Felipe Alamos
Bhargavi Ganesh
12/11/19

Is Twitter a Proxy for Public Opinion?

Comparing tweets to survey results, a case study on environmental issues.

Abstract

In this paper, we consider whether Twitter is a reliable proxy for public opinion. To do so, we conduct topic modelling and sentiment analysis on Twitter data and compare the results to responses to traditional surveys. We make an implicit assumptions that surveys are an accurate representation of public opinion. Our topic modelling results show that a surprisingly large number of topics that emerged from topic modeling on Twitter data overlapped with the topics that had the highest responses in the survey. The sentiment analysis revealed that the sentiment of tweets was on average concentrated in extremes (either positive or negative), whereas the survey results were more moderate-negative. In conclusion, based on our results, we establish that Twitter can be an adequate proxy for public opinion when the goal is to understand which topics people care more about, but it is less accurate at representing the sentiments of the masses with respect to certain topic.

Introduction

As use of social media platforms continues to become more widespread, social scientists have started analyzing these platforms' data in order to ascertain human behaviors and trends. An increasing number of journal articles contain analysis of social media data, commonly using methods like topic modeling and sentiment analysis to understand public opinion on a given topic. The most commonly used platform for this type of analysis is Twitter. The widespread use of Twitter across the world, and the fact that the platform is mostly text-based, makes it a strong candidate for natural language processing techniques. As Steinhert-Threkheld notes in his book [1], *Twitter as Data*, papers using Twitter data started appearing after 2010, and span a range of topics. For example, Gohil, et. al published a study which surveyed the various different methods used to analyze health care sentiment on Twitter[2]. Other papers have analyzed the validity of using Twitter sentiment to predict outcomes, such as the results of an election[3], or fluctuations in the stock market[4].

The growing use of Twitter data in research papers begs the question of whether Twitter data is actually an accurate measure of public opinion. There is reason to believe that Twitter data is indeed demographically skewed on a few dimensions. A survey[5] on U.S. Twitter users conducted by the Pew Research Center suggested that U.S. Twitter users tended to be younger, and have higher levels of household income and educational attainment relative to the

[1] Zachary C. Steinert-Threlkeld. 'Twitter as Data', 2018.
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5938573/
[3] https://ieeexplore.ieee.org/abstract/document/6355554
[4] https://ieeexplore.ieee.org/abstract/document/7955659
[5] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

general adult population. Additionally, Twitter tends to be dominated by a smaller group of impassioned users. The Pew study notes that approximately 10% of Twitter users in their sample were responsible for 80% of the tweets produced. These findings suggest that Twitter may not always be an accurate proxy of public opinion.

In this paper, we validate Twitter data using traditional survey results on the topics of the environment and climate change. Traditional surveys typically have rigorous survey sampling techniques which ensure that the data is a reflection of the population at large. Therefore, we make the assumption that survey data is a proper source to use to validate our results. First, we conduct topic modeling on the Twitter data to assess whether the main topics that emerge overlap with the top survey responses from a global survey on the environment. Second, we conduct sentiment analysis on tweets related to climate change and compare the results to a survey question on climate change.

Data

For both the Twitter data and survey data, we conducted topic modelling for the United States and England. We chose these countries because we are dealing with text data, and thought that it would be helpful to pick countries that predominantly speak English. In addition, both the US and England have been big players in global discussions around the environment and climate change, so a good amount of tweets on these topics were expected. For the sentiment analysis section, we only conducted our analysis on the U.S.

Twitter Data

The Twitter data for this report was collected using the python library GetOldTweets3. The Python library scrapes Twitter's advanced search mechanism. The GetOldTweets library allowed us to specify the keywords, timeframe, country, and radius we would like to search within[6]. We chose to use this Python library because the official free Twitter API only allows developers to download tweets that were posted 5-7 days prior to accessing the API.

We searched for tweets that contained the hashtags #environment and #climatechange for the topic modelling and sentiment analysis sections, respectively. We initially downloaded tweets from 2009 in order to match up the timeframe of the tweets with the date range in which the survey was conducted. However, the sample size for tweets containing the desired hashtags was relatively low for 2009, consisting of just 60 tweets. To increase our sample size, we extended the search timeframe from 2009-2013, ultimately capturing 2,000 tweets for each hashtag. The results in this report are specific to survey results and tweets from the U.S. Results related to England are presented in Appendix 2.

[6] Scraping Twitter data with this kind of queries is legal according to Twitter terms of use.

Data Pre-Processing

Before creating topic models and conducting sentiment analysis, we cleaned the tweets using the process described below.

In the first stage, we cleaned the tweets using traditional cleaning methods: transforming words to lowercase, removing stop words and numbers, removing punctuation, and removing whitespaces.

Next, we removed the keywords from our collected tweets, in order to focus on the subtopics that emerged. In addition, we removed retweets and repeated tweets. We also had to remove usernames that were showing up in the body of tweets.

Survey Data

For the survey data, we chose to use the 2010 International Social Survey Programme's survey module[7] on the environment. The survey contains 50,437 observations, across 36 countries.

The survey asks a series of detailed questions on attitudes towards environmental issues. In particular, we focused on three questions in the survey. Two of the questions are presented below.

1) *Here is a list of some different environmental problems. Which problem, if any, do you think is the most important for your country as a whole?*
2) *Which problem, if any, affects you and your family the most?*

The survey respondent is asked to choose from one of the following responses:

- *Air pollution*
- *Chemicals and pesticides*
- *Water shortage*
- *Water pollution*
- *Nuclear waste*
- *Domestic waste disposal*
- *Climate change*
- *Genetically-modified foods*
- *Other*
- *None of these*

[7] https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5500&db=e&doi=10.4232/1.11418

- *Can’t choose*
- *No answer*

Additionally, the survey contains the following question on climate change:

In general, do you think that a rise in the world's temperature caused by climate change is (Please tick one box only.)

- *Extremely dangerous for the environment*
- *Very dangerous*
- *Somewhat dangerous*
- *Not very dangerous*
- *Not dangerous at all for the environment*
- *Can't choose*
- *No answer*

In the next section, we report the topic modeling results for Twitter data, and compare the topics that emerge with the most frequent survey responses. We follow this up by reporting the sentiment analysis results, and comparing these to the survey responses on the question regarding climate change.

Analysis and Results

Comparison of Twitter Topic Modelling to Survey Responses

We conducted topic modelling on tweets related to the environment within the U.S. The topic modeling technique we chose was Latent Dirichlet Allocation, also known as LDA. LDA is a specific type of probabilistic topic model controlling the assignment of words to topics. It discovers the latent topics in a text, and clusters the text around these topics. We chose this method because it was commonly used in previous literature related to Twitter data. Since our goal is to explore whether Twitter is a reasonable proxy for public opinion in social science research, we decided to use and validate the methods commonly used by researchers.

Exploration of term frequencies

Before conducting the topic modelling analysis, we did a short analysis on term frequencies related to the tweets that contained the keyword of “ #environment”.

Below, we present a wordcloud representation and a table of results.

Table 1

Term	Frequency
Green	79
Water	74
Day	55
Great	47
Health	45
Climatechange	41
Energy	41

Table 1 shows the words that appear the most frequently across all of our tweets. Already, we can see some overlap between the most frequent terms used on Twitter and the most frequent survey responses. Particularly worth noting are the terms climate change, green, energy, and water. To understand what the latent topics are, we conducted a topic modelling analysis below.

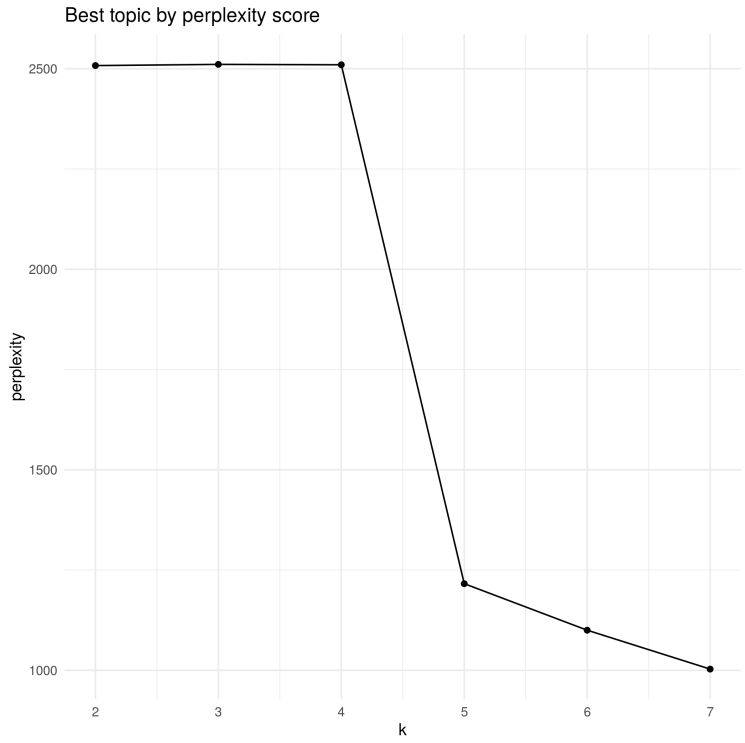
Optimal number of topics

The first part of our topic modelling analysis involved figuring out an appropriate number of topics to choose. First, we created a document-term matrix with our pre-processed words, which consists of tweets in rows and terms in columns, where each cell shows a term from each tweet.

To decide on the optimal amount of topics to choose for our data, we analyzed both perplexity and cohesion scores. The perplexity score is a measure of how well a model predicts a sample. The lower the perplexity score, the better a model predicts a sample. We know that the perplexity score is always decreasing in the amount of topics (k). That said, we had to choose a k with low perplexity without sacrificing model clarity, in the sense that if we have too many topics, it could become very hard to extract clear distinctions between them. We calculated perplexity and cohesion scores for models containing a number of topics ranging from 2 to 15 topics.

We present the perplexity score for the first 7 topics in Figure 1 below.

Figure 1

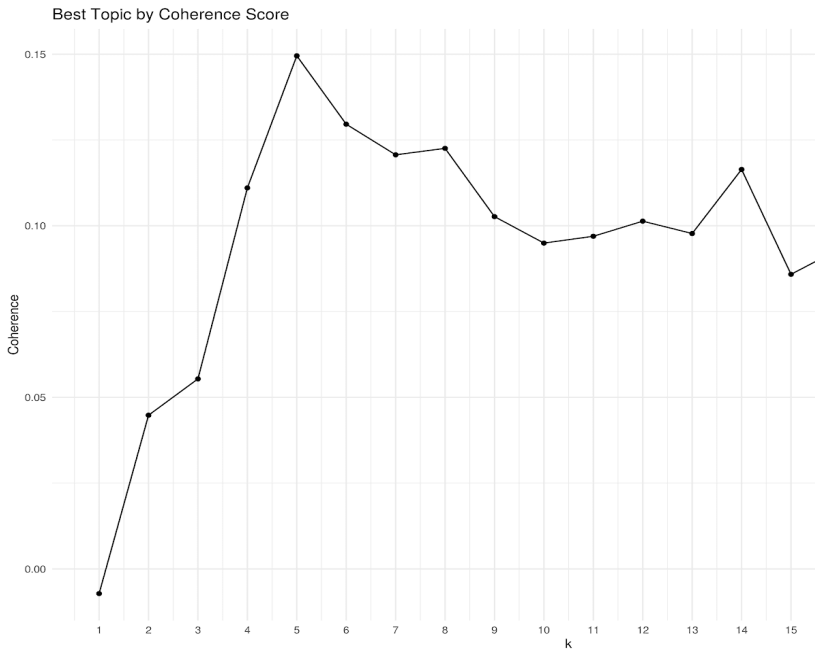


k	perplexity
2	2508
3	2511
4	2510
5	1216
6	1100
7	1003

We can observe that when moving from k=4 to k=5, there is a substantial decrease in the perplexity value. After that, perplexity decreases at a more regular rate. Next, we computed the coherence score for each model. The coherence score is a probabilistic measure of how well terms within a topic fit together. In other words, the coherence score measures the quality of the topics being produced. Overall, the higher the coherence score, the better.

As Figure 2 shows, the best coherence score is also achieved at k=5.

Figure 2

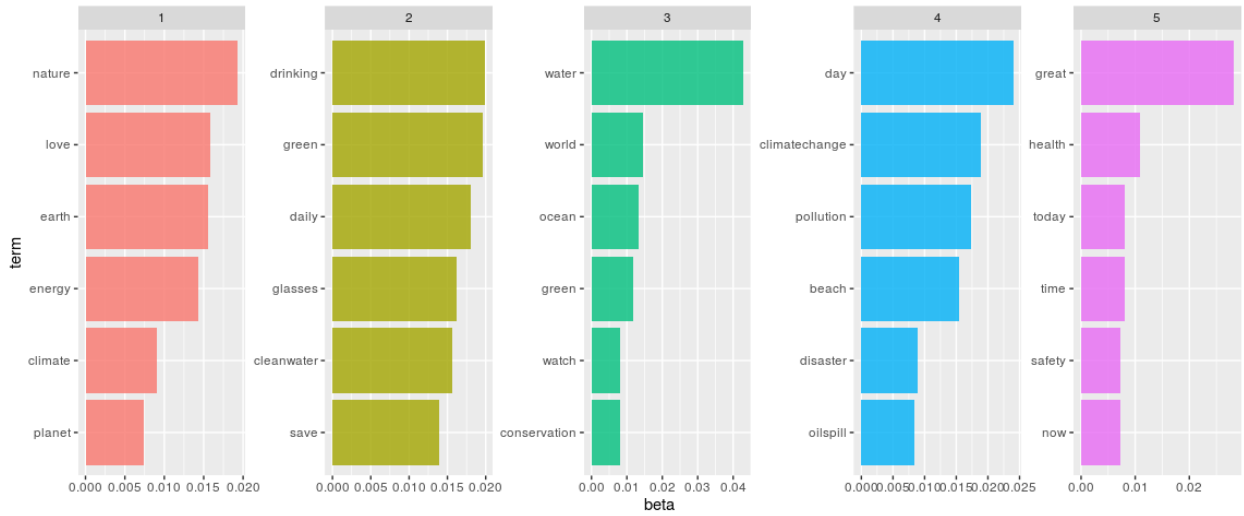


As a result of both the perplexity and coherence score analysis, we decided that the best number of topics for our analysis was 5.

Analysis of topics

After training our LDA model at k=5, we obtained the following topics. A visual representation of the topics that emerged can be seen in Figure 3 below. The number in the gray box above each graph represents the topic number. For each topic, we presented the six words most likely to appear in the topic. This likelihood was based on beta values, which tell us the probability of a given term appearing in a topic.

Figure 3



By visual inspection, topics 1-3 appear to be focused on the environment in a global/country sense, and the positive actions one could take to protect or save the environment. The word “glasses” seems to be the biggest outlier as far as what we would expect to see for these topics. Topic 4 appears to be focused on specific negative climate-related events such as disasters and oil spills. Topic 5 appears to be focused on health and safety. The presence of words such as “now” and “today” suggests an emphasis on immediate and tangible day-to-day concerns related to the environment.

Comparing Twitter topics and survey responses

We compared our topic modelling results for the #environment tweets with responses to a survey question related to the environment, highlighted in our Data section. The question from the survey is presented below:

Here is a list of some different environmental problems. Which problem, if any, do you think is the most important for your country as a whole?

To formalize our comparison of Twitter topics (topics resulting from topic modeling on Twitter data) and survey topics (answers to the survey question), we created the following similarity rule. We first looked at the survey topics that represented more than half of the survey responses. We will call them ‘popular survey topics’. We defined a similarity rule in the following way:

- If less than 50% of the popular survey topics show up in the Twitter topics, then we say that the Twitter results and survey results are **not similar**. More precisely, we say that a survey topic is counted as showing up in a Twitter topic if the survey topic (or a synonym) appears in the top 6 terms (ranked by betas) present in the Twitter topic.
- If 50-70% of the popular survey topics show up in the top 6 terms in any of the Twitter topics, then we say that the Twitter results and survey results are **moderately similar**.
- If more than 70% of the popular survey topics show up in the top 6 terms, we say that the Twitter results and survey results are **very similar**.

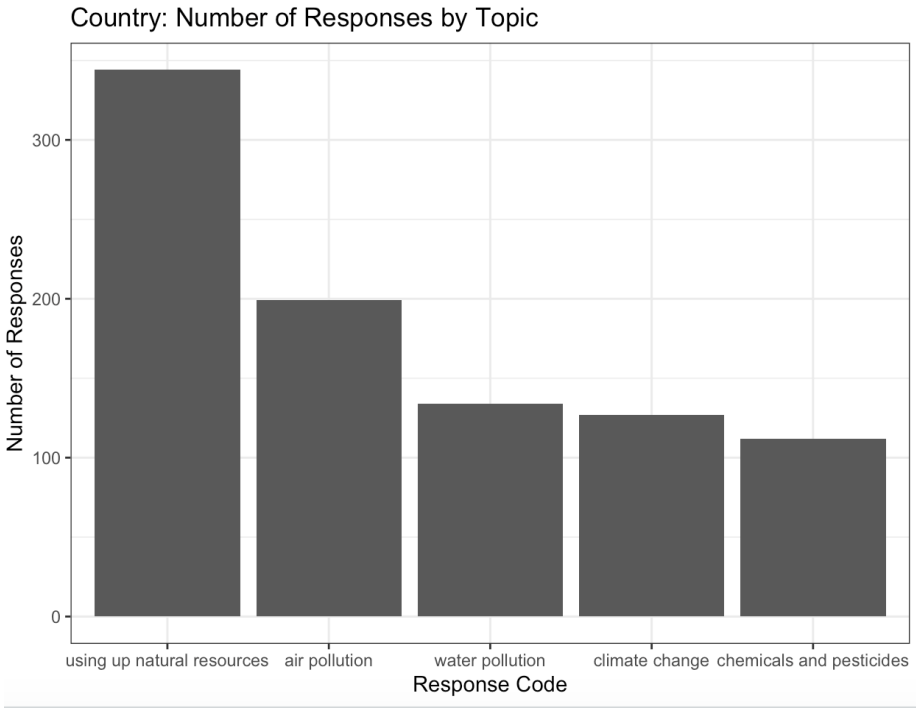
This methodology accounts for the fact that there is a wider range of possible Twitter responses than the pre-determined survey answer choices. Additionally, we are only looking at the top 6 terms for each topic because the beta values are fairly low after the 6th term (below beta 0.01 in general).

Our similarity measure is just one approach that could be used for this purpose. We can imagine many different possible ways of measuring similarity. It is also important to note that the unit of measurement differs slightly for tweets compared to surveys. In our Twitter data, we represent the top terms using beta values/probability, whereas our survey results are presented in terms of the frequency of responses. In the caveats/future work section, we discuss these shortcomings and propose that a more formalized distance-based metric might provide more precise results.

Results for Country-related question

In Figure 4 below, we present a histogram of the most frequently selected answers to the country-related question. The chart below shows the top 5 most frequent responses (excluding the “can’t choose” option, which we explain in the caveats/future work section). We have ordered the columns by frequency.

Figure 4



The full table of results is presented in Table 2 below:

Table 2

Response	**Count**
Using up natural resources	344
Air pollution	199
Can’t choose	181
Water pollution	134
Climate change	127
Chemicals and pesticides	112
Water shortage	95
Nuclear waste	88
Genetically modified foods	61
Domestic waste disposal	60

None of these	22
No answer	7

More than 50% of the responses fall into the following categories: using up our natural resources, air pollution, water pollution, and climate change. We now look for these terms in the Twitter topic models.

We understood “green energy” to be a good synonym for “using up our natural resources”. We find “energy” in the first Twitter topic, and “green” in Twitter topics 2 and 3. The word “pollution” shows up in Twitter topic 4. “Water” shows up in the Twitter topic 3. Lastly, “climate change”, as well as “climate”, show up in Twitter topics 4 and 1, respectively. By our defined metric we can say that the Twitter results are very similar to the survey results.

Lastly, as a side note, It is interesting to observe that Twitter topic 5 (health/safety concerns) is not a possible survey choice. In addition to validating our Twitter results using the survey, it is useful to point out that survey responses could be limiting, and that perhaps Twitter can be used to fill in some of those gaps.

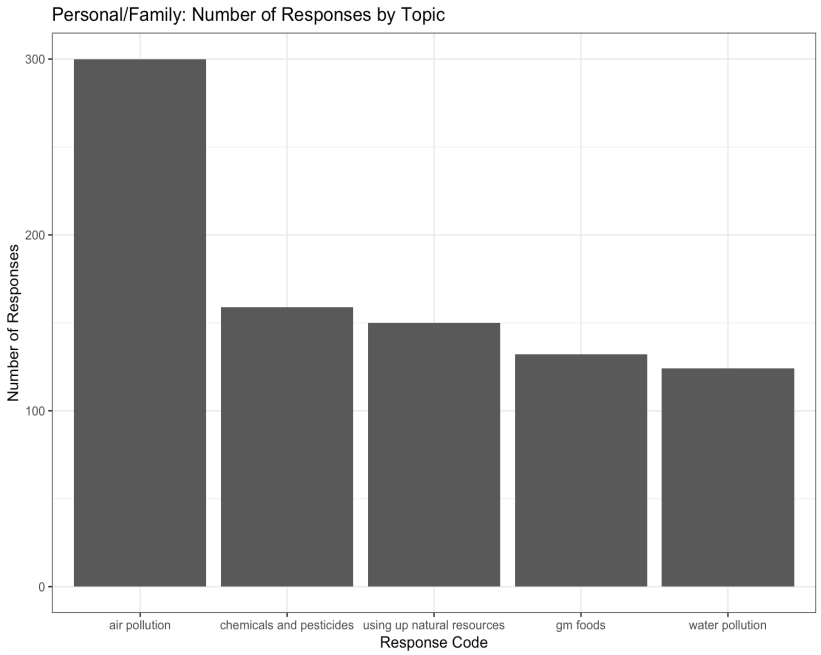
Results for Family/Personal related question

As mentioned before, the same question about the environment was asked with respect to an individual/family. We thought it would be interesting to see if the nature of opinions expressed on Twitter are more related to personal concerns or a country/global view. The question is phrased in the following manner:

Which problem, if any, affects you and your family the most?

We present the survey responses, ranked by frequency in Figure 5 below :

Figure 5



We can observe that, much like the country-related question, “air pollution” and “using up natural resources” show up in the top 5 most frequent survey responses. As we mentioned before, these terms are present in our Twitter topic models. However, the “chemicals and pesticides” and “genetically modified foods”, do not show up in our Twitter topic models. In addition, it is interesting to note that the answer ‘climate change’, which was ranked 4th in the country-related question and was very present in the Twitter topics, is not ranked highly for the family-related question.

Using the same similarity measure as used for the country-related question, we can determine that the Twitter results are only moderately similar to the survey results. This is an interesting finding, because it shows us that Twitter is likely better at measuring public opinion on issues from a more country-specific/global standpoint, compared to a personal/family perspective.

Topic Modelling Analysis for England

The topic modelling analysis was repeated for tweets and survey responses from England. This analysis can be found in Appendix 2. As a general trend, we observe the same pattern: topics emerging from Twitter topic modelling overlap with the most frequent survey responses for the country related question, but do not overlap as much for the personal/family related question.

Comparison of Twitter Sentiment Analysis to Survey Responses on Climate Change

Twitter Sentiment Analysis

For the second part of our analysis, we wanted to understand the extent to which sentiments on Twitter could be compared to results from a survey. For this part of the analysis, we decided to use a survey question which had a range of responses that could be more easily be scaled to represent sentiments on a certain issue. The issue we chose for this analysis was climate change. Accordingly, we downloaded tweets by searching for the keyword #climatechange on Twitter, and specifying the country as the U.S.

Sentiment Dictionaries

Typically, sentiment analysis is conducted by using a dictionary of words to labeled sentiments, and joining this to the existing text data to determine what the average sentiment of a text would be. In this part of our analysis, the trickiest aspect was finding a way to compare the scale of sentiment dictionaries with the scale of survey responses.

Therefore, we used three different sentiment dictionaries and scaled them in different ways to compare the output. The three dictionaries used were: Afinn, NRC, and Vader. These dictionaries label words according to different metrics. The Afinn dictionary labels words on a scale from -5 to +5, where negative scores indicate more negative sentiments and positive scores indicate more positive sentiments. The NRC dictionary, on the other hand, labels each word according to a certain emotion such as disgust, anger, sadness, or joy. The Vader dictionary assigns a score to each tweet based on the aggregate positive or negative words present in the tweet.

We initially chose the Afinn dictionary for our analysis because we thought we could truncate the scale from -5 to +5 to be roughly similar to the 5-point scale of survey responses related to climate change. We then compared the types of sentiments expressed to those shown by the NRC dictionary. Finally, we used the Vader dictionary on our tweets, because this dictionary is optimized to be used with social media data. Importantly, the Vader dictionary has been used in previous literature on conducting sentiment analysis with Twitter data.

Afinn Sentiment Results

First, we determined a sentiment score for each tweet, labelling them on a range from negative to positive sentiment. The Afinn dictionary already has a scale from -5 to 5 which accomplishes this. For each tweet, we assigned a score using either the mean or mode of the scores for each

word in that tweet. Below, we show the number of tweets that fall in each score boundary. An important consideration is that the sentiment analysis results are presented by tweet, whereas the survey results are presented by individual frequencies. This difference is a point we expand upon in the caveats/future work section.

Figure 6

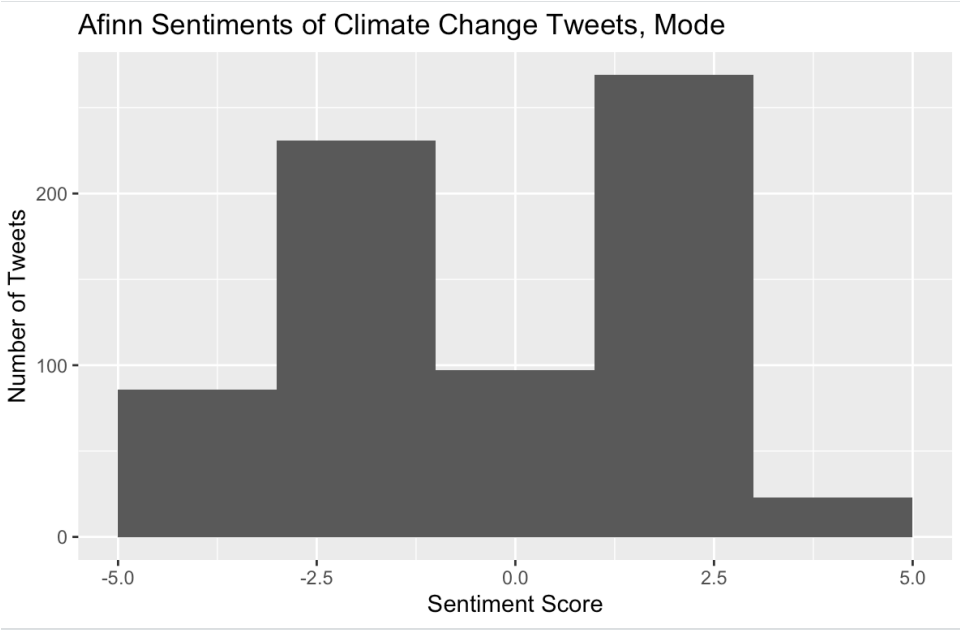


Figure 7

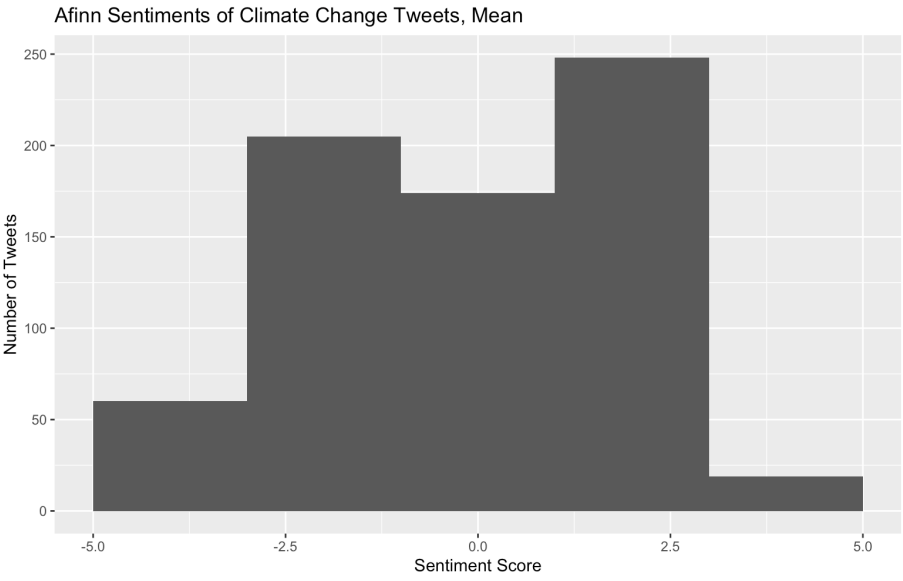
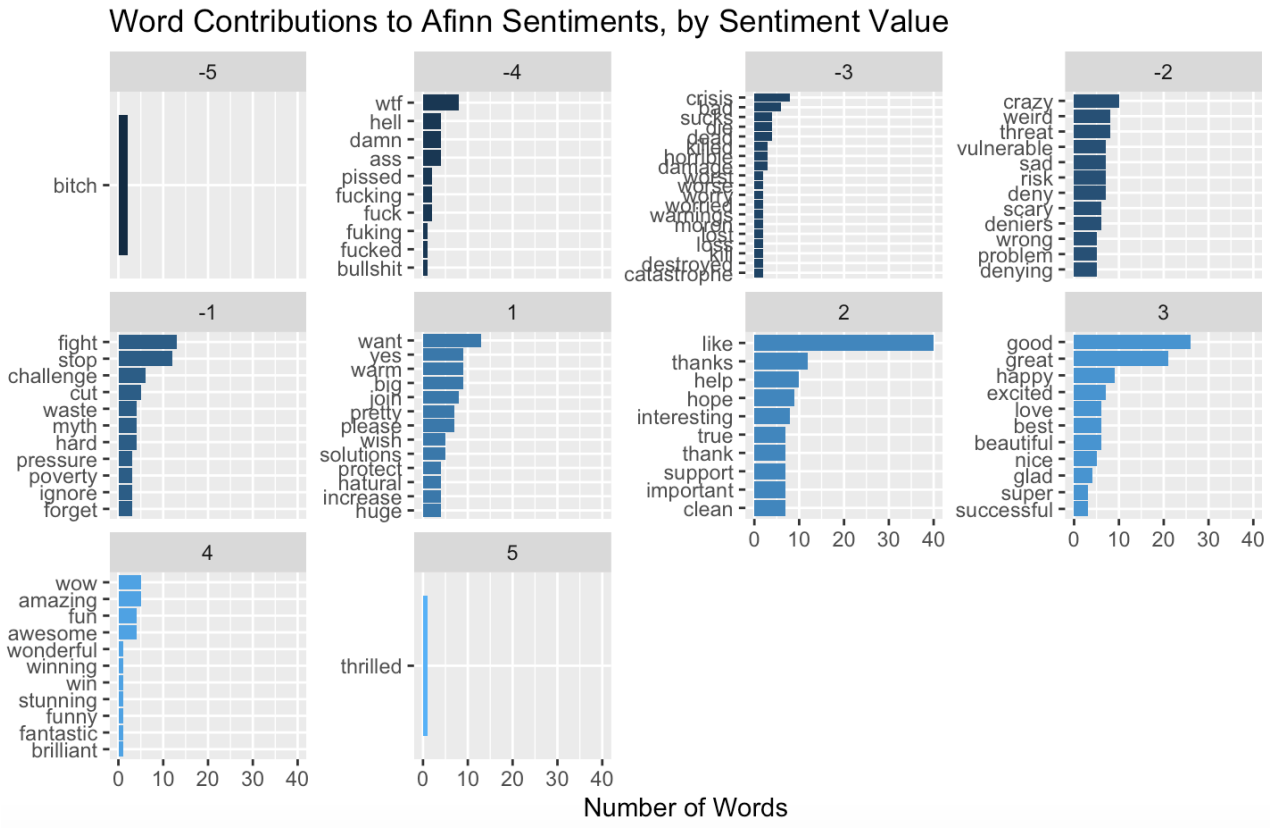


Figure 6 above shows that when we find a sentiment score for each tweet based on the mode, it appears that tweets are on average more polarized, with more tweets falling on the extremes than in the neutral. Some of this is averaged out when we find the sentiment score based on the mean, as we see in Figure 7.

In Figure 8 below, we show the words contributing to each of the Afinn sentiment values.

Figure 8



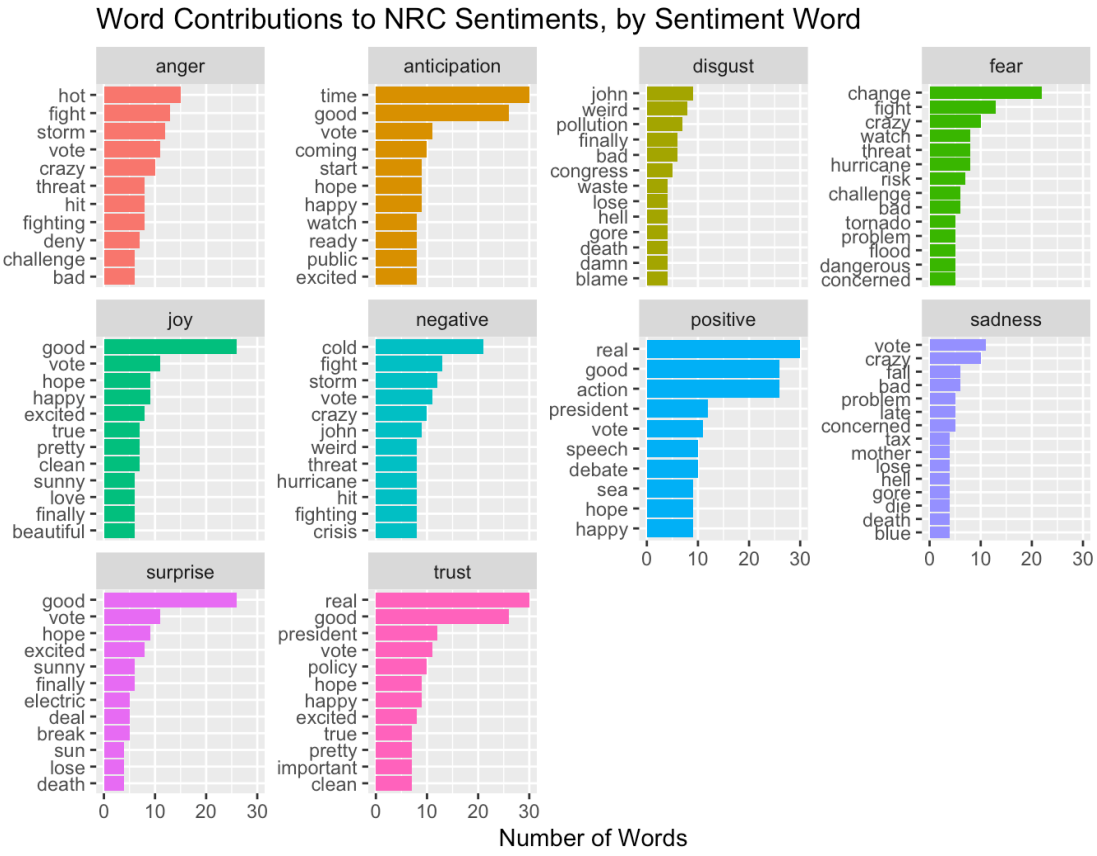
The climate change-related survey question asks whether a respondent considers a rise in temperatures caused by climate change to be dangerous for the environment, on a scale from extremely dangerous to not dangerous at all. The negative and positive sentiments seem roughly in line with the spirit of the survey question, with some caveats (see caveats/future work section).

In the next section, we look at the NRC dictionary sentiments to get more of a sense of what constitutes a positive or negative sentiment in our tweets.

NRC Sentiment Results

In Figure 9 below, we can see the Twitter words that are associated with each of the categories presented within the NRC dictionary.

Figure 9



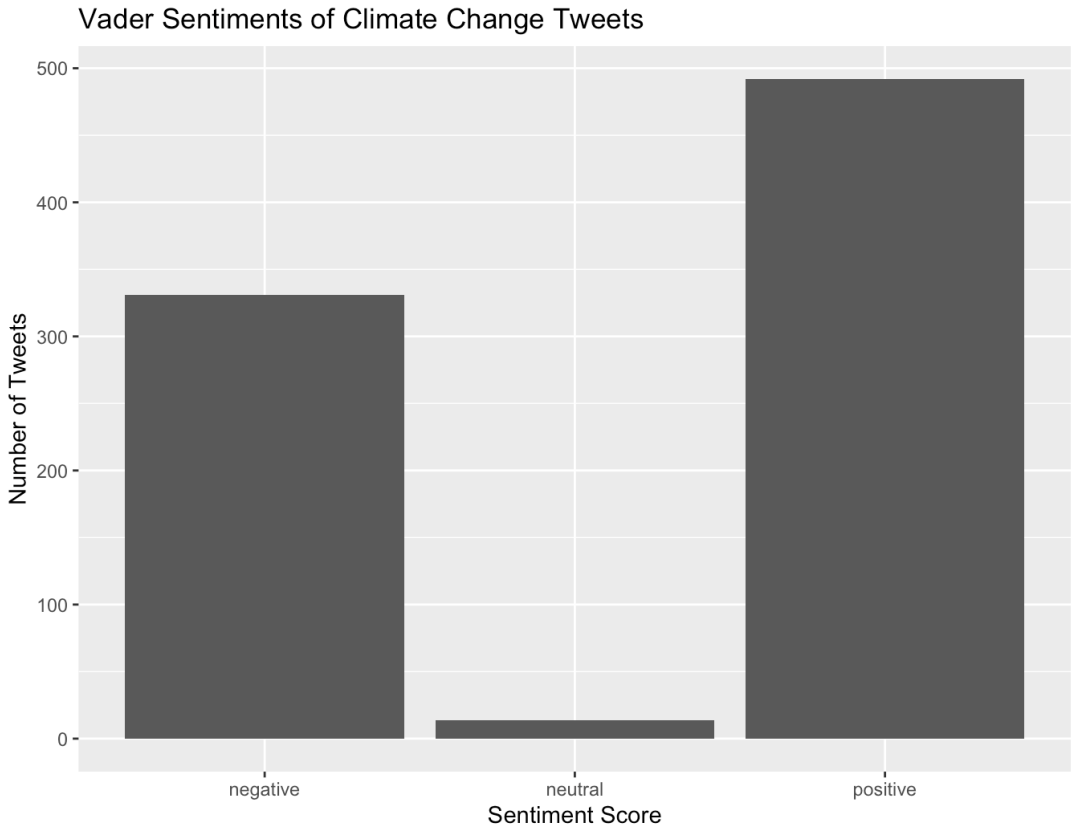
Similar to the Afinn results, more positive sentiments appear to express people's hope that action is being taken towards fixing the issue of climate change, whereas the more negative sentiments show fear and sadness about climate change.

Vader Sentiment Results

The Vader dictionary uses a normalized, weighted composite score to rate a given sentence. The convention for compound Vader scores is that a compound score is zero if the word is not present in the Vader dictionary. If the compound score is greater than or equal to 0.05, the sentence has a positive sentiment. If the compound score is between -0.05 and 0.05, the sentence has a neutral sentiment. If the compound score is less than or equal to -0.05, then the sentence has a negative sentiment.

The resulting scores for each tweet are plotted in Figure 10 below:

Figure 10



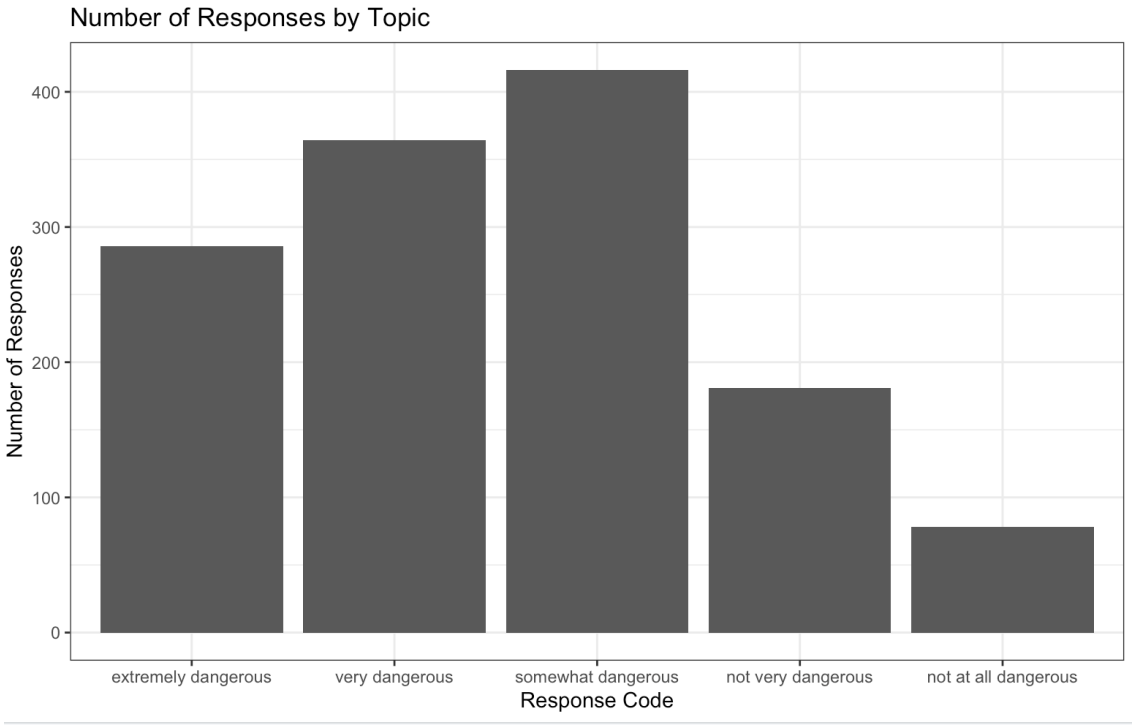
When we use the Vader sentiment dictionary, there appear to be more extreme tweets than picked up by the Afinn dictionary. The Vader results show us even more clearly how polarized Twitter is.

One surprising finding to note is that more of the tweets appear to be positive than negative. We address potential reasons for this in our caveats/future work section.

Survey Responses

The climate change-related survey question asks whether a respondent considers a rise in temperatures caused by climate change to be dangerous for the environment, on a scale from extremely dangerous to not dangerous at all. In Figure 11 below, we present the survey responses by frequency. The columns are ordered from negative to positive sentiment.

Figure 11



We can note from the results above that most of the responses fall in the moderate-negative categories. This provides a significant contrast to the sentiment analysis results, especially when compared to the extreme results from using the Vader dictionary.

Conclusion

This paper uses both topic modeling and sentiment analysis to compare results from Twitter data to traditional surveys. The results presented in the main body of the paper are based off of tweets downloaded in the U.S. and survey responses filtered to include only U.S. responses. The techniques used to compare Twitter and survey data are exploratory in nature.

Our topic modelling results show that a surprisingly large number of topics that emerged from topic modeling on Twitter data overlapped with the topics that had the highest responses in the survey. The sentiment analysis revealed that the sentiment of tweets was on average concentrated in extremes (either positive or negative), whereas the survey results were more moderate-negative. In conclusion, based on our results, we establish that Twitter can be an adequate proxy for public opinion when the goal is to understand which topics people care more about, but it is less accurate at representing the sentiments of the masses with respect to certain topic.

Caveats/Future Work

For topic modelling, we constructed a similarity measure and demonstrate that a number of the terms that show up with high probability in our topic modelling results are also frequent survey responses. While this comparison is a starting point, it is not exact. First off, the top terms that appear in a topic model are determined probabilistically, whereas the top responses in the survey are determined by frequency. Second, our measure for determining how similar tweets are to survey responses is mostly based on arbitrarily determined cutoffs. Therefore, we assert that future work could focus on creating a formalized distance metric to compute the differences between LDA topic modelling results and survey responses. One potential approach would be to use multidimensional scaling to build an underlying dimension of environmental concerns or climate preferences. This approach would standardize both Twitter and survey data so that they were on the same scale and the two data sources could be more formally compared.

For sentiment analysis, we compared the sentiments of words in tweets to the frequency of survey responses to a question with a range from negative to positive assertions. This comparison was done specifically on the issue of climate change. In general, we find that Twitter sentiments are more extreme/polarized than survey responses, with a few caveats. First off, the scales differ for sentiment analysis when compared to the survey data. While the Afinn measure comes close to the 5-point scale in the survey, it is still not exactly the same. The Vader dictionary only has a 3-point scale, so our comparison to this measure is based on rough groupings of the 5-point scale in the survey responses. Second, the results from the Afinn and Vader dictionary are somewhat surprising and point to a potential interpretability issue when comparing the two data sources. Particularly with the Vader dictionary, there are far more positive words than negative words associated with climate change. Looking at the NRC and Afinn word groupings, it is possible to see that words such as “hope” and “support” show positive sentiment. This meaning could be confusing in that one might hope that issues related to climate change would get better, but that doesn’t necessarily mean that they have a positive view on the current state of the climate. This caveat further indicates some of the challenges of interpreting sentiment analysis in general, and nuances in the English language that are hard to account for. Finally, there is a key difference in the unit of measure for frequency in Twitter data versus frequency in survey responses. In the survey responses, each unit of measure is an individual, whereas on Twitter, each unit of measure is a tweet. This means that if one person tweeted many times, we would inaccurately be counting each tweet as being equivalent to a survey response. One potential remedy for this would be to isolate usernames and measure the overall sentiment per username. This is something we could attempt in future work.

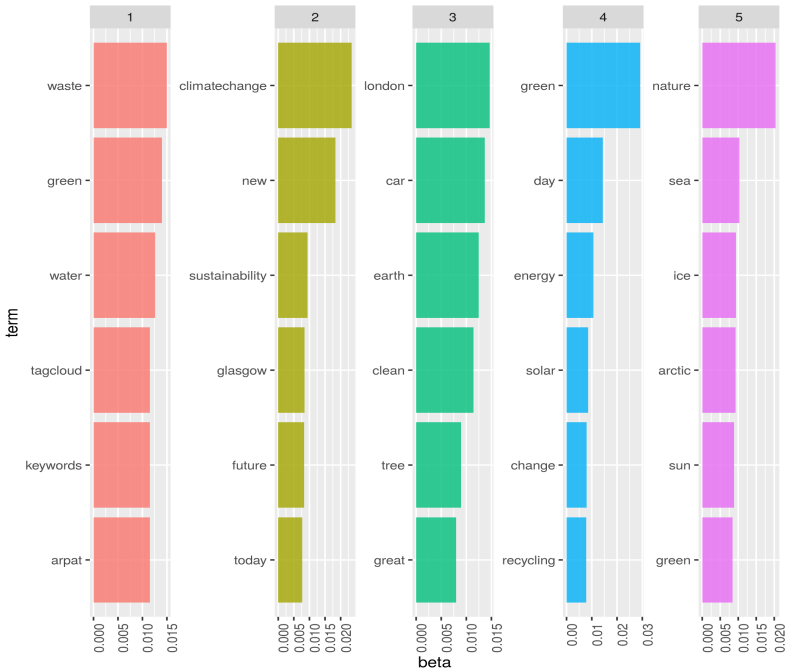
Appendix

Appendix 1: Code

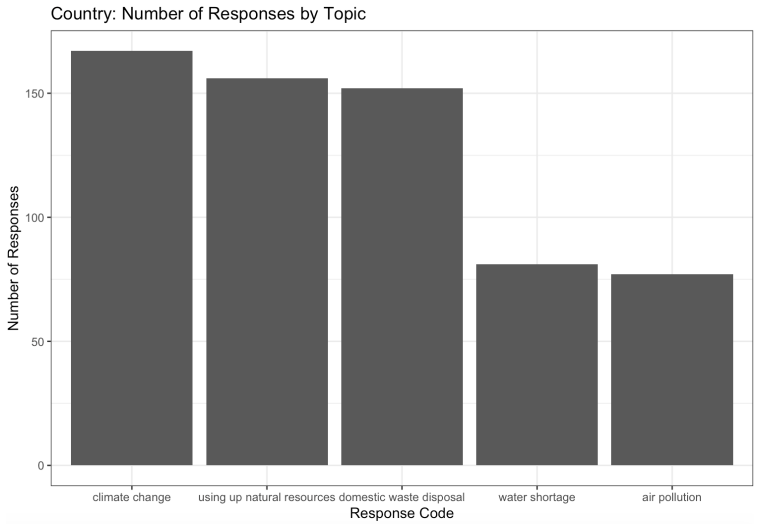
The code for our project can be found in
https://github.com/fhalamos/topics-sentiments-twitter-data

Appendix 2: Topic modelling analysis for England

After fitting LDA models to England tweets, we observe the following topics:

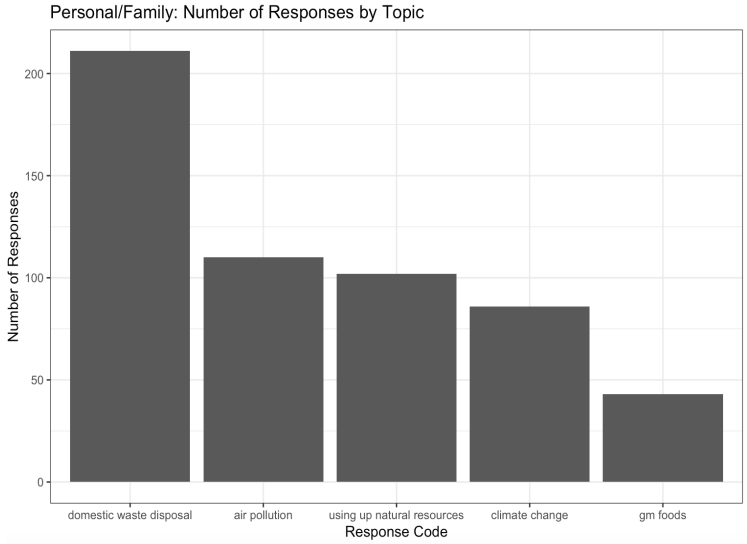


The survey responses histogram for the country-related question are:



We observe that the between the most popular survey topics, “climate change” and “using up natural resources” are present in the Twitter topics. In addition, pollution had fewer survey responses compared to the U.S., and did not show up in our Twitter topics. However, “domestic waste disposal”, which shows up as a frequent survey response, is not detected by the Twitter topics.

Survey responses histogram for the personal/family-related question are presented below:



In this case, “domestic waste disposal” gets most of the survey responses, followed by “air pollution”, but neither of them can be found in the Twitter topics. Only “using up natural

resources” can be found in the twitter data.

We can observe a similar pattern to that of the U.S. The most popular responses in the survey for the country-related question can be adequately captured by the Twitter topics, but the same does not happen for answers to the personal/family-related question.