# Methodological Appendix for Blog Post

## The YouGov Data

We borrow data from a YouGov survey, which was in the field between September 16 and October 20 (N = 5945). In this survey, we asked respondents which of the following issues would be important in deciding how they would vote in the 2020 US presidential election: the economy, health care, immigration, abortion, terrorism, the federal budget deficit, medicare, the rights of the LGBTQ+ population, the environment, taxes, law and order, the rights of racial and ethnic minorities, gun policy, social security, trade policy, drug use/opioid addiction, and the COVID-19/Coronavirus pandemic. In a second question, we asked which of these issues would be the most important.

## The Twitter Data

Our analysis and collection of the Twitter data was guided by three dictionaries of key-terms, categorised by issue and sub-issue. We refer to these dictionaries as data dictionaries. Each data dictionary includes a unique collection of terms.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | The Issue Data Dictionary | Contains terms relating to policy issues | 958 terms in total |
| 2 | The Election Data Dictionary | Contains terms relating to the election | 276 terms in total |
| 3 | The Hashtag Data Dictionary | Contains election, candidate and party-related hashtags which appeared on Twitter between October 6 and November 10, 2020 | 598 terms in total |

The terms in these dictionaries are first categorised by theme or actor, and then by date. This is the date when the term was added to the list. The search terms in these data dictionaries were used to scrape Twitter for relevant Tweets, between October 6 and November 10, 2020. We do not include all word forms in these dictionaries. For example, when we search for the term ‘tax’, we also automatically catch words such as ‘tax cut’ and ‘tax bill’. In the below table, we list the categories included in these dictionaries.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Meta-category | Sub-category |
| 1 | The Issue Data Dictionary | Economy; Housing; Energy costs; Conspiracy; Education; Defense; Healthcare; Immigration; Abortion; Terrorism; International issues (general); International security; international issues(specific); international security; international issues; LGBT rights; Rights of racial and ethnic minorities; Women’s rights; Environment; Law and Order; Social security; Trade; Drugs/Opioid crisis; COVID-19; election integrity; postal voting | Economy; Austerity; Budget Deficit; Cost of Living; Employment; Unemployment; GDP; Tax; Interest Rates; Recession; Businesses; Consumer Debt; Banks; Finance; Poverty; Housing; Energy Costs; Conspiracy; Schools; University; Military; Medicare; Healthcare; Immigration; Abortion; Terrorism; ISIS; Far right extremism; Far left extremism; International issues; International security; Ebola; Afghanistan; Russia; Syria; Ukraine; China; North Korea; EU; Iran; Middle East; Israel; Iraq; Somalia; Morocco; LGBT; Racism; Indigenous rights; Sexism; Environment; Crime; Courts; Supreme Court Nomination; Police; Fraud; Prison; Gangs; Gun control; Social security; Pension; Trade; Drugs; COVID-19; Election integrity; Postal voting |
| 2 | The Election Data Dictionary | Election; Political actors; Parties | Election general; Election outcome; Election campaign; Election action; Electoral process; Debate; VP Debate; Debate moderator; President; Donald Trump; Mike Pence; Melania Trump; Trump Family; Joe Biden; Jill Biden; Hunter Biden; Jo Jorgensen; Spike Cohen; Howie Hawkins; Angela Walker; Kamala Harris; Douglas Emhoff; Greens; Democratic Party; Leading Democrats; Libertarian; Republican Party; Leading Republicans; Republican staff or allies; Democrat staff or allies |
| 3 | The Hashtag Data Dictionary | Election; Election action; Political actors; Parties | Election general; Election action; Electoral process; Vote counting; Post-election; Debate; VP Debate; Election integrity; Supreme Court Nomination; President; Vice-President; Donald Trump; Mike Pence; Melania Trump; Trump Family; Joe Biden; Jill Biden; Hunter Biden; Jo Jorgensen; Spike Cohen; Howie Hawkins; Angela N. Walker; Kamala Harris; Douglas Emhoff; Democratic Party; Leading Democrats; Libertarian; Republican party; Leading Republicans; Greens |

These data dictionaries were added to iteratively, on Thursday or Friday for every week during this period (between October 6 and November 10), to include new terms which reference the policy issues, aspects of the election and hashtags and otherwise would have been missed. No terms were added to this dictionary beyond 11 December, three days before the confirmation of Joe Biden’s victory on December 14, 2020. We gathered data daily using the Twitter API, and we collected 17,060,042 tweets in total.

## Our Measures (Methodological specifics)

Using the Twitter Search API, we attempted to collect a maximum of 500 tweets for each of the terms. This was done on a daily period across the campaign timeframe (6th October – 10th November). The Search API moves backwards in time for up to seven days. In most cases these tweets were found within the first few hours if not minutes of the search timeline. We used the creation times of the first and last collected Tweet for each term to calculate a Tweet rate for each term in Tweets/minute. The higher the rate the more popular the term. In the final analysis we aggregated the rates across the subcategories and only used the top 10 terms from each subcategory. This eliminated the terms where we were unable to collect 500 tweets and consequently would have been subject to collecting the same tweet across multiple days.

In an attempt to restrict the Tweets to those originating in the U.S we used the Twitter API approach of providing a centre point and a radius to encompass the whole of the US. We did not restrict on language, nor were we able to check the actual country of origin. In most cases this did not cause a problem except for the search term ‘un’ (United Nations). Here we found that almost all the tweets collected were in Spanish including the Spanish word ‘un’. We disregarded all these Tweets.

The Tweets collection took place daily at midnight (GMT). This equates to 4pm PST and 7pm EST. The collection process typically took 5-6 hours. On 24th October (so collecting Tweets from 23rd) the automatic process failed and was run manually at approx. 9am GMT. The effect of this was the collection times in the US were largely overnight or the early hours of the morning. The result of this was that although we could still collect the tweets the Tweet rate was noticeably smaller across the board for this date.

The table below indicates shows how many Tweets were collected per day.

|  |  |
| --- | --- |
| **Date** | **Tweets collected** |
| 06/10/2020 | 468034 |
| 07/10/2020 | 467450 |
| 08/10/2020 | 467486 |
| 09/10/2020 | 465914 |
| 10/10/2020 | 520776 |
| 11/10/2020 | 520552 |
| 12/10/2020 | 520580 |
| 13/10/2020 | 520506 |
| 14/10/2020 | 520336 |
| 15/10/2020 | 520386 |
| 16/10/2020 | 519318 |
| 17/10/2020 | 438308 |
| 18/10/2020 | 433142 |
| 19/10/2020 | 432235 |
| 20/10/2020 | 434457 |
| 21/10/2020 | 435325 |
| 22/10/2020 | 435299 |
| 23/10/2020 | 431893 |
| 24/10/2020 | 457580 |
| 25/10/2020 | 457552 |
| 26/10/2020 | 455507 |
| 27/10/2020 | 456215 |
| 28/10/2020 | 456889 |
| 29/10/2020 | 452075 |
| 30/10/2020 | 451729 |
| 31/10/2020 | 480990 |
| 01/11/2020 | 480142 |
| 02/11/2020 | 474650 |
| 03/11/2020 | 473931 |
| 04/11/2020 | 470342 |
| 05/11/2020 | 470494 |
| 06/11/2020 | 464634 |
| 07/11/2020 | 509839 |
| 08/11/2020 | 505660 |
| 09/11/2020 | 496887 |
| 10/11/2020 | 492929 |
| **Total** | **17060042** |

## Contact

To obtain the data dictionaries as CSVs and documentation, please contact the DiCED team.