



International Journal of Web Information Systems

Measuring climate change on Twitter using Google's algorithm: perception and events

Ahmed Abdeen Hamed Alexa A. Ayer Eric M. Clark Erin A. Irons Grant T. Taylor Asim Zia

Article information:

To cite this document:

Ahmed Abdeen Hamed Alexa A. Ayer Eric M. Clark Erin A. Irons Grant T. Taylor Asim Zia , (2015), "Measuring climate change on Twitter using Google's algorithm: perception and events", International Journal of Web Information Systems, Vol. 11 Iss 4 pp. 527 - 544

Permanent link to this document:

<http://dx.doi.org/10.1108/IJWIS-08-2015-0025>

Downloaded on: 17 March 2016, At: 11:09 (PT)

References: this document contains references to 34 other documents.

To copy this document: permissions@emeraldinsight.com

The fulltext of this document has been downloaded 149 times since 2015*

Users who downloaded this article also downloaded:

Colin Dey, (2002), "Methodological issues: The use of critical ethnography as an active research methodology", Accounting, Auditing & Accountability Journal, Vol. 15 Iss 1 pp. 106-121 <http://dx.doi.org/10.1108/09513570210418923>

Andréa Kalendová, (2000), "Spinels – reactive pigments affecting the diffusion of corrosive substances through the coatings", Pigment & Resin Technology, Vol. 29 Iss 4 pp. 215-221 <http://dx.doi.org/10.1108/03699420010339326>

Bill Merrilees, Don Getz, Danny O'Brien, (2005), "Marketing stakeholder analysis: Branding the Brisbane Goodwill Games", European Journal of Marketing, Vol. 39 Iss 9/10 pp. 1060-1077 <http://dx.doi.org/10.1108/03090560510610725>

Access to this document was granted through an Emerald subscription provided by emerald-srm:234603 []

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

*Related content and download information correct at time of download.

Measuring climate change on Twitter using Google's algorithm: perception and events

Ahmed Abdeen Hamed

EPSCoR, University of Vermont, Burlington, Vermont, USA

Alexa A. Ayer

Rubenstein School, University of Vermont, Burlington, Vermont, USA

Eric M. Clark

*Department of Computer Mathematics, University of Vermont,
Burlington, Vermont, USA, and*

Erin A. Irons, Grant T. Taylor and Asim Zia

EPSCoR, University of Vermont, Burlington, Vermont, USA

Measuring
climate
change on
Twitter

527

Received 24 August 2015

Revised 30 August 2015

Accepted 30 August 2015

Abstract

Purpose – The purpose of this paper is to test the hypothesis of whether more complex and emergent hashtags can be sufficient pointers to climate change events. Human-induced climate change is one of this century's greatest unbalancing forces to have affected our planet. Capturing the public awareness of climate change on Twitter has proven to be significant. In a previous research, it was demonstrated by the authors that public awareness is prominently expressed in the form of hashtags that uses more than one bigram (i.e. a climate change term). The research finding showed that this awareness is expressed by more complex terms (e.g. "climate change"). It was learned that the awareness was dominantly expressed using the hashtag: #ClimateChange.

Design/methodology/approach – The methods demonstrated here use objective computational approaches [i.e. Google's ranking algorithm and Information Retrieval measures (e.g. TFIDF)] to detect and rank the emerging events.

Findings – The results shows a clear significant evidence for the events signaled using emergent hashtags and how globally influential they are. The research detected the Earth Day, 2015, which was signaled using the hashtag #EarthDay. Clearly, this is a day that is globally observed by the worldwide population.

Originality/value – It was proven that these computational methods eliminate the subjectivity errors associated with humans and provide inexpensive solution for event detection on Twitter. Indeed, the approach used here can also be applicable to other types of event detections, beyond climate change, and surely applicable to other social media platforms that support the use of hashtags (e.g. Facebook). The paper explains, in great detail, the methods and all the numerous events detected.

Keywords Web mining, Applications of Web mining and searching, Web structure/linkage mining, Web design metrics, Web media

Paper type Research paper

This research is partly funded by Vermont EPSCoR award number EPS-1101317. We acknowledge the Vermont Advanced Computing Core which is supported by NASA (NNX 06AC88G), at the University of Vermont for providing High Performance Computing resources that have contributed to the research results reported within this paper. We particularly thank Dr Peter Dodds and Dr Chris Danforth for the Twitter garden-hose historical dataset.



International Journal of Web
Information Systems
Vol. 11 No. 4, 2015

pp. 527-544
© Emerald Group Publishing Limited
1744-0084
DOI 10.1108/IJWIS-08-2015-0025

Introduction

Climate change is a real fact (Neil Adger *et al.*, 2005). The governmental efforts spent for climate change gravitate toward focusing on providing public goods and growing the public awareness (Callaway, 2004). Inevitably, measuring the public awareness on climate change has become an important indicator for governments and policy-makers (Whitmarsh, 2009). In this pursuit, numerous efforts around the globe have strived to capture the awareness using different methods and approaches; Lorenzoni *et al.* (2007) studied how the general public in the UK perceive climate change challenges the barriers to engaging with it. The authors based their work on qualitative data that provided in-depth understanding of how the UK public makes sense of climate change. Curry (2004) designed a survey to study people's attitudes toward climate change mitigation. His study particularly focused on the public opinions concerning energy and carbon capture by answering 17 survey questions. Semenza *et al.* (2008), also designed a survey to study the climate change public awareness and concerns. The survey targeted individuals who live in Portland, Oregon, and Houston, Texas, in the period from June and September 2007. Sampei and Aoyagi-Usui (2009) analyzed Japanese newspaper coverage of global warming from January 1998 to July 2007, and how the public opinion was impacted by such coverage.

With social media exploding in popularity, public opinions have become more accessible than ever. This has ignited various efforts to take to Twitter feed as sensors for such climate change perceptions. Particularly, Kirilenko *et al.* (2015) studied the linkage of how local temperature affects the awareness on Twitter. From the Twitter data they collected, the study confirmed that the Twitter populations indeed recognize extreme temperature exceptions which is expressed in complex terms such as "Climate Change" or "Global Warming". Similarly, Hamed and Zia (2015) used Google's Algorithm, as a surrogate that demonstrates the public awareness, to rank climate change terms revealed on Twitter. Though, recent studies have proven beneficial, more perspectives are yet to be investigated, for instance, #ActOnClimate, which is the event that signals President Obama's action plan to respond to Climate Change, were out of scope.

This research takes a deeper look at the emergent hashtags to capture the current climate change events (events) communicated by the public on Twitter. The premise of this paper is to mine and investigate newly emerging hashtags that signal events related to climate.

Data description

Words are powerful means in which they make up the literature that we currently possess. At the time of writing, it was not feasible to get access to the Climate Change Ontology vocabulary (Esbjörn-Hargens, 2010). We turned to text mining climate change literature as an interesting alternative. Literature expresses a tremendous range of knowledge in the text within. This fact, though useful, also represents a computational challenge in terms of extracting highly representing word features (Hearst, 1999). Using means of text mining becomes necessary to acquire those features. Constructing feature-based networks of such features establishes a notion of ground-truth to compare against. Mining significant word features, designing data models and algorithms are evident (Jing *et al.*, 2002; Hearst, 1999; Forman, 2003; Sebastiani, 2002). Text mining is popular in solving many computational problems in biology, medical science, social

sciences and earth sciences (Ye, 2003; Jensen *et al.*, 2006; Marsi *et al.*, 2014; Macintyre *et al.*, 2014; Pang and Lee, 2008). Here, we extracted word features from literature gathered from the *Nature Climate Change* journal. This computational task is a step in measuring the public awareness of climate change on Twitter.

Two different data sets were collected independently:

- (1) *Literature source*: We searched the Web portal of *Nature Climate Change* for the keyword “climate” between year 1990 and 2014. The search produced 3,263 articles from which we extracted the titles.
- (2) *Twitter source*: We harvested 185 million tweets using the Twitter streaming API, which provides a 1 per cent free sample from all the public tweets, in the period of July 2014 and August 2015.

We also harvested over five million tweets mentioning a climate-related key word from an exhaustive search from the 10 per cent “garden hose” random sample of Twitter spanning 2012 through 2014.

Term-ranking method

As for the preprocessing for the literature titles, we applied a noise removal task to eliminate the stop words. Titles are rich sources of keywords and have been used in various text-mining tasks (Kolchinsky *et al.*, 2010; Pardalos *et al.*, 2008; Do *et al.*, 2006). Titles are lengthy, and they often contain noise words that are not relevant to the study. Such words (e.g. and, this, were, that) do not contribute to the analysis and they have to be eliminated. After applying a preprocessing step for noise elimination, we then analyzed the pure titles to select the features that best describe the data set. These features are what we used to construct the ground-truth network (GTN) that is described in sections to follow. The data also contained duplicates, which we ended up removing so that the features selected are not biased by such duplications. This, in turn, reduced the number of titles to 2,084. A sample of the titles gathered from *Nature*’s Web portal when searching for the keyword “climate” is as follows:

- Blanket peat biome endangered by climate change.
- Breaking the climate change communication deadlock.
- Largest teach-in ever focuses US on climate change.
- A blind spot in climate change vulnerability assessments.
- Atmospheric chemistry: A new player in climate change.
- A meta-analysis of crop yield under climate change and adaptation.
- Adapting to climate change through urban green infrastructure.
- Assessment of the first consensus prediction on climate change.
- Climate-society feedbacks and the avoidance of dangerous climate change.
- High Arctic wetting reduces permafrost carbon feedbacks to climate warming.
- Physically based assessment of hurricane surge threat under climate change.

Here we provide raw titles, as they were originally published, and without the noise-removal step. These titles show some of the significant features extracted using TFIDF (e.g., climate, carbon).

Using the common TFIDF Information Retrieval scheme, we can select the best features that describe the corpus of titles. The TFIDF measure (Salton and Buckley, 1988) is a term-weighting scheme that is used to describe the product of term frequency and the inverse document frequency. This measure is considered one of the most commonly used term-weighting metric in Information Retrieval (Aizawa, 2003), text mining and summarization (Zhang *et al.*, 2011; Radev *et al.*, 2004). The TFIDF scheme can be described as follows:

$$tf - idf_{i,d} = tf_{i,d} \times idf_i = tf_{i,d} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

When applying the TFIDF measure against the data set, we gathered 2,260 word features; many of these features were duplicates. Filtering out the redundant features reduced the number of unique features to 712 instead. From these features, we collected the top two features of each title. This fosters a meaningful word-pairs mechanism that facilitates the construction of networks. This pairing process was further constrained but adding this following condition: for each title in the corpus, a word-pair must be adjacent (a bigram) to be consider in the GTN network. Domain bigram networks have proven informative (Xie *et al.*, 2011) and have been used to model textual content in various natural languages processing studies (Kam *et al.*, 2008; Levenbach, 1998). Bigrams are a much richer model than the (“Bag of Words”) (Bekkerman and Allan, 2003; Tan *et al.*, 2002) which has been previously used to construct Twitter K-H networks (Hamed *et al.*, 2015).

The process of constructing the bigrams generated 364 bigrams with 283 unique features. Such features were stored as a Comma Separated Value file and was imported as a network using the igraph Python library to construct the GTN as follows: for each bigram, we treated its sides, left hand side (LHS) and right hand side (RHS), as vertices connected by a directed edge from LHS to RHS, preserving the sequence of the words as it appears in the title. As for the edge weight, we assume a weight of 1, giving it full credit, as this is the ground-truth/benchmarking knowledge we calculated from literature to compare against. By repeating this process for all bigrams, we achieved a networks of a 283 vertices and 726 directed and weighted edges. Sample literature bigrams computed using TFIDF generated from the literature titles retrieved from the *Nature Climate Change* journal are: (climate,change), (global,warming), (target,power), (power,plants), (global,ocean), (ocean,warming), (2008,cooling), (2020,emissions), (sustainable, future), (fight,climate), (blind,spot), (solar,power), (challenging,political), (political, climate), (cloudy,picture), (extreme,heat), (climate,risks), (climate,justice), (conservation, measures), (global,forest), (forest,carbon), (carbon,balance), (flood,risk), (challenging, climate), (carbon,budget), (climategate,inquiry), (past,revealed), (sea,ice), (marine,invaders), (consensus,prediction), (summer,heat), (dry,southern), (aerosols,impacts), (cool,polar), (summer,extremes), (biomass,burning), (burning,black), (ocean,acidification), (carbon, feedbacks), (snow,albedo), (change,communication), (emissions,gap), (cities,resilience).

As a step of verification of how good the bigrams collected, we used a historical tweet data set from 2012-2014 to see whether there is traffic or trends around the pair features. In Figure 1, the accounts in our data set are binned and plotted on a log-log scale as a function of their total keyword related tweets. Over 2.3 million (80.76 per cent) individuals in our data set tweeted only one climate-related keyword tweet, which

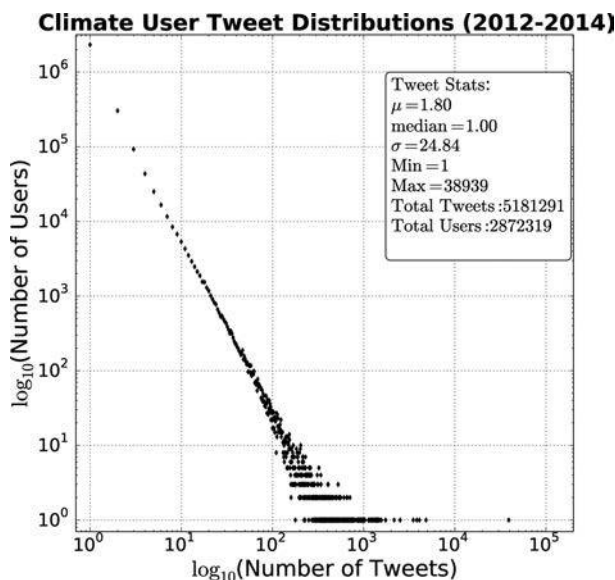


Figure 1.
 Binned user climate
 keyword tweet
 distribution
 (2012-2014)

corresponded to 44.77 per cent of all of the tweets collected. This is evidence that our data captures a broad number of individuals. In fact, 98.64 per cent of the accounts in this dataset had no more than ten climate related tweets, which corresponded to 74.62 per cent of the total tweets collected.

The average number of tweets per user was 1.80, with a standard deviation of 24.84 and a maximum of 38,939. This particular high-frequency tweeter is an example of an automated weather bot, an account that tweets weather updates for a particular region at scheduled intervals. The tail end of this distribution mainly corresponds to weather bots, news bots and other forms of automated tweeting machines. The vast majority of tweets are from a wide variety of individuals, helping strengthen our results.

Further, we analyzed for the bigram monthly frequencies. Figure 2 shows the monthly frequency of four high-impact bigrams that are plotted as a function of time. The tweets were converted into their local post time using the timezone metadata provided by Twitter's streaming API. Each of the distributions exhibited spikes in 2014. The largest notable frequency spikes came from the ("global warming") bigram during January of 2014 (38,283 tweets) and from "climate change" during September of 2014 (70,558 tweets). These particular bigrams support the models ability to discover high impact phrases to search for on the Twittersphere. Figure 3 shows the network that we constructed from the bigrams. The network exhibits a reasonably large portion of connected terms. The remaining terms, though connected as pairs, are isolated from the connected larger part of the network. These pairs such as ("blind spot" and "cloudy picture") do not have an associated contributing rank to them. They are not pruned here, however, to test whether they receive better attention when analyzing the tweet data set. It is also clear to see that the terms ("climate") and ("carbon") appears to be pivotal in the

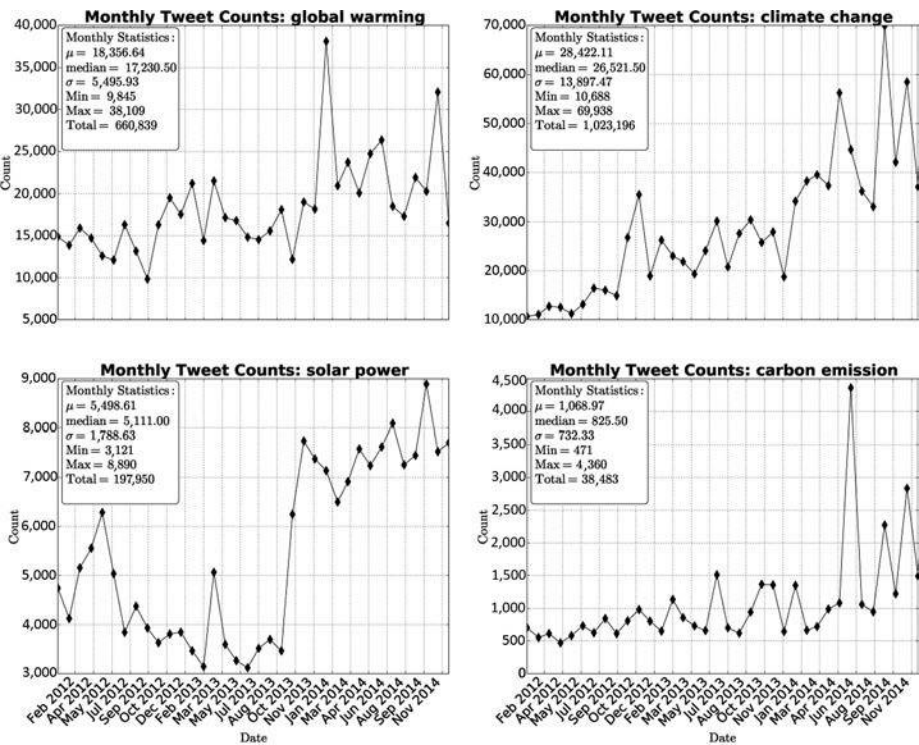


Figure 2. Frequency time series of tweets containing four selected high-impact bigrams in monthly bins spanning 2012 through 2014

Notes: (a) Monthly Tweet Counts: global warming; (b) Monthly Tweet Counts: climate change; (c) Monthly Tweet Counts: solar power; (d) Monthly Tweet Counts: carbon emission

network, which indeed demonstrate an indication of how significant these terms are when measuring their corresponding PageRank.

To obtain a rich understanding of the climate change awareness on Twitter, we curated 83 million tweets gathered between July 2014 and August 2015. Each tweet was processed in the same manner that was applied against the titles, and was also encoded in the “UTF-8” standard. After the noise removal was completed, we extracted all the bigrams using the same features that were computed for the GTN. This process produced $\approx 321,000$ non-unique bigrams; some are more frequent than others. We used frequencies to consolidate similar bigrams, and assigned as weight to the edges that connect the two sides of the bigrams. From the unique set of bigrams, we constructed a Twitter bigram network which we refer to as TBN. By constructing the network this way, we have means of comparing the rank of each term in both networks (i.e. GTN vs TBN). This comparison also guarantees the identification of the top-ranked terms in each network and measures the difference in ranking for each term. By knowing the top terms in the literature, we can then easily investigate whether the same terms receive similar attention and score a similar ranking on Twitter. This process indeed constitutes the notion of public awareness that we seek to establish in the experiment section below.

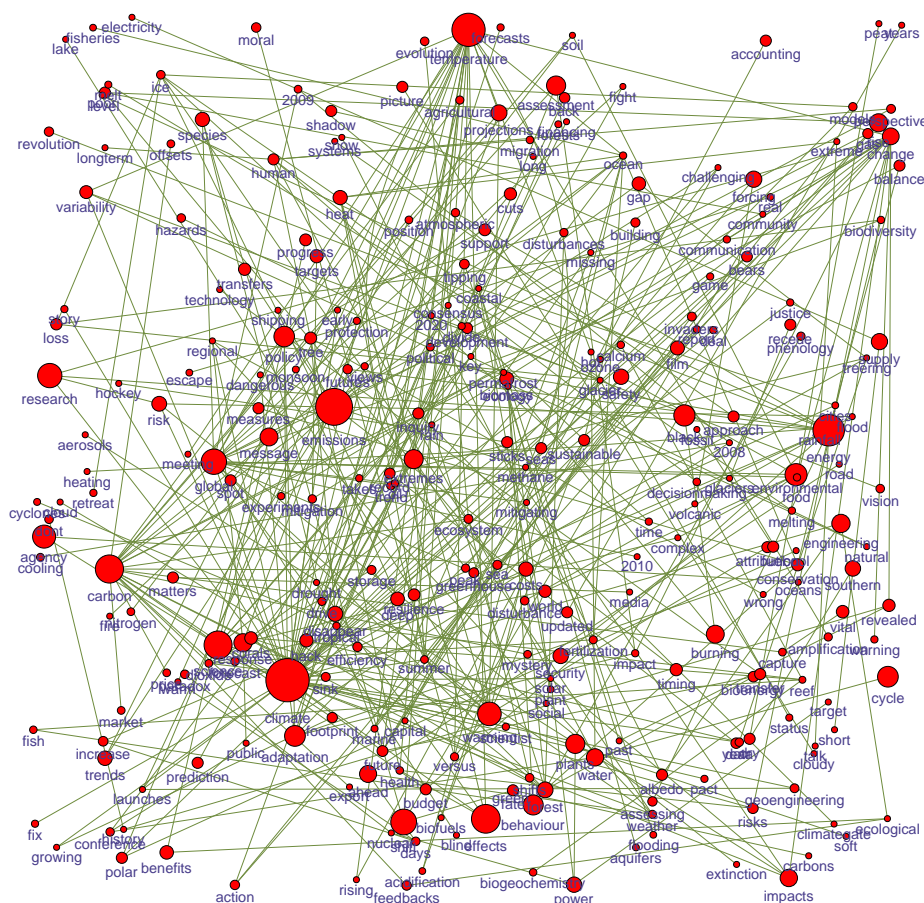


Figure 3. The network of bigrams computed from the literature titles

Notes: It is worth noting that the terms “climate” and “emission” have the most connected term in this network

Previously, we explored the climate change perception on Twitter using a means of Information Retrieval and PageRank analysis of multiple types of networks. The climate change awareness problem has been explored in a large magnitude of studies. However, it has never been explored using social media data such as Twitter in the way we approach it here computationally using data mining and network analysis methods. There are formal climate change vocabulary (e.g. ontology), we are still unable to access such resource yet. As a workout solution for this issue, we selected a set of bigram features from literature, using commonly used Information Extractions methods. We gathered the most relevant article titles that are publicly available from using the search query “climate” in the period between 1990 and 2014. The bigrams selected naturally offered a network mechanism which

we used as a ground-truth and also offered network centrality measures that helped to establish the notion of public perception.

GTN vs TBN PageRank network analysis

After constructing the GTN and the (TBN), we now can compare the PageRank of each term in both networks. As stated above, we used the bigram frequencies as an edge weight (i.e. a score of how frequent each pair of terms were linked) of the TBN. The weight value can be assigned to each pair p in the set of all bigrams P as follows:

$$\forall p \in P: l_w = \frac{p_f}{n} \tag{2}$$

where l_w is the weight assigned for each link, p_f is the number of occurrences for each bigram and n is the total number of bigrams derived from the TBN. Table I show the top 30 ranked terms in the GTN. The table also shows how these terms were ranked in the TBN network constructed from the tweet data set. Here we observe that the top terms are Climate and Carbon, which also confirm the observation we made on the visualization in Figure 3. These terms, however, did not show to be highly ranked in the TBN (19, 70), as the table demonstrated. The term “Weather” scored the exact rank at 30 in both literature and Twitter. Figure 4 shows two WordClouds, the one to the left is a

Feature source	Climate	Carbon	Temperature	Global	Change	Emissions	Energy	Warming
Literature	1	2	3	4	5	6	7	8
Twitter	19	70	98	37	1	42	16	51
	Forest	Costal	Ocean	Tropical	Science	Ice	Human	Summer
Literature	9	10	11	12	13	14	15	16
Twitter	57	127	75	120	3	34	58	23
	Adaptation	Policy	Reef	Food	Future	Dioxide	Power	Glacier
Literature	17	18	19	20	21	22	23	24
Twitter	87	21	124	11	39	104	14	122
	Record	Response	Ecosystem	Impacts	Sustainable	Weather	–	–
Literature	25	26	27	28	29	30	–	–
Twitter	13	N/A	113	N/A	101	30	–	–

Table I.

The order of the top 30 terms found in the GTN and their corresponding orders in the Twitter network

Notes: The first row shows the actual words, and the second row show corresponding ranks in the GTN (literature) network. The third row shows the ranks of the same words occurring in the TBN (Twitter) network. Due to the page limitation, we only included the top 30 terms. The analysis shows that the term (“Climate”) is ranked number 1 while it was found to be ranked number 19. Some terms did not make it to TBN (“Impact”) and (“Response”), which entails that such terms were not paired during the bigram process



Figure 4.
GTB WordCloud on
the left side
compared with TBN
WordCloud on the
right side

Notes: The size of the word corresponds to the order so the bigger words represents higher ranks. In the GTN: the top-ranked terms are Climate, Carbon, Emission and Temperature. In the TBN: WordCloud Science, Technology, Health and Change are the top ranked. This highlights that Twitter words are more broad compared with the highly technical terms in the literature

representation of the GTN, while the one to the right is a representation of the TBN. The term size reflects the importance in the ranking.

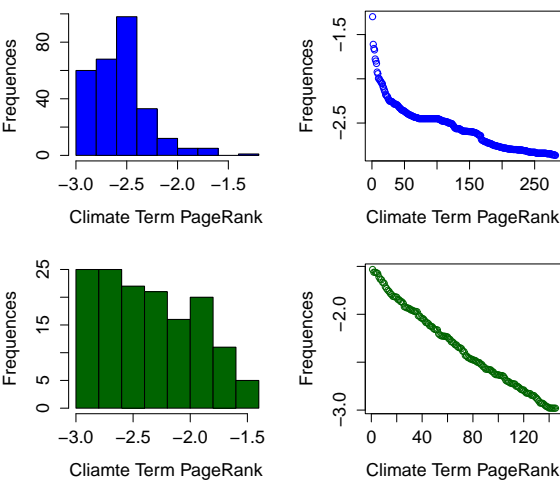
Figure 5 also shows the distribution of PageRanks of network. The GTN appears to be normally distributed, with the maximum value near the middle of the distribution. The TBN, on the other hand, shows a closer ranking distribution where the frequencies were established in a frequency range between 5 and 25. Additionally, the TBN logs show linearity which may suggest a power-law distribution as opposed to the distribution of GTN that seems to be converging around a value of zero.

GTN vs K-H PageRank network analysis

Hashtags are special informational devices that people use on social media to express an opinion, describe an entity or mention a location. There are endless types and forms of hashtags (Hamed *et al.*, 2014). While a hashtag can be made of a single word (e.g. #Hot), other hashtags can be made of a few words, numbers or even symbols (e.g. #BusinessIntelligence, #420 and #HealthCareReform).

It is important to consistently compare networks with similar characteristics to achieve the highest level of fair comparison possible. However, when hashtags are involved, it is not possible to construct word–hashtag bigrams, as hashtags can be placed anywhere in the tweet and they do not necessarily follow any grammatical rules. In fact, it is more accurate to assume informal hashtag constructs instead. Because we previously computed ground-truth features and bigrams, we handled this issue in two steps:

- (1) Comparing hashtags that are identical to such features constitutes a fair ground of comparison. Accordingly, we searched the tweet data set for hashtags that are identical to the features and bigrams: (Carbon → #Carbon) and (“Global Warming” → #GlobalWarming).
- (2) The other issue is related to connecting the ground-truth word features to the tweet hashtags, which was handled using Association Analysis experiments.



Notes: We here present the histogram to display how the term frequencies are distributed. The top GTN histogram drawn in blue shows the highest values are distributed near the middle while there are some missing terms around the end. The histogram, demonstrated in green, shows that the frequencies decreased as the ranks increased. The ranks in the TBN are also much lower than the GTN counterparts. While the maximum rank frequency was 25 in the TBN, it was observed near 90 in the GTN. Additionally, the PageRank behavior of the TBN appears to be more linear, while the PageRanks of GTN appear to be converging toward a specific value. This demonstrates significant rank differences between the literature terms and what their Twitter counterparts have. This explains that fact that the social language on Twitter is not as technical or scientific as the language used in the climate literature, which is expected. This also explains why some of the bigrams in the GTN network did not hold in the tweet data set

(a) GTN Term PageRank Logs; (b) GTN Term PageRank Logs; (c) FTN Term PageRank Logs; (d) FTN Term PageRank Logs

Figure 5.
Comparison of the
GTN Term
PageRanks logs with
TBN counterparts

The approach lends itself naturally to a type of network that guarantees directional and weighted edges. This was done by parsing each tweet in the data set, searched for each of the known features we previously gathered (e.g. carbon, warming, emission). Each feature we found was paired up with all the hashtags that existed in the tweet. A tweet such as this: (“#Florida officials ban the term climate change”) produces the following pairs (climate, #Florida) and (“change, #Florida”). Using *a priori* [Agrawal and Srikant](#)

(1994), an Association Analysis algorithm, we generated rules from the transaction pairs with their actual confidence values which we naturally used as a weight for the edge that connects the LHS of each rule with its RHS.

Association Analysis experiments depend on a minimum confidence configuration parameter that must be set at the beginning of each experiment. Due to the novelty of this research, it is not known which confidence level to set this parameter. Therefore, we performed a series of experiments with different minimum confidence configurations to test which network is a better representation for the comparison task as follows: (minConf: 0.10, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95, 1.0). The rules that were generated from each experiment were then used to make up a network of Keyword–Hashtag associations that can be compared with the GTN. Table II shows experiments, configurations and a number of rules produced for each configuration.

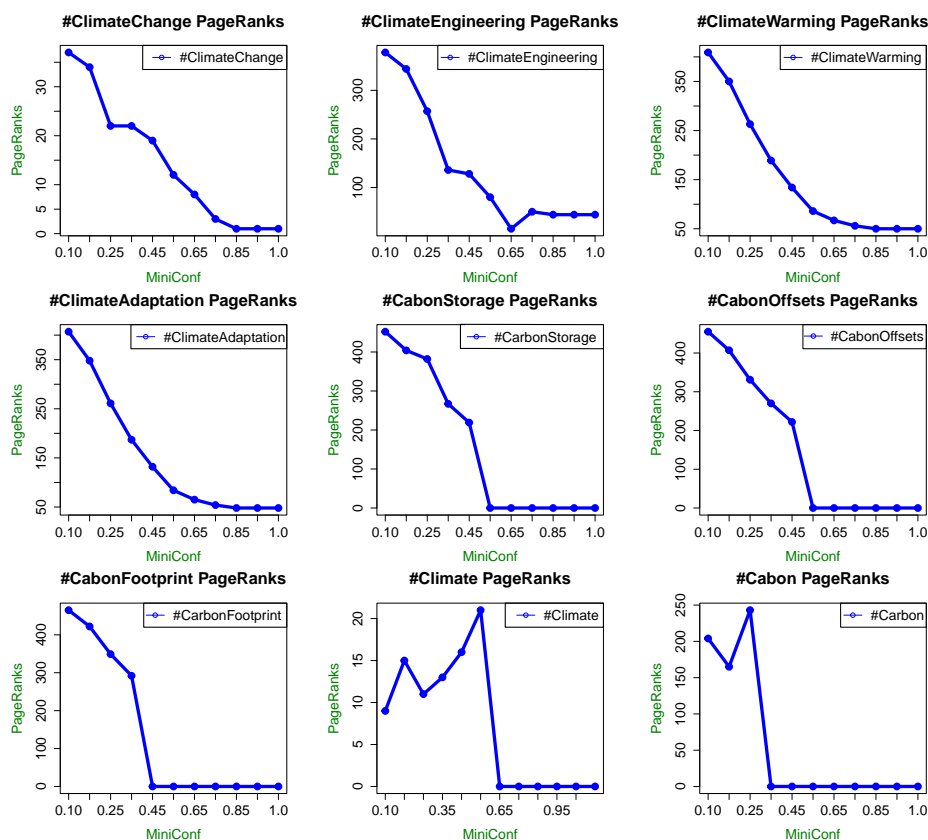
Climate change events identification

In the previous section, we looked at hashtags that may be identical to one or more unigram (word feature). In this section, we look at the hashtags that are emergent (i.e. may or may not overlap with one unigram, at most) (Figure 6). Such hashtags may signal climate change events that we aim to capture. To make the process of event detection less noisy, we only considered association transactions that contain the token (“climate”). A transaction is made of a pair of (unigram, hashtag). Therefore, a transaction that does not contain the (“climate”) token (e.g. carbon), will likely to be comprised of a hashtag that contains the token “climate”. If the bigram is the token (“climate”), it is likely to have a hashtag that is entirely emergent. This makes these transactions candidates for a climate change events which may or may not be significant. While the 2014 data set produced 5,084,741 transactions, the 2015 produced 5,556,991 transactions. The minimum confidence configuration parameter was set to 70 per cent. The experiments produced 48 climate change rules with minimum support that is a perfect confidence of 100 per cent. As for the data set collected in 2015, the data

Experiments no.	Mini confidence	No. of rules generated	Unique terms/hashtags
Exp: 1	0.10	1,232	561
Exp: 2	0.15	785	550
Exp: 3	0.25	448	456
Exp: 4	0.35	258	269
Exp: 5	0.45	161	246
Exp: 6	0.55	89	150
Exp: 7	0.65	61	111
Exp: 8	0.75	45	85
Exp: 9	0.85	38	74
Exp: 10	0.95	35	74
Exp: 11	1.0	35	74

Notes: The first column shows the experiment ID used. The second column shows the minimum confidence parameters used in each experiment. The third column shows the number of rules generated by the experiment using such configuration, and the last column shows the total number of unique terms found in the rules generated. The table gives a better understanding of how the change of minimum confidence can eventually affect the network construction and the ranking of terms

Table II.
The association
analysis experiment
configurations with
number of rules
returned from each
experiment using a
priori algorithm



Notes: For each hashtag, various association analysis experiments were performed to determine the actual rank of each term using different minimum confidence configurations. The plots show several observations that are noteworthy: (1) Single-hashtag terms score low ranks when comparing with their corresponding ranks in the GTN using a very low confidence score. Specifically, (#climate, and #carbon) hashtags has a rank of 9 and 200, respectively. However, the traffic dies out as the minimum confidence score increase, and hence the sudden drop in theranking of their specific plots; (2) Complex hashtags such as #Climate Change, #Climate Engineering and #Climate Adaptation show a rising trend as the minimum confidence reaches 100 per cent. This demonstrates how reliable such terms in measuring the awareness of climate change compared with single word hashtags #climate; (3) The most exciting observation that must be highlighted is that the analysis shows that the #ClimateChange hashtag scored the top term in the K-H Network when the minimum confidence approached 100 per cent. This is indeed in line with the term ranking specified by the GTN and shows striking evidence of a great deal of climate change awareness on Twitter; (4) The carbon-related hashtags demonstrate that the general public are aware of climate change; however, they might not be as aware of what it entails. All carbon-related terms score a much lower rank, given that carbon was the second highest ranked in the ground-truth

Figure 6.
The various
climate-related
hashtags and
corresponding ranks
on Twitter derived
from the dataset

produced 55 climate change rules with a 100 per cent confidence. Table III shows the outcome of analyzing the dataset into two partitions: one that occurred in 2014 and the other partition of tweets that occurred in 2015. Table III shows the highly associated events identified that are related to climate change. From further exploring the various semantics of each hashtag event, we hand-selected the most influential events signaled by the hashtags globally and domestically in the USA. A list of valid transactions that were captured from the two different datasets is as follows:

- (global,#climatechange), (sea,#climatechange), (fossil,#climate), (carbon,#climatechange), (solar,#actonclimate), (energy,#actonclimate), (picture,#climatevoter), (global,#climatechange), (solar,#climatechange), (atmospheric,#climatechange), (consensus,#climatechange), (polar,#climate), (price,#climate2014), (carbon,#climate2014), (coastal,#climate), (oil,#climate), (climate,#scribdocs), (world,#climatefact), (bears,#climate) (A sample of transactions collected from the twitter datasets from 2014, and 2015.

A transaction is a pair of a bigram (i.e., a feature selected using TFIDF) and hashtags that may or may not contain the token ["climate"].

Discussion and conclusions

In this paper, we explored the climate change perception and event identification on Twitter using data mining and network analysis of multiple types of networks. The climate change perception problem has been explored in a large magnitude of studies. However, only recently it has been explored using social media data such as Twitter. At the time of performing the analysis of this paper, we were unable to secure a formal source of climate change vocabulary (e.g. ontology). This created the need of generating a set of word features from literature (e.g. *Nature Climate Change*). We gathered the most relevant article titles that were publicly available from the web portal using the search query "Climate" in the period between 1990 and 2014. Using the bigram features selected, we constructed various types of networks (e.g. ground-truth and hashtag-based networks).

The paper demonstrated in detail the various types of experiments performed. For the purpose of measuring the perception of the general public, we performed a comparative analysis that compared the words PageRanks (a.k.a., Google's Algorithm) in the GTN against the Twitter words counterparts networks TBN. The analysis demonstrated that though there is indeed a great deal of awareness of climate change,

Data set	No. of records	No. of associations	No. of significant events
2015	5556991	55	9
2014	5084741	48	2

Notes: Each data set is represented by a row in a table. The columns list the number of records generated from the dataset that contains either the word climate in the tweet or the token climate in a hashtag. The number of associations column refers to the number of rules generated by each data set, while the number of significant events refers to the hashtags that climate change events during that year. It is clear that there are many more events in 2015 vs 2014. This could be due to the face that the tweets were collected toward the end of 2014

Table III.
The basic statistics
generated from data
sets collected in 2014
and 2015

Table IV.
The emergent events that took place in both 2015 and 2014, which was associated with the word (“climate”)

the social-language used on Twitter, expressed in mere words, is not as common as literature. Clearly, people on Twitter have other alternatives to express the same idea. Additionally, using slang and metaphorical expression is very common on social platforms. The word rank analysis showed a large gap in the ranking. This becomes clear in observing the top terms in the GTN network and contrasting their ranks with the same terms in the TBN. As shown in Table I, while the top term is climate, the term comes as number 19 in the TBN. Similarly, the second and third top terms (i.e. carbon and temperature) scored 70 and 98 in the TBN. This clearly demonstrates a large gap in the level of awareness if only words were considered (Table IV) (Figure 7).

Dataset	Event	Event highlight
2015	#askPotus	Ask the President of USA
2015	#earthDay	Earth Day, 2015
2015	#action2015	International Coalition Organization
2015	#CFCC15	Common Future Under Climate Change
2015	#paris15	UN Climate Change Conf., 2015
2015	#kickBigPollutersOut	Petition to protect COP21 against polluters
2015	#MSGTreePlantationDrive	Tree Plantation Drive Promotion
2014	#cop21	sustainable Innovation Forum, 2015
2014	#hulkSmash	Clean Petition on Avaaz.org
2014	#NoKXL	Reject Keystone XL

Notes: While some events were found internationally influential such as #EarthDay in 2015, others were only domestic. The hashtag #askPotus, was launched when President Obama of the USA answered questions in real time on Twitter, which included questions about climate

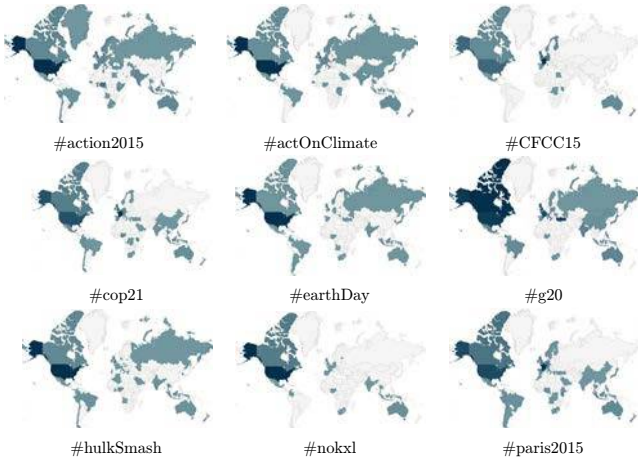


Figure 7.
The most significant hashtags that signaled climate change events by geolocation of tweets where the hashtags were widely socialized

Notes: The darker the color, the more frequent socialization of the hashtag. It is clear that hashtags such as #action2015, #actOnClimate and #earthDay were widely socialized across the globe

On the extreme contrary of word, hashtags are common textual constructs in place of grammatically correct contexts. Hashtags are powerful devices because of their ability to group the tweets in common and the wide range of knowledge they may represent. To ensure the fairness of the analysis done here, we incorporated the hashtags that linked any of the unigram that made up the GTN. Using the modern K-H network, we conducted a series of association analysis experiments that indeed lead to demonstrate a surprising amount of awareness of the topic. It turned out that the compound hashtags that are made of more than one word are more powerful than single word hashtags. For example, the (#Climate) hashtag has shown to be less significant in capturing climate change expressions on Twitter. On the other hand, a more complex hashtags such as (#ClimateChange, #ClimateEngineering and #ClimateAdaptation) are shown to be more dominant and everlasting given the increase of the minimum confidence parameter.

One of the most interesting observations found is that (#ClimateChange) was identified as the top-ranked hashtag among all hashtags and words that made up the K-H network. Indeed, this reflects positively on the amount of awareness that persists on social media in general and Twitter in particular. However, it is important to point out that the awareness traffic was slim and not persistent for more scientific or technical terms. Though the term (“Carbon”) is the second highest ranked from the scientific point of view, we find little awareness about the term. This can be explained as the general public lack a deeper understanding of the various dimensions of climate change. It can also be associated with the current climate change events that took place at the time of collecting the data. It is possible to find more general hashtags such as (#ClimateChange or #Climate 2014) during the UN Climate Summit than finding more specific terms.

Delivering a full picture of the perception motivated the mining and the analysis of all hashtags included in the data set. As a preprocessing step, we captured all tweets that contained either the work (“climate”) or hashtags that contained the token (“climate”). Combined with all other bigrams, this generated association analysis transactions such as (global, #climateChange), (solar, #climateFact), (development, #actOnClimate) and (climate, #algae). It is clear to observe that some of these hashtags may or may not overlap with the set of features we collected. For instance, while the (#actOnClimate) hashtag contains the token climate and also among the bigrams, the (#algae) hashtag does not correspond to the features collected. Performing association analysis experiments have honed the significant hashtags that pointed to events.

Among the most significant events that were identified in the data set were worldwide events which were merely signaled using hashtags. For instance, Earth Day of 2015 was signaled using the hashtag (#EarthDay). Such hashtag was included in tweets from all over the world and from each of the six continents. Other events started domestically somewhere in the world and spread such as (#MSGTreePlantationDrive). This hashtag was a promotion for planting trees as a positive response to climate change, which originated in India. However, the event spread worldwide and was captured in the association analysis experiments. This demonstrates our strong intuition on how hashtags can indeed signal climate change events, and how it can be identified using association analysis experiments.

In the future direction of this paper, we will continue to investigate the climate change event identification which will also include the temporal dimension lacked in this paper. This requires acquiring a large set of historical tweets that we currently do not possess. The temporal analysis will also shed light on the extreme event patterns that occurred in the past (e.g. Hurricane Irene) and how such events are discussed on Twitter from a climate change point of view. Another significant aspect of this research, which will necessarily capture a precise measure of the awareness, is the study of the evolution of the climate change social language. This idea can be more formalized by the development of an ever-evolving climate change social ontology. This socio-formal ontology may lead to understanding how the climate change may impact the development of disease, water quality [as it has already been signaled in the tweet dataset by the transaction (climate and #algae)] and food industry.

References

- Agrawal, R. and Srikant, R. (1994), "Fast algorithms for mining association rules in large databases", *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94*, Morgan Kaufmann Publishers, San Francisco, CA, available at: <http://dl.acm.org/citation.cfm?id=645920.672836>, pp. 487-499.
- Aizawa, A. (2003), "An information-theoretic perspective of tfidf measures", *Information Processing and Management*, Vol. 39 No. 1, pp. 45-65, available at: www.sciencedirect.com/science/article/pii/S0306457302000213
- Bekkerman, R. and Allan, J. (2003), "Using Bigrams in text categorization", *Microsoft Academic Search*, University of Massachusetts, Amherst.
- Callaway, J.M. (2004), "Adaptation benefits and costs: are they important in the global policy picture and how can we estimate them?", *Global Environmental Change*, Vol. 14 No. 3, pp. 273-282, available at: www.sciencedirect.com/science/article/pii/S0959378004000366, the Benefits of Climate Policy.
- Curry, T.E. (2004), "Public awareness of carbon capture and storage: a survey of attitudes toward climate change mitigation", PhD thesis, MA Institute of Technology, Cambridge, MA.
- Do, T.D., Hui, S.C. and Fong, A.C.M. (2006), "Associative feature selection for text mining", *International Journal of Information Technology*, Vol. 12 No. 4.
- Esbjörn-Hargens, S. (2010), "An ontology of climate change", *Journal of Integral Theory and Practice*, Vol. 5 No. 1, pp. 143-174.
- Forman, G. (2003), "An extensive empirical study of feature selection metrics for text classification", *The Journal of Machine Learning Research*, Vol. 3 No. 1, pp. 1289-1305.
- Hamed, A.A., Wu, X., Erickson, R. and Fandy, T. (2015), "Twitter K-H networks in action: advancing biomedical literature for drug search", *Journal of Biomedical Informatics*, Vol. 56, pp. 157-168, available at: <http://dx.doi.org/10.1016/j.jbi.2015.05.015>
- Hamed, A.A., Wu, X. and Rubin, A. (2014), "A twitter recruitment intelligent system: association rule mining for smoking cessation", *Social Network Analysis and Mining*, Vol. 4 No. 1, available at: <http://dx.doi.org/10.1007/s13278-014-0212-6>
- Hamed, A.A. and Zia, A. (2015), "Mining climate change awareness on twitter: a pagerank network analysis method", *Computational Science and Its Applications - ICCSA 2015 - 15th International Conference*, Banff, AB, 22-25 June, Part I, pp. 16-31, available at: http://dx.doi.org/10.1007/978-3-319-21404-7_2

- Hearst, M.A. (1999), "Untangling text data mining", *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, Association for Computational Linguistics, College Park, Maryland, pp. 3-10.
- Jensen, L.J., Saric, J. and Bork, P. (2006), "Literature mining for the biologist: from information retrieval to biological discovery", *Nature Reviews Genetics*, Vol. 7 No. 2, pp. 119-129.
- Jing, L.P., Huang, H.K. and Shi, H.B. (2002), "Improved feature selection approach tfidf in text mining", *Proceedings 2002 International Conference on Machine Learning and Cybernetics*, IEEE, Vol. 2, pp. 944-946.
- Kam, X.N.C., Stoyaneshka, I., Torniyova, L., Fodor, J.D. and Sakas, W.G. (2008), "Bigrams and the richness of the stimulus", *Cognitive Science*, Vol. 32 No. 4, pp. 771-787, available at: <http://dx.doi.org/10.1080/03640210802067053>
- Kirilenko, A.P., Molodtsova, T. and Stepchenkova, S.O. (2015), "People as sensors: mass media and local temperature influence climate change discussion on twitter", *Global Environmental Change*, Vol. 30, pp. 92-100, available at: www.sciencedirect.com/science/article/pii/S0959378014001952
- Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A.A. and Rocha, L.M. (2010), "Classification of protein-protein interaction full-text documents using text and citation network features", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7 No. 3, pp. 400-411, available at: <http://dx.doi.org/10.1109/TCBB.2010.55>
- Levenbach, G.J. (1998), "A Dutch bigram network", *Word Ways*, Vol. 21 No. 11, available at: <http://digitalcommons.butler.edu/wordways/vol21/iss3/11>
- Lorenzoni, I., Nicholson-Cole, S. and Whitmarsh, L. (2007), "Barriers perceived to engaging with climate change among the UK public and their policy implications", *Global Environmental Change*, Vol. 17 No. 3, pp. 445-459.
- Macintyre, G., Jimeno Yepes, A., Ong, C.S. and Verspoor, K. (2014), "Associating disease-related genetic variants in intergenic regions to the genes they impact", *PeerJ*, Vol. 2, p. e639, available at: <https://dx.doi.org/10.7717/peerj.639>
- Marsi, E., Oztürk, P., Aamot, E., Sizov, G. and Ardelan, M.V. (2014), "Towards text mining in climate science: extraction of quantitative variables and their relations", *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, Reykjavik, Iceland.
- Neil Adger, W., Arnell, N.W. and Tompkins, E.L. (2005), "Successful adaptation to climate change across scales", *Global Environmental Change*, Vol. 15 No. 2, pp. 77-86.
- Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2 Nos 1/2, pp. 1-135.
- Pardalos, P., Boginski, V.L. and Vazacopoulos, A. (2008), *Data Mining in Biomedicine*, Springer, US, Vol. 7.
- Radev, D.R., Jing, H., Sty, M. and Tam, D. (2004), "Centroid-based summarization of multiple documents", *Information Processing and Management*, Vol. 40 No. 6, pp. 919-938, available at: www.sciencedirect.com/science/article/pii/S0306457303000955
- Salton, G. and Buckley, C. (1988), "Term-weighting approaches in automatic text retrieval", *Information Processing and Management*, Vol. 24 No. 5, pp. 513-523, available at: www.sciencedirect.com/science/article/pii/0306457388900210
- Sampei, Y. and Aoyagi-Usui, M. (2009), "Mass-media coverage, its influence on public awareness of climate-change issues, and implications for japans national campaign to reduce greenhouse gas emissions", *Global Environmental Change*, Vol. 19 No. 2, pp. 203-212.

- Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys (CSUR)*, Vol. 34 No. 1, pp. 1-47.
- Semenza, J.C., Hall, D.E., Wilson, D.J., Bontempo, B.D., Sailor, D.J. and George, L.A. (2008), "Public perception of climate change: voluntary mitigation and barriers to behavior change", *American Journal of Preventive Medicine*, Vol. 35 No. 5, pp. 479-487.
- Tan, C.M., Wang, Y.F. and Lee, C.D. (2002), "The use of bigrams to enhance text categorization", *Information Processing and Management*, Vol. 38 No. 4, pp. 529-546, available at: [http://dx.doi.org/10.1016/S0306-4573\(01\)00045-0](http://dx.doi.org/10.1016/S0306-4573(01)00045-0)
- Whitmarsh, L. (2009), "Behavioural responses to climate change: Asymmetry of intentions and impacts", *Journal of Environmental Psychology*, Vol. 29 No. 1, pp. 13-23.
- Xie, X., Jin, J. and Mao, Y. (2011), "Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks", *BMC Evolutionary Biology*, Vol. 11 No. 1, p. 242, available at: <http://dx.doi.org/10.1186/1471-2148-11-242>
- Ye, N. (2003), *The Handbook of Data Mining*, Lawrence Erlbaum Associates, Mahwah, NJ, Vol. 24.
- Zhang, W., Yoshida, T. and Tang, X. (2011), "A comparative study of tf*idf, {LSI} and multi-words for text classification", *Expert Systems with Applications*, Vol. 38 No. 3, pp. 2758-2765, available at: www.sciencedirect.com/science/article/pii/S0957417410008626

Corresponding author

Ahmed Abdeen Hamed can be contacted at: ahamed@uvm.edu