Social Network Analysis and Mining (2019) 9:24
https://doi.org/10.1007/s13278-019-0568-8

ORIGINAL ARTICLE

Topic modeling and sentiment analysis of global climate change tweets

Biraj Dahal[1,2] · Sathish A. P. Kumar[3] · Zhenlong Li[4]

Received: 5 September 2018 / Revised: 22 May 2019 / Accepted: 27 May 2019 / Published online: 10 June 2019

Abstract
Social media websites can be used as a data source for mining public opinion on a variety of subjects including climate change. Twitter, in particular, allows for the evaluation of public opinion across both time and space because geotagged tweets include timestamps and geographic coordinates (latitude/longitude). In this study, a large dataset of geotagged tweets containing certain keywords relating to climate change is analyzed using volume analysis and text mining techniques such as topic modeling and sentiment analysis. Latent Dirichlet allocation was applied for topic modeling to infer the different topics of discussion, and Valence Aware Dictionary and sEntiment Reasoner was applied for sentiment analysis to determine the overall feelings and attitudes found in the dataset. These techniques are used to compare and contrast the nature of climate change discussion between different countries and over time. Sentiment analysis shows that the overall discussion is negative, especially when users are reacting to political or extreme weather events. Topic modeling shows that the different topics of discussion on climate change are diverse, but some topics are more prevalent than others. In particular, the discussion of climate change in the USA is less focused on policy-related topics than other countries.

1 Introduction

Big data analytics, text mining, and data mining have been applied in several contexts and applications in the literature (Alampalayam and Kumar 2004; Xu and Kumar 2014, 2015a, b; Xu et al. 2016; Alampalayam and Natsheh 2008; Raj and Kumar 2017). This paper looks at a climate change-related text mining perspective. According to the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report, continued climate change will have severe and irreversible impacts for people and ecosystems all over the world. Some of the impacts already attributed to climate change with high confidence or likelihood are increased frequency and intensity of daily temperature extremes, negative impacts on crop yields, sea-level rise, and ocean acidification (IPCC 2014). Different steps can be taken to mitigate the future impacts of climate change, but the implementation of these steps in public policy depends on the public opinion on climate change. A March 2017 poll of the public's views on global warming in the USA showed that 66% of Americans worry either a great deal or a fair amount about global warming (Saad 2017). However, traditional polling methods do not take advantage of the increasing popularity of social media and the climate change discussions found within (Kirilenko and Stepchenkova 2014).

The traditional approach of measuring public opinion about climate change is through surveys. However, in recent years social media and microblogging sites have become very popular discussion forums for many users around the world. Social media usage among American adults has increased almost tenfold since 2005 (Perrin 2015). Twitter is one of the most popular social media and microblogging sites. Twitter users post up to 280-character messages called tweets containing their thoughts and opinions, and these tweets are affected both by real-world events and the trends of other messages posted in social media (Zubiaga et al. 2014). By 2017, over 500 million tweets were posted per day (Omnicore 2018). In response to this explosive growth in usage, social scientists have taken advantage of

✉ Sathish A. P. Kumar
skumar@coastal.edu

[1] Department of Computer Science, Clemson University, Clemson, SC, USA

[2] Department of Mathematical Sciences, Clemson University, Clemson, SC, USA

[3] Department of Computing Sciences, Coastal Carolina University, Conway, SC, USA

[4] Department of Geography, University of South Carolina, Columbia, SC, USA

this newfound dataset such as Mandel et al. (2012) on the topic of public opinion of Hurricane Irene and Tumasjan et al. (2010) on the topic of election result prediction. On a similar vein, Twitter can provide a great data source for discerning public opinion and risk perception on climate change.

There are many ways to study and quantify public opinion on climate change. van der Linden (2017) discusses the various influences climate change risk perception and considers the impact that social networks and influences has on developing risk perception and overall opinion. To that end, this study considers the Twitter as a repository for analyzing the public perception on climate change in a social network setting.

Specifically, this study evaluates the topics and opinions of climate change discussion found on Twitter using text mining, a process of deriving useful information from a collection of texts, on a corpus of tweets that have been time stamped and geotagged. Several statistical techniques can be useful for this purpose, such as topic modeling and sentiment analysis. Topic modeling refers to any technique that discovers the hidden semantic structure in a corpus which provides insights into the different themes present in the texts (Blei 2012). Topic modeling will be used to determine the sub-topics of discussion on Twitter about climate change. Sentiment analysis is the process of identifying the emotions and opinions expressed in a particular text (Medhat et al. 2014). Sentiment analysis will be used to determine the different levels of positive and negative opinion on climate change present in the dataset (Koto and Adriani 2015). Employing these text mining techniques would provide great insight into the types of discussion that can be found in a climate change Twitter corpus.

The contributions of this study are to: (1) visualize the level of awareness about climate change globally and (2) understand the geospatial trends in topics and sentiment of the tweets related to climate change.

The rest of the paper is organized as follows. Section 2 discusses some related works. Section 3 describes the data and methodology used for the experimentation and analysis. Section 4 presents the volume analysis, topic modeling, and sentiment analysis experimentation results. Section 5 discusses the sentiment analysis results. Section 6 discusses the topic modeling results. Lastly, Sect. 6 provides the summary and conclusion.

2 Related works

Recent studies have effectively applied topic modeling to glean information in a variety of different contexts. Shi et al. (2016) used topic modeling to measure the proximity or relatedness of businesses using business description text. They processed unstructured text describing a variety of firms to compute business proximity between pair of firms in a simple scalable topic modeling approach. Lee et al. (2016) used topic modeling to calculate a similarity measures between users in a location based social network. Topic modeling was used on user biographies to calculate this similarity measure, and then used those results to model social network formation. The use of topic modeling for processing large amounts of unstructured text to analyze the dataset as a whole in those studies is similar to the goal of topic modeling in this study.

Similar work has been done on the subject of text mining of climate change-related tweets. Cody et al. (2015) applies sentiment analysis to a dataset of climate change-related tweets, showing how the sentiment has changed over time. Our study goes beyond this analysis by conducting a spatial analysis in addition to temporal analysis. Kirilenko and Stepchenkova (2014) conducted a geospatial and temporal study on climate change-related tweets, analyzing the differences between regions and different times; however, more complex text mining techniques such as sentiment analysis were not used.

To our knowledge, there has not been work in the literature that applies topic modeling or sentiment analysis to do a geospatial and temporal analysis on climate change-related tweets on a global level. There are other studies outside the domain of climate change that examine Twitter from a spatiotemporal perspective. Ghosh and Guha (2013) applied topic modeling to a corpus of geotagged obesity-related tweets and generated spatial visualizations in the USA, validating the topics produced by the topic model spatially. Lansley and Longley (2016) did an in-depth analysis on a corpus of geotagged tweets from London. Topic modeling was applied, and the topics produced were analyzed by location, time posted, and user profile.

3 Experimental methodology

3.1 Methodology

The goal of this study is to perform volume analysis, sentiment analysis, and topic modeling to a collection of tweets and compare these results over space and time. To this end, several different steps must be performed sequentially for data preparation, and then the three different methods analysis must be performed and then analyzed.

First, a clean dataset of tweets containing geospatial and temporal values must be obtained. This is done in three steps: data collection, data cleaning, and then data preparation. After those steps, the dataset is then processed for three separate analysis techniques: volume analysis, topic modeling, and sentiment analysis. The results from these different analysis steps are then discussed and evaluated. The volume analysis is done to establish basic geospatial and temporal facts about the dataset. Topic modeling is done to determine the topics present in the dataset. Sentiment analysis is done to determine the sentiment alignment or value of every tweet. Volume analysis is done first so that the results of topic modeling and sentiment analysis can be normalized and put into context. The results from each method of analysis will then be evaluated and discussed.

The purpose of the experimentation is to identify any significant results from the volume analysis, topic modeling, or sentiment analysis steps and discuss the potential reasons for these results.

Figure 1 illustrates the overall workflow of this study. The remainder of the section describes the specific techniques used in every single step, presenting the challenges faced as well as the identified solutions.

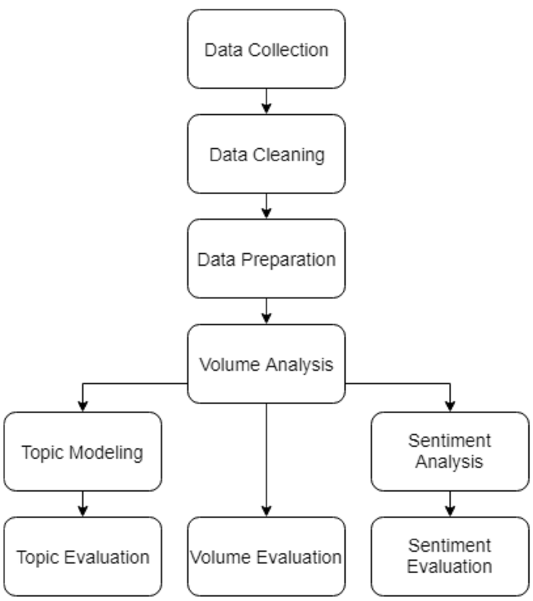


Fig. 1 Overall research workflow

3.2 Data collection and preparation

The dataset consists of 390,016 tweets from July 1, 2016, to February 28, 2018, collected using the Twitter Stream Application Programming Interface (API). This dataset consists of geotagged tweets posted in that period that contained at least one of the following keywords or some variant of them: climate change, carbon dioxide, fossil fuel, carbon footprint, emissions. Each geotagged tweet has a longitude and latitude pair associated with the location the user was in when the tweet was posted. The data for each tweet have the following relevant variables: tweetid (the unique code given to the tweet), userid (the Twitter code of the user that posted that tweet), postdate (the time and day the tweet was posted), latitude, longitude, and message (the body of the tweet).

The Twitter Stream API gives access to a random sample representing about 1% of all tweets posted in the time period. This was done to collect approximately 2 billion geotagged tweets from July 1, 2016, to February 28, 2018. A keyword search for the words listed above was used to extract construct the dataset used in this study. Table 1 displays a few tweets in the dataset.

The dataset needs to be cleaned up before analysis. This is done by manually examining a small sample of the dataset to see which kinds of false positives exist and then removing many false positives of that type from the entire dataset programmatically. The biggest offender among the keywords was “emission,” which can be found in many tweets that have nothing to do with climate change, especially tweets written in the French language.

The main step in data preparation is reverse geocoding which is the process of obtaining a readable place name or address from a latitude/longitude point. This would facilitate the process of pulling every tweet from a particular region at

Table 1 Sample of tweets in the dataset. The city, region, and country fields were calculated and added to the original dataset. The tweeted and userid are other data variables that are not shown in the table to protect the identities of the Twitter users

Postdate	Latitude	Longitude	Message	Country	Region	City
2017-01-30 02:19:21	−41.2529983521	174.754333496	What is the future for #NZ forests? @sciblogsnz: “Forest health in a changing climate” https://t.co/X8oE9kjMPi #climatechange #environment	NZ	Wellington	Kelburn
2017-10-29 11:35:54	35.309047699000004	−98.7169952393	@DaveLeeC3 @JacobAWohl @realDonaldTrump So are we giving Obama credit for the hurricanes too? What about Global Warming?	USA	Oklahoma	Weatherford
2017-12-05 08:33:26	60.18334198	24.6751346588	.@miapetrakumpula: Climate change is real and we need European action and commitment to change the future. #FutureofEurope	FI	Uusimaa	Espoo

once, which is important when for example comparing the volume of tweets posted between countries. Reverse geocoding is performed by associating each tweet with nearest city from the tweets latitude/longitude values. This is done using the *reverse_geocoder* Python library (Thampi 2016). Once a city is determined, region and country names can be obtained. Three new data fields were calculated and added to every tweet in the dataset: country name, region name (if the city belonged to some political region), and city name.

Another concern with Twitter data is the influence of bot users that post the same tweet at high volumes on particular days. These bot or spam users can greatly influence the analysis especially temporally, so they must be removed. One particular Twitter user was responsible for 6632 tweets, all in Australia. This was easily the most prolific tweeter, posting more than five times as many tweets as the second highest tweeter. The number of duplicate tweets per user was calculated to detect other bot users, but no other user posted more than a few tweets per day. Thus, the aforementioned user was determined to be the only bot user, and it was the only user removed.

After all of these steps, the size of the final cleaned dataset was 366,244 tweets.

3.3 Volume analysis

Examining the volume of tweets posted in a particular area or at a particular time is an important first step in exploring the data and can provide some interesting results by itself. The raw number of tweets posted in each country is examined. To reduce bias, the number of tweets per country is normalized by both the population and the total number of geotagged tweets from that country. The number of unique Twitter users in the dataset can also be examined in a similar way. In addition to geospatial analysis, the number of tweets per day can also be analyzed to see whether spikes in Twitter activity on a given day coincide with any real-world events that occurred on that day.

3.4 Topic modeling

There are a variety of different subjects of discussion in the dataset, and isolating them is not a trivial task. Given the large size of the dataset, it is not possible to read all of the documents in the corpus individually, since each tweet is one document in the corpus. Since none of the tweets are labeled with topics, an unsupervised machine learning technique known as topic modeling is employed to determine which topics are present and in what quantities. The topics produced are probabilistic mixtures of words that represent word co-occurrence trends in the dataset.

A good topic model should produce human-interpretable topics that are distinct from each other. The top words in the topic should coherently belong to some unified concept without too much overlap with the concepts in other topics. These standards are chosen so the “best” topic model is one that corresponds to distinct subjects of discussion in the dataset.

3.4.1 Data preparation

Before any topic modeling can be done, each tweet needs to be adjusted to a format that will facilitate good topics. The documents or tweets will be converted to a bag-of-words (BOW) corpus, where each document in the corpus is separated into a list of the words the tweet contains. In a BOW corpus, the order between words does not matter. To do this conversion, all hyperlinks and non-alphabetic characters such as punctuation and numbers will be removed from each tweet message, and then the message will be split by space to obtain the list of words.

After this, a list of stop words, which are common words in the English language that have no inherent meaning such as articles and conjunctions, as well as the keywords “climate,” “change,” “climatechange,” “global,” “warming,” and “globalwarming” are removed. This is done to ensure that the topics produced by topic modeling are meaningful and not dominated by the same top words. The phrase “amp” appears whenever a mobile user includes an ampersand (&) in their tweet, so it has no inherent meaning and was removed as well. The keywords were found to be in virtually every tweet, so they are filtered out to improve topic quality. In addition, to further reduce noise in the topics inferred, words that rarely occur in the corpus are removed. Preliminary experimentation was done to determine the optimal threshold in which words that occur less frequently than the threshold should be removed. Setting the threshold too high would remove many words with meaningful value from the corpus, and setting the threshold too low would pollute the corpus with words occurring too rarely to be processed well topic modeling. Preliminary experimentation showed that a threshold of 200 was a good compromise, so words that occur less than 200 times in the entire corpus were filtered out.

3.4.2 Latent Dirichlet allocation (LDA)

One of the most popular topic models is latent Dirichlet allocation (LDA) (Blei et al. 2003). In this paper, LDA is referred to both as an algorithm that will be used to train a topic model and as a topic model trained from the aforementioned algorithm. In LDA, a topic is essentially just a probability distribution over every word found in the corpus. The number of topics must be chosen before LDA is run. The main assumptions of the model are that each document in the corpus is a probabilistic mixture of topics, and each topic is a probabilistic mixture of terms. LDA works by looking at the word co-occurrences within documents, assuming that words that occur in the same document are more likely to be on the same topic than words that are not, and that documents that contain the same words are more likely to contain the same topics than documents that do not. In this context, a document is simply a tweet.

LDA is a generative model, meaning that the model can be used to generate a corpus. Figure 2 shows a plate diagram of LDA's generative process. Suppose our corpus has M documents each with N_m (m from 1 to M) words with a total of distinct W words in the corpus overall, and suppose there will be k topics. α and η are hyperparameters for Dirichlet distributions that produce the k-dimensional document/topic (θ) vectors and W-dimensional topic/word (β) vectors, respectively. θ and β will act as parameters for categorical distributions from which topics and words will be selected (sampled).

The generation process is as follows: (1) For each topic j, β_j is sampled. (2) For each document m, θ_m is sampled. (3) For each word position n in document m (so this process is repeated N_m times), a topic z is sampled from the categorical distribution parameterized by θ_m. (4) Finally, a word is sampled from the categorical distribution parameterized by β_z. The goal of training an LDA model is to determine θ and β such that the probability of generating the actual corpus is maximized (Blei 2012).

LDA takes α, η, and k as parameters and randomizes all other values (other than w). Then, each iteration slowly improves these values, using the guiding principles that words that occur in the same document are likely to be on the same topic and documents that contain the same words likely have some of the same topics. After many iterations, a fully trained LDA model is obtained, the main objects of interest being the document/topic and topic/word matrices.

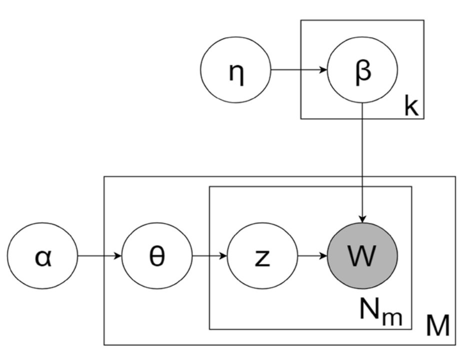


Fig. 2 LDA generative process. Circles represent variables, and rectangles (plates) represent repetition among documents, words, and topics. The shaded circle is the only variable that is actually visible in the corpus, the rest are latent in the model

3.4.3 LDA parameters

While LDA is typically used on longer documents and the performance may suffer on shorter documents such as tweets, Mehrotra et al. (2013) shows that the performance of LDA (i.e., the performance of the topic models produced by LDA) on Twitter data is significantly improved when tweets are aggregated or pooled together by some common factor to produce pseudo-documents for the corpus. With that in mind, the tweets are pooled by author (based on userid) before LDA is performed. This means every document in the corpus includes the messages of every tweet posted by a particular twitter user in a BOW format. This was chosen because preliminary experimentation showed that pooling by author produced the best topics. This method will be referred to as author-pooled LDA.

The most important parameter of LDA is the number of topics the model should infer from the corpus, k. It is not clear how many topics the dataset should be divided into. Too few topics could lead to an incomplete analysis as several different distinct topics could potentially be merged together, but too many topics could result in several different topics that represent a cohesive subject together but are individually confusing. To tackle this issue, author-pooled LDA is run with 5 topics, 20 topics, and 80 topics, and the quality of the inferred topics is compared to determine the optimal number of topics. As for the other parameters (α and η), they are set to the default of 1/number of topics, which is the default value in *gensim*, the Python software library that will be used to conduct LDA (Rehurek and Sojka 2010).

Author-pooled LDA is performed on the entire world, and then on each of the top four countries (Australia, Canada, UK, USA) to see how the topics differ by country. Figure 3 displays the workflow for topic modeling.

Finally, the performance of author-pooled LDA will be compared with other recent topic models.

3.5 Sentiment analysis

While topic modeling and volume analysis can be used to analyze what discussion is taking place in different regions or time periods, sentiment analysis is useful for determining the emotional state or opinion that is exhibited in the dataset. A numeric score is given to each tweet in the dataset: a positive number corresponding to positive sentiment or a happy emotional state toward the subject being discussed, and a negative number corresponding to negative sentiment or an unhappy emotional state toward the subject being discussed. A sentiment score close to zero indicates a neutral sentiment or emotional state.

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a simple sentiment analysis model that was created for social media sentiment analysis (Hutto and Gilbert

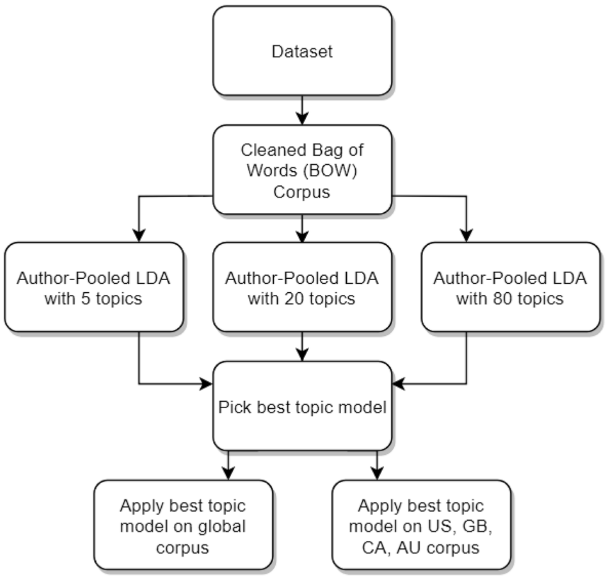


Fig. 3 Topic modeling workflow

2014). The model examines the lexical features of a document, i.e., the words that appear, to determine a preliminary sentiment score and then applies five different rules based on general syntactic and grammatical conventions to modify that score, with the final score falling between − 1 (strongly negative) and 1 (strongly positive). For example, the rules treat exclamation points and capitalization as sentiment amplifiers, scaling the positive or negative sentiment of a document. In addition, the rules are also able to handle contrastive conjunctions (e.g., the word "but") and negations (e.g., the word "not"), which means the model is able to differentiate the above opposite tweets. Table 2 shows the sentiment score that VADER would assign for a few tweets.

However, given the short and often sardonic nature of tweet messages, it is questionable whether VADER can accurately score tweets. For example, VADER assigned a sentiment score of 0.153 to a tweet with the message "The idiotic climate-change deniers hijacked the EPA & are ruining the environment for profit. Like the good ol days." That tweet should be assigned a negative score, but VADER assigned it a positive one. This shows that the VADER method produces false positives.

One way to evaluate a sentiment analysis method is to test its performance against human labeling. In this research, 500 tweets from the dataset are examined, and each tweet is manually classified as having "negative," "neutral," or "positive" sentiment. Then, VADER is performed on those 500 tweets, and the differences between VADER and the manual classification is evaluated. Tweets that were manually labeled "neutral" are not considered for this evaluation; the two questions are about how well VADER can identify positive tweets and how well VADER can identify negative tweets.

The default classification strategy is to consider tweets with sentiment scores less than 0 as negative and greater than 0 as positive. However, a model may be biased toward assigning positive sentiment or negative sentiment, so thresholds other than 0 can be used. Adjusting the threshold will affect precision and recall values, which are metrics that represent a classification model's ability to minimize false positives and false negatives, respectively (Manning et al. 2008).

Preliminary experimentation shows that VADER with a threshold of 0.25 has the best values of precision and recall, with an improvement in both metrics for both classification directions (positive and negative) over VADER with a threshold of 0, so that threshold will be used.

Sentiment analysis is conducted in the following steps. (1) VADER is used to calculate VADER scores for each tweet in the dataset. (2) Tweets with VADER scores greater than 0.25 will be classified as positive (+ 1), tweets with scores between 0 and 0.25 are classified as neutral (0), and all other tweets with scores less than 0 are classified as negative (− 1). (3) The classification scores, which will subsequently be referred to as sentiment score, of tweets are aggregated by tweet location and tweet postdate to obtain geospatial and temporal distributions of overall (net) sentiment. In particular, the net sentiment is compared temporally between the top four countries (Australia, Canada, UK, USA) to see whether spikes in positive or negative sentiment can be associated with real-world events.

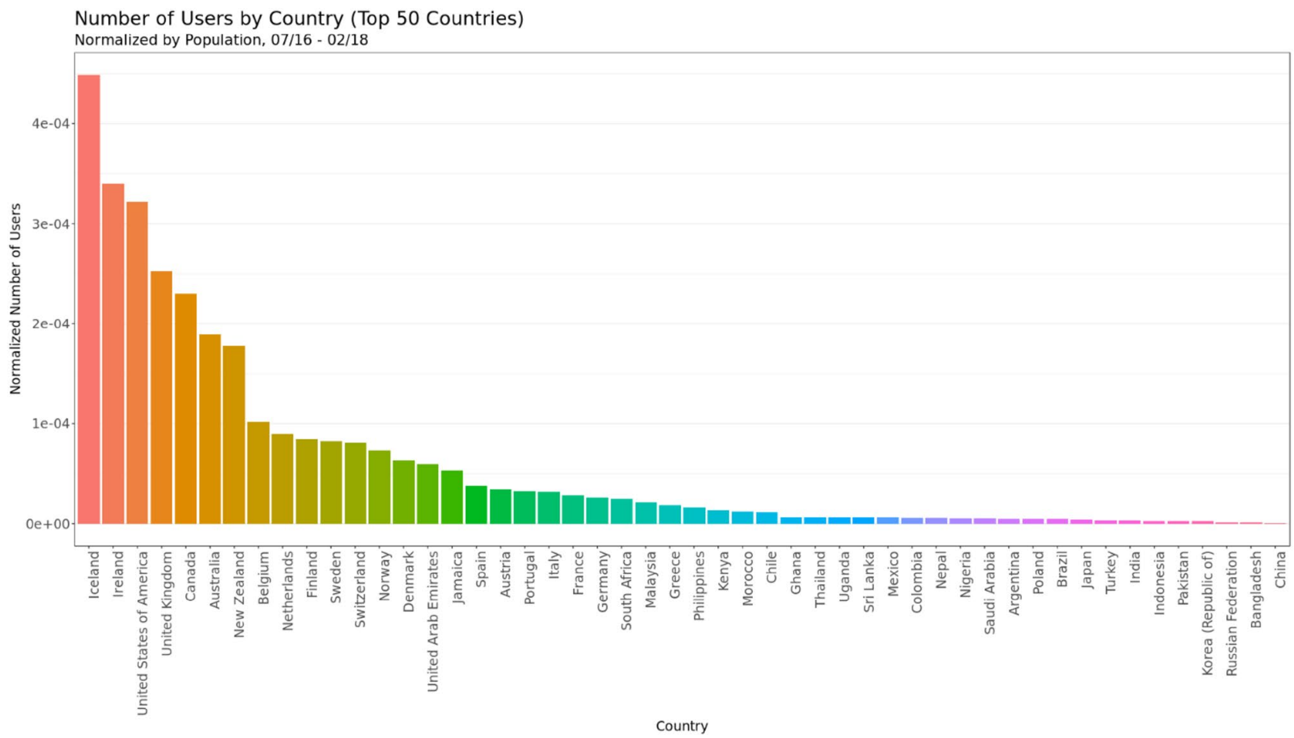
One limitation of VADER is that the dictionary used only considers English words. However, since the tweets in the dataset were selected by English keywords, the vast majority of them are written completely or mostly in English. For

Table 2 Sample of VADER's sentiment scoring (with URLS removed and mentions censored)

Tweet	Sentiment score
"I feel at home at #Barbuda and I want to be a part of the solution. These places have virtually no #carbon footprint"	0.1027
Climate change deniers; regardless of how many non-scientists "agree with him;" have no place at the @*****	− 0.2023
"@***** Oh and by the way. Climate Change is a HOAX."	− 0.4278
"We can use sustainable agriculture to create a more robust food system that does not contribute to #climatechange #LifestyleMed2016"	0.5849

tweets that are primarily in a foreign language, VADER will return a sentiment of 0 or neutral. Thus, the influence of these foreign language tweets will be negligible on the overall results, so this is not a major limitation.

4 Experimental results and discussion



4.1 Volume analysis

4.1.1 Spatial

Figure 4 displays a dot distribution plot for the dataset where one translucent dot was placed at the location of every tweet. The number of tweets per country is plotted in Fig. 5. These figures indicate that countries with high numbers of English-speaking population are more prevalent in the dataset. This is expected because the dataset was constructed by keyword search. The top four countries are the USA with 207,726 tweets, the UK with 34,102, Canada with 21,250, and Australia with 15,400.

The number of climate change-related tweets in a region was normalized by the total number of tweets from that region. Figure 6 shows the normalized number of climate change-related tweets by country for the top fifty countries in terms of tweet volume. Figure 7 is similar to Fig. 6 except that it is normalized by the total population of the country instead of the total number of tweets. The analysis was conducted through normalization by both tweets and population. Normalization by the total number of tweets in a country calculates of the percent or ratio of tweets relating to climate change (in English) out of all tweets from that country, whereas normalization by population calculates the average number of English climate change-related tweets per person. This is done to validate the rationale that the tweets under consideration are the representation of the population in a given region. Note that only the top fifty countries in terms of tweet volume were shown; otherwise there would be significant bias for small countries or islands that are typically tourism destination because a disproportionate number of tweets would be posted in those countries for their very small citizen population.

The number of users per country is another variable of interest. Some users (3%) posted tweets in multiple countries; those users will contribute to the counts of each of those countries. Figure 8 shows the number of users per country normalized by population. Again, only the top 50 countries in terms of number of users were included in the graph.

Many countries with high raw volume of tweets in the dataset have much lower normalized counts. The SUA and the UK have similar normalized counts, and they fall behind many other European, African, and Asian countries, many with small English-speaking populations. One explanation for

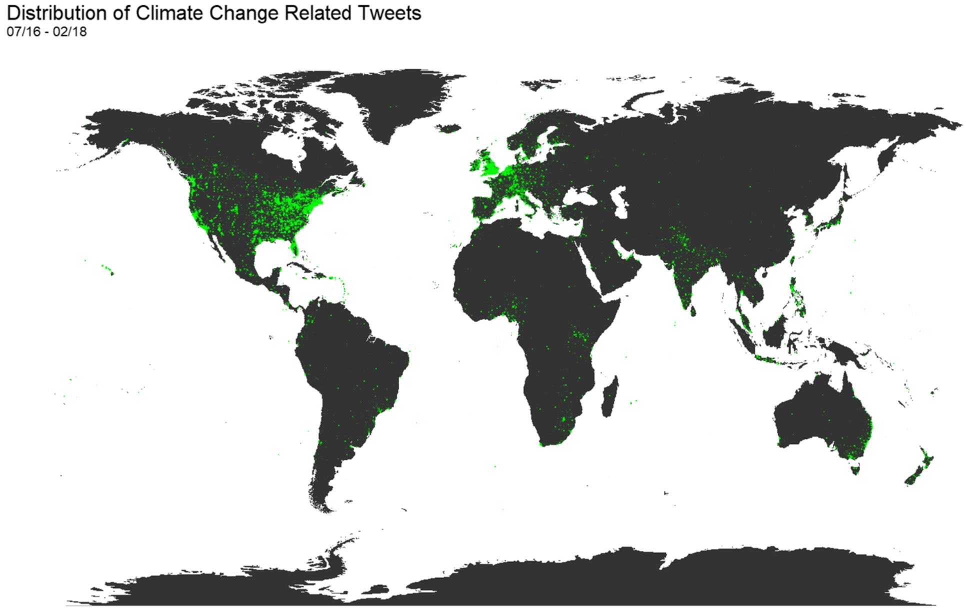


Fig. 4 Tweet dot distribution map

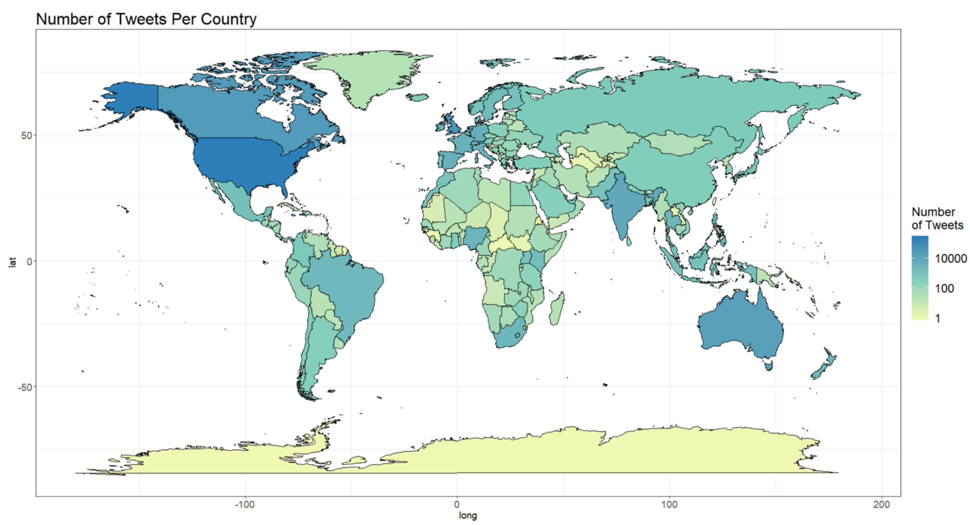


Fig. 5 Number of tweets per country

this is that since climate change is a global issue, Twitter users from those countries are more likely to post tweets to join the global discussion on climate change compared to other topics.

Normalization by population is biased against countries with few tweets in the English language. As one would expect, English-speaking countries are at the top of these plots, and near the bottom are countries without high English-speaking populations. There is also a very high correspondence between the normalized number of tweets and the normalized number of authors per country except Iceland which tops the users chart but does not appear in the volume chart. This is caused by the cutoff point of top 50 countries in terms of raw volume.

Over half of the tweets in the dataset came from the USA. In fact, the total tweet counts of many individual US states exceed the total tweet counts of most countries. For example, the number of tweets from California (33,158) exceed the number of tweets from Australia, which is the country with the third highest total tweet count. This indicates that investigation on the state level for the USA is necessary for a thorough analysis. Figure 9 shows the normalized number of climate change-related tweets by state in the USA.

Within the USA, the northeast has a relatively high amount of climate change discussion. It appears that most of the southeastern coast is falling behind on their discussion of climate change compared to other regions. This is surprising, given the hyperactive 2017 hurricane season that mostly affected that coast. This suggests that the southeastern coast is lacking in its conversation about climate change compared to the northeast which has much higher percentages of discussion. This could be caused by the cultural and political differences between the regions. The differences in climate

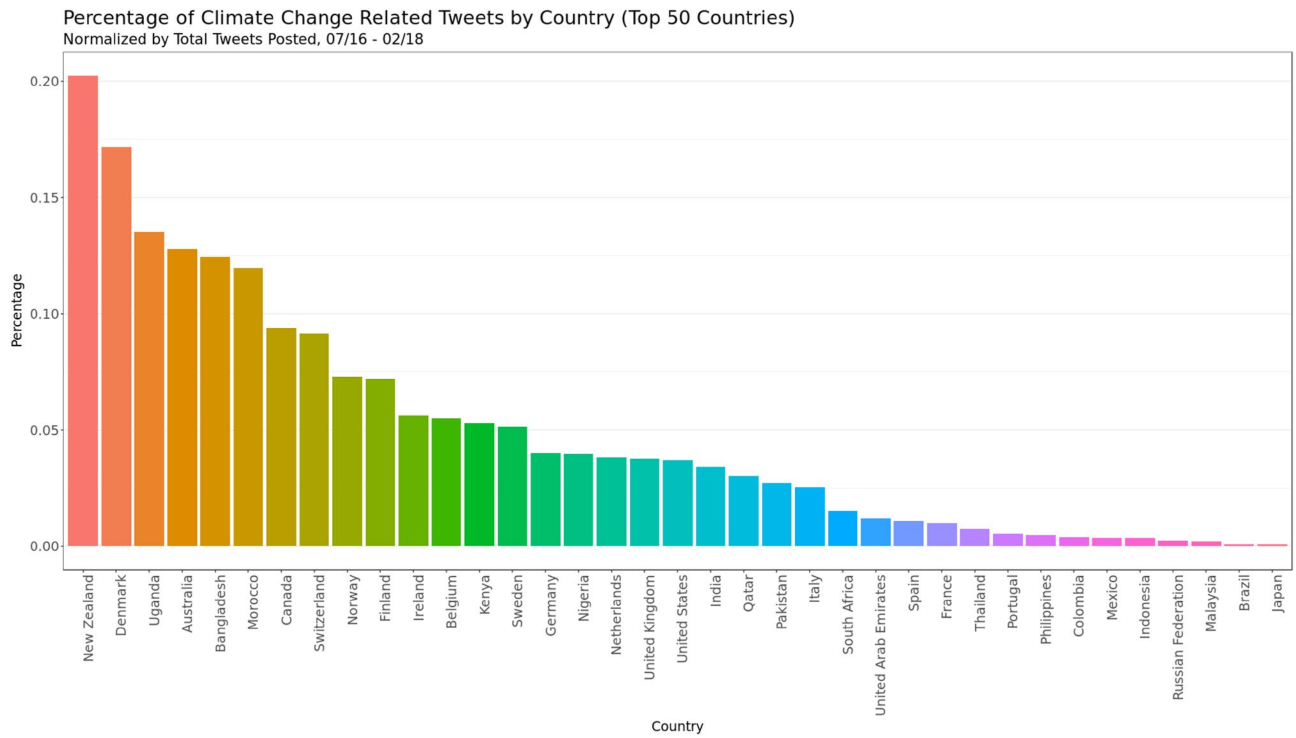


Fig. 6 Normalized percentage of tweets by country

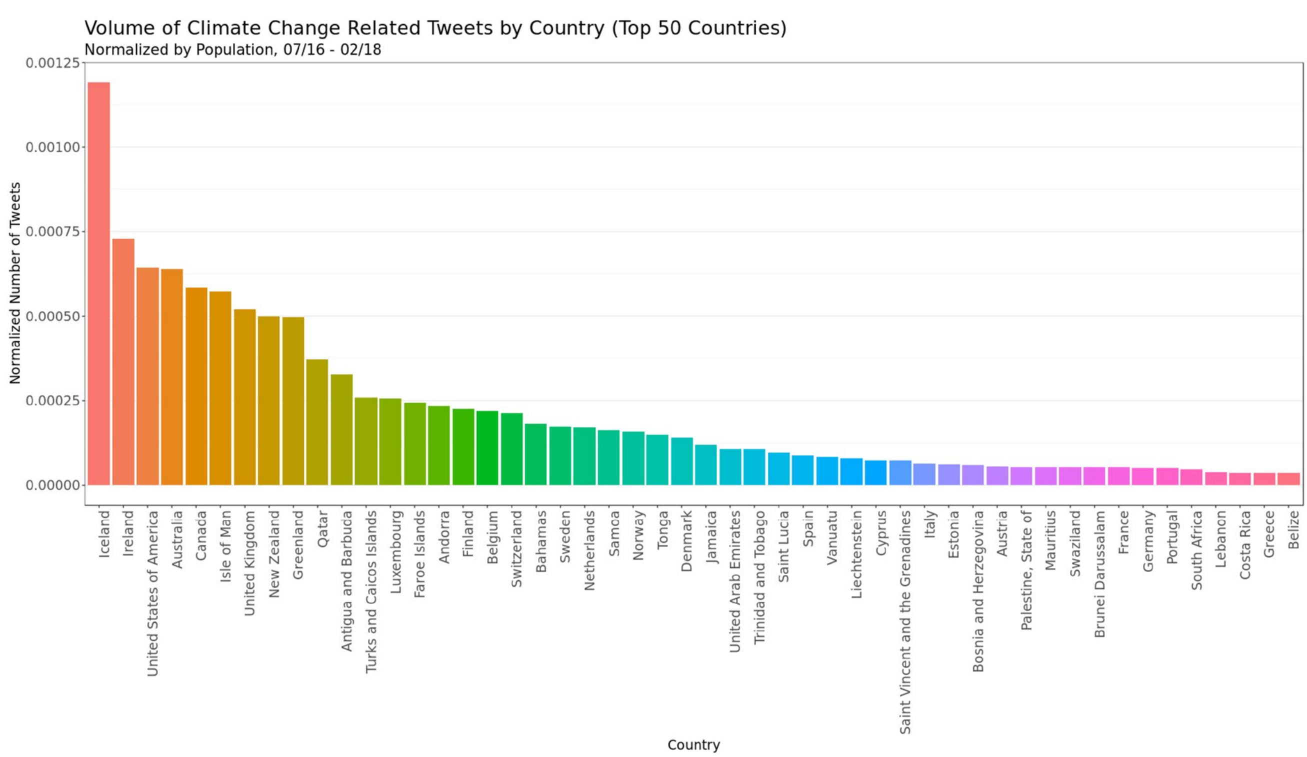


Fig. 7 Volume of tweets per country, normalized by population

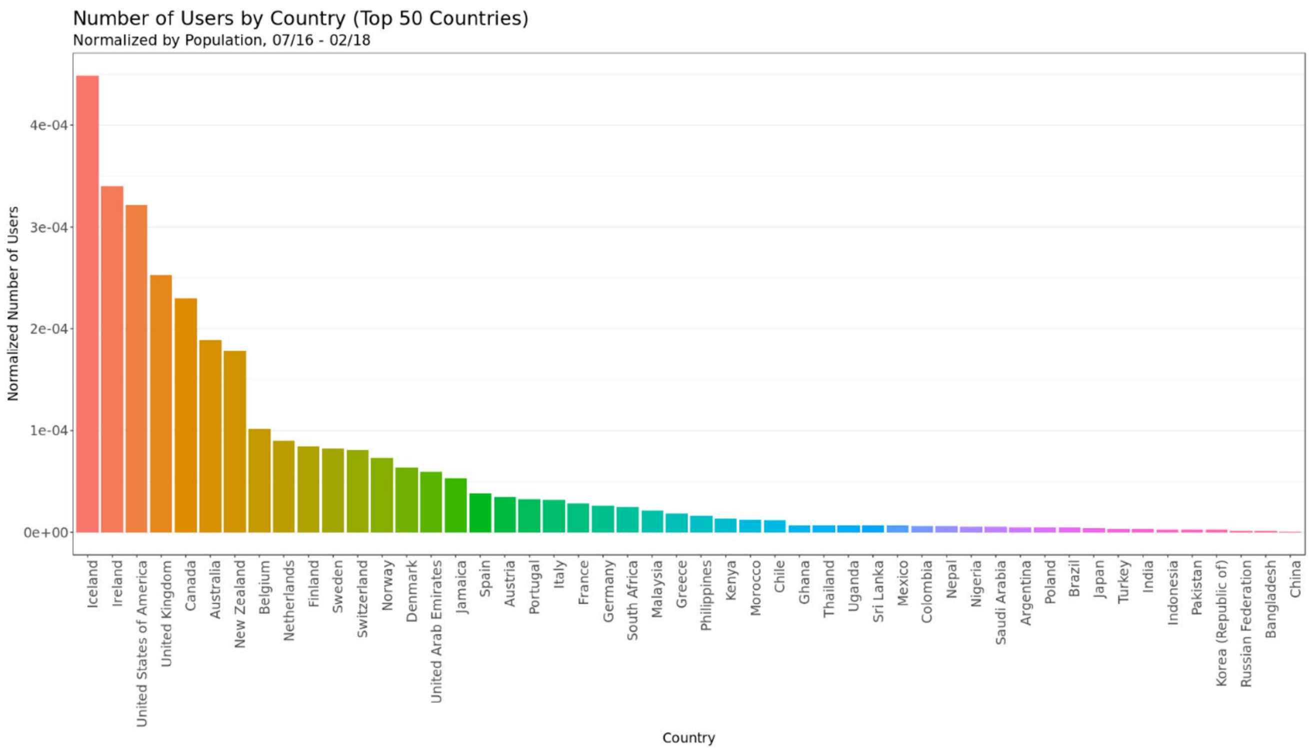
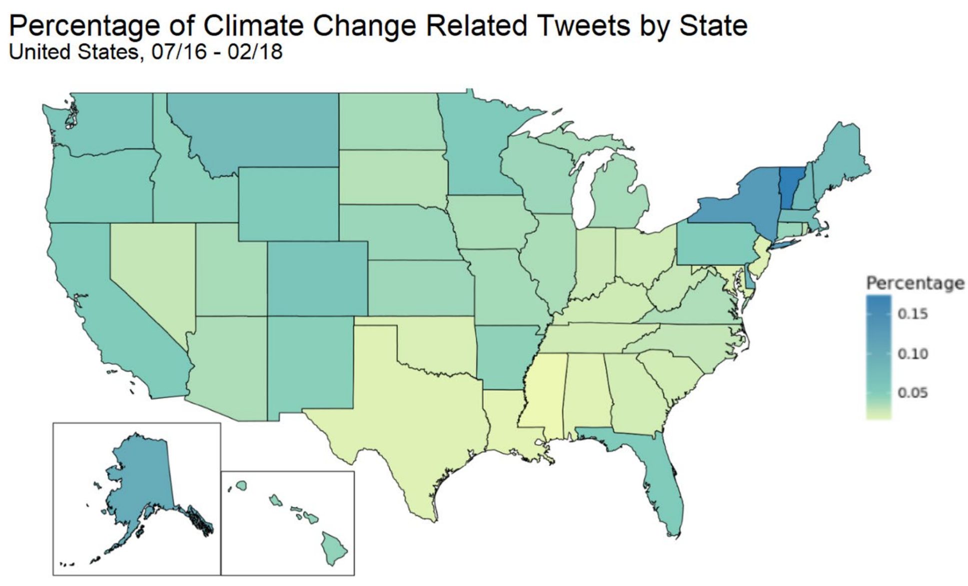


Fig. 8 Number of users per country, normalized by population

Fig. 9 Normalized number of tweets by US state



are another explanation; the southeast is a much warmer area, while the northeast is much more prone to cold temperatures and snow. Global warming may be discussed on twitter on particularly warm days in the northeast, whereas the warming on southeast states may be more difficult to detect and therefore there is less of a Twitter reaction.

When normalized by the total number of geotagged tweet, the percentage of climate change discussion (out of all geotagged tweets) for coastal states is 0.0369 while the percentage of inland states is 0.0350. This is not a huge difference despite coastal states being more directly affected by sea-level rise and storm surges, two major projected effects of climate change (IPCC 2014).

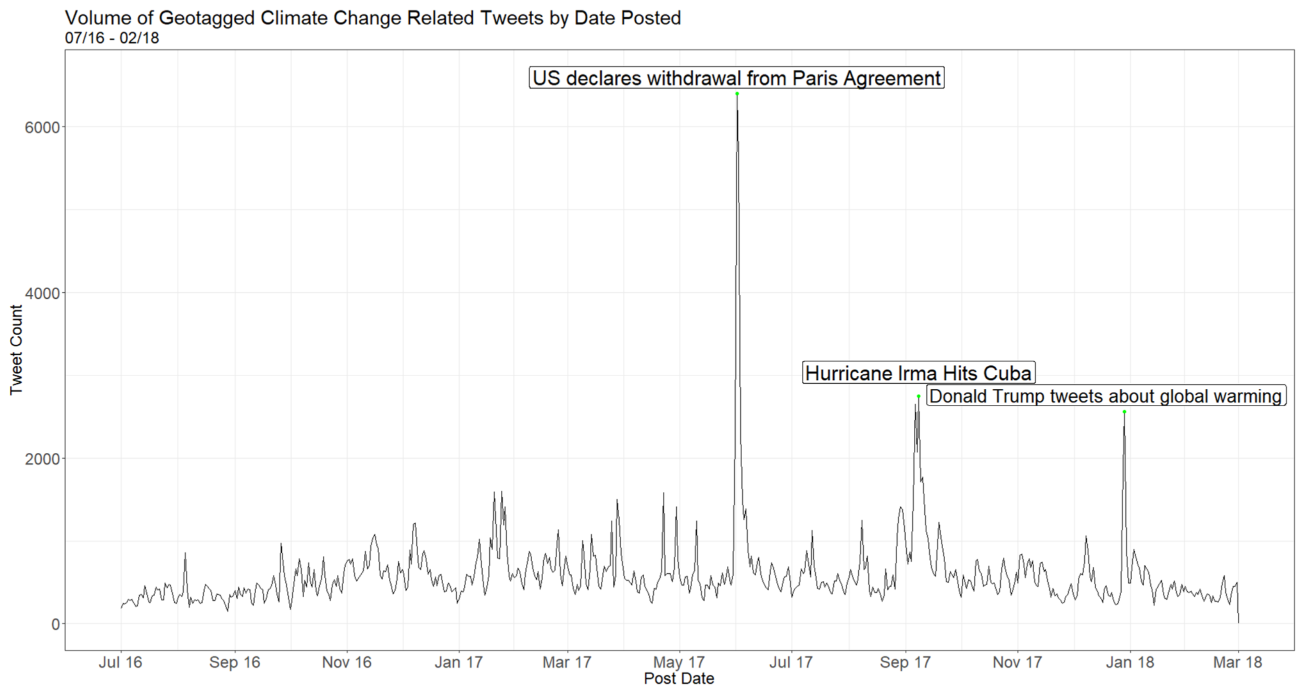


Fig. 10 Number of tweets per day

4.1.2 Temporal

Figure 10 shows the number of tweets in the dataset per day. Several peaks in the dataset are labeled with significant climate change-related events that occurred on the same day.

The spikes in tweet volume coincide with the US declaration of intent to withdraw from the Paris Climate Accords on June 1, 2017, Hurricane Harvey's formation in the second half of August 2017, Hurricane Irma's formation early September 2017, and a divisive tweet posted by US President Trump on the subject of global warming on December 28, 2017.

Many of the spikes in tweets were associated with events in the USA or events that most affected the USA. This makes sense given that over half of the dataset was from the USA. When a significant event occurs, many US Twitter users post tweets commenting on the event. Twitter users outside the USA can see these tweets and reply to them, or they may have heard about the event in the news and chose to voice their opinion through a tweet.

On the other hand, the most significant spike was in reaction to the US's withdrawal from the Paris Climate Accords, which was both a domestic event in that it affected US policy going forward and an international event since the agreement was signed by the vast majority of countries. In addition, Hurricane Irma affected many other countries in the Atlantic such as Cuba, so it was more of an international event. This suggests that while climate change Twitter discussion is heavily influenced by events affecting the USA, those events that are also international have even more discussion.

4.2 Sentiment analysis

The total sentiment per day can be calculated by summing the sentiments (+ 1, 0, or − 1) scores of every tweet on a particular day. Figure 11 shows a time series plot of the total sentiment per day over the entire world with some labeled extrema.

Note that many peaks in positive sentiment are on days with warmer than average weather which is often seen positively in cold regions in autumn or winter, such as in September 2016 and September 2017. The largest peak occurred on March 25, 2017, the day of Earth Hour in 2017, a worldwide environmental moment encouraging people to turn off many electric lights for 1 h. Many tweets on that day were speaking positively about Earth Hour and the environment, which explains the spike in positive sentiment. Most spikes in negative sentiment coincide with spikes in the volume of tweets as shown in Fig. 10, which means that when there is significant climate change discussion on twitter, the discussion is usually negative. Since climate change is treated as a political issue by many Twitter users, this result makes sense. Looking at the individual tweets themselves, the debate on whether climate change is actually occurring is intense on either side. Many tweets that support or deny climate change are filled with derogatory and demeaning language toward the other side.

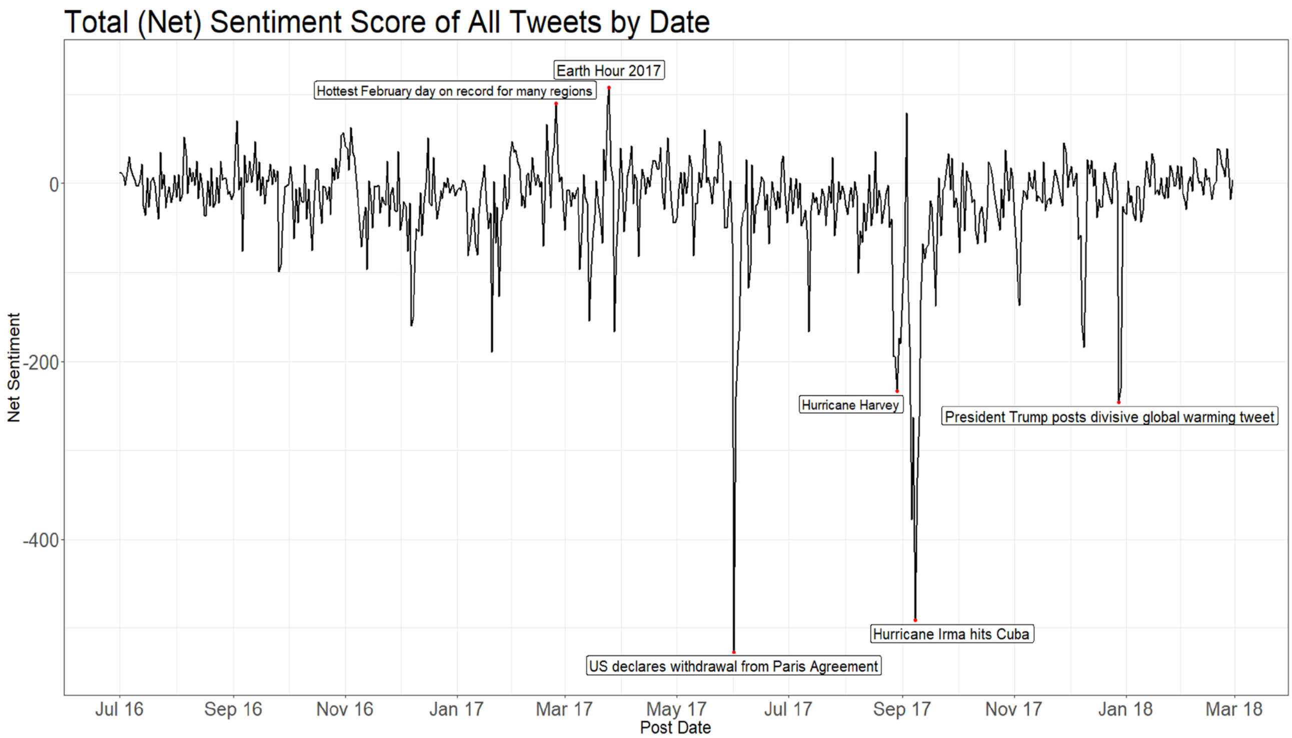


Fig. 11 Total sentiment of tweets by day

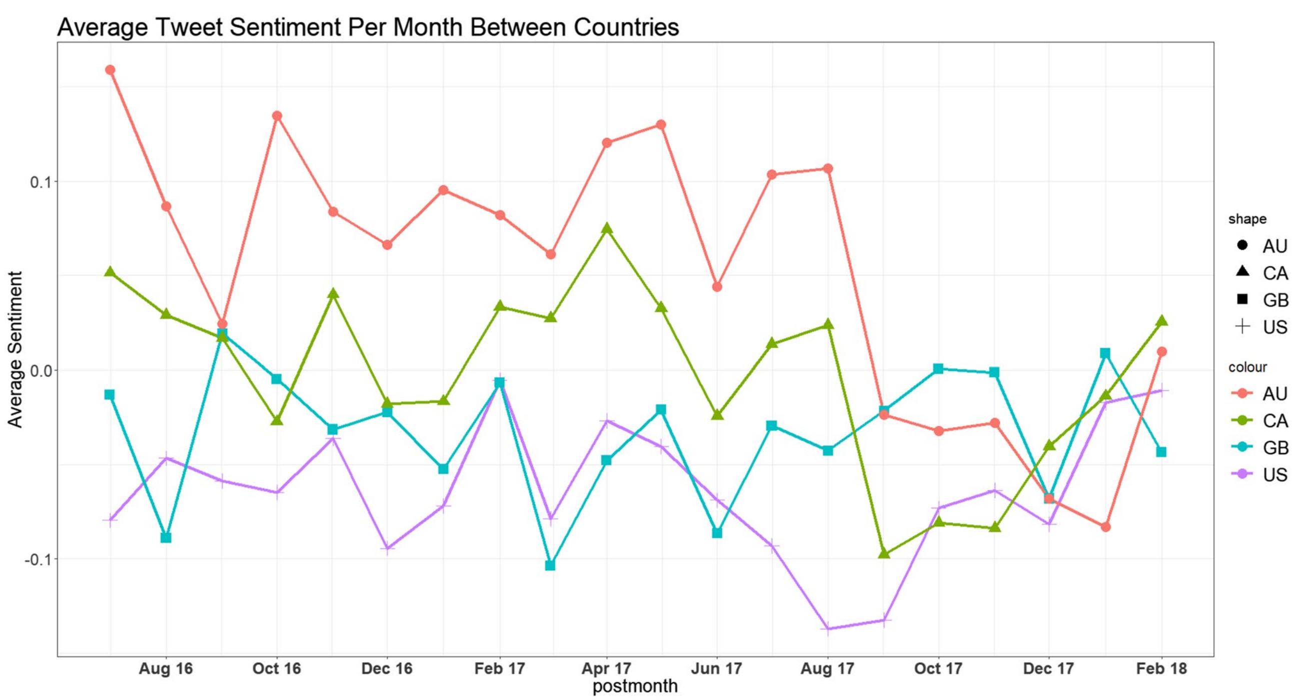


Fig. 12 Average sentiment per month between top 4 countries

The largest peak in positive sentiment occurred during Earth Hour, but this did not result in any of the largest spikes in overall tweet volume. This suggests that compared to planned, top-down advocacy movements, the public reacts more to drastic political, weather, or social media events. This is similar to the findings of Leas et al. (2016) which observed that climate change-related Twitter engagement after Leonardo DiCaprio's Oscar acceptance speech citing climate change spiked much higher than during Earth day or the 2015 United Nations Climate Change Conference. The Twitter masses seemed to be stirred to discuss climate change more in reaction to significant one-time events than planned top-down approaches for climate or green advocacy.

Figure 12 shows the average (mean) sentiment of all tweets per month by country among the top 4 countries in terms of total volume of tweets—Australia, Canada, the UK, the USA. The mean was taken to facilitate a direct comparison between the different countries. The average sentiment per country (for all months) is 0.0643 for Australia, −0.0061 for Canada, −0.0366 for the UK, and −0.0697 for the USA.

Many peaks and troughs are shared between countries. The most notable exceptions to this are around September 2017, when the sentiment of all countries was on a decline from August with the exception of the UK. This suggests that the Twitter conversation in the Great Britain (GB) about climate change was least affected by the 2017 hurricane season which seemed to greatly affect the conversation the other countries. Another exception where the UK sentiment increased but all other countries decreased is on September 2016.

4.3 Topic modeling

Experiments showed that using 20 topics produced the best topic models for the dataset. In addition, choosing 20 topics is a manageable number that also allows for a wide variety of subjects to appear in the different topics.

Section 4.3.4 contains a comparative evaluation of the performance of the author-pooled LDA technique with 20 topics against some other recent topic modeling techniques.

4.3.1 Author-pooled LDA with 20 topics

Many different topic models were produced by running author-pooled LDA with 20 topics with different random seeds. One sign of a good topic model is when the topics can be given a label that corresponds to the top words in that topic. Table 3 displays one good author-pooled LDA topic model with topic labels. Figure 13 displays the same topic model but with word clouds. The top 15 most likely words per topic are displayed. The labels were chosen by looking at the top 50 words per topics with their probabilities and examining the tweets that are most likely per topic. Some cells have the same topic; tweets that fall into either topic should be grouped together in the analysis. Cells with incomprehensible topics are not labeled.

While the topic model shown in Table 3 seems to have coherent, meaningful topics, it is important to note that any one topic model may not be able to describe the dataset. What is more important are the parameters used. Since LDA begins with a degree of randomness, it produces a slightly difference topic model every time, and the quality of the produced topic models can vary wildly. However, in general the topics produced are similar, and most iterations of author-pooled LDA with 20 topics had topics related to "energy," "carbon footprint," "weather," "politics," and "belief."

The relevant topics that could be labeled are transportation, international agreement, politics/hoax, nature extremes (hurricane)/belief, humanitarian, carbon footprint, solar panels/sustainability, environment, weather/belief, energy, weather, and fossil fuel industry. This is a very diverse collection of topics that covers politics, belief, economics, and environment.

Many of the topics in Table 3 are not related to climate change. This shows that topic modeling can be done to filter out irrelevant documents of tweets that may have been detected by the keyword search but have nothing to do with climate change. In this case, all tweets that with 50% or higher probability to be in either the unrelated or the French topic can be automatically filtered out as unrelated to climate change.

A few other topics seem to be incomprehensible. LDA is fundamentally a statistically trained model; it works to minimize some important statistics that do not always directly translate to human interpretability. In addition, there will inevitably be words with relatively little meaning, especially in a Twitter dataset, but LDA must assign them to some topic. The word cannot be discarded, so the topic it ends up being assigned to lose some meaning. However, most of the topics in Table 3 are indeed meaningful, which was the desired result.

Some of the subtle differences between topics are not clear when viewed from a human interpretation perspective. However, they all have a basis in the word co-occurrence statistics in the dataset. For example, the "Weather/Belief" and "Weather" topics in Table 3 seem very similar. However, the former topic includes words relating to belief that may have frequently co-occurred with weather words, such as "don't," "belief," whereas the latter topic was inferred from tweets that did not have too many of those belief words. The first topic was likely influenced by uses that expressed their belief or disbelief in climate change in response to weather, and the second topic may have been formed from users simply commenting on the weather and perhaps offhandedly mentioning climate change.

Given the number of tweets posted during the 2017 hurricane season, it is interesting that there was no topic completely dedicated to hurricanes in the model as seen in Table 3; all hurricanes fall under the nature extremes/belief category. A dedicated "hurricane" topic did appear in some of the other LDA trials, but the topic's presence is inconsistent. It makes sense that tweets on extreme natural

Table 3 Labels and top words for the 20 topics for one author-pooled LDA topic model

lol, years, year, cn, thats, believe, ago, control, people, climatechangeisreal, hey, changed, remission, time, good, old, record, word, science, left	Transportation: emissions, new, emission, zero, tax, pollution, sea, s, levels, air, york, vehicle, diesel, low, x, carbon, green, article, reduction, rising	International agreement: doesnt, climatechangeisreal, Trump, believe, Paris, exist, God, Parisagreement, gt, denial, like, president, America, Trumps, agreement, deal, world, USA, action, planet	Politics/hoax: Trump, hoax, man, scientists, al, gore, Donald, like, science, u, says, Chinese, Obama, know, dont, s, caused, humans, president, news
Nature extremes/belief: real, isnt, science, deniers, dont, people, Earth, Trump, fuck, said, think, fucking, tell, hurricane, believe, thing, like, nature, mother, right	Humanitarian: world, need, help, support, fight, talk, changes, women, children, time, people, economics, countries, solution, Africa, watch, adaptation, impact, planet, effects	Carbon footprint: carbon, emissions, footprint, dioxide, reduce, emission, says, reducing, threat, national, gms, low, study, security, help, plan, new, increase, Trump, cost	Solar panels/sustainability: carbon, solar, panel, stop, footprint, build, day, food, issue, use, reduce, UAE, methodology, issues, way, dont, homes, like, great, pay
Environment: environment, o, epa, dioxide, carbon, oxygen, effects, eu, e, pruitt, situation, hand, trees, air, scott, great, feel, environmental, globally, p	People, love, going, Earth, dont, want, im, money, like, truth, care, know, planet, polar, deny, need, cause, think, feel, world	Weather/belief: dont, believe, think, people, like, real, debate, yall, thing, degrees, involved, gonna, know, theres, shit, aint, wow, crazy, hot, November	Energy: energy, emissions, fossil, fuel, methodology, coal, power, clean, ideology, need, capitalism, fuels, nuclear, solar, renewable, industry, future, economy, jobs, policy
Emissions, stop, climateaction, acton climate, car, conference, trndnl, vw, die, climatechangeinwords, scandal, heat, vegan, Nigeria, test, fr, world, documentary, cars, s	French (unrelated): la, lemissionpolitique, le, en, di, e, emission, il, que, les, et, el, emissioni, non, y, l, cest, je, animal, pas	Weather: im, weather, like, cold, winter, fake, right, news, snow, ice, hot, today, degrees, sure, getting, day, know, February, yes, warm	Cancer (unrelated): remission, great, today, art, peace, thanks, future, work, chat, thank, day, cop, lets, join, im, health, check, cancer, action, pm
Fossil fuel industry: fossil, fuels, fuel, oil, industry, burning, companies, like, stop, coal, need, big, solar, reason, planet, f, time, auspol, divest, world	sbi, denier, u, water, lol, surviving, dr, guys, ginoclock, understand, Texas, real, mission, travel, s, thoughts, meat, difference, hes, arctic	Emissions, gas, greenhouse, time, human, UK, youre, called, methane, population, hoax, Tesla, contribution, world, caused, higher, na, activity, effect, cause	Rights, res, good, white, job, house, anti, pa, rescue, w, save, world, hate, website, logic, mammal, lgbt, r, abortion, human

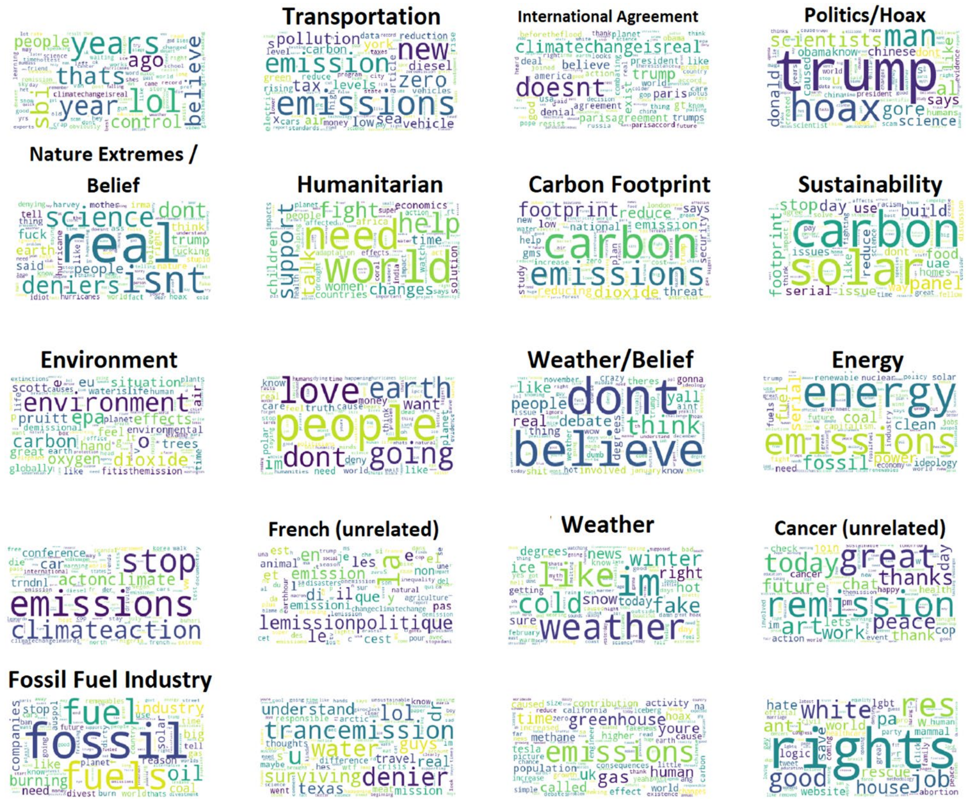


Fig. 13 Word clouds for an author-pooled LDA topic model. The size of a word within a topic's word cloud is proportional to the probability of the word within the topic

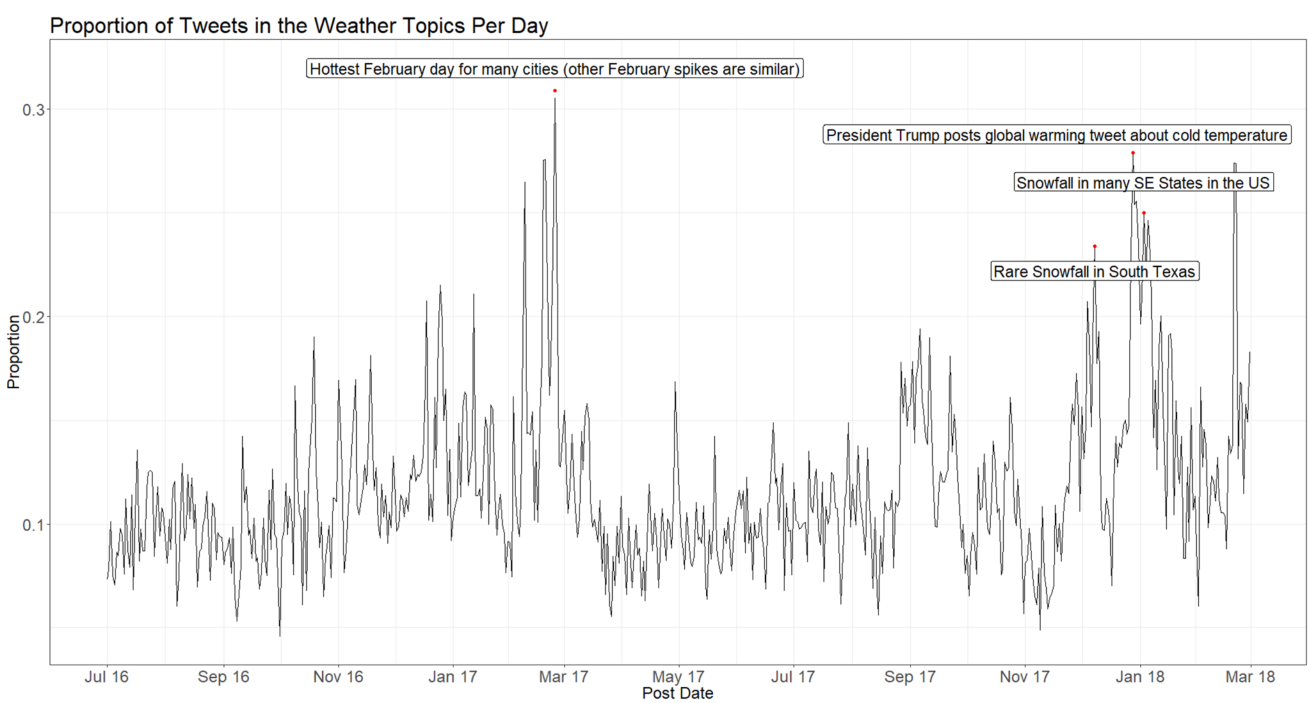


Fig. 14 Proportion of "weather" topic discussion per day, including topics 10 and 14

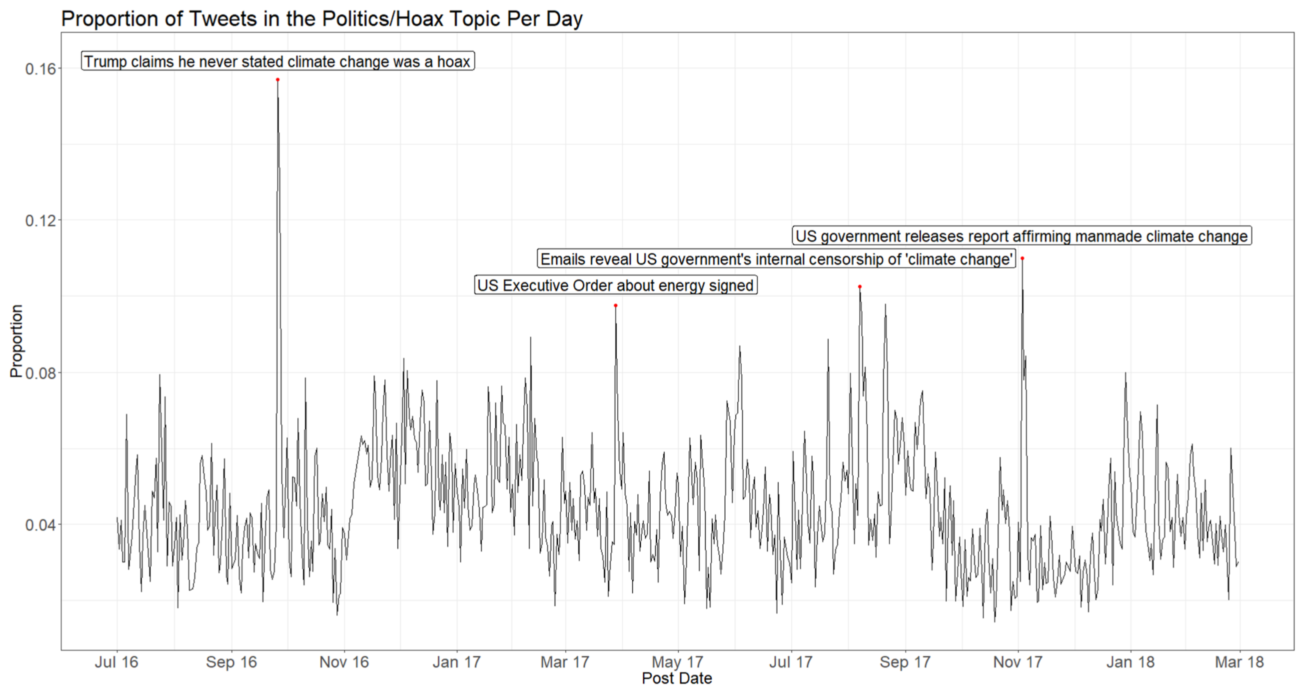


Fig. 15 Proportion of "politics/hoax" topic discussion per day

events, hurricane or otherwise, would be grouped together, but words relating to "belief" seem to dominate the topic. One possible explanation for this phenomenon is selection bias. For example, one hurricane-related tweet has the message "Harvey; Irma; now Jose two hurricanes and a tropical storm and people still think climate change isn't real?" Note that this tweet brings up many hurricane-related terms, but also discusses the truth of climate change. Most of the hurricane-related tweets in the dataset share similar features to this message. If the part about climate change was not included, then the tweet would not be included in the dataset. Thus, it is expected that many hurricane-related tweets also discuss the user's belief (or lack thereof) in climate change. Vocabulary about the validity of climate change or global warming will then populate the topic, meaning the topic will not purely contain hurricane-related words.

4.3.2 Temporal analysis of the topics

Each tweet in the dataset was assigned to its most probable topic. By doing this, individual topics can be isolated for further study. The percentage of topic discussion can be examined for any topic or group of topics. Figure 14 displays a time series plot of the proportion of tweets belonging to the weather topics ("weather" and "weather/belief") for the topic model shown in Table 3 out of all tweets by day. Spikes in discussion of the weather topics coincide with extreme weather events. Many of these specific days coincided with extreme weather in the USA; this is expected since the majority of tweets in the dataset were from the USA.

Figure 15 displays a similar time series plot of the proportion of tweets belonging to the "politics/hoax" topic. Note that virtually all of the spikes coincide with climate change-related political events in the USA. This suggests that this topic has a heavy US bias. 6.06% of all USA tweets was in the politics/hoax topic, compared to 4.13% UK, 5% Canada, 3.91% Australia. This confirms that the topic was most prevalent in the USA, but the discussion was not too low for the other top 4 countries.

4.3.3 Spatial analysis of the topics

Topics among the countries are examined. The topics produced from running author-pooled LDA with 20 topics for each of the top countries were similar. The models usually contained "energy," "footprint," "politics," and "belief" topics of some form. The topic model on the US corpus was almost identical to the topic model on the world corpus. The topic model on the Great Britain (GB) corpus was somewhat similar as well, though the quality and meaningfulness of the topics was diminished. Please note that The International Organization of Standardization (ISO) defines codes for the names of countries in ISO 3166. ISO 3166-1 alpha-2 is two-letter country codes, and in that standard the "United Kingdom of Great Britain and Northern Ireland" is assigned the code GB. Canada had many topics that included words on the subject of economics or urbanization. Australia had

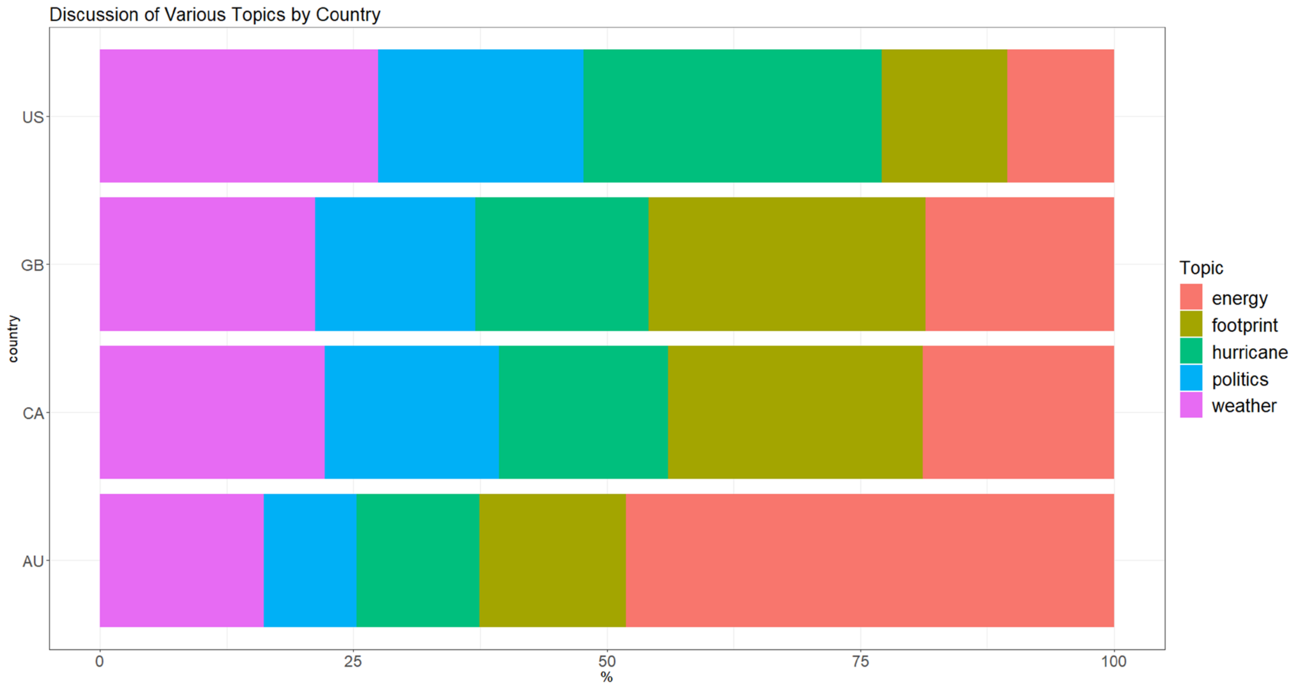


Fig. 16 Topic discussion for select topics by country

many energy-related topics with seemingly little difference between them.

However, the models on the GB, Canada, and Australia corpuses were poor overall; they had significant topic overlap and much less topic interpretability. This is likely caused by the small corpus size of those countries compared to the USA and the whole world corpuses. There were often multiple topics that were virtually indistinguishable from each other. In addition, many topics seemed to encompass multiple subjects at once. It was not feasible to isolate any subjects of discussion and calculate the proportion of discussion with that subject, because that would require including multiple "topics" in the topic model per subject that each have significant noise. Thus, the approach of calculating the proportion of topic discussion by running LDA on country corpuses is not optimal.

To address this issue, just one topic model is trained on the entire corpus, and the topics produced from this model are then used to analyze the percent of topic discussion per country.

This approach was done using the topic model shown in Table 3. For each of the four countries, tweets in topics 3, 4, 6, 11, and 14 were isolated and the percentage of tweets in those topics, corresponding to "Politics/Hoax," "Nature Extremes (hurricanes)/Belief," "Carbon Footprint," "Energy," and "Weather," respectively, was calculated. Figure 16 displays these percentages. Note that for each topic, the percentage discussion is defined as the volume of tweets on that topic divided by the volume of tweets from all five of the topics times 100.

Notice that the weather, politics, and hurricane topics dominate the US discussion compared to other countries. This indicates that tweets from the USA are more reactionary toward weather or political events than other countries, and this is likely caused by the significant 2017 hurricane season and change in political administration. The carbon footprint and energy topics were chosen because tweets and words in that topic were less reactionary in nature and often referred to future policies or plans. This suggests that users in the USA are behind in the discussion of policy or plans to address climate change compared to the other top 4 countries.

4.3.4 Topic model performance evaluation

To evaluate the quality of the model produced by topic modeling through author-pooled LDA, the metric of UMass coherence will be examined. The UMass coherence measure is a word co-occurrence statistic that assesses topic quality by looking at how frequently words within a topic co-occur in the corpus, a measure of whether a topic has human-identifiable semantic coherence. A topic with a good UMass coherence score contains words that are consistent with each other, i.e., the words are more likely to co-occur in the same documents, whereas a topic with a bad UMass coherence score will contain words that do not occur in the same documents together. See Röder et al. (2015) for the formula used to calculate UMass coherence as well as more details on the metric itself.

The quality of author-pooled LDA will be compared to three other techniques used in recent papers employing topic modeling. Karami et al. (2018) applies LDA without any pooling to tweets to characterize comments related to diet and exercise. This is standard LDA.

Yan et al. (2013) proposes a biterm topic model (BTM) for short texts that, unlike LDA, models the generation of word co-occurrence patterns through the generation of biterms (pairs of words) rather than individual words. Like LDA, each document has a topic distribution, but the topic distributions are used to generate biterms rather than individual words themselves. This is the same as applying LDA to a corpus where each document is converted from a bag of words to a bag of biterms (pairs of words).

Lastly, Steinskog et al. (2017) introduces the idea of pooling tweets that share the same hashtag. A hashtag is a word or phrase preceded by a hash symbol (#) that Twitter users use to align their tweets with a specific topic. 132,782 of the tweets in this study's corpus contain hashtags. Each document in this hashtag-pooled model represents a hashtag and contains the text of every tweet containing that hashtag.

The three techniques in the preceding paragraph will be compared against author-pooled LDA (the technique used in this paper). Five topic models each will be produced using the following techniques: author-pooled LDA, LDA (unpooled), BTM, and hashtag-pooled LDA. The same hyperparameters and corpus cleaning methods will be done for each of the models. Repeated here for convenience, this means removing all stopwords, keeping only words that occur more than 200 times in the corpus, producing 20 topics, and using 0.05 for the value of the symmetric hyperparameters α and η in the model inference process. Note that the filtering of words occurring more than 200 times is done before the conversion of the bag of words to bag of biterm in BTM.

Table 4 shows the average UMass coherence score for each topic modeling technique. To obtain the UMass coherence of a topic model, the UMass coherence scores of each topic are summed. Thus, this value represents the aggregate quality of a topic model. These results show that author-pooled LDA outperforms both LDA and BTM. However, hashtag-pooled LDA has a somewhat better UMass coherence score. Despite this, author-pooled LDA is a better topic model for the dataset because only 36% of the Twitter climate dataset was accounted for in hashtag-pooled LDA, as only a fraction of tweets contained hashtags. Hashtag-pooled LDA requires abandoning most of the dataset for only marginal improvement in coherence. This comparison suggests that hashtag-pooled LDA is probably a better choice for a dataset of tweets all containing hashtags, but for the climate dataset author-pooled LDA is the better choice as every tweet in the dataset is accounted for in the model.

Table 4 Comparison of UMass coherence between topic models produced by different topic modeling techniques. Five topic models were produced for each technique, and the coherence was averaged

Topic modeling technique (20 topics)	Average UMass coherence
Author-pooled LDA	−3.5494
LDA	−6.3745
BTM	−12.4762
Hashtag-pooled LDA	−2.5932

5 Limitations and future work

One major limitation of this study is the nature of the Twitter dataset itself. Many tweets in the dataset are indecipherable, making both topic modeling and sentiment analysis ineffective on those tweets. Additionally, many tweets include hashtags and mention other users in place of using plain English to describe their message. This makes it difficult to perform text mining, as the important context of the tweet (the conversation) is not available. Future data collection could potentially preserve the context information of a tweet so it could be used for analysis. This information can improve the performance of topic modeling and sentiment analysis because the qualities of other tweet in the conversation context can provide information to classify the tweet in question.

One other limitation is that only tweets in English could be examined. Though topic modeling was able to identify the tweets written in French, the sentiment analysis method is forced to classify those tweets as neutral since it was developed with English in mind. One way to overcome this limitation is to translate the tweets into English. If enough tweets in foreign languages were available, a supervised sentiment analysis technique could be trained that can handle multiple languages.

One last limitation is the issue of performing topic modeling on a Twitter corpus, especially one that is pre-filtered. While pooling by author before performing LDA greatly improved the topic models produced, the models produced are not as good as the topic models produced with corpuses from other domains. However, specific features of Twitter such as the hashtag system and the fact that tweets are usually only on one topic, in addition to using the knowledge about the keywords used in the pre-filtering can be harnessed to improve the performance of topic modeling.

As a future research topic, it is possible that topic modeling and sentiment analysis could be performed together to produce topics that have some sentiment alignment (positive or negative) associated with them. This could further differentiate topics based on not only word choice but also overall sentiment of the tweet. For example, it could be possible that the word "hoax" is used more often in tweets with negative sentiment, and the word "science" is used more often in tweets with positive sentiment, and this fact could be used to put those words in different topics.

In addition, considering the time of the day that the tweet was posted could be an additional step done for sentiment analysis to determine the time periods within the day that users tend to be more negative or positive during a special social or political event. This could be done independently or paired with topic modeling to show the evolution of sentiment minutes or hours after a special event. This analysis, or any other temporal analysis done in this paper, can be combined with a spatial analysis in some complex analysis. This paper only discussed the two aspects separately, but a joint spatiotemporal analysis could have some value.

6 Conclusion

This study analyzed the discussion of climate change on Twitter from geospatial and temporal perspectives. Topic modeling was used to automatically extract the different subjects of discussion, and sentiment analysis was used to classify tweets as positive, neutral, or negative. These techniques were especially useful for event detection, as spikes in topic discussion or total sentiment were in direct correspondence with significant real-world events.

The discussion of climate change on Twitter is negative overall; it seems that users discuss climate change primarily as part of a usually negative reaction to current events. Both sides of the debate on whether climate change was occurring used demeaning language to refer to the other side, indicating that the debate is heated, potentially due to its political nature. The only regular peaks in positive sentiment are during warm autumn and winter days. This suggests that as the climate changes over time, Twitter users will speak happily when there are warmer days in the winter but speak more negatively during most other events.

Topic modeling with LDA inferred a wide variety of topics, covering many different aspects of the climate change discussion including politics, economics, weather, belief, and the environment. These topics were shared between countries, but the quantity differed. Compared to the UK, Canada, and Australia, the USA is behind in its discussion of topics related to policies and plans to address climate change. Within the USA, the southeast coast has proportionally low discussion of climate change, potentially due to the politics or warm climate in the area. For the development of mitigation policies, Twitter users in the USA should discuss more about their opinions on policy or potential plans. Otherwise, the USA, especially its southeastern states, is at risk of falling behind other countries in climate change mitigation.

Acknowledgements This work was supported by National Science Foundation (Grant No. AGS 1560210) and South Carolina Research Foundation, University of South Carolina (USA) (Grant No. ASPIRE-I 13540-17-44820).

References

Alampalayam SP, Kumar A (2004) Predictive security model using data mining. In: Global telecommunications conference, 2004. GLOBECOM'04, vol 4, pp 2208–2212. IEEE

Alampalayam SK, Natsheh EF (2008) Multivariate fuzzy analysis for mobile ad hoc network threat detection. Int J Bus Data Commun Netw (IJBDCN) 4(3):1–30

Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Cody EM, Reagan AJ, Mitchell L, Dodds PS, Danforth CM (2015) Climate change sentiment on Twitter: an unsolicited public opinion poll. Plos One. https://doi.org/10.1371/journal.pone.0136092

Ghosh D, Guha R (2013) What are we "tweeting" about obesity? Mapping tweets with topic modeling and geographic information system. Cartogr Geogr Inf Sci 40(2):90–102. https://doi.org/10.1080/15230406.2013.776210

Hutto CJ, Gilbert EE (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international conference on weblogs and social media (ICWSM-14), Ann Arbor, MI

IPCC (2014) Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Climate change 2014: synthesis report. IPCC, Geneva, Switzerland

Karami A, Dahl A, Turner-McGrievy G, Kharrazi H, Shaw G (2018) Characterizing diabetes, diet, exercise, and obesity comments on Twitter. Int J Inf Manag 38(01):1–6

Kirilenko AP, Stepchenkova SO (2014) Public microblogging on climate change: one year of Twitter worldwide. Glob Environ Change 26:171–182. https://doi.org/10.1016/j.gloenvcha.2014.02.008

Koto F, Adriani M (2015) A comparative study on Twitter sentiment analysis: which features are good? In: Natural language processing and information systems. NLDB 2015. Lecture notes in computer science, pp 453–457. https://doi.org/10.1007/978-3-319-19581-0_46

Lansley G, Longley PA (2016) The geography of Twitter topics in London. Comput Environ Urban Syst 58:85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Leas EC, Althouse BM, Dredze M, Obradovich N, Fowler JH et al (2016) Big data sensors of organic advocacy: the case of Leonardo DiCaprio and climate change. PLoS ONE 11(8):e0159885

Lee GM, Qui L, Whinston AB (2016) A friend like me: modeling network formation in a location-based social network. J Manag Inf Syst 33(4):1008–1033

Mandel B, Culotta A, Boulahanis J, Stark D, Lewis B, Rodrigue J (2012) A demographic analysis of online sentiment during Hurricane Irene. In: Proceedings of the second workshop on language in social media

Manning C, Raghavan R, Schütze H (2008) Introduction to information retrieval. Retrieved from https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 5(4):1093–1113

Mehrotra R, Sanner S, Buntine W, Xie L (2013) Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval-SIGIR 13. https://doi.org/10.1145/2484028.2484166

Omnicore Agency (2018) Twitter by the numbers (2018): stats, demographics & fun facts. Retrieved from https://www.omnicoreagency.com/twitter-statistics/

Perrin A (2015) Social network usage: 2005–2015. Pew Research Center. Retrieved from http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/

Raj P, Kumar SA (2017) Big data analytics processes and platforms facilitating smart cities. Smart cities: foundations, principles, and applications. Wiley, pp 23–52

Rehurek R, Sojka P (2010) Software framework for topic modeling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, pp 45–50

Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of ACM international conference on web search and data mining

Saad L (2017) Global warming concern at three-decade high in U.S. Gallup. Retrieved from https://news.gallup.com/poll/206030/global-warming-concern-three-decade-high.aspx

Shi Z, Lee GM, Whinston AB (2016) Toward a better measure of business proximity: topic modeling for industry intelligence. MIS Q 40(4):1035–1056

Steinskog A, Therkelsen J, Gambäck B (2017) Twitter topic modeling by tweet aggregation. In: Proceedings of the 21st nordic conference on computational linguistics, pp 77–86

Thampi A (2016) Reverse geocoder [computer software]. Retrieved from https://github.com/thampiman/reverse-geocoder

Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Proceedings of the fourth international AAAI conference on weblogs and social media

Van der Linden S (2017) Determinants and measurement of climate change risk perception, worry, and concern. In: Nisbett M (ed) Oxford research encyclopedia of climate science. Oxford University Press, Oxford. https://doi.org/10.1093/acrefore/9780190228620.013.318

Xu BH, Kumar SA (2014) Big data analytics framework for improved decision making. In: Proceedings on the international conference on internet computing (ICOMP) (p 1). The steering committee of the world congress in computer science, computer engineering and applied computing (WorldComp)

Xu B, Kumar S (2015a) A text mining classification framework and its experiments using aviation datasets. In: 25th workshop of information technology and systems (WITS), Dallas, TX, 2015 SSRN

Xu B, Kumar SA (2015b) Big data analytics framework for system health monitoring. In: 2015 IEEE international congress on big data (bigdata congress). IEEE, pp 401–408

Xu J, Wang S, Su S, Kumar SA, Wu C (2016) Latent interest and topic mining on user-item bipartite networks. In: 2016 IEEE international conference on services computing (SCC). IEEE, pp 778–781

Yan X, Guo J, Lan Y, Cheng X (2013) A biterm topic model for short texts. In: Proceedings of the 22nd international conference on world wide web, pp 1145–1456

Zubiaga A, Spina D, Martínez R, Fresno V (2014) Real-time classification of Twitter trends. J Assoc Inf Sci Technol 66(3):462–473. https://doi.org/10.1002/asi.23186