Climate change on Twitter: Content, media ecology and information sharing behaviour

Abstract

This paper presents a study of the content, use of sources and information sharing about climate change analysing over 60,000 tweets collected using a random week sample. We discuss the potential for studying Twitter as a communicative space that is rich in different types of information and presents both new challenges and opportunities. Our analysis combines automatic thematic analysis, semantic network analysis, and text classification according to psychological processes categories. We also consider the media ecology of tweets, the external web links that users shared. In terms of content, the network of topics uncovered presents a multi-dimensional discourse that accounts for complex causal links between climate change and its consequences. The media ecology analysis revealed a narrow set of sources with a major role played by traditional media, and that emotionally arousing text was more likely to be shared.

Keywords

Twitter, climate change, semantic graphs, media ecology

1. Introduction

Climate change is a major challenge facing society and a problematic issue to communicate: it has complex causes and consequences largely beyond people's biographical horizons - few will directly experience its consequences (Schäfer, 2012). The media have been identified as an especially important agent in the formation of common sense knowledge about climate change (Carvalho, 2010; Moscovici, 2000), leading much research to examine the climate change discourse in various mass media, mostly print newspapers and television (Moser, 2010). Yet, as stakeholders from scientists to policy-makers increasingly turn to social media to disseminate information about climate change and mobilise support for (in)action on climate change and members of the public increasingly use social media (Schäfer, 2012), the climate change discourse on social media becomes a priority research area.

While the existing body of research on the public understanding of science (PUS) offers well-established ways of studying traditional mass media as well as public and policy-makers' issue perceptions (via content analysis, surveys/experiments and case studies, respectively) (Suerdem, Bauer, Howard, & Ruby, 2013), little agreement exists as to what methods can be employed to reliably study social media and what insights can be achieved. The absence of methodological guidelines can be attributed to the challenging, hybrid nature of social media as both 'media' (information sources) and data about people's behaviour. This hybrid nature means that its study can potentially provide unprecedented access to people's behaviour and the communicative space around an issue.

The hybrid nature of social media also suggests a mix of different but complementary theoretical frameworks may need to be harnessed to guide analysis. In particular, the notion of a 'personal public' (Schmidt, 2014) is fundamental to understanding the role and use of social media such as Twitter. Personal publics can be considered an ideal type of communicative space characterised by three main features. Information is selected and displayed according to

criteria of personal relevance (rather than following journalistic news factors); information is addressed to an audience consisting of network ties and is made explicit (in contrast to being broadcast to an unknown mass audience), and information is often disseminated in a conversational rather than unidirectional way.

In the context of social scientific research, Twitter can offer: information about the technological affordances of this platform; details about the social and textual relationships of its users and of user-generated content; and, insights into the shared rules and expectations within the community of users (Weller et al, 2014). In the context of PUS research, Twitter data represents an opportunity to study several aspects related to the reception and use of information about science and technology. Twitter data contains information about: content generated by users; news and information selected from other sources and published on Twitter; conversations held on the topic with other users and their network relationships; and information that receives more attention and is circulated the most via sharing. The possibility of studying these aspects and their interaction is unique to social media and, we believe, represents an important opportunity to research further the public understanding of science.

The current challenge is establishing a theoretically meaningful and methodologically viable use of Twitter data for social scientific research in general and in PUS studies in particular. In this study, we present an approach that makes use of the richness of Twitter data; it does not exploit all the information available, the reasons for which we discuss further in the conclusions. Three theoretical approaches and methodologies are applied to unpack Twitter data about climate change: text mining and semantic networks on the content of tweets generated by users; a psychological-processes-based classification of tweets to learn about the nature of the conversations on climate change conducted on this platform; and content analysis of the external webpages included in tweets about climate change and their sharing on Twitter.

2. Theoretical framework

The three selected approaches account for the different levels of analysis that can be carried out using Twitter data. The different levels refer to the fact that Twitter users can both share content from other sources and express their views about a topic.

Algorithmic-based content analysis such as automatic thematic identification and semantic networks aim to identify social representations emerging in the 'Twittersphere'. Because Twitter is also about users' comments and conversation, identification of general psychological processes and sentiment inferred by the classification of user-generated text is an additional level of investigation worth adding to thematic identification. In addition, content analysis of web links is another important level of analysis in order to acknowledge another crucial function of Twitter: as a medium for information sharing as demonstrated by Kwak, Lee, Park, and Moon (2010).

First, the overall content of tweets can be analysed through the lens of social representations theory. In social representations theory or SRT (Moscovici, 2000), the aim was to identify representations of a given scientific or technological issue, their adoption by social groups and their role in the formation of common sense knowledge. SRT posits that: 1) there are competing definitions of issues in the public sphere referred to as social representations (Gaskell, Bauer, & Durant, 1998); 2) definitions are a matter of framing that aims to impact opinion/attitude formation and legislation; 3) the framing battle 'is being waged in the arena of language, as much as that of science' (Ogden, 2001, p. 340).

Communication plays a crucial role forming social representations. In this context and in line with previous studies, social representations are conceptualised as user-generated semantic networks at the aggregate level (Veltri, 2013a, 2013b). Social representations theory can thus be employed to map the semantic spectrum of issues discussed in relation to climate change on Twitter.

Second, the content of tweets can also be studied from the perspective of psychological processes associated to word use, in order to infer them from tweets about climate change. The rationale is that tweets are also users' opinions and public statements[1] shared with their 'personal publics'. Strategic use of words and their classes (how something is said) can be as revealing of underlying psychological processes as explicit statements (what is said) (Fiedler & Mata, 2014). Research has, for example, uncovered relationships between 'style' or 'function' words (e.g. pronouns, quantifiers, adverbs) and psychological processes (see Pennebaker, 2011). The original context of LIWC use is when texts are associated to psychological characteristics of the text producers and therefore used for profiling and therapeutic aims. In this context, a tweet is not enough for this aim: we deal with multiple users. However, an alternative use is to apply LIWC's dictionary of psychological categories and related words to analyze texts without inferring about the psychological status of the text producer because the unit of analysis is not the individual (tweeter) but the overall 'tweetscape' about, in this case, climate change.

LIWC's analysis can also provide more nuanced information about emotional states inferred from text compared to sentiment analysis. The emotional component plays an important role in the formation of attitudes related to climate change as shown by Hoijer (2010), who also considers the role of emotions in the context of social representations theory.

Third, Twitter data can be analysed with 'media ecology' in mind, in other words, considering what content is posted on Twitter and shared by other users. Here, we move from user-generated content analysis to selection of online news. A recent shift from traditional 'push media' to emerging 'pull media' has implications for how the public consumes science news and information (Anderson, Brossard, & Scheufele, 2010). By virtue of increased control and choice over media content afforded by the Internet, understanding the media ecology of tweets means better understanding sources of content that users re-publish through their Twitter accounts. On Twitter, information sharing can be conceptualised in two stages or

degrees: 1) selecting and posting from a range of different sources and publishing a tweet (first degree of sharing) 2) retweeting a tweet (second degree of sharing) thus triggering conversations and information diffusion. While there is an increasing body of research on 'retweeting' (e.g. Bogdanov, Busch, Moehlis, & Singh, 2013) and even some literature in the more specific case of climate change news (Hansen, Arvidsson, Nielsen, & Colleoni, 2011; Segerberg & Bennett, 2011), the first degree of sharing has received little attention.

The closest existing study on this issue is a recent study on the emailing of articles from *The New York Times* (Berger & Milkman, 2012). The authors found a strong link between positive affect and 'virality' (sharing an article via email). Based on psychological theories they concluded that this relation is universally valid. However, this study was limited to the particular configuration of analysing information shared from one source (*The New York Times*) via email, a social network not necessarily used in the same way as social media (see the critique of Hansen et al., 2011). The NYT carried out the other existing study on motivations and sharing behaviour (actual choices).

In this study, we analysed the first degree of sharing in the context of climate change related tweeting activity focusing on two aspects: first, we carried out a content analysis of the most frequently posted web links in tweets to understand what kind of content is users select and publish on Twitter. Second, we rated the content of these web links on a two-dimensional emotional scale to test if it correlated with how frequently a web link was shared. Hence, the content analysis represents a classification of sources of the first degree of sharing and thus an exploration of the 'media ecology' around the tweets on or related to climate change. Previous studies have stressed the importance of emotions in determining the sharing of online news in general (Berger & Milkman, 2012) and in the case of climate change's tweets as well (Hansen et al., 2011). We test the importance of emotional value in the first degree of sharing (what people post on Twitter rather than tweets shared on Twitter).

3. Climate change on social media

There is a nascent body of research on climate change discourse in social media, which has focused on Twitter. This focus comes as no surprise given the widely-shared agreement that studying Twitter in particular is necessary for multiple reasons, from the fact that it is simply too important now to ignore with its global reach and growing number of users and posts, to it being able to provide a window on various aspects of society (Weller et al., 2014). This body of research has studied how the emotion and affective valence of a tweet may influence its 'virality' in terms of the probability of a retweet (Hansen et al., 2011); use of web links in climate change-related tweets and sources behind those web links (e.g. mass media, personal websites) and tweet content (topics) (Segerberg & Bennett, 2011); or structure (who links to whom/who talks to whom) and tweet content (topics) (Pearce, Holmberg, Hellsten, & Nerlich, 2014).

To give an overview of these findings, after analysing two samples of tweets – tweets about the Copenhagen Climate Change Conference (COP15) and a random sample – Hansen et al. (2011) found that sentiment differentially influenced retweet probability of tweets containing news web links and non-news tweets. Negative sentiment promoted retweeting in news-driven tweets from both samples, positive sentiment promoted retweet probability of non-news tweets. From this analysis Hansen et al. (2011) gained insights about the similarities of news-tweeting and the logic of traditional news media: the emergence of negative content as a strong promoter of retweeting in news-driven tweets resonates with the classic theory of selection and diffusion in traditional news media according to which negative affect is a key contributor to propagation.

Next, Segerberg and Bennett (2011) analysed random samples of tweets from two hashtag streams associated with two climate change protest marches that were among a number of protests leading up to COP15. Their analysis of web links showed that alternative media

websites (as opposed to, for example, mass media websites) accounted for the greatest share of web links in both streams. The analysis of tweets' content showed that tweets mostly described the marches but were less likely to share logistical information, leading the authors to gain insights into the orchestration of the marches. Finally, Pearce et al. (2014) analysed tweets about the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment AR5 WG1 report, published in September 2013, in which they identified 'science', 'geographical discussions', and 'societal concerns' (e.g. geoengineering) as the most prominent topics.

4. Research aims

In the light of this review, our study can best be described as an exploratory exercise in method and theory bridging to analyse an issue of major societal significance – climate change on an important social media platform – Twitter. In contrast to current research we do not focus on a period that overlaps with a specific event (e.g. COP15), but our research questions resonate with some of the research aims of existing research (e.g. content (topic), web links, sentiment of tweets). Once again, our chief aim is to present one way of conducting research on Twitter, which we hope will stimulate further discussion, and more theoretically-informed, methodologically-robust research in this (and other) areas of Twitter analysis.

According to each level of analysis that can be carried out on tweets, we ask the following research questions: RQ1) What is the content of climate change-related tweets?; RQ2) What are the psychological categories of words underlying tweets about climate change?; RQ3) What are the sources of information about climate change that are shared on Twitter in terms of first degree sharing?; and RQ4) Does the emotional value of content play a significant role in predicting if (and how often) it will be shared in a tweet?

The research questions address different levels of analysis and each of the latter requires a different theoretical framework (Table 1). As mentioned in section 2, social

representations theory provides a framework about climate change-related content, Pennebaker's approach (Pennebaker & Chung, 2013) about the content emotional value and psychological context, and 'media ecology' and sharing behaviour is discussed in terms of which sources and what characteristics of external information increase the likelihood of it being shared in a tweet.

INSERT TABLE 1

5. Methodology

There is an on-going debate about how standard content analysis should be adapted to web content (Herring, 2010); this paper adopts a mixed methods research design. It uses metadata about general patterns of information sharing about climate change and an automatic text analysis of the content of tweets. In addition, it performs an analysis identifying general emotional charge (the negative or positive valence) of the corpus, providing information about the general psychological mechanisms associated to the tweets' content.

5.1 Data collection

The sampling procedure adopted the criteria of data collection based on the content of the tweets and their time of publication. The study tracked seven randomly selected days of tweets between the 1st of March and 30th of June 2013, collecting tweets containing keywords and hashtags 'climate change' and 'global warming'. Duplicates were eliminated via parsing using tweets' ID tags, the unique number identifying each tweet. A random week generation technique[2] was adopted to select the seven days for data collection.

The unit of analysis was a tweet, which had a major impact on sampling due to the large computing power needed to process a very large number of tweets. The following strategy was adopted: in 2013 two tests were carried out, monitoring one week in March and one in April, and the tests produced an average of 8,300 tweets per day about climate change. It

is important to clarify that we refer mainly to tweets in English, therefore, this is not an estimate of tweets about climate change in total. The corpus was restricted to English-language tweets in order to be analysed using the lemmatization and automatic coding described in the next section. Data collection was limited to a week in order to stay within the computational limit of the software used, in particular in the case of T-LAB and LIWC2007[3].

The source of data was the public API of Twitter using a free (for non-commercial use) version of the proprietary service of tracking public tweets based through the DMI-CAT platform (Borra & Rieder, 2014). It is written mostly in PHP and runs in a webserver (LAMP) environment. Using the search terms 'climate change' or 'global warming', 60,122 tweets in total were collected during the 7 days of tracking. The tracking algorithm includes the following information: content of the post, web link in the tweet, username and number of followers. Web links information was extracted using Twitter's 'SpiderDuck' URL Fetcher. For the analysis using LIWC2007 (Pennebaker, Booth & Francis, 2007), tweets were stripped of web links and 'titles of web links' so that they contained only user inputted text (re-tweets were included). Hence, user-generated content constitutes the corpus of the classification. The software tool employed to carry out the analysis of the corpus was the software LIWC2007 (Pennebaker et al., 2007; Tausczik & Pennebaker, 2010).

5.2 Procedure for automatic thematic analysis and semantic network analysis

The first step was to aggregate all tweets in a corpus in order to analyse this using text-mining techniques. The corpus was prepared for quantitative text analysis using the software T-Lab 9, a text mining software used in the social sciences (Lancia, 2012). The output of T-Lab was a multidimensional scaling (MDS) map in which keywords are represented on a two dimensional scale in terms of their proximity reflecting their co-occurrences in the corpus (more details about the statistical procedure are reported in the Appendix). In the last stage, latent semantic analysis was carried out (Deerwester et al, 1990).

Semantic network analysis originates in the cognitive science literature, which proposes the existence of a structural meaning system in human memory (Collins & Quillian, 1972). Later works, notably 'Coding Choices for Textual Analysis' (Carley, 1993), used map analysis to extract the main concepts from texts and relations between them. Along this line of thought, semantic network theorists have argued that frequency, co-occurrence, and distances between words and concepts allow researchers to explore meanings embedded in texts (Danowski, 2010; Doerfel, 1998). Examples of this approach include Chung and Park (2010) on political speeches, Zywica and Danowski (2008) on Facebook open-ended survey responses, and more recently, Yuan, Feng and Danowski (2013) on discussions about privacy on the Chinese microblogging platform Weibo (a Twitter equivalent).

In procedural terms, we followed three steps. First, we extracted the most frequent keywords using the T-Lab TDF-IDF algorithm. Second, we computed the co-occurrences between keywords. The generation of networks is based on Carley's approach to coding texts as cognitive maps (Carley & Palmiquist, 1992) and Danowski's approach to proximity analysis (Danowski, 1993). Semantic networks translate text into networks of concepts and the links between them, where a concept can be a word or a phrase (i.e., n-gram) (Popping, 2003). Links between concepts form via co-occurrence. Third, we computed centralities measures and clustered keywords using the modularity technique (Newman, 2006).

We later analysed the semantic network using Pajek (Batagelj & Mrvar, 2014) in order to compute centrality measures. Another step was to import the network including the centrality values in the software Gephi (Bastian, Heymann & Jacomy, 2009) that has the most options in terms of visualization for the semantic network and can also perform community detection using Newman's algorithm (Newman, 2006).

There are a number of well-known measures of centrality in a network including: betweenness centrality, closeness centrality, out-degree, in-degree, PageRank, hubs and authorities. Each

can be used to capture different aspects of the relationship between concepts. The concept of prominence refers to the properties of central locations of main concepts. In social network analysis it refers to degree, betweenness and closeness centrality (see Appendix for formulas used). Central concepts are close if they have minimum steps relating to all other nodes. Freeman's degree centrality (Freeman, 1979) has been one of the most frequently utilised indices to represent the notion of importance.

Words with high degree centrality help identify the most salient topics. Words with high betweenness centrality identify topics that are active mediators in the communication bridging between topics. Words with high in-closeness centrality identify topics that function nuancing and defining central topics (Doerfel, 1998; Yuan et al., 2013). Translation of the network roles is possible because the relevant (corresponding) nodes essentially inherit the functionally prominent roles: a node with the highest betweenness centrality takes on the role of a mediator of communication, as it is required to represent itself explicitly to bridge different clusters of words. By comparison, an object with a high in-closeness is the eventual result of communicative interaction that positions itself closest to the centre of reference playing a 'connotative' role or in other words, a 'qualifier' or 'contextualizer'.

5.3 Procedure for classification of textual psychological processes

In classifying and analysing the corpus using LIWC, we selected only tweets containing users' statements because we wanted to analyse only user-generated content, not external sources such as web links or quotes. After going through all words in the corpus, LIWC calculates the percentage of each LIWC category present. For example, we might discover that 5.67% of the words in a given body of text are about family and 3.38% are auxiliary sadness. The LIWC output thus lists all LIWC categories and shows how much each category was used in the given text.

Dictionaries are central in the LIWC programme because a dictionary refers to the collection of words that define a particular category. The dictionaries in LIWC2007 have gone

through a long period of testing, resulting in the current version (Pennebaker et al, 2007). Most importantly, the 80 language categories in LIWC have been linked in hundreds of studies to interesting psychological processes as previously discussed. LIWC output can be checked with baseline information provided by the developers, the means and grand means produced by many datasets used to develop this software. Checking differences between our own corpus and such benchmarks is our way to help interpret the LIWC output. We also have at our disposal a similar sized dataset of tweets, about nanotechnology previously analysed by Veltri (2013a, 2013b) as a further benchmark value.

5.4 Information sharing and media ecology

We will briefly describe the procedure used to analyse the web links contained in tweets. Web links were extracted using Twitter's 'SpiderDuck' URL Fetcher. Tweets were manually checked to ensure they were in English. Subsequently, we selected web links that were shared at least ten times, i.e., links present in at least 10 tweets. This choice was related to the practical need to reduce the corpus of tweets to carry out a content analysis of the webpages of web links.

Content analysis is a quantitative research tool for systematic and inter-subjective description of communication content starting from existing conceptual categories. Categories describing the properties of media content relevant to our research questions are reported in the Appendix. We explored the role of emotional arousal elicited by the articles using the Self-Assessment Manikin (SAM) method (Bradley & Lang, 1994). SAM is a non-verbal pictorial assessment technique (in which pictures represent answers rather than statements) directly measuring the pleasure, arousal and dominance associated with a person's affective reaction to a wide variety of stimuli. The SAM method measures both valence (positive-negative emotion) and arousal (intensity).

The content analysis was validated using Krippendorf's alpha inter-coder agreement measure (Krippendorff, 1987) that scored 0.92 using a random sample of 76 web links (10% of total

corpus).

6. Findings

This section first presents the outcome of the analyses performed on the entire corpus of tweets (*N*=60,122): latent semantic analysis, semantic networks and psychological processes classification. In the last part of this section we report findings from the content analysis carried out on the subsample of web links contained in the tweets.

6.1 Latent semantic analysis

The algorithmic grouping of keywords in tweets in semantic domains according to the procedure described in the methodology section revealed a total of four thematic areas (Figure 1). The first corresponds to the top left corner of the MDS graph indicating the theme of call for action about climate change with an emphasis on its consequences. The latter themes are predominant in the second thematic space corresponding to the top right corner of Figure 1. In this case, climate change is associated with consequences and, in particular, rising sea levels and extreme weather.

INSERT FIGURE 1

The third theme in the bottom right corner of the MDS representation refers to the policy dimension of climate change, referring to regulations, discussions by committees and political actors (e.g., the President of the United States). The fourth thematic space includes local (geographically specified) news associated with climate change such as the news that Mumbai will be under threat of extreme weather if global temperatures continue to rise, and floods in Europe. In addition, this thematic cluster also contains tweets about climate change deniers and sceptics. Overall, the spectrum of thematic areas in the tweets covered four areas: 1) calls for action and increasing awareness of climate change; 2) discussions about the consequences of climate change such as extreme weather and representing a risk discourse; 3) policy debate about climate change and energy; and 4) local events associated with climate change. These

results are in line with the further details that the semantic network analysis, presented next, provides. In the Appendix, table 4 reports examples of tweets of each theme.

6.2 Semantic network analysis

Overall, the semantic network of tweets' content, excluding web links, confirmed insights provided by the latent semantic analysis. Network visualisations have been filtered to retain only the words with above-average thresholds for each centrality measure; this is necessary to have a 'readable' network. Degree centrality indicates the most salient topics expressed by words, which apart from 'climate change' included 'awareness', 'flood', 'action', and 'energy' – indicating clearly in what terms climate change was discussed on Twitter[4]. Compared to degree centrality, a node with the highest betweenness centrality takes up the role as a mediator of communication, as it is required to represent itself explicitly to bridge different clusters of words. Words with high betweenness (specifically high in-betweenness) represent macro topics that are then discussed in sub-networks (Figure 2). Community detection among nodes (words), indicating their clustering, is carried out using Newman's modularity method (Newman, 2006).

INSERT FIGURE 2

From Figure 2, we can see the macro-topics: 'energy', 'awareness', 'Earth', 'action', 'carbon', 'call' and 'agree'. 'Earth' and 'action' are about the same context of a call for action and mobilisation. 'Awareness' represents a macro-topic in which weather events and specific issues are related to climate change to increase the latter's salience (e.g., floods, Mumbai, the threat of extreme weather due to global temperatures rising). The macro-topic 'agree' relates to the scientific debate and relative consensus about the existence of climate change. It can be considered as evidence that climate change is still an object of discussion rather than taken for granted. Similar to this bridging concept, we have the node 'cause' which is both linked to climate change causes and to their effect (the closest nodes to cause are 'human' and 'effects').

‘Energy’ is linked mainly to policy items such as ‘committee’, ‘government’, ‘Obama’. It refers to energy policies and debate around this issue within the context of climate change. The presence of items related to ‘Obama’ and the United States (US) government is unsurprising given the prevalence of North Americans among Twitter’s English-language speaking users. Related but distinct sub-networks refer to ‘carbon’ and ‘call’ both concerned with the reduction of carbon emissions and the call for action about meeting targets about such reductions. In the latter case, countries such China and the United Kingdom (UK) frequently emerge as linked nodes.

The last set of nodes considered refers to words with high closeness centrality. We interpret an object with a high in-closeness centrality as one that positions itself close to the centre of reference playing a ‘connotative’ role or in other words, a ‘qualifier’. These qualifiers are words such as: ‘warm’, warm temperatures; ‘video’, the video content of some web links in the tweets; ‘tornado’, a case of extreme weather related to climate change in terms of their increased frequencies; ‘time’, as in ‘time for action’; ‘turn’, in terms of having reached a turning point about global warming. In addition, there is a set of verbs related to climate change such as ‘threaten’, ‘check’, ‘know’, ‘mean’ and ‘tackle’ that correspond to some of the macro-topics previously discussed, such as tackling climate change, explaining its threats and causes.

The semantic network extracted from the body of tweets presents a similar picture of the latent semantic analysis in identifying the most present topics (degree centrality) and their grouping (betweenness and closeness centralities).

6.3 Tweet classification and psychological categories

In this section, two different kinds of analysis will be discussed: first, the outcome of the language style classification performed using LIWC2007 to gain insights about the psychological processes associated with the tweets and second, sentiment analysis.

Starting with language style classification, as mentioned in the methodology section, only texts inputted by users were considered. The top part of Figure 3 presents the results of the language classification performed on the tweets' corpus.

INSERT FIGURE 3

The first chart presents the classification in terms of linguistic objects. Figure 3 presents the relative weight of psychological related categories of words in terms of percentages in the climate change dataset and two further datasets for benchmarking. The first of these comparative datasets is part of LIWC development documentation and consists of 168 million words from a wide range of sources (newspapers, blogs, novels, technical articles and text produced by people in experiments designed for benchmarking the software). This dataset provides the base rate values (grand means) of each language category. The second dataset is composed of over 24,000 tweets about nanotechnology from a previous study (Veltri, 2013b). The use of this second benchmarking dataset is helpful because it contains data from Twitter (while the base rate dataset does not) about a scientific issue in the public domain.

Figure 3 presents parallel plots comparing datasets (climate change N=974,053 words; nanotech N=317,286; base rate dataset N=168,354,504) for each linguistic category. Comparison reveals that the climate change dataset is characterised by rather high percentages of words related to: causation, 6.7% vs. 1.55% vs. 2%, $z=201.76$ $p < 0.001$ against baseline, $z= 135.04$, $p < 0.001$, against nanotechnology database; and, in particular, anger, 0.99% vs. 0.33% vs. 0.29%, $z=65.12$, $p< 0.001$ against baseline, $z=50.53$, $p< 0.001$ against nanotech. The climate change dataset also has a higher percentages of words related to motion, 6.38% vs. 2.3% vs. 1.3%, $z=162.91$, $p< 0.001$ against baseline, $z= 163.12$, $p< 0.001$ against nanotech. Regarding causation, this confirms the previous analysis and emphasises the importance of the 'causality discourse' about climate change on Twitter.

INSERT FIGURE 4

Second, the sentiment analysis carried out on the corpus reveals that tweets about climate change were classified mainly as neutral followed by roughly the same amount of positive and negative tweets (Figure 4). In terms of the emotional content, anger was most frequently identified. Overall, analysis revealed that tweets about climate change contained many words about causation, motion and anger.

6.4 Information sharing and media ecology

In this section, we report the findings from the content analysis performed on web links combined with relevant web metrics on how many times each link was shared on Twitter. First, the number of shares ($\acute{X}$ =2.98, *SD*=8.3, *N*=15914) in our corpus was not normally distributed as demonstrated by many previous studies.

The majority of web links in tweets about climate change were from professional news organisations such as newspapers or public broadcasting companies (67%), followed by nonprofessional blogs (8%) and non-governmental organisations (NGOs) (9% combining environmental NGOs and others). Links from other social media represent only 5% of the overall amount of web links in tweets.

Regarding content type, the majority of web links pointed at news articles reporting anything related to climate change (74%) followed by news articles that discussed a specific new scientific study related to climate change (14%). The third type of content was links to videos related to climate change (4%). Videos were from professional news sources (40% of the videos), social media (25%), and nonprofessional blogs (18%). Seventy-eight per cent of the analysed web links were 'descriptive' news articles about climate change and the remaining 22% had an 'affirmative'/'call for action' frame. ANOVA analysis was performed on both media source categories and content type regarding the number of shares received by web links but no significant effect was found – perhaps due to the very low number of shares for several categories.

The last step was to explore the relationship between the number of shares that a web link received and their score in terms of emotional valence (negative-positive) and arousal (bland-intense) measures using the SAM non-verbal scale. A statistically significant positive and mild correlation was found only for arousal and number of shares, r (766) = .14, $p < .01$, indicating that web links with content classified as more emotionally arousing were usually shared more often than others.

7. Discussion of findings

Findings from the different analyses offer an intriguing picture of the representations of climate change in the social medium Twitter. Analysing the semantic dimensions of tweets, we have identified a sophisticated discourse that includes multiple issues and angles. The four thematic clusters that emerged are related to calls for action and awareness of climate change, its consequences and causes, and the policy debate about climate change and energy. Of particular interest is the cluster about consensus and the causal relationships between the causes and effects of climate change. The emphasis on tweets' content related to climate change 'causation issue' revealed by the LIWC's analysis supports the idea of a social grounding process that is not fixed.

The media ecology analysis reveals a dependence upon professional sources of information such as newspapers and public broadcasting. In this case, Twitter users relied on traditional sources more than anything else and there is little ecological diversity of sources from the World Wide Web. Hence, findings related to the degree of sophistication of the discourse may simply reflect the traditional media discourse. However, on Twitter traditional sources were selected with remarkable sophistication, with users behaving rather like 'curators'. The explanation for this sophistication might be that we have tapped into a 'topical influence backbone' (Bogdanov et al., 2013), a sub-network of Twitter users who are followers of science and technology-related news. In any case, the lack of diversity of sources is notable

and in sharp contrast with common expectations about the diversity of communication channels online.

Regarding information sharing behaviour, a modest positive correlation was found between the number of shares and emotional arousal levels of the web links' content. Given that the large majority of tweets were classified as neutral, the weak relationship is to be expected. The result, however, is in line with previous findings that content with high, arousing emotional value, independently from its valence, has a greater chance of being shared online (Berger & Milkman, 2012). The emotional tweets we uncovered were mainly characterised by anger and sadness, as suggested by the LIWC classification.

Finally, we must draw attention to some limitations of this study. The generalizability of this study is limited by two restrictions affecting its research design. First, it was limited to tweets in the English language and second, monitoring covered a relatively short time period (although 7 days has been fairly standard in similar studies). Both limitations stem from the lack of necessary computing power to process and analyse a very large number of tweets. For example, tracking climate change over one year might have resulted in over twenty million tweets in English alone, and there is no way to predict how large this number might be if other languages were included.

8. Conclusions

This study proposed theoretical and methodological approaches for the study of an issue of major societal significance: climate change on the social media platform Twitter. The study combines different levels of analysis related to tweets. It combines a set of analytic techniques related to the content of tweets including semantic clustering, semantic networks, psychological processes classification, sentiment analysis and content analysis. Twitter offers a large amount of information that allows researchers simultaneously to study the content of tweets and interactions with tweets' content. Users publish their content and comment and

share (re-tweet) other users' tweets. Audiences on Twitter are often composed of articulated social ties and therefore social networks are also crucial for understanding the content that is discussed or commented on.

More generally, this study exemplifies the analytical stepping stones to a promising approach in public opinion and public understanding research based on social representation theory. If social representations are operationalised as collectively generated semantic networks together with the availability of data from online user-generated-content, we have a theoretical and analytical framework to analyse the formation and evolution of socio-semantic networks. Bauer & Gaskell (1999) outline a paradigm for research on the dynamics of social representations with a network approach. The minimal system involved in representation is the triad: two persons (subject 1 and subject 2) who share a concern with an object (O). The triangle of mediation [S-O-S] is the basic unit for the elaboration of meaning. Meaning is not an individual or private affair, but always implies the 'other'. To this triangle of mediation, a time dimension, capturing past and future, is added to denote the project (P) linking the two subjects and the object. There has been considerable progress in the analysis of the so-called 'two-modes' networks, analysing the co-evolution of semantic and social networks combining topics and users (Roth & Cointet, 2010). This approach is very well suited to move from cross-sectional measurements that failed to capture the fluid dynamics of public opinion discussions.

Another area of interest is the role of emotions in the anchoring and objectification of climate change as exemplified by the work of Birgitta Höijer (2010). To detect emotionally loaded discussion and opinions about an issue using just user-generated text has been and remains challenging. However, application of the method operationalised in the software LIWC goes beyond simple negative-positive sentiment analysis. This possibility opens up a new way of looking at social representations of scientific issues in which collective emotions can added to the picture in both theoretical and analytical terms.

Furthermore, we analysed the 'media ecology' surrounding a tweet, in other words, the analysis of the primary sources shared in a tweet, an issue of the utmost importance as users are more prone to complex use of information sources and information foraging behaviour (Pirolli, 2007). The relationship between different media as much as the selection and sharing of online science related news will play an increasing role in the context of studying public opinion dynamics related to public understanding of science: for example, one application is the study of selective exposure to science news online (Yang, 2015).

In summary, both the analysis of semantic networks and the psychological processes text classification yielded interesting results and are well suited to analyse social media data. There are still substantial limitations in this kind of analysis. The first is that semantic information about narrative structures are discarded by the use of co-occurrences based text mining techniques. Narrative structures underlying discourses related to the topic under investigation are an important factor on how social representations are organised. While the identification of themes and subthemes can be reliably obtained by automatic procedures and applied to a multi-languages corpus of large size, higher order structure of meanings such narratives and arguments\claims are much harder to automatically extract.

Social media constitute an increasingly vast pool of potential data for analysing public opinion dynamics regarding public understanding of science. It is important that social scientists take the opportunity to use these data; currently the corporate sector leads in this field. There is a need for a theoretically informed, methodologically robust, and critical – as well as ethical – use of online data. We hope this paper can trigger more discussion and research to fulfil this need.

References

Anderson, A. A., Brossard, D., & Scheufele, D. A. (2010). The changing information environment for nanotechnology: Online audiences and content. *Journal of Nanoparticle Research, 12*, 1083-1094.

Bastian, M., Heymann, S., & Jacomy M. (2009). *Gephi: An open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media. San Jose, US: San Jose Mc Enery Convention Center.

Batagelj, V., & Mrvar, A. (2014). Pajek - Program for large network analysis. Available at: http://vlado.fmf.uni-lj.si/pub/networks/doc/pajek.pdf.

Bauer, M. W., & Gaskell, G. (1999). Towards a paradigm for research on social representations. *Journal for the Theory of Social Behaviour*, *29*(2), 163-186.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192-205.

Bogdanov, P., Busch, M., Moehlis, J., & Singh, A. K. (2013). *The social media genome: Modeling individual topic-specific behavior in social media*. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining ASONAM. Niagara, Canada: University of Calgary.

Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Managemen*t, *66*(3), 262 - 278.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49-59.

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Social Methodology, 23*, 75-126.

Carley, K. M., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models, *Social Forces*, *70*(3), 601-636.

Carvalho, A. (2010). Climate change as a 'grand narrative'. *Journal of Science Communication*, *9*(4), C03.

Chung, C. J., & Park, H. W. (2010). Textual analysis of a political message: The inaugural addresses of two Korean presidents. *Social Science Information*, *49*(2), 215–239.

Collins, A. M., & Quillian, M. R. (1972). How to make a language user. In E. Tulving, & W. Donaldson (Eds.), *Organization of memory* (pp. 309-351). New York: Academic Press.

Danowski, J. (1993). Network analysis of message content. In W. Richards, & G. Barnett (Eds.), *Progress in Communication Science* (pp. 197-222). Norwood: Ablex.

Danowski, J. A. (2010). Inferences from word networks in messages. In K. Krippendorff, & M. Bock (Eds.), *The content analysis reader* (pp. 421-430). London: Sage Publications.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, *41*(6), 391-407.

Doerfel, M. L. (1998). What constitutes semantic network analysis? A comparison of research and methodologies. *Connections*, *21*(2), 16-26.

Fiedler, K., & Mata, A. (2014). The art of exerting verbal influence through powerful lexical stimuli. In J. P. Forgas, O., Vincze, & J., László, (Eds.), *Social cognition and communication* (pp. 43–62). New York: Psychology Press.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification Centrality in social networks conceptual clarification. *Social Networks*, *1*, 215–239.

Gaskell, G., Bauer, M., & Durant, J. (1998). The representations of biotechnology: Policy, media and public perceptions. In J., Durant, M., Bauer, & G., Gaskell (Eds.), *Biotechnology in the public sphere: A European source book* (pp. 3-15). London: Science Museum Press.

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., & Colleoni, E. (2011). Good friends, bad news-affect and virality in Twitter. *Future Information Technologies*, *185,* 34-43.

Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J., Hunsinger, M., Allen, & L. Klastrup (Eds.), *The International Handbook of Internet Research* (pp. 233-249). Berlin: Springer Verlag.

Höijer, B. (2010). Emotional anchoring and objectification in the media reporting on climate change. *Public Understanding of Science*, *19*(6), 717–731.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, *111*(24), 8788–8790.

Krippendorff, K. (1987). Association, agreement, and equity. *Quality and Quantity*, *21*, 109-123.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* The 19th International Conference on World Wide Web. Raleigh, US: Raleigh Convention Center.

Lancia F. (2012). The logic of the T-LAB tools explained. Available at: http://www.mytlab.com/textscope.pdf.

Lansdall-Welfare, T., Sudhahar, S., Veltri, G. A., & Cristianini, N. (2014). *On the coverage of science in the media: A big data study on the impact of the Fukushima disaster*. 2014 IEEE International Conference on Big Data. Washington, US.

Moscovici, S. (2000). *Social representations: Explorations in social psychology.* Cambridge: Polity Press.

Moser, S. C. (2010). Communicating climate change: History, challenges, process and future directions. *WIREs Climate Change*, *1*, 31-53.

Newman, M. (2006). Modularity and community structure in networks. PNAS, *103*(23), 8577–8582.

Ogden, S. (2001). The language of agricultural biotechnology: Terminate or be terminated. *Organization & Environment, 14*, 336–340.

Pearce, W., Holmberg, K., Hellsten, I., & Nerlich, B. (2014). Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 Report. *PLoS ONE*, *9*(4), doi:10.1371/journal.pone.0094785.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: A text analysis program (Version LIWC2007)*. Austin, TX: LIWC.net.

Pennebaker, J. W., & Chung, C. K. (2013). Counting little words in big data: The psychology of individuals, communities, culture, and history. In J. P., Forgas, V., Orsolya, & J. László (Eds.), *Social Cognition and Communication* (pp. 25-42). New York: Taylor and Francis.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford: Oxford University Press.

Popping, R. (2003). Knowledge Graphs and Network Text Analysis. *Social Science Information*, *42*(1), 91–106.

Roth, C., & Cointet, J. P. (2010). Social and semantic coevolution in knowledge networks. *Social Networks*, *32*(1), 16–29.

Schäfer, M. S. (2012). Online communication on climate change and climate politics: A literature review. *WIREs Climate Change,* doi:10.1002/wcc.191.

Schmidt, J. (2014). Twitter and rise of personal publics. In K., Weller, A., Bruns, J., Burgess, M., Mahrt, & C., Puschmann (Eds.), *Twitter and society* (pp. 3-14). New York: Peter Lang.

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. *The Communication Review*, *14*(3), 197–215.

Suerdem, A., Bauer, M. W., Howard, S., & Ruby, L. (2013). PUS in turbulent times II - A shifting vocabulary that brokers inter-disciplinary knowledge. *Public Understanding of Science*, *22*(1), 2–15.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Veltri, G. (2013a). Viva la Nano-Revolucion! A Semantic Analysis of the Spanish National Press. *Science Communication*, *35*(2), 143–167.

Veltri, G. (2013b). Microblogging and nanotweets: Nanotechnology on Twitter. *Public Understanding of Science*, *22*(7), 832-849.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society*. New York: Peter Lang.

Yang, J. (2015). Effects of popularity-based news recommendations ('most-viewed') on users' exposure to online news. *Media Psychology*, doi:10.1080/15213269.2015.1006333.

Yuan, E. J., Feng, M., & Danowski, J. A. (2013). 'Privacy' in semantic networks on Chinese social media: The case of Sina Weibo. *Journal of Communication*, *63*(6), 1011-1031.

Zywica, J., & Danowski, J. (2008). The faces of Facebookers: Investigating social enhancement and social compensation hypotheses; Predicting Facebook and offline popularity from sociability and self-esteem, and mapping the meanings of popularity with semantic networks. *Journal of Computer-Mediated Communication*, *14*(1), 1–34.

Endnotes

Table 1 Cross tabulation of available data in Twitter and related theoretical and methodological frameworks

Object of analysis	Level of Analysis	Theoretical Framework	Methodology	Research Aim
All content of tweets	Opinion mining of topics	Social Representations Theory	Quantitative thematic text mining; Semantic Network Analysis	Mapping of semantic clusters to explore semantic spectrum related to climate change (RQ1)
Only user generated content in tweets	Psychological process involved in the user-generated content	Psychological study of language	Computer assisted analysis of function words and linguistic category	Explore psychological processes involved in climate change tweets (RQ2)
Web links in tweets and related metadata (e.g. number of shares)	Information sharing behaviour	Personal Public and First degree of sharing	Content Analysis and Statistical Modelling	Explore first degree of sharing and media ecology related to climate change (RQ3); Model the role of emotion in predicting share count (RQ4)

Figure 1 Multidimensional scaling of co-occurrences of lexical units from the tweet's content (N=60,122). MDS is a set of data analysis techniques that allow us to analyse similarity matrices in order to provide a visual representation of the relationships among the data within a space of reduced dimensions, in this case to represent the relationships among the lexical units.

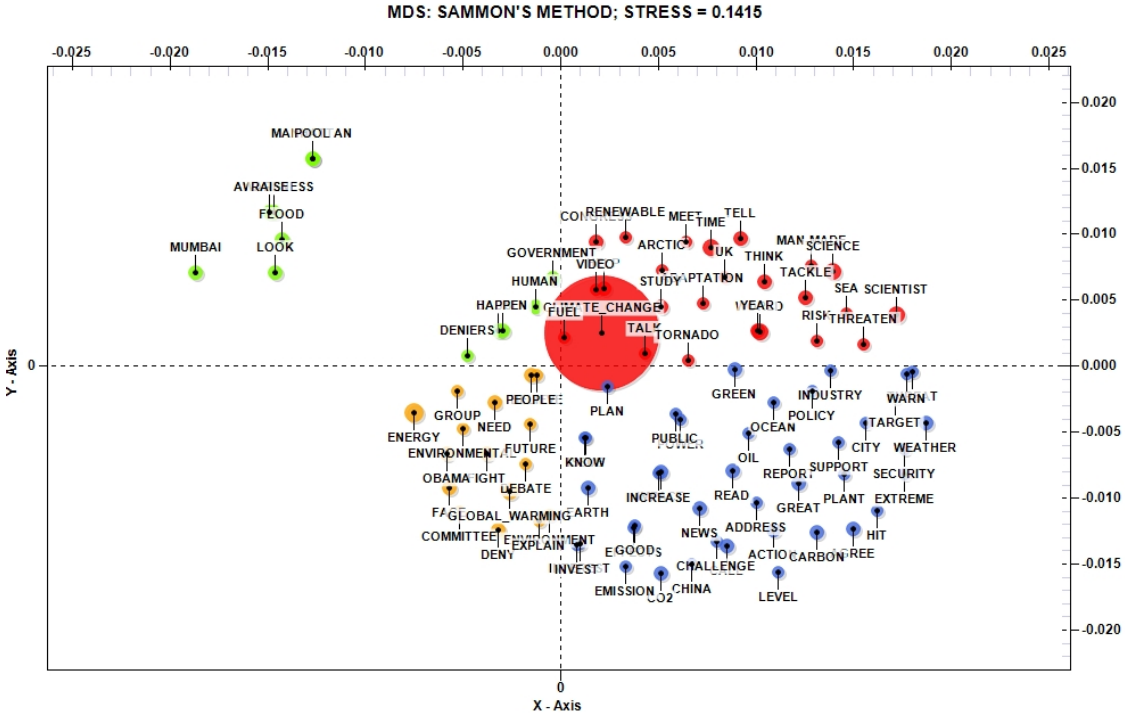


Figure 2 Semantic network of lexical units' co-occurrences from tweets related to climate change. Size of nodes indicates their betweenness centrality and colour their modularity class (group). Layout use is YifanHu Multilevel.

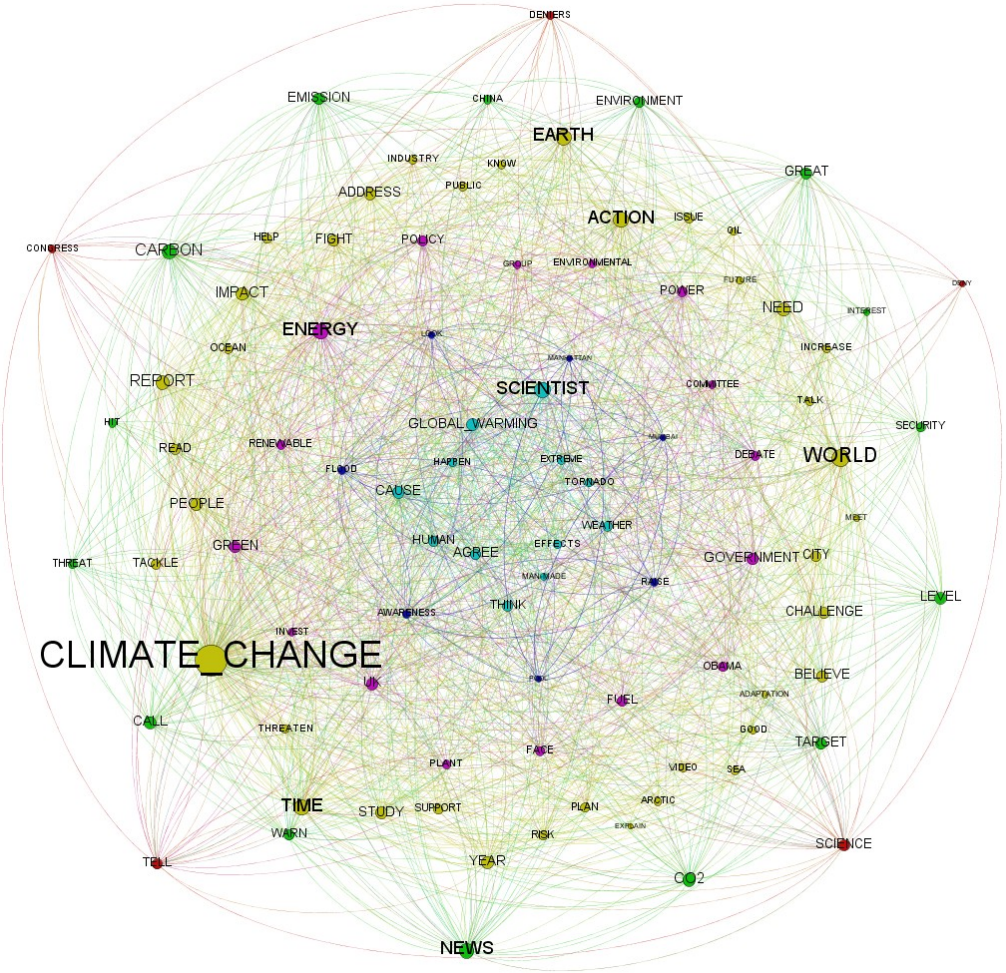


Figure 3 After classifying three different datasets, including the climate change on twitter one, Figure 3 compares the LIWC categories percentages (Y axis) of lexical units across the base rate, climate change and nanotech on the X axis.

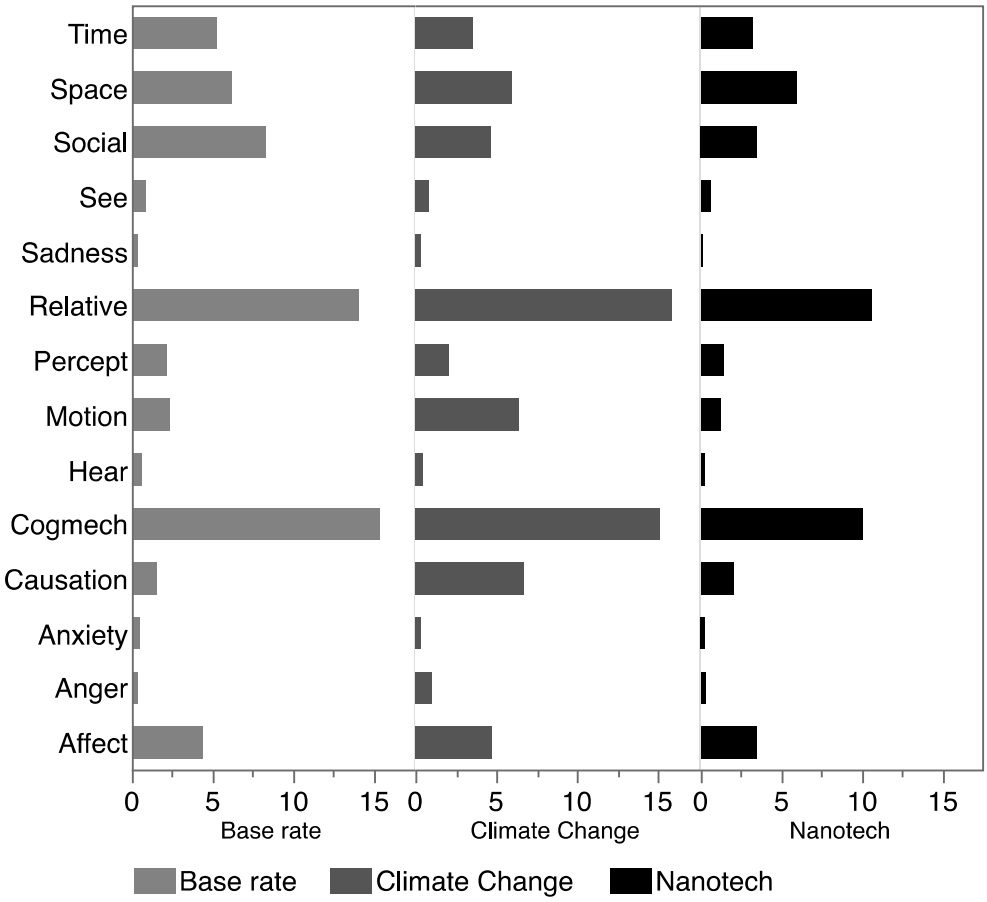
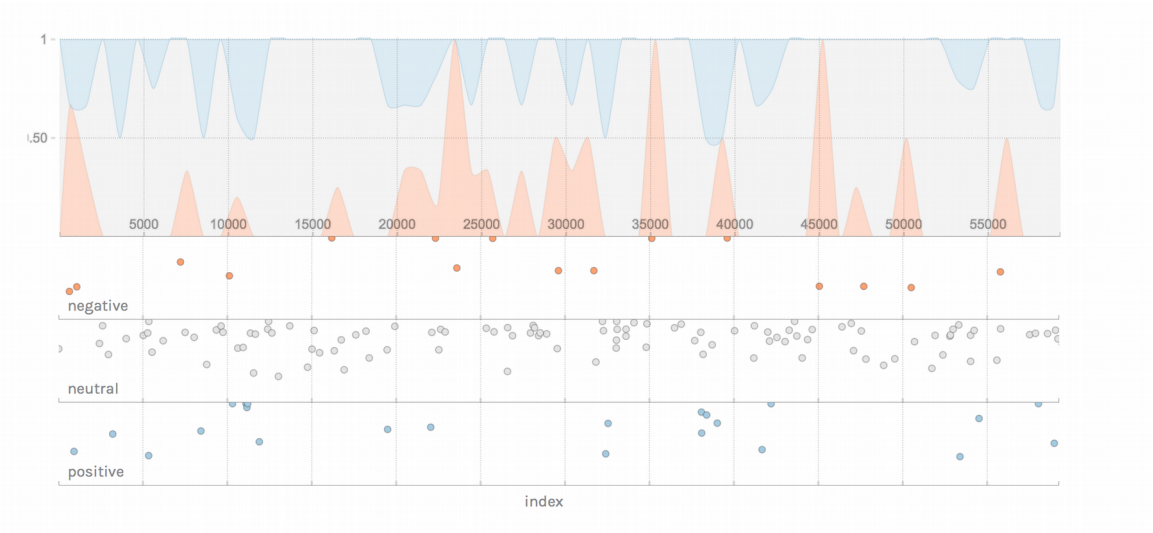


Figure 4 Distribution of sentiment classification of tweets across the entire corpus (N=60,120). Red dots stand for negative tweets, grey for neutral and blue for positive ones.



1 Personal statements on Twitter are limited to 140 characters and the effect of such technological affordance on user-generated content is something to consider.
2 The online random generator http://www.random.org/calendar-dates/ generated the 7 days. Weekdays were obtained from: http://www.random.org/calendar-dates/. Days selected were: 28/03/2013;15/05/2013;02/04/2013;11/06/2013; 17/05/2013;10/03/2013;24/04/2013.
3 In conversations with the developer of T-Lab (Prof Lancia) and LIWC (Prof Pennebaker) , we established that a corpus within the 100,000 words was the safest solution to perform the study's analysis.
4 Full reporting of centralities values is available in the Appendix, Tables 1 to 3.