Predicting COVID-19 outbreaks with Twitter data

Arran J. Davis, Cole B. J. Robertson, James Carney, Seán G. Roberts, and Dermot Lynott

Contents

1 Introduction

Twitter data is increasingly being used to predict public health outcomes. Previous has shown that mentions of illnesses and illness symptoms by Twitter users can be used to predict future infection rates during epidemics and pandemics (Lopreite, Panzarasa, Puliga, & Riccaboni, 2021). Other research has used Twitter data to estimate regional variations in use of future-referring grammatical markers (e.g. *will*, *could* or *might*), showing that these regional variations predict risk-taking behaviours and disease outcomes (Ireland, Schwartz, Chen, Ungar, & Albarracín, 2015).The studies presented here work to extend this research in several key ways.

1.1 Study 1: Predicting COVID-19 outbreaks using a 'symptom talk' classifier on tweets from the UK in early 2020

Study 1 will build a "symptom talk" machine learning classifier that can predict the likelihood that a tweet is about a user's illness symptoms. This study extend the results of Lopreite et al. (2021), who showed that the frequency of tweets about 'pneumonia' and 'dry cough' were significantly higher across Europe in the months before these symptoms were declared as associated with COVID-19 by the WHO (as compared to the same months in previous years). They also found that changes in tweet frequency related to 'pneumonia' and 'dry cough' were greatest in areas that were initially the hardest hit by the pandemic (e.g., Lombardy, Italy). Study 1 will move past simple key word searches for symptoms related to COVID-19 (e.g., 'dry cough') through building a symptom talk classifier that can identify general illness symptoms, thereby making it of possible use in monitoring the spread of diseases with yet unknown symptom profiles.

1.1.1 Replicating the datasets used by Lopreite et al. (2021)

Lopreite et al. (2021) created an initial database containing all tweets containing the keyword "pneumonia" in the seven most spoken languages in the European Union (i.e., English, German, French, Italian, Spanish,

Polish, and Dutch; (Ginsburgh, Moreno-Ternero, & Weber, 2017)) and posted on Twitter form 1 December 2014 to 1 March 2020.

This dataset was then subsetted substantially: "We then identified the users that cited pneumonia in the selected European countries between 15 December 2019 and 21 January 2020, and compared them with the total number of users that cited pneumonia in the same weeks of the previous year." (p. 5). The date of 21 January 2020 was chosen as the end date because it is the day on which the COVID-19 became a "Class B notifiable disease"; World Health Organization (2020b)]. Tweets from the period between 15 December 2019 and 21 January 2020 were then compared to tweets from the same days from the previous year (and, as a robustness check, to the same days from all previous years since 2014).

Regarding geographic subsetting, the authors used only tweets posted by users in the countries where the main language is one of the seven most-spoken languages in the European Uniion (i.e., the United Kingdom, Germany, France, Italy, Spain, Poland, and the Netherlands). How the authors determined Twitter users' geographic locations is not entirely clear (leaving open the geolocation method for the current study), however, once users' geolocations were determined, they were assigned to NUTS1 European regions (p. 5).

1.2 Study 2: Predicting COVID-19 outbreak "hot spots" using regional variations in Twitter users' time perspectives

This study woud extend previous research linking use of future-referring grammatical markers with negative health outcomes related to risky behaviours. For example, Ireland et al. (2015) showed that use of use of future-referring grammatical markers negatively predicted HIV rates across US counties, and that county-level variation in future-referring grammatical markers was related to mentions of 'risky leisure activities' (e.g., 'hitting the bong'). The authors suggest a positie link between propensity to engage in risky behaviours and HIV infection. COVID-19 infection is also related to risk taking, with, for example, mask-wearing being negatively related to COVID-19 infection (Armstrong-Mensah, Tetteh, & Tetteh, 2021).

Study 2 will extend this research by (a) using more appropriate methods for measuring Twitter users' future-referring grammatical markers and (b) testing whether previously published future-referring grammatical marker effects held in a new country and with a new outcome measure that is also related to risky behaviour (i.e., COVID-19 infection rates in the UK).

1.3 Study hypotheses

1.3.1 Study 1: Predicting COVID-19 outbreaks using a 'symptom talk' classifier on tweets from the UK in early 2020

This study would test two hypotheses related to changes in the frequency of tweets about illness symptoms (identified by a 'symptom talk' machine learning classifier).

1.3.1.1 The Increased Symptom Talk Hypothesis

The frequency of tweets about illness symptoms is higher from 1 December 2019 to 21 January 2020 than it is over the same date span in previous years. COVID-19 became a "Class B notifiable disease" (World Health Organization, 2020b) on 21 January 2020; tweets after this date should be excluded from this analyses, as there "would be no obvious way to disambiguate messages [tweets] concerned with genuine local cases [here, symptom talk] ... from messages elicited by mass media coverage of the outbreak" (Lopreite et al., 2021, p. 2). Thus, the date span of 1 December 2019 to 21 January 2020 was chosen as the date span to compare to previous years.

Difference tests could be as simple as t-tests, although they should probably use multilevel regression models that account for regional variations in symptom talk more generally, or methods similar to those of Lopreite et al. (2021). These analyses could also be broken down into European NUTS1 regions, or by country,

as done by Lopreite et al. (2021). Comparisons should also look at *number of users* and not raw tweet frequencies, following the methods of (Lopreite et al., 2021).

Data: Tweet symptom talk frequencies from 15 December 2019 to 21 January 2020 would be compared to tweet symptom talk frequencies from the same weeks in the three previous previous years. The comparisons could be made using tweets from the countries where the main language is one of the seven most-spoken languages in the European Uniion (i.e., the United Kingdom, Germany, France, Italy, Spain, Poland, and the Netherlands); English (the most common) and native language (i.e., German in Germany) tweets from each country would be used. Thus, English and native language tweets from 15 December 2019 to 21 January 2020 in each country would be compared to English and native language tweets from the same date span in preceeding years. The actual date scrapes should be for the weeks between 1 Decemeber and 1 March, for each year; the Lopreite et al. (2021) dataset begins 1 December 2014. However, given differences in tweet volumes (Lopreite et al. (2021) only collected tweets mentioning 'pnemonia' whereas the current study would collect all tweets), the current study will collect tweets from all relevant countries beginning on 1 December 2016 (i.e., 1 December 2016 to 1 March 2017, etc.).

1.3.1.2 The Regional Variation in Symptom Talk Hypothesis

European locations that were particularly hard hit by COVID-19 during the first wave of the pandemic would show greater pre-pandemic to early pandemic changes in the frequency of tweets about illness symptoms. These analyses could use a form of regression to test whether changes in the frequency of tweets about illness symptoms predicted COVID-19 outcomes at the regional level. Thus, change in symptom talk frequencies (pre-pandemic years to pandemic years) for each region would be the predictor variable, and the outcome would be the COVID-19 cases or hospitalisations.

Data: Tweet symptom talk frequencies from the influenza seasons of previous years (1 December to 21 January) would be compared to the same time period at the beginning of the pandemic. This could be done for each country proposed above.

1.3.1.3 The Temporal Variation in Symptom Talk Hypothesis

Tweet symptom talk frequencies can be used to predict COVID-19 case frequencies over time and by region This could be a form of multilevel regression analysis that uses the frequency (at the regional level) of symptom talk in tweets at time T_0 to predict COVID-19 case loads or hospitalisation rates (at the regional level) at time T_0 + ~14 days.

Data: Tweet symptom talk frequencies from two months before the first confirmed COVID-19 case in the UK (31 January 2020) to a year after the WHO officially declared a pandemic on 11 March 2020 (i.e., 1 December 2019 - 1 April 2021). Although 11 March 2021 marked a year since the WHO officially declared a pandemic, 1 April 2021 would likely be a better end date for these analyses,

1.3.1.4 Considerations

- Sensorimotor norms will be used to give descriptive statistics for tweets identified as 'symptom talk' - this will be done to give readers a better understanding of the types of language being classified by the symptom talk classifier. Sensorimotor norms could be compared across time to identify changes in the types of symptoms that are being tweeted about.
- The symptom talk classider should be trained on *pre-pandemic* tweets.
- The frequency of illness talk should probably be at the individual level; the percentage of individuals using 'symptom talk' for a given time period (although overall frequency - ignoring users - may better estimate the severity of symptoms).
- General trends in symptom talk from flu season to flu season should also be examined to test whether changes in symptom talk are a result of broader historical trends (it could be that as time goes on people are generally more likely to use symptom talk on Twitter).

- Lopreite et al. (2021) cut-off their tweet sample on 21 January 2020, when the WHO announced COVID-19 and its symptoms, because there "would be no obvious way to disambiguate messages concerned with genuine local cases of pneumonia from messages elicited by mass media coverage of the outbreak" (p. 2). The symptom talk classifier somewhat avoids this problem, since it would be trained to identify when people are talking about any kind of symptom, but it is unclear to me what the correct methodology with regard to the date ranges. On the one hand, we would ideally use tweets from before media coverage affected peoples' behaviour and tweeting (i.e., those from January 2020 and earlier). On the other hand, focusing on tweets from January 2020 and earlier gives very little outcome data (COVID-19 infections) to work with, since cases in the UK are less than 10/day for all of February 2020.

1.3.2 Study 2: Predicting COVID-19 outbreak "hot spots" using regional variations in Twitter users' time perspectives

This study would test a hypothesis related to the relationship between time perspectives (i.e., future orientation) and COVID-19 outcomes; future-orientation has been shown to predict a range of outcomes related to risky behaviours (e.g., Ireland et al., 2015).

1.3.2.1 The Future-(un)Certainty Hypothesis

UK regions which use a higher proportion of low-certainty future-shifting modal constructions (e.g. *could*, *may*, *might*, *should*, *possibly*, *probably*, etc.) will exhibit higher rates of COVID-19 outbreaks. Regions which use a higher proportion of high-certainty future-referring terms (e.g. *certainly*, *definitely*, *will*, *be going to*, *shall*, etc.) will exhibit lower rates of COVID-19 infection. We refer to this hypothesis as the future-(un)certainty hypothesis.

Data: Time perspectives will be assessed using the tweets from the three previous pre-pandemic flu seasons (i.e., the data made available from Study 1); when future orientation is measured should not exactly matter, as long as it is not *during* the pandemic.

1.3.2.2 Considerations

- It could also be interesting to see whether the linguistic future time reference patterns of tweets changed from before to during the pandemic (potentially as a result of stress levels), although this is may be another study (which could use the data collected for both Study 1 and Study 2.

2 Methods

2.1 Tweet collection

The Twitter API (v2) for academic research was used to collect tweets for both studies. For all time periods in both studies, tweets were collected from the entirity of each country using the `place_country` field in the Twitter API query (Twitter, 2021). For example, the UK country code (`GB`) was used in the `place_country` field to collect tweets from the UK; this returns all tweets with meta-data about tweet-specific locations in the UK (i.e., Twitter "Places" that are in the UK and specific GPS latitude/longitude point coordinates that can be associated with a Twitter Place that is in the UK). The same procedure was used for the remaining countries (i.e., Germany, France, Italy, Spain, Poland, and the Netherlands).

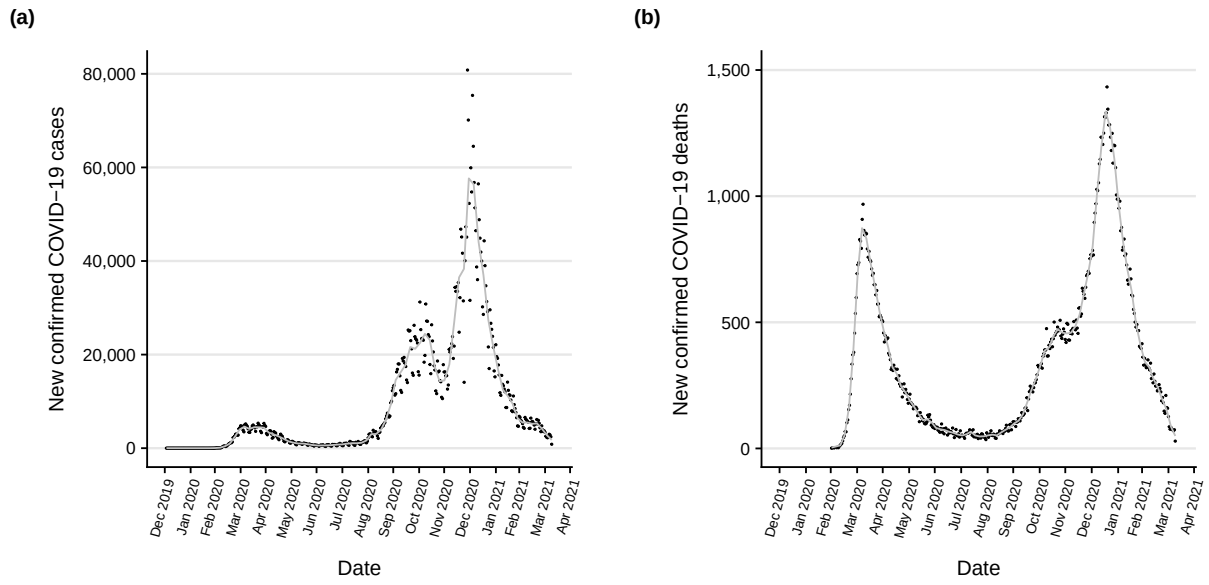

Figure 1: (a) Confirmed COVID-19 cases and (b) deaths attributed to COVID-19 in the United Kingdom from 1 December 2019 to one year after the WHO declared COVID-19 a pandemic (11 March 2021).

2.1.1 Tweet collection dates

Tweet collection dates were meant to capture the 'three waves' of the COVID-19 pandemic in the United Kingdom (see Figure 1), as documented by Public Health England's data on COVID-19 cases and deaths (Public Health England, 2021), as well as the final 'pre-pandemic' flu season spanning 1 December 2019 to 21 January 2020 (the day on which the COVID-19 became a "Class B notifiable disease"; World Health Organization, 2020b). Tweet collection dates also included the three preceding (i.e., pre-pandemic) flu seasons.

Flu seasons were defined as lasting from 1 December in year `x` to 1 April in year `x + 1`. Data suggest that case loads for influenza and other respiratory viruses generally peak in the UK during these four months (Public Health England, 2019). Using previous flu seasons as a comparison is crucial to analyses involving the symptom talk classifier, as this classifier was also likely measure influenza-related symptom talk; using pre-pandemic flu seasons as a benchmark helps to contorl for increases in symptom talk for reasons other than COVID-19 (e.g., influenza).

All tweets sent from the UK were thus collected for four seperate time periods:

- **1 November 2019 to 1 April 2021:** This time period encapsulates the first documented case of COVID-19 (thought to be in early November 2019) and the first year of the pandemic (the WHO officially declared a pandemic on 11 March 2020; World Health Organization, 2020a). The first case in the UK was recorded on 31 January 2020, and new cases exceed 1,000/day by the end of March 2020, and then peak (first wave) at over 5,000/day by mid April 2020. The second and third waves in the UK peak in November 2020 and around 1 January 2021, respectively.
- **1 December 2018 to 1 April 2019:** The most recent pre-pandemic flu season.
- **1 December 2017 to 1 April 2018:** The second most recent pre-pandemic flu season.
- **1 December 2016 to 1 April 2017:** The third most recent pre-pandemic flu season.

Given that tweet collection from countries other than the UK was done in an attempt to replicate and advance the findings of (Lopreite et al., 2021), the data spans for tweets from Germany, France, Italy, Spain, Poland, and the Netherlands differed from that of the UK. Lopreite et al. (2021) focused on the 1 December 2019 to 21 January 2020 date span. The authors explained that avoiding focusing analyses on tweets created after 21 January 2020 would remove any confounds created by media coverage of COVID-19 (which became a "Class B notifiable disease" on 21 January 2020; World Health Organization, 2020b). However, they did collect data on the 'winter seasons' more generally, which according to the published plots, encompassed mid-November until 1 March.

Thus, this dates span (15 November to 1 March) was used to collect tweets from Germany, France, Italy, Spain, Poland, and the Netherlands for the following dates:

- **15 November 2018 to 1 March 2019:** The most recent pre-pandemic 'winter season'.
- **15 November 2017 to 1 March 2018:** The most recent pre-pandemic 'winter season'.
- **15 November 2016 to 1 March 2017:** The most recent pre-pandemic 'winter season'.

2.1.2 Additional dataset reduction

Data from the tweet collection dates and locations described above were then subsetted based on the following rules:

- Only tweets in English and in the native language of the country of origin were used (e.g., English and Dutch in the Netherlands, and only English in the United Kingdom).
- All tweets (and corresponding users) that cited news with a direct URL were removed [Lopreite2021].
- All tweets from users with over 2,000 users (this seems arbitrarily defined) were removed to eliminate the effects of "press agencies and celebreties" (Lopreite et al., 2021, p. 5).
- All tweets that contained either "Coronavirus", "COVID", "COVID-19", or "China" were removed.

3 References

Armstrong-Mensah, E., Tetteh, A. K., & Tetteh, G. R. (2021). COVID-19 pandemic: Face mask mandates, hospitalization, and infection rates in the United States. *International Journal of Translational Medical Research and Public Health*, *5*(2), 113–124.

Ginsburgh, V., Moreno-Ternero, J. D., & Weber, S. (2017). Ranking languages in the European Union: Before and after Brexit [Journal Article]. *European Economic Review*, *93*, 139–151.

Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L. H., & Albarracín, D. (2015). Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, *34*(S), 1252.

Lopreite, M., Panzarasa, P., Puliga, M., & Riccaboni, M. (2021). Early warnings of COVID-19 outbreaks across europe from social media. *Scientific Reports*, *11*(1), 1–7.

Public Health England. (2019). Surveillance of influenza and other respiratory viruses in the UK: Winter 2018 to 2019. *Annual Flu Reports*.

Public Health England. (2021). *Coronavirus (COVID-19) in the UK*. Retrieved from https://api.coronavirus.data.gov.uk/v1/data

Twitter. (2021). *Filtering tweets by location*. Retrieved from https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location

World Health Organization. (2020a). *Coronavirus disease 2019 (COVID-19) Situation Report – 51.*

World Health Organization. (2020b). *Report of the WHO-China Joint Mission on Coronavirus Disease 2019.*