). Political Beliefs affect Compliance with COVID-19 Social Distancing Orders. *SSRN Electronic Journal*. Retrieved January 8, 2021, from https://www.ssrn.com/abstract=3569098 (cit. on p. 3)

Perrin, A., & Anderson, M. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Pew Research Center. Washington, DC. Retrieved January 6, 2021, from https://www.pewresearch.org/wp-content/uploads/2019/04/FT_19.04.10_SocialMedia2019_topline_methodology.pdf. (Cit. on p. 2)

Porter, M. F. (1980). An algorithm for suffix stripping. *Program* (cit. on p. 7).

Qazi, U., Imran, M., & Ofli, F. (2020). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. *ACM SIGSPATIAL Special*, *12*(1), 6–15. Retrieved January 13, 2021, from https://dl.acm.org/doi/10.1145/3404820.3404823 (cit. on pp. 2, 4, 5)

SafeGraph, Inc. (2020). *Social Distancing Metrics*. SafeGraph. Retrieved January 7, 2021, from https://docs.safegraph.com/docs/social-distancing-metrics. (Cit. on pp. 2, 6)

Simonov, A., Sacher, S., Dubé, J.-P., & Biswas, S. (2020). *The Persuasive Effect of Fox News: Non-Compliance with Social Distancing During the COVID-19 Pandemic* (Working Paper No. 27237). National Bureau of Economic Research. Cambridge, MA. Retrieved January 7, 2021, from http://media.mediapost.com.s3.amazonaws.com/uploads/FoxNewsPaper.pdf. (Cit. on p. 3)

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, *118*, 26–40. Retrieved January 12, 2021, from https://linkinghub.elsevier.com/retrieve/pii/S0047272714000929 (cit. on p. 3)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288 (cit. on p. 9).