Microblogging and Nanotweets: Nanotechnology on Twitter

Giuseppe A. Veltri¹

Abstract

The social web represents a new arena for local, national and global conversations and will play an increasing role in the public understanding of science. This paper presents an analysis of the representations of nanotechnology on Twitter, analysing over 24,000 tweets in terms of web metrics, latent semantic and sentiment analysis. Results indicate that most active users on nanotechnology are distributed according to a power law distribution and that web metric indicators suggest little conversation on the topic. In terms of content, there is a remarkable similarity with previous studies of nanotechnology's representations in other media outlets. Related to content is the sentiment analysis that indicates predominantly positively loaded words in the corpus. Negative sentiments mainly took the form of uncertainty and fear of the unknown rather than open hostility.

Keywords: Social Web, Twitter, Nanotechnology, opinion mining, latent semantic analysis.

Introduction

The portrayal of scientific and technological objects in the mass media plays a fundamental role in our understanding of how such issues are discussed and

¹ G.Veltri@uea.ac.uk

interpreted in the public sphere and constitute common sense knowledge (Moscovici, 2000). There is a wide scientific literature on this topic that combines communication theories and studies of mass media influences (e.g. agenda setting theory, McCombs and Shaw, 1993; second level agenda setting, McCombs & Ghanem, 2001; cultivation theory, Gebner and Gross, 1976; agenda building, Lang and Lang, 1983; second level agenda building, Fahmy, Wanta, Johnson, and Zhang, 2011) with public understanding of science research. In particular, studying how mass media frame an emerging technology is important for observing definitions and associated meanings that are legitimised or stigmatised. Science in the media can be described in terms of both the kind of medium analysed (e.g. printed media, television, movies, Internet and the World Wide Web) and of their theoretical approach to the study of the role of media in the public perception of a new technology. Clearly, the choice of the media analysed also influences the theoretical and methodological approaches. For example, years after the advent of Television as the main mass media, printed media were used comparatively more frequently because textual data could be analysed in both quantitative and qualitative approaches following consolidated methodologies. This is an exploratory study of the use of the social web platform Twitter¹ to share and discuss news on a 'new' technology: nanotechnology. It aims to contribute to the literature as one of the first study on the use of the social web in the context of the public understanding of science.

The earliest studies on science in the media obviously focussed on printed media. The focus was on non-fiction scientific journalism and it produced a great deal of

research about science journalists and the ‘quality and quantity’ of their media output. Many studies focused on ‘objective’ measures of media content, such as accuracy and readability (see e.g. Bader, 1990; Dunwoody and Scott, 1982). These studies reinforced the idea that ‘simplifications’ had been made in the way science was covered and implied that this was the reason for the low level of accurate scientific knowledge among the lay public, and also for negative attitudes towards science in general. Since many of these studies were based on a linear communication diffusion model (Bucchi, 2002) of scientific information, most of them concluded that not enough information was published and that what was published was not provided in sufficient quantity or detail to have been useful for the public in forming an opinion about the scientific or technological fact discussed. The result of this approach to the study of the role of mass media in science communication was a focus upon the accuracy of the scientific information published. Hence, the mass media was considered a ‘dirty mirror’ (Friedman, 1989) of science, or sometimes even as the direct cause of a hostile public reception as the result of the negative coverage of a given scientific issue. However, the deficit model approach and linear model of communication was criticised for a simplistic view of the processes of human communication and interpretation. Several studies argued that media coverage of controversies cannot simply be ‘improved’ by better ‘dissemination’ of scientific or technological information because it is shaped by structural relationships within communities and cultural contexts (Wilkins and Patterson, 1991). Other studies included the cultural contexts in which the framing of a new technology took place in the mass media and discussed the communication

processes involved. Some examples of this approach are: Gamson and Modigliani (1989) on nuclear power and cultural resonance of symbolic resources; Nelkin (1994), on the use of metaphors by scientists; and Neidhardt (1993), who looked at the public as a communication system.

A different approach was taken in one of the studies based on social representations theory (Moscovici, 2000), in which 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.

Communication plays a crucial role in social representations formation. Studies in this area are focused on the idea that in the public sphere there are competing definitions, in what is a complex game played for the control of semantics in the public sphere (Gaskell, Bauer and Durant, 1998). Considering that definitions are not just technical issues, but are a matter of framing for the purpose of opinion and attitude formation and for regulation, competing representations in the media is a field where the battle 'is being waged in the arena of language, as much as that of science' (Ogden, 2001, p. 340). The latter theoretical framework has informed several studies on media representations of scientific issues through the notion of 'anchoring' using both qualitative and quantitative methods or a combination of both (Veltri and Suerdem, 2011; Veltri, 2012).

Other media and Internet

In parallel to the theoretical shift from a linear diffusion model to a more interpretative approach, after the analysis of printed media, a new emphasis was placed on other mass media, mostly to television, as well as film, novels, etc. (e.g

Gopfert, 1996; Jones, 1997; Holliman, 2004). Researchers analysed television news/documentary shows and their representations of scientific and technological issues facing the difficulties in defining the sampling techniques and the absence of a largely shared and approved analytical method that could be applied to different types of data. Yet, there is a wide range of studies on the televisual framing of a scientific or technological issue, using either adapted content and thematic analysis from printed media (e.g. Leon, 2008) or adopting more interpretative approaches (e.g. Hoijer, 2010).

Another important change occurred in the media with the advent of the Internet and the World Wide Web, both attracted the attention of scholars aiming to study the role of websites in the dissemination of scientific knowledge (e.g. Eveland & Dunwoody, 1998; Byrne, Patrick et al. 2002). Initially, much emphasis was placed on a linear and mono-directional dissemination of information that conceptually resembled the way the 'old media' worked, but later the interactive and social features of Web 2.0 were taken in consideration (Waldrop, 2008). Recently, online media have started to be considered for their potential as public arenas for discussion among individuals in the forms of mailing lists, chat rooms and discussion forums (Delbourne et al., 2011; Triunfol, 2004; Rogers and Marres, 2000). This goes back to the theoretical approach to interpersonal conversation about science and technology and the processes of negotiations of meanings.

The incredibly fast diffusion of social web services complicates the roles that the social web plays in the ways individuals seek and process information in contemporary society. As virtually costless and extremely versatile channels of diffusion, as tools for coordinating and mobilizing social movements and their capacity for ‘crowd-sourcing’ (Shirky, 2010) large cognitive tasks are just some examples of characteristics of these new media. Social web platforms like Facebook and Twitter are often considered as ‘conversational hubs’. In the case of Twitter, Zappavigna (2011) argues that there is a shift to a ‘searchable’ discussion. Similarly, there have been an increasing number of studies using Facebook as a source of data and as a medium to study the sharing of information other than social interaction. Lewis, Kaufman et al. (2008) discussed the potential use of Facebook data in the context of merging ‘virtual’ and ‘real life’ variables and the identification of specific network behaviours. Examples of this are more and more frequent, for example Waters, Burnett et al. (2009) study on non-profit organizations and their use of Facebook, or Westling (2007) on the impact upon political communication and campaigns. In addition, there are increasing studies on the role of the social web in the diffusion and understanding of information related to health (Jain, 2009; Hawn, 2009).

There is also evidence of an increasing role of the social web in the information foraging behaviour of news online. Social networks are becoming increasingly more popular as venues for sharing content with friends than blogs, e-mail or message boards, according to a report by AOL and Nielsen Online (Pew Research Center, 2010). The analysis of 10,000 social media messages found that 92% of Web users used social sites to share content with their friends, with 60% of the

material shared were links to published content such as news and media sites.

With the rise of the participatory web and social media and resulting proliferation of user-generated content, the public potentially plays a far larger role in all stages of knowledge translation, including information generation, filtering, and amplification.

Giving its dual nature of information sources and conversations enablers of social media, this study explores the role of Twitter in the sharing and discussion of information on nanotechnology bridging the relevant theoretical and methodological literature from new media studies. According to Marwick & Boyd (2011), in the context of posting information about social issues the main reasons for re-tweeting are to spread or emphasise tweets to new audiences, to comment on someone else's tweet by re-tweeting and adding new content and to publicly agree with someone.

Social media and in particular Twitter is used as a source of data as new "infoveillance" methods such as mining, aggregating, and analysing online textual data in real-time are becoming available (Pang and Lee, 2008). Twitter is potentially suitable for longitudinal text mining and analysis. The brief (<140 characters) text status updates ("tweets") users share with "followers" (e.g., thoughts, feelings, activities, opinions) contains a wealth of data. Mining these data provides an instantaneous snapshot of the public's opinions and behavioural responses. Longitudinal tracking allows identification of changes in opinions or responses. In addition to quantitative analysis, the method also permits qualitative exploration of likely reasons why sudden changes have occurred (e.g.,

a widely read news report) and may indicate what is holding the public's attention.

An example of applying this methodology to scientific issues is Chew and Eysenbach (2010) in which the authors monitored the spread of information on the H1N1 (or 'swine flu') in Twitter using it as a real time tracking tool of public attention on the pandemic. However, there have been very few studies analysing the use of Twitter to represent a new technology or to 'anchor it'.

Representations of Nanotechnology

Studies about nanotechnology in the mass media have mainly analysed printed media and there are few studies about this new technology's representations in other media. Gorss and Lewenstein (2005) tracked nanotechnology in the American press from 1986 to 2004. According to the authors, media attention to nanotechnology seems to mimic the coverage of biotechnology in its early stages as a public issue: it started with low levels of attention, which then rose sharply as the issue spread from elite media to general outlets. As for biotechnology, the initial coverage of nanotechnology was largely positive and focussed on progress or the potential economic benefits and with relatively little discussion of risk. Similarly, Scheufele and Lewenstein (2005) also made predictions about key themes of coverage: foci on the economic potential or the scientific promise of nanotech early in the issue cycle. Friedman and Egolf (2005) analysed the US and UK coverage of nanotechnology's risks between 2000 and 2004 using the most prominent national newspapers and newswire services and concluded that the UK had a slightly more negative coverage than the US and placed a larger

emphasis on societal risks. This is in line with the findings of Gaskell, Eyck et al. (2005) that compared public attitudes towards nanotechnology in the UK and US and also provided an analysis of nanotechnology's coverage in one British and one American national newspaper (*The Independent* and *The New York Times*). Their research showed that the American newspaper covered the potential benefits of nanotechnology more than the British one, it also showed that risks' coverage appeared later compared to benefits.

Although Europe, or at least the UK, produced more negative coverage than the US, this does not mean that it was overall negative. Anderson, Allan et al. (2005) analysed the framing of nanotechnology in the British press, and suggested a considerable amount of uncertainty about the nature of nanotechnology with a tendency for being optimistic and highlighting the potential benefits more than the risks.

In the few other European countries studied, nanotechnology coverage has been overall positive indeed, for example in Norway (Kjølberg, 2009) and Denmark (Kjærgaard, 2010). The main themes emerged in these two countries were : 'nanotechnology as important for the future' and 'nanotechnology is under control' and 'benefits outweigh risks'.

In the German national press, nanotechnology also appears to be predominantly framed in positive terms with an emphasis on medical and economic benefits (Donk, Metag, Kohring, & Marcinkowski, 2012) very much in line with the above described studies. A recent study on nanotechnology's coverage in Slovenia (Groboljsek & Mali, 2012) highlights a similar pattern although there was very little coverage by local media.

Outside the analysis of newspapers' coverage, (Landau, Groscurth, Wright, & Condit, 2008) studied the impact of visual images of nanotechnology identifying a specific visual domain of science images that are organized according to polarities affecting the volatility of attitudes towards the technology under scrutiny. Similarly, Hanson (2012) analysed the evolution of Nanotechnology's images and their tropes that relies increasingly on computer generated images. Schummer (2005) instead astutely combined network analysis and content analysis using Amazon.com data to identify the networks of books that the public reads about nanotechnology in the US. His findings suggest that the broader public connects nanotechnology to visions about dissolving the human/machine distinction.

However, there are no studies on how nanotechnology's information is shared on the social web. The increasing importance of the Web for information seeking indicates the significance of studying this public arena and presents enormous potential in data collection as much as difficult methodological challenges.

Methodology

There is an on-going debate on how standard content analysis should be adapted to Web content (Herring, 2010) and this paper adopts a mixed research design. It uses web-metrics data about general pattern of information sharing on nanotechnology and an automatic text analysis on the content of tweets. In

addition, it performs ‘sentiment analysis’ to identify the general emotional charge – in terms of the negative or positive valence - of the corpus.

This study explores the use of social media to share and discuss information about nanotechnology and provide data to the following research questions:

- Research question 1: What are the strategies to gather and share information by Twitter about nanotechnology?
- Research question 2: What are the main drivers of these conversations,
- Research question 3: How is nanotechnology anchored in this social medium?
- Research question 4: What is the overall ‘sentiment’ (where sentiment indicated the overall polarization of tweets in terms of negative or positive emotions) of the tweets?

Sampling

The sampling procedure adopted the criteria of data collection based on their content feature and their time of publication. The study tracked for 60 days randomly selected tweets in 2011 with the procedure of collecting tweets containing the use of the keywords and hashtags ‘nanotechnology’; ‘nano’; ‘nanotechnologies’. The unit of analysis was a tweet and a major influenced the sampling strategy constrain: the computing power needed to process a very large number of tweets. The following strategy was adopted: in 2010 two test were carried out, monitoring one week in September and one in December, both tests produced an average of 2,600 tweets on nanotechnology per month. The upper computational limit of the two software used in the analysis roughly

corresponded to 30,000 tweets. Hence, 60 days in 2011 were selected using a random calendar date generator.

A hashtag is a tag embedded in a message posted on Twitter, consisting of a word within the message prefixed with a hash sign (#). Although the use of these previously mentioned hashtags occasionally captured tweets that were not written in English, the corpus was restricted to English tweets in order to be analysed using the lemmatization and automatic coding described in the next section. 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 on a Python script. The total amount of tweets collected during the 60 days of tracking is of 24634 ($N=24634$). The tracking algorithm includes the following information: content of the post, username and number of followers.

Procedure

Data were first analysed using descriptive web metrics in terms of basic characteristics of the tweets and the user's overall behavior. The second step was to aggregate all tweets in a new and distinct corpus in order to analyse it using text-mining techniques. The corpus was prepared for quantitative text analysis using the software T-Lab 7.3 and IBM Text analytics. Automatic lemmatization was performed with a custom dictionary, thus the vocabulary of the corpus was checked to disambiguate homographs. In addition, the resulting keywords and those excluded were checked to avoid false positives. An important decision regarded the text segmentation strategy and therefore the selection of the unit of context used in the analysis. Three options were available: sentences, chunks or

paragraphs. The unit ‘sentence’ was considered the best given the short nature of ‘tweets’. More precisely, an elementary context as a sentence was every sequence of words interrupted by full stop and carriage return, whose dimensions are up to 1000 characters (the limit for a tweet is 140).

The analysis proceeded in the following ways. First, it constructed a data table composed of context units by lexical units with presence/absence values.

Secondly, it applied TF-IDF (term frequency–inverse document frequency) normalization and scaling of row vectors to unit length (Euclidean norm). Then it clustered context units using as measure cosine coefficients and as method bisecting K-means². For each of the obtained partitions, it constructed a contingency table of lexical units by clusters and it applied a chi square test to all the intersections of the contingency table. The last step was to perform a correspondence analysis of the contingency table of lexical units by clusters.

Because the algorithm used (bisecting K-means) produces a hierarchical clustering, there several potential solutions in terms of cluster partitions. Thus, another important decision required setting a maximum number of clusters to be obtained. These are actually constrains in the sense that allowing for twenty clusters does not necessary mean that all of them will be produced because the algorithm will always minimize the number of clusters. After testing with two limits, ten and twenty, it obtained the same results indicating that the clusters obtained were stable. Subsequent to computing the quantitative semantic analysis, semantic clusters were interpreted going back to articles’ text fragments referred by most highly associate lemmas.

Sentiment Analysis

The sentiment polarization classification was operationalized according to the following procedure using IBM Text Analytics.

The first step was to create ‘concepts’. During the extraction process, the text data is scanned and analysed in order to identify interesting or relevant single words or phrases in the text. These words and phrases are collectively referred to as ‘terms’. Using the linguistic resources, the relevant terms are extracted and then similar terms are grouped together under a lead term called a ‘concept’.

The grouping is based on two techniques: concept root derivation and concept inclusion. The ‘Concept Root Derivation’ technique creates categories by taking a concept and finding other concepts that are related to it by analysing whether any of the concept components are morphologically related, or share roots. This technique is very useful for identifying synonymous compound word concepts, since the concepts in each category generated are synonyms or closely related in meaning.

The concept inclusion technique builds categories by taking a concept and, using lexical series algorithms, identifies concepts included in other concepts. The idea is that when words in a concept are a subset of another concept, it reflects an underlying semantic relationship. For example, if you have the concept ‘continental breakfast’, which has the component set {breakfast, continental}, and you have the concept breakfast, which has the component set {breakfast}, the algorithm would conclude that continental breakfast is a kind of breakfast and group these together.

When the concept root derivation or the concept inclusion techniques are applied, the terms are first broken down into components (words) and then the components are de-inflected. When a technique is applied, the concepts and their associated terms are loaded and split into components based on separators, such as spaces, hyphens, and apostrophes. For example, the term ‘system administrator’ is split into components such as {administrator, system}.

However, some parts of the original term may not be used and are referred to as stop words (e.g. and, as, by, of, etc).

A second order of identification proceeds in creating ‘types’ out of concepts.

Types are semantic groupings of concepts. When concepts are extracted, they are assigned a type to help group similar concepts. Several built-in types are delivered with IBM SPSS Text Analytics, such as <Location>, <Organization>, <Person>, <Positive>, <Negative> and so on. For example, the <Location> type groups geographical keywords and places. This type would be assigned to concepts such as Chicago, Paris, and Tokyo. Concepts that are not found in any type dictionary but are extracted from the text are automatically typed as <Unknown>.

The sentiment types are semantic groupings of concepts based on their emotional valence. Such emotional valence is associated using *Sentiwordnet* 3.0 (Esuli and Sebastiani, 2006), an enhanced lexical resource explicitly devised for supporting sentiment classification and opinion mining applications (Pang and Lee, 2008).

Findings

This section will first present a set of web metrics values followed by the latent semantic analysis and the sentiment analysis performed on the entire corpus of tweets ($N=24634$).

Webmetrics

The type tweets collected in the corpus were 92% new messages, 7% 'retweets' (a message forwarded to each of the user's followers) and 1% 'mentions' (or quotes, usually associated with replies). Perhaps unsurprisingly, 94% of tweets contained a link to a website. The high number of links to external websites indicates a trend for 'forwarding' behaviour rather than producing only single comments. On Twitter this kind of behaviour is not unusual. Two concepts are important in evaluating tweets and users: reach and exposure. Reach is the total number of unique Twitter users that received tweets about the search term. Exposure is the total number of times tweets about the search term were delivered to Twitter users. These two metrics can be applied to users as well. The exposure of a user is his/her number of followers, the reach is the number of unique followers compared to other users.

Table 1 summarize the aforementioned metrics for the corpus on nanotechnology. The total number of posts was 24,634 made by 10,030 users, the total exposure was 6,685,493 while the reach was 3,335,870.

Totals	Count	
Total Reach	3,335,870	
Total Posts	24,634	
Total People	10,030	
Users by Posts/Person		Number of users
	Number of posts	
	1	6,272
	2 to 3	2,234
	4 to 5	681
	> 5	843
Users by # Followers		
	Number of followers	Number of users
	0 to 100	7,649
	100 to 500	1,503
	500 to 1000	373
	1000 to 2500	282
	> 2500	223

Table 1 Breakdown of number of posts, reach and users.

Using these values we can compute the reach/exposure ratio that represents the depth of penetration of tweets about a topic. A lower reach/exposure ratio suggests that people are seeing tweets about a topic over and over, while higher ratio numbers suggest broad but shallow penetration of that topic.

A moderate or normal ratio will be anywhere between 0.2 and 0.59 (Bakshy, Hofman, Mason & Watts, 2011). This suggests a power-law distribution of tweets, re-tweets, and amplification. In this case, some people are tweeting multiple times; some influencers are tweeting to lots of followers; and most people are

tweeting once or twice to their smaller set of followers. In the case of the nanotechnology corpus, the ratio was 0.498 indicating a moderate ratio and a Pareto cumulative distribution³, this is a feature common to small-world networks and the very nature of World Wide Web (Albert, Jeong and Barabasi, 1999).

Figure 1 Top 25 users according to activity (number of posts about nanotechnology) and Top 25 users according to number of followers (exposure)

Figure 1 presents the top 25 most active users showing a typical power law distribution in the number of tweets per user. The list comprises a mix of

dedicated individual users and of twitter channels dedicated to nanotechnology news. However, the most active users appear to be individuals rather than the official channels of scientific institutions or organizations. It may be noted that Figure 1 confirms what has been described before about the nature of the tweets' distribution in terms of user activity, the pattern represents a Pareto's distribution in which a few users are very active while a large number of users posted only one or two tweets.

Figure 1 presents the top 25 users that tweeted on nanotechnology in terms of their exposure (or number of followers). The top two tweeters in terms of number of followers and therefore those with the highest exposure are the twitter channels of a TV channel dedicated to science – ScienceChannel- and an algorithmically constructed channel (tw_top_science) that selects and retweets science news from various official sources (such as Nature, Scientific American, NASA, etc.). The next two are science fiction fans with a large number of followers - 'Duncan Jones' is also the director of two science fiction movies and the fifth is the tweet channel of the well known science fiction writer William Gibson. As in the case of the most active users, the number of followers is distributed according a Pareto's distribution.

The next step is to analyse the content of the tweets using latent semantic analysis to identify the main semantic domains present in the overall corpus.

Latent Semantic Analysis

The algorithmic grouping of tweets in semantic domains according to the procedure described in the methodology section revealed a total of seven clusters (Figure 2).

The seven clusters were labelled as: 'Fight Cancer'; 'News/Conference'; 'Mobile Phone'; 'Science Projects', 'Applications', 'Opposition' and 'Residual'.

The semantic cluster 'Fight Cancer' accounted for 9.89% of the corpus. The cluster 'News Conference' was 18.6%, the cluster 'Mobile Phone' 15.5%, 'Science Projects' 15.5%, nanotechnology and a wide range of other applications for 9.2%. The only semantic cluster containing opposition to nanotechnology (Opposition) was 15.2%. The residual cluster was 16% of the corpus and contained all the elementary contexts that could not be grouped in a statistically significant way by the algorithm of the software.

The cluster 'Fight Cancer' - as suggested by its label - regards the application of nanotechnology to fight and cure cancer. The presence of a specific cluster indicates a significant emphasis on the medical application of nanotech and specifically for drug-delivery targeting tumour cells. Most of the tweets in this cluster were about news on potentially breakthrough medical treatments based on nanotechnology and this is the reason for the semantic proximity with 'news' cluster. Examples are 'Breast Cancer: How Far Have We Come? nanotechnology o rads 1,000x more sensitive than mammogram <http://t.co/oN2mcAs4> via @huffingtonpost'; 'EGEN, Nanotechnology Characterization Laboratory Partner to Accelerate Human Clinical Trial of Brain Cancer Therapy <http://t.co/dtFoCSJS>'; 'Cancer Nanotechnology: Methods and Protocols

(Methods in Molecular Biology): Early detection of cancer at the c...

<http://t.co/Xo6uZEIK>'.

Figure 2 Clusters obtained from the detection algorithm of elementary contexts from the overall corpus (lemmas, $N = 9199$). Geometric proximity/distance stands for semantic proximity or distance.

The second cluster 'News/Conference' contains tweets about public events related to nanotechnology, this is the biggest cluster indicating that a large number of tweets are about promoting these conferences and public meetings. Examples of these sorts of tweets are: 'REMINDER: Conference on applying #nanotechnology to save #environment in Taiwan <http://t.co/NqJ6oVI1> #Taiwan'; 'Hysitron will be exhibiting at the NanoTechnology for Defense Conference in Bellevue, Washington on 24-27 October. Visit us at Booth #212'; 'The Nanotechnology

Forthcoming in the journal *Public Understanding of Science*

Conferences <http://t.co/FsJ2xOSd> #FutureConcepts'; 'Trends in Nanotechnology (TNT2011) International Conference - Canary Islands (Spain): call for papers open - <http://t.co/GcSy2dfB>'; 'Indian form body to boost nanotechnology <http://t.co/q7d6l549>'.

The third cluster is labelled 'mobile phone' and it concerns nanotechnology's application to mobile phones, in particular smartphones. As with the potential medical application of nanotechnology (see the previous cluster on cancer), its application on mobile phones was at the centre of a substantial amount of tweets standing out among the other gadget/goods oriented applications. Examples of these tweets are: 'HzO Nanotechnology Seal Keeps Smartphone from Drowning %oÛÒ 2012 CES Press Preview in NYC: <http://t.co/DwEHL5XK>'; 'Nokia HumanForm Nanotechnology Smartphone Concept video Social Media News and viral video PkCrunch Mobile Phone ... <http://t.co/xkBrU3Y4>'; Dial N for Nano -- how nanotechnology will improve smartphone battery life <http://t.co/QYZ6qBcL> #tech'. The following cluster is 'Science Projects' concerning current nanotechnology research and also a broad discussion of its potential. Examples of tweets in this group are: Mapping Nanotechnology Innovations and Knowledge: Global and Longitudinal Patent and Literature Analysis (Integr... <http://t.co/Ja2ioIJL>

Nanotechnology is Driver to Take U.S. Out of Economic Doldrums, Says Nanofilm CEO, Scott Rickert; Sustainability for nanotechnology: NanoSustain: The NanoSustain project is investigating the life-cycle of sever... <http://t.co/O6Uyu6Ga>; Nanotechnology Now - Press Release: 'EU project IMOLA starts ... <http://t.co/7Dlsws1z>.

The cluster 'Application' refers to applications other than medical and mobile phones applications. It contains a wide array of potential applications with a noticeable amount of agricultural applications. Here are some example from the corpus:

Nanotechnology is Changing the Rules in Ceramics, says Nanofilm CEO Scott Rickert <http://t.co/7qEqWBTj>; Clear insulation coatings allow easy inspection of pipes. Possible w/ #nanotechnology <http://t.co/E7g3z7AV>;

Nanotechnology Applications for Agriculture & Water Safety in Developing Countries. <http://t.co/43LhEq1t>; Agronomist Alberto Popper tells <http://t.co/iRXvr2ZR> about what makes nanotechnology tick and how it is implemented. <http://t.co/BrzzdsxR>; #Geek #News: Nanotechnology used to create a superhydrophobic surface for clothes. No stains and it's just kinda cool. <http://t.co/pwAGYlGH>.

The last cluster is about the risk and opposition to nanotechnology. Essentially there are three themes present in the cluster: the potential toxicity of nanotechnology for food and consumer goods, its military application and the 'pandora's box' scenario of a technology that might run out of control. These are common themes in the risk discourse of nanotechnology found in many previous studies. Examples of tweets in in this cluster are: Dark side of nanotechnology <http://twitter.com/NanotecYPM/status/136178184235393025>; Why mankind must face and destroy the growing robot menace? Humans scared of robots? Nanotechnology is far more scary. <http://t.co/5DHVFu56>; Good things come in small packages? A chat about nanotechnology and food safety: Photo: Titanium

Forthcoming in the journal *Public Understanding of Science*

dioxide nanopar... <http://t.co/yOVFLMhd>; Consumer Fact, Not Science Fiction:
The Need To Keep Nanotechnology Products Safe

<http://twitter.com/NanotecYPM/status/13615485873778688>;

Risk vs. Reward: Why it's important to thoroughly assess the toxic risks of
nanotechnology <http://t.co/MdDvXKWr> #science #nanotechnology;

Video: Nanotechnology Ushers in Cloak of Invisibility | Truth Is Scary
<http://t.co/IRGvYswV>.

The next step is to outline the findings from the sentiment analysis to see the
overall emotional valence of the tweets contained in the corpus.

Sentiment analysis

As described in the previous methodology section, clusters were created in terms
on the emotional valence of the words, such as adjectives and nouns that have
either an average positive or negative value, of tweets and their co-occurrences.
Figure 3 shows the semantic network of categories associated with an overall
positive sentiment.

Figure 3 Semantic network of nanotechnology in terms of contexts (semantic domains) and sentiments (negative and positive). Asterisks indicate the strength of sentiment polarization, three asterisks being the strongest. The figure expands the details of the semantic network of ‘excellent’ in terms of contexts (semantic domains) and sentiments.

As shown by the network, positive feelings about nanotechnology are related in particular to business, medicine and science in general. As discussed in the latent semantic analysis, the application for medical treatments (in particular a remedy for cancer) and for commercial goods (smartphones) showed a major salience in the corpus. Positive markers such as ‘excellent’ and ‘well’ suggest positive feelings.

In Figure 3 we explore in more details the semantic network linked to excellent. From the rather dense network, it is possible to discern several contexts associated with the positive type excellent such as ‘medicine’, ‘apparel/clothes’, ‘chemistry’, ‘computer/computing’ ‘coating’, ‘physics/carbon nanotubes’, ‘home improvement’ etc. The semantic network of positive terms related to

nanotechnology confirms and substantiates the previous interpretation of the latent semantic analysis: a wide range of nanotechnology applications are represented as positive. It can be noticed that the strongest link is between the excellent and ‘almost improving’ category, indicating that nanotechnology is firmly within the inevitable scientific and technological progress narrative.

Next, we examine the semantic network associated with negative sentiments represented in Figure 3. The context is that negative tweets about nanotechnology were few in number. Most negative sentiments were expressed in the form of: ‘unknown problem’ (meaning unknown foreseeable problem); ‘not safe’; ‘needs more’ understanding; ‘mad/crazy’ referring to the dystopian and self-destructive scenario into which humankind is approaching; ‘hype’ a lower tone of critique regarding the unjustified hype around nanotechnology and its promises (or claimed promises). There is also a strong link with ‘would be good’, which relates to the scepticism surrounding the virtually risk free applications of nanotechnology.

Overall, the semantic network of negative sentiments reflects the content of the semantic cluster identified with the latent semantic analysis labelled as ‘Opposition’ in the previous analysis. The negative feelings in tweets about nanotechnology appear to be mainly associated with the unknown risks of nanotechnology, in particular the toxicity of nanoparticles, and side effects of its application to consumer goods rather than the most catastrophic scenarios that are nevertheless present in the corpus.

Discussion

The findings from the web metric analysis of the tweets on nanotechnology indicate that, overall, some people are tweeting multiple times, some influencers are tweeting to many followers, and most people are tweeting once or twice to their smaller set of followers. The most active users appear to be individuals rather than the official channels/representatives of scientific institutions or organizations, however, the latter are, in general (and perhaps unsurprisingly) the most followed users. In both cases, the most active and followed users are described by a Pareto's distribution. The low percentage of 'retweets' and 'mentions' indicate that nanotechnology is not an object of conversation on Twitter but rather another channel of diffusion (Boyd, Golder and Lotan, 2010). Moving to the content of the tweets, there is a remarkable similarity with the findings of previous studies on the discussion and representation of nanotechnology in the national press of different countries (e.g. Anderson, Allan et al., 2005; Kjølberg, 2009; Kjærgaard, 2010; Donk, Metag, Kohring, & Marcinkowski, 2012; Groboljsek & Mali, 2012; Veltri, 2012). Perhaps is not surprising, given the nature of the medium, that the large majority of tweets were dedicated to reporting news of events or applications for medical research and consumer goods of nanotechnology. Little emphasis was placed on resources to understand the science and technology involved. The tweets about concerns and opposition to nanotechnology were limited (around 15% of the corpus) and mainly focused the issue of potential toxicity of nano-materials and the military applications of nanotechnology (for example 'the cloak of invisibility' for soldiers

and spies). Related to content is the sentiment analysis that indicates a predominantly positively loaded words in the corpus. Negative sentiments mainly took the form of uncertainty and fear of the unknown rather than open hostility. The generalizability of this study is limited by three main restrictions that impacted upon its research design: limiting the sample of tweets to the English language only, the relatively short period of monitoring and tracking (60 days) - although this is fairly standard with these sorts of studies, and the characteristics of Twitter users in comparison to the general public or the average Internet user. Both limitations are caused by the lack of necessary computing power to process and analyse an extremely large number of tweets. For example, tracking over one year might have resulted in more or less than 300,000 tweets in English alone, there is no way to tell how large this number might be if other languages were also included.

Data from the numerous waves of data collection conducted by the PEW Internet Research Center (2010) indicate some of the groups who are notable for their relatively high levels of Twitter use include:

- Internet users ages 18-29 are significantly more likely to use Twitter than are older adults.
- Urban residents are roughly twice as likely to use Twitter as rural dwellers.
- Women and the college-educated are also slightly more likely than average to use the service.

On a more general note, individuals that use ICT tend to have larger and more diverse social networks, providing evidence against fears of socially isolated individuals. On average, the size of people's discussion network - those with

whom they discuss important matters- is larger among ICT users compared to non-users (Hampton, Sessions, Ja Her and Rainie, 2009).

Java, Song, Finin & Tseng (2007) investigated motivations behind tweeting use computational linguistics techniques confirmed that Twitter is mainly used to communicate daily activities and to seek or share information about a given topic.

Conclusions

This study combines a mature technique of semantic clustering with sentiment and web metrics analysis. The choice and combination of such methodologies mirrors the research questions defining the boundaries of this study. The overall approach is original because focuses on the 'representational space', the implicit frames and the social representations of nanotechnology diffused through a social media rather than investigating only the medium, its users and their structural properties, which is usually the case in many recent studies about the social web. The semantic clustering represents the main analysis with the web metrics and sentiment analysis as additional and contextual data. This strategy of analysis aims at answering the main research question: How is nanotechnology anchored in this social medium? Additional analysis informed by web metrics on the use of Twitter as medium represents useful information to understand nanotechnology news diffusion in the social web. Web metrics provide insights about users and the structure of Twitter's 'nanotechnology network' but an exhaustive study of the latter goes beyond the scope of this analysis. For example, a social network analysis could determine the network roles and characteristics of users (e.g. centrality) in the sharing and discussing of a topic. In this context, the main

research question would be: Who are those that spreads and collects nanotechnology's information on Twitter (or whichever social media) and what are their network's characteristics? However, a social network analysis of this kind poses several methodological challenges, for example it requires a longitudinal dimension of data collection and related computing power that are not easy to access.

Beside web metrics, sentiment analysis provides additional information on emotional polarisation associated to the semantic clusters and therefore present in the sharing and discussing of nanotechnology on Twitter. This information adds a further evaluative dimension of the semantic space, something important because the social web is above all a medium of opinions exchange. As much as for the network structure, emotional polarisation could easily be a study in itself with a more refined sampling strategy and in depth analysis. This sort of studies are a promising perspective in the studying of public opinion, giving that there are little doubts on the role of emotions in guiding the assessment of technologies and their risks and benefits (Roeser, 2010).

Findings suggest that nanotechnology is not an object of conversation on Twitter but rather information is shared and diffused by a limited number of 'power users', highly 'followed' and 'linked'. The content of the tweets indicates a semantic space that mirrors largely the one of traditional media outlets such newspapers. However, the combination of semantic and sentiment analysis

reveals that the negative emotions are more related to uncertainty rather than hostility.

The methodology and analysis applied in this study could monitor the status of the conversation of nanotechnology (or other technologies) on a given social media in order to detect the origin, growth and diffusion of semantic constructs and relative emotional polarizations. Adding a geospatial tagging could also help the precision and interpretation of such monitoring and even lead to predictions combined with other social variables (see the large-scale example of Leetaru, 2011).

This study represents a first attempt to use the social web as a source of data in public understanding of science research, offering a discussion about its potential and on the necessity of establishing a solid set of methodologies.

There can be few doubts, we believe, as to the growing importance of the social web in the study of the public understanding of science. These new arenas constitute a vast pool of potential data for analysing public opinion dynamics regarding the public understanding of science and risk perception. Social media are now a key source of data as new opinion-tracking methods such as mining, aggregating, and analysing online data in real-time have become available.

In the near future, the first objective will be to develop a standard methodology capable of dealing with the complexity and scope inherent to this kind of analysis, while the long-term goal is to analyse specific topics using online data and to examine real-time reactions to specific events. As mentioned above, it is possible to apply recent techniques of online opinion mining and social network analysis

of the diffusion routes and evolution of opinions and ideas (memes). This approach could lead to the development of a memetics of public understanding of science and risk perception (Weng et al., 2012). There are exciting opportunities ahead that need thorough theoretical and methodological debate, as is always the case when new opportunities arise.

¹ Twitter is a microblogging service that was founded in early 2006 to enable people to share short text messages—“tweets”—with other members. Because the system was originally designed for tweets to be shared via SMS, the maximum length of a tweet is 140 characters. Twitter combines elements of social network sites and blogs, but with a few notable differences. Like social network sites, profiles are connected through an underlying articulated network, but these connections are directed rather than undirected; participants can link to (“follow”) others and see their tweets, but the other user need not reciprocate. Like blogs, participants’ Twitter pages show all of their tweets in reverse chronological order, but there is no ability to comment on individual posts. User profiles are minimal and public, but users can make their tweet stream public or protected (a.k.a. private); the default and norm is public.

² TF-IDF (term frequency–inverse document frequency) formula is as follows:

$$w_{i,j} = tf_{i,j} \times idf_i$$
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Where:

$tf_{i,j}$ = number of occurrences of the i term in j document

df_i = number of documents containing i

N = total number of documents

Term Frequency ($tf_{i,j}$) value can be normalized as follows:

$$tf_{i,j} = tf_{i,j} / \text{Max}(f_{i,j})$$

where $\text{Max}(f_{i,j})$ is the maximum frequency of i (any term) in the j (document).

³ The Pareto’s distribution is a power law probability distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena.

References

- Albert, R., Jeong, H., & Barabasi, A. (n.d.). æL., 1999, The diameter of the worldwide web. *Nature*, 401, 130.
- Anderson, A., Allan, S., Petersen, A., & Wilkinson, A. (2005). The framing of nanotechnology in the British Newspaper Press. *Science Communication*, 27, 200-220.
- Bader, R. (1990). How science news sections influence newspaper science coverage: a case study. *Journalism Quarterly*, 67 (1), 88-96.
- Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.
- Boyd, D, Golder, S. and G. Lotan. 2010. “*Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.*” HICSS-43. IEEE: Kauai, HI, January 6.
- Bucchi, M. (2002). *Science in society: an introduction to social studies of science*. London: Routledge.
- Byrne, P. F., Namuth, D. M., Harrington, J., Ward, S. M., Lee, D. J., & Hain, P. (2002). Increasing public understanding of transgenic crops through the World Wide Web. *Public Understanding of Science*, 11(3), 293–304. Sage Publications.
- Chew, C., Eysenbach G., (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE* 5(11): e14118.

Forthcoming in the journal *Public Understanding of Science*

Delborne, J.A., Anderson, A.A., Kleiman, D.L., Colin M. and Powell, M. (2011).

Virtual deliberation? Prospects and challenges for integrating the Internet in consensus conferences. *Public Understanding of Science* 20 (3), 367-384.

Donk, A., Metag, J., Kohring, M., & Marcinkowski, F. (2012). Framing Emerging Technologies: Risk Perceptions of Nanotechnology in the German Press. *Science Communication*, 34(1), 5-29.

Dunwoody, S. & Scott, B. T. (1982). How scientists as mass media resources. *Journalism Quarterly*, 59, 52-59.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), p. 417-422, Genova, IT.

Eveland, W. P., & Dunwoody, S. (1998). Users and navigation patterns of a science World Wide Web site for the public. *Public Understanding of Science*, 7(4), 285-311. Sage Publications.

Fahmy, S. S., Wanta, W., Johnson, T. J., & Zhang, J. (2011). The path to war: Exploring a second-level agenda-building analysis examining the relationship among the media, the public and the president. *International Communication Gazette*, 73, 322-342.

Friedman, S. M. (1989). TMI: The media story that will not die. In: L.M. Walters, L. Wilkins and T. Walters (Eds.), *Bad tidings: Communication and catastrophe*. Hillsdale, NJ. Lawrence Erlbaum.

Friedman, S.M., and Egolf, B.P. 2005. Nanotechnology: risks and the media. *IEEE Technology and Society Magazine* 24, 5-11.

Forthcoming in the journal *Public Understanding of Science*

Gamson, W. A. and A. Modigliani (1989) Media discourse and public opinion on nuclear power: a constructionist approach, *American Journal of Sociology*, 95, 1-37.

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

Gerbner, G., & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication*, 26(2), 172-199.

Gopfert, W. (1996). Scheduled science: TV coverage of science, technology, medicine and social science and programming policies in Britain and Germany. *Public Understanding of Science* 5 (4), 361-374.

Gorss, J. and Lewenstein, B. (2005), The Salience of Small: Nanotechnology Coverage in the American Press, 1986-2004", *International Communication Association*, 1-22.

Groboljsek, B., & Mali, F. (2012). Daily Newspapers' Views on Nanotechnology in Slovenia. *Science Communication*, 34(1), 30-56.

Hampton, K., Session, L., Ja Her E. and Rainie, L. (2009). Social Isolation and New Technology. Pew Internet & American life Project. Available at <http://pewresearch.org/pubs/1398/internet-mobile-phones-impact-american-social-networks>.

Hanson, V. L. (2012). Amidst Nanotechnology's Molecular Landscapes: The Changing Trope of Subvisible Worlds. *Science Communication*, 34(1), 57-83.

Forthcoming in the journal *Public Understanding of Science*

Hawn, C. (2009). Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. *Health Affairs* 28 (2), 361-368.

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.

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

Holliman, R. (2004). Media Coverage of Cloning: A Study of Media Content, Production and Reception. *Public Understanding of Science* 13 (2), 107-130.

Jain, S.H. (2009). Practicing medicine in the age of Facebook. *The New England journal of medicine* 361 (7), 649-651.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56-65.

Jones, R. A. (1997). The Boffin: a stereotype of scientists in post-war British films (1945-1970). *Public Understanding of Science* 6 (1), 31-48.

Kjærgaard, R. S. (2010). Making a small country count: Nanotechnology in Danish newspapers from 1996 to 2006. *Public Understanding of Science*, 19, 80-97.

Kjølberg, K. L. (2009). Representations of nanotechnology in the Norwegian newspapers: Implication for public participations. *Nanoethics*, 3, 61-72.

- Landau, J., Groscurth, C. R., Wright, L., & Condit, C. M. (2008). Visualizing nanotechnology: the impact of visual images on lay American audience associations with nanotechnology. *Public Understanding of Science*, 18(3), 325-337.
- Lang, G. E., & Lang, K. (1983). *The battle for public opinion: The president, the press, and the polls during Watergate*. New York, NY: Columbia University Press.
- Leetaru, K. H. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, Volume 16, Number 9 - 5 September 2011. Available at <http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/rt/prINTERfriendlY/3663/304>
- Leon, B. (2008). Science related information in European television: a study of prime-time news. *Public Understanding of Science*, 17(4), 443-460.
- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4), 330-342.
- Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114-133.
- McCombs, M., & Ghanem, S. (2001). The convergence of agenda-setting and framing. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life: Perspectives on media and our understanding of the social world* (pp. 67-81). Mahwah, NJ: Lawrence Erlbaum.

Forthcoming in the journal *Public Understanding of Science*

McCombs, M.E., and Shaw, D.L. (1993). The Evolution of Agenda-Setting Research: Twenty-Five Years in the Marketplace of Ideas. *Journal of Communication*. 43(2), 58-67.

Moscovici, S. (2000). *Social Representations: Explorations in Social Psychology*. Cambridge: Polity Press.

Nelkin, D. 1994. 'Promotional metaphors and their popular appeal'. *Public understanding of science*, 3 (1), 25-31.

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

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1/2), 1–135.

Pew Research Center (2010). Global Publics embrace social networking. Report available at: <http://pewresearch.org/pubs/1830/social-networking-computer-cell-phone-usage-around-the-world>.

Roeser, S. (2010). Emotions and risky technologies. Delft: Springer.

Rogers, R., and Marres, N. (2000). Landscaping climate change: a mapping technique for understanding science and technology debates on the World Wide Web. *Public Understanding of Science*, 9(2), 141.

Scheufele, D. A., & Lewenstein, B. (2005). The public and nanotechnology: How citizens make sense of emerging technologies. *Journal of Nanoparticle Research*, 7, 659-667.

Forthcoming in the journal *Public Understanding of Science*

Schummer, J. (2005). Reading nano: the public interest in nanotechnology as reflected in purchase patterns of books. *Public Understanding of Science*, 14(2), 163-183.

Shirky, C. (2010). *Cognitive surplus: Creativity and generosity in a connected age*. New York: Penguin Press.

Triunfol, M. L. (2004). Dynamics of List-Server Discussion on Genetically Modified Foods. *Public Understanding of Science*, 13(2), 155-175.

Veltri, G. A. (2012). Viva la Nano- Revolucion! A Semantic Analysis of the Spanish National Press. First published online 22nd of March 2012, *Science Communication*.

Veltri, G. A., & Suerdem, A. (2011). Worldviews and discursive construction of GMO-related risk perceptions in Turkey. First published online 15th of November 2011, *Public Understanding of Science*.

Waldrop, M. M. (2008). Science 2.0: great new tool or great risk? *Scientific American* May 2008. Available at <http://www.scientificamerican.com/article.cfm?id=science-2-point-o>

Waters, R. D., Burnett, E., Lamm, A., & Lucas, J. (2009). Engaging stakeholders through social networking: How non-profit organizations are using Facebook. *Public Relations Review*, 35(2), 102-106.

Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports* 2, 335.

Westling, M. (2007). Expanding the Public Sphere: The Impact of Facebook on Political Communication. *Society* 28, 835-860.

Forthcoming in the journal *Public Understanding of Science*

Wilkins, L., & Patterson, P. (Eds.). (1991). *Risky business: Communicating issues of science and risk, and public policy*. New York, Greenwood.

Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society*, 13(5), 788–806.