Computational Social Networks
a SpringerOpen Journal

RESEARCH Open Access

Sentiment leaning of influential communities in social networks

Borut Sluban[1]*, Jasmina Smailović[1], Stefano Battiston[2] and Igor Mozetič[1]

*Correspondence:
borut.sluban@ijs.si
[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
Full list of author information is available at the end of the article

Abstract

Social media and social networks contribute to shape the debate on societal and policy issues, but the dynamics of this process is not well understood. As a case study, we monitor Twitter activity on a wide range of environmental issues. First, we identify influential users and communities by means of a network analysis of the retweets. Second, we carry out a content-based classification of the communities according to the main interests and profile of their most influential users. Third, we perform sentiment analysis of the tweets to identify the leaning of each community towards a set of common topics, including some controversial issues. This novel combination of network, content-based, and sentiment analysis allows for a better characterization of groups and their leanings in complex social networks.

Keywords: Social networks; Communities; Sentiment analysis; Influence

Introduction

Environmental and sustainability issues are among the major societal concerns today. The formulation of environmental policies is often a result of the interaction between antagonistic interest groups, including policy makers (governments and international organizations), advocacy groups representing the interest of specific industry sectors, and civic activists. The motivation for this research is to contribute to a better understanding of the dynamics of advocacy and activism around policy issues. We expect that the results will help policymakers in monitoring the response of various interest groups to the proposed regulations and policy targets.

The explosive growth of social media and user-generated contents on the Web provides a potentially relevant and rich source of data. This work is based on data from Twitter [1], a social networking and micro blogging service with over 270 million monthly active users, generating over 500 million tweets per day.

We collect a broad range of tweets related to the environmental issues and address the following research questions:

- Can one identify influential communities and environmental topics of interest?
- Are there differences in their leanings towards various environmental topics?

Our results indicate that there are observable differences in sentiment leanings towards various environmental issues between the major communities.

There are several aspects of Twitter data analysis that are relevant for this research. On the one hand, Twitter is a social network, and several types of networks can be constructed from the data, e.g., followers, mention, or retweet networks. Network analysis algorithms then yield interesting network properties, such as communities, modularity, various, and centralities. On the other hand, Twitter data can also be analyzed for its contents, by applying text mining and sentiment analysis algorithms. A novelty of our research is that we combine both types of analysis. We detect influential communities, identify discussion topics, and assign sentiment of the communities towards selected topics.

There are three different ways how users on Twitter interact: 1) a user follows posts of other users, 2) a user can respond to other user's tweets by mentioning them, and 3) a user can forward interesting tweets by retweeting them. Based on these three interaction types, Cha et al. [2] define three measures of influence of the user on Twitter: *indegree influence* (the number of followers, indicating the size of his audience), *mention influence* (the number of mentions of the user, indicating his ability to engage others in conversation), and *retweet influence* (the number of retweets, indicating the ability of the user to write content of interest to be forwarded to others). They find that mention and retweet influence are correlated, but that indegree alone reveals little about the user's actual influence. This is also known as *the million follower fallacy* [3]. Instead of the number of followers, they show that it is more influential to have an active audience who mentions or retweets the user. Suh et al. [4] analyze factors which have a positive impact on the number of retweets: URLs, hashtags, the number of followers and followees, the age of the account, but not the number of past tweets. Bakshy et al. [5] quantify the influence on Twitter by tracking the diffusion of URLs through retweet cascades. They find that the longest retweet cascades tend to be generated by the most influential users in the past.

Closely related to our research is the work by Conover et al. [6], albeit applied to the problem of political polarization. They construct both retweet and mention networks from political tweets and apply community detection. It turns out that the retweet network exhibits clear community segregation (to the left- and right-leaning users), while the mention network is dominated by a single community. In [7], they compare the predictive accuracy of the community-based model to two content-based (full text tweets and hashtags-only) models. The community-based model constructed from the retweet network clearly outperforms the content-based models (with the accuracy of 95 vs. 91 %).

The above research indicates that the retweet influence seems to be the most promising measure of influence on Twitter, and that community detection in the retweet network will likely yield the most influential communities. However, in the environmental domain, the community segregation is not as clear as in the political domain. We therefore characterize communities not only by their influential members, but also by their prevalent discussion topics and sentiment.

Sentiment analysis has been applied to Twitter in several domains [8], most notably for stock market predictions [9], and in political elections. There has been some controversy whether Twitter analysis can be used to predict the outcome of elections—Gayo-Avello gives a survey of various studies [10]. We have successfully applied Twitter sentiment analysis to monitor Slovenian presidential election in 2012 and Bulgarian parliamentary elections in 2013 [11]. Most of the other approaches are based on tweet volume or simple sentiment analysis by counting positive and negative sentiment words in tweets. In contrast, we apply supervised machine learning, the SVM classification in particular [12]. The training data comes either from manually annotated tweets (which are problem-specific and of high quality, but expensive in terms of resources needed), or from generic, smiley-based tweets [13] (which are of lower quality, but very extensive).

This paper is based on our preliminary work, presented in a workshop proceeding [14], and, in several aspects, extends the proposed methodology. First, the experiments capture 1 year of Twitter data and hence analyze twice the original amount of data. Second, the structural properties of most prominent communities discussing environmental topics are examined. Third, content filtering is enhanced by similarity calculation in a multi-dimensional vector space. Finally, a custom sentiment model, trained on manually labeled domain-specific tweets, is applied to produce better sentiment classification results.

The paper is organized as follows. In the "Methodology: discovering influential communities and their sentiment" section, we present the network and content analysis employed in our work. We describe the Twitter data acquisition and construction of the retweet network. We use a standard community detection algorithm and define the Twitter user and community influence measures. A standard text mining approach is used to identify topics discussed by the major communities. For sentiment analysis, we construct a binary SVM classifier with neutral zone, from three different sets of training data. The "Results and discussion" section describes the outcomes of the experiments. First, we analyze the structural properties of the most influential communities, in terms of their internal and external influence, and balance of the influence distribution. We identify categories of influential communities (e.g., environmental activists, news media, skeptics, celebrities) and the topics of their interests. Sentiment classification is applied to the tweets of different communities, and sentiment leaning of the communities towards different topics is analyzed. We highlight interesting findings and some unexpected results. We conclude with plans for future work.

Methodology: discovering influential communities and their sentiment

We have monitored Twitter for a period of the entire year 2014. We use the Twitter Search API and define a wide range of queries to select tweets related to environmental and energy topics (see see Table 6 in Appendix for the full list of queries). The collected environmental tweets are then used to construct a social network and identify influential users and communities, as well as their topics of interest and sentiment. The process of identifying community interests and their leanings consists of three steps. First, the network of users retweeting each other is constructed, and the densely connected communities are detected. Second, the content published by these communities is analyzed to reveal the communities' interests, and finally, sentiment analysis is performed to asses the sentiment leaning of the communities with respect to different topics of interest.

Network structure and influence measures

We explore which Twitter users share similar content on environmental topics. To model this phenomenon, we construct a retweet network, connecting users who are in a retweet relation, i.e., an undirected edge between two users indicates either one user retweeted the other or vice versa. The network is constructed from 30.5 million tweets about environmental topics, acquired between January 1, 2014 and December 31, 2014. The network consists of 3.7 million users (nodes) linked by 9.7 million retweet relations.

The largest part of the network consists of one large connected component of 3.4 million users, the rest are components of size smaller than 1000 users. In the largest component, we want to find groups of users that share similar views on environmental topics. If we assume that retweeting is a proxy of expressing agreement on the published content, the retweet network can be regarded as consisting of the connections between users who agree on a certain topic. Therefore, the problem translates into partitioning the network in the so-called communities. In the field of complex networks, the notion of "community" corresponds, loosely speaking, to a subset of nodes that are more densely connected among themselves than with the nodes outside the subset. Several definitions of community and methods to detect communities have been proposed in the literature (see [15] for a review).

We apply a standard community detection algorithm, the Louvain method [16], to our retweet network. The method partitions the network nodes in a way that maximizes the network's modularity. Modularity is a measure of community density in networks. It measures the fraction of edges falling within groups of a given network partitioning as compared to the expected fraction of edges in these groups, given a random distribution of links in the network [17]. Among the available detection algorithms in the optimization-based class, the Louvain method is one of the few methods that are suitable: (i) to analyze large networks with good scalability properties and (ii) to avoid ex-ante assumptions on their size [18].

Further, we propose an approach to identify the most influential users in the network, i.e., users whose content is apparently approved and shared the most. Let the retweet network be represented as a directed graph G, with edges $E(G)$. A directed edge $e_{u,v}$ from the user u to the user v indicates that contents of the user u have been retweeted by the user v. Let $w(e_{u,v})$ be the weight of the edge $e_{u,v}$ indicating the number of times that the user v retweeted the contents of the user u. Then *user influence* $I(u)$ is defined as

$$I(u) = \sum_{e_{u,v} \in E(G)} w(e_{u,v}) \tag{1}$$

The differences in the structure of the detected communities $C_1, \ldots, C_n$ are examined through the influence of the users of a particular community C_k. We address this by measuring the *intra and inter-community influence* of each community, as well as by measuring the distribution of influence among the community's users.

Community influence is defined as the cumulative influence of all its users,

$$I(C) = \sum_{u \in C} I(u) = \sum_{u \in C} \left(\sum_{e_{u,v} \in E(G)} w(e_{u,v}) \right) \tag{2}$$

It can be divided into the influence that the community users have within their own community and the influence they exert outside their community. Hence, we define *intra-community influence* I_{in} and *inter-community influence* I_{out} as:

$$I_{in}(C) = \sum_{u \in C} I_{in}(u) = \sum_{u \in C} \left(\sum_{\substack{e_{u,v} \in E(G) \\ v \in C}} w(e_{u,v}) \right) \tag{3}$$

$$I_{out}(C) = \sum_{u \in C} I_{out}(u) = \sum_{u \in C} \left(\sum_{\substack{e_{u,v} \in E(G) \\ v \notin C}} w(e_{u,v}) \right) \tag{4}$$

The ratio between these two measures I_{out}/I_{in} reveals the extent to which a community is influential outside its "borders" versus its internal content exchange.

Furthermore, to measure the distribution of user influence within a community, we use the *Herfindahl-Hirschman index* (HHI), commonly used in economics to measure the amount of competition among leading companies in an industry with respect to their market share [19]. When applied in the context of community structure, we look at the N leading users $u_i,\ i \in \{1, \ldots, N\}$, in a community C in terms of their normalized intra-community influence $r_i = I_{in}(u_i)/\sum_{j=1}^{N} I_{in}(u_j)$. Hence, the Herfindahl-Hirschman index is defined as

$$HHI(C) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} \left(\frac{I_{in}(u_i)}{\sum_{j=1}^{N} I_{in}(u_j)} \right)^2 \tag{5}$$

The squared sum of influence ratios ranges from $1/N$ to 1, where lower values indicate a dispersed and more balanced influence distribution, whereas higher values reflect the community influence being concentrated only on few strongly influential users.

Content identification and filtering

The retweet relation can be considered as the agreement between users on the published content. Hence the retweet network reveals which users support similar interests, without looking into the actual content. On the other hand, to identify the content and to see what are different groups of users talking about, we adopt a standard text mining approach as follows.

1. For each group of users g_i, $i \in \{1, \ldots, N\}$, create a document d_i that aggregates all the content which the users of the group g_i have published.
2. The vocabulary (i.e., the set of terms) used by groups $\{g_1, \ldots, g_N\}$ is obtained from the documents $\{d_1, \ldots, d_N\}$. *Term frequency* $TF_i(t)$ denotes the number of appearances of a term t in a document d_i.
3. For each term t from the vocabulary, *document frequency* $DF(t)$ is the number of documents in which t appears.
4. For each of the documents, $\{d_1, \ldots, d_N\}$ construct a bag of words (BoW) vector where each term value in the vector is the TFiDF value of the term t from the vocabulary:

   $$TFiDF_i(t) = TF_i(t) \cdot \log \frac{N}{DF(t)} \tag{6}$$

   Term frequency-inverse document frequency (TFiDF) is a standard and widely used measure of importance of a term t to a document in a collection of documents [20].

We use this adopted text mining approach to identify the terms that are the most distinctive and therefore the most characteristic for the content tweeted by different groups of users. More specifically, we use the detected retweet communities as the groups of users. Next, we employ the above procedure to summarize and represent the most characteristic topics in the content of each community. Such content identification and representation is done by displaying only the selected number of the highest *TFiDF* ranked terms from a BoW vector constructed for a selected community. In this way, we are able to get a readable and reliable overview of the specific interests and topics discussed in the observed communities.

On the other hand, for the purpose of identifying the leaning of different communities towards specific topics of interest, we have to retrieve the individual tweets forming a certain topic. We employ a filtering procedure based on document similarity, to obtain tweets that revolve around a specified topic (query). In this case, each tweet from the dataset is treated as an individual document and is transformed into a BoW vector. Hence, the filtering works as follows.

1. The vocabulary V of a specific domain is obtained from all unique tweets acquired for the targeted domain. From V the base of the document vector space is constructed by standard text preprocessing (stemming, stop-word removal, n-grams) resulting in terms $t_1, \ldots, t_n$.
2. For each tweet tw_i, $i \in \{1, \ldots, m\}$, from the dataset D, a BoW vector $\boldsymbol{v_i}$ of *term frequencies* $TF_i(t)$ for each term t in tw_i is constructed and normalized.
3. A BoW model of the examined domain can be represented by a matrix M with rows $\boldsymbol{v_i}$ for each $tw_i \in D$.
4. The dataset D is filtered according to a query that is transformed into a normalized BoW vector $\boldsymbol{q}$.
5. Similarity between query q and tweets $tw_i \in D$ is calculated as $\boldsymbol{s} = M \cdot \boldsymbol{q}$:

$$\begin{bmatrix} s_1 \\ \vdots \\ s_m \end{bmatrix} = \begin{bmatrix} m_{1,1} & \cdots & m_{1,n} \\ \vdots & & \vdots \\ m_{m,1} & \cdots & m_{m,n} \end{bmatrix} \cdot \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix} \tag{7}$$

where s_i, $i \in \{1, \ldots, m\}$, is the cosine similarity[1] between the query vector $\boldsymbol{q}$ and $\boldsymbol{v_i}$ representing tweet tw_i, and $m_{i,j}$ is the (normalized) term frequency of term t_j in tweet tw_i.

Given a query q and the calculated similarity vector $\mathbf{s}$, the filter returns tweets tw_i for the indices i where s_i is greater than a given threshold. Note that, since the number of terms (n) and especially the number of tweets (m) can be very large, in practice the computations are performed with sparse representations of vectors and matrices.

Sentiment analysis

Our goal is to measure the collective attitude of a Twitter community towards a certain topic. The first step is to measure the sentiment of each individual tweet posted by the community. To perform Twitter sentiment analysis, we construct a sentiment classifier from the training data. We employ the Support Vector Machine (SVM) algorithm [12], and in particular its SVM^{perf} [21–23] implementation. The SVM algorithm requires a labeled collection of instances to build a model. We have collected three labeled Twitter datasets which differ in terms of size, discussion topics, and labeling method. We have trained three corresponding sentiment models and compare their performance on the same testing set. The best sentiment classification model is then used in the rest of our analyses.

The first dataset consists of 1.6 million positively and negatively labeled tweets collected by the Stanford University [13][2]. The labeling of the tweets is based on the presence of positive (e.g., ":)") or negative (e.g., ":(") emoticons, which were then removed from the dataset for training. Although such approach does not provide the highest labeling quality, it is a reasonable and inexpensive substitute for manual tweet labeling [24]. The tweets in this dataset are general and not focused on any specific domain.

The second dataset consists of general English tweets too, but the tweet labels were obtained by manual annotation. In this dataset, there are 25,721 positive, 23,250 negative, and 37,951 neutral hand-labeled tweets.

The tweets in the third dataset are a uniformly sampled subset of our environmental tweets, therefore highly domain-specific. This dataset consists of 2,850 positive, 5,569 negative, and 11,439 neutral hand-labeled tweets, from January to December, 2014. We randomly choose 20 % of these tweets (preserving the labeling distribution of the whole dataset) as a test set, used for evaluating the trained sentiment models. The rest of the 80 % tweets from the domain-specific dataset were used for training the domain-specific sentiment model.

Sentiment models are built only from the positive and negative tweets. However, the classification covers three categories: positive, negative, and neutral as well. A tweet is classified as positive (negative) if its distance from the SVM hyperplane is higher than the average distance of positive (negative, respectively) training examples from the hyperplane. Otherwise, i.e., if it is too close to the hyperplane, it is classified as neutral. Similar approaches to adapting the binary SVM classifier to the three-class setting were already applied in our previous studies [24, 25].

Twitter messages are adequately preprocessed, using both standard and Twitter-specific techniques. Standard preprocessing [26] includes tokenization, stemming, unigram and bigram construction, removing terms which do not appear at least twice in the corpus, and construction of term frequency (TF) feature vectors.[3] Additionally, Twitter-specific preprocessing [8, 13, 24] transforms usernames, hashtags, and collapses repetitive letters.

We build three sentiment models (smiley-labeled general, hand-labeled general, and hand-labeled domain-specific) using the corresponding preprocessed positive and negative tweets, and tested their performance on the separate test set described above. In Table 1, we report the results in terms of macro-averaged error rate [27] and in terms of macro-averaged F-score of positive and negative classes [28]. We are particularly interested in the correct classification of the positive and negative tweets.

As can be seen from Table 1, the best performing sentiment model is the hand-labeled domain-specific one as it achieved the lowest error rate and the highest macro-averaged F-score on the test set. Note that this model is trained on only 6,735 tweets, while the other two models employed substantially more tweets (1.6 million for the smiley-labeled general model and 48,971 for the hand-labeled general model). Therefore, the results

Table 1 The evaluation results of smiley-labeled general, hand-labeled general, and hand-labeled domain-specific sentiment models on the test dataset in terms of the macro-averaged error rate and the macro-averaged F-score of positive and negative classes

Sentiment model	M error rate (%)	F_{avg}
Smiley-labeled general	61.3	0.20
Hand-labeled general	59.3	0.25
Hand-labeled domain-specific	52.9	0.39

indicate that the high-quality domain-specific tweets produce better sentiment models even if the number of such tweets is lower. For the rest of our study, we use the hand-labeled domain-specific sentiment model trained using the complete hand-labeled domain-specific dataset.

The sentiment of different communities regarding a specific topic is calculated as follows. First, for each community, the tweets posted by its users are selected. Second, the sentiment of each tweet is determined and weighted by its retweet count. Third, the weighted negative and positive sentiment of tweets is aggregated for each user and summed over all users in the community. Finally, the leaning of a community towards a specific topic is computed as the *polarity* of the aggregated weighted sentiment multiplied by the ratio of sentiment carrying tweets (*subjectivity*) of the respective community. The polarity and subjectivity measures are adapted from [29]. The pseudo-code for community sentiment computation is presented in Algorithm 1.

Algorithm 1 Computing community sentiment

```
Require: $\mathcal{C}$ : community,
         $T_S$ : sentiment annotated tweets,
         $\bar{D}_P$ : avg. distance of positive training examples,
         $\bar{D}_N$ : avg. distance of negative training examples

  function CommunitySentiment ($\mathcal{C}$, $T_S$):
      $pos = 0$
      $neg = 0$
      $all = 0$
      for $user$ in $\mathcal{C}.users$ do
          $userTweets = T_S$.byUser($user$)
          for $tw$ in $userTweets$ do
              if $tw.sentiment > \bar{D}_P$ then
                  $pos$ += $tw.retweetCount$
              else if $tw.sentiment < \bar{D}_N$ then
                  $neg$ += $tw.retweetCount$
              end if
              all += $tw.retweetCount$
          end for
      end for
      $polarity = \frac{pos-neg}{pos+neg}$
      $subjectivity = \frac{pos+neg}{all}$
      return $polarity \times subjectivity$
  end function
```

Results and discussion

We present the results of the proposed methodology for identifying interest groups and their leaning towards different environmental topics, in terms of network and community structure, content categorization and identification, and sentiment analysis.

Network and community structure

We analyze a retweet network of 3.7 million users linked by 9.7 million retweet relations. In Fig. 1 we present the distribution of out-degree and influence I (as defined by Equation 1) for the nodes of the network. Community detection results in over 125,000 communities. Their size distribution is presented in Fig. 2. Notice that both plots are in *log–log* scale and therefore even only by eye inspection we can say that the distribution displays a fat tail in the sense that it deviates strongly to the right from a Gaussian distribution. This means that, in line with the empirical literature on social networks, nodes with very high degree and communities with a very large size occur with frequency much larger than in a Gaussian scenario.

We focus our analysis on communities of considerable size, which also produced a sufficient amount of tweets for meaningful content identification and sentiment analysis. This results in 12 communities, each with more than 50,000 users, and with at least 10,000 unique tweets.

The analysis in terms of community influence and its distribution among their users reveals significant structural differences among the largest communities. Results are presented in Table 2. The ratio between the inter- and intra-community influence, $I_{out}(C)$ and $I_{in}(C)$, shows that the majority of communities are greatly introverted, as their influence outside their "borders" presents less than a quarter of the impact they have. However, there are two communities ($k = 1$ and 4) that have almost a third of their influence outside the community, and one where its external influence is almost as high as its internal influence ($k = 5$).

The distribution of influence within communities, as measured by the Herfindahl-Hirshmann index (HHI), also shows interesting differences among communities. The lowest values of HHI are around 0.03, for communities $k = 6, 9, 10$, and 11. Hence, these are the communities that have the lowest inequality in terms of I_{in} among their 50 most influential users. Whereas communities $k = 8$ and 12 have the highest inequality between their 50 most influential users. It is interesting to notice that community $k = 6$ with the lowest inequality is also the second most introverted. Other than that, we find no obvious relation between HHI and the relative inter-community influence.

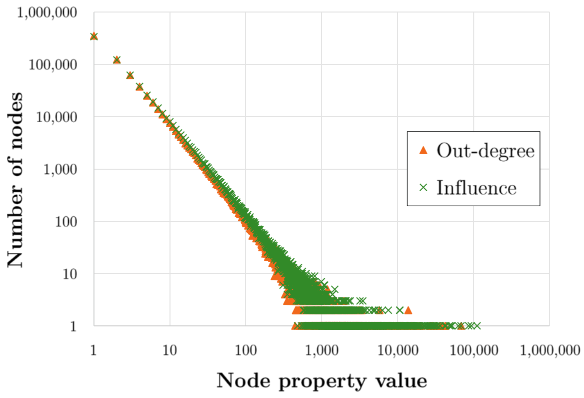

Fig. 1 Out-degree (orange) and influence (green) distribution in the retweet network

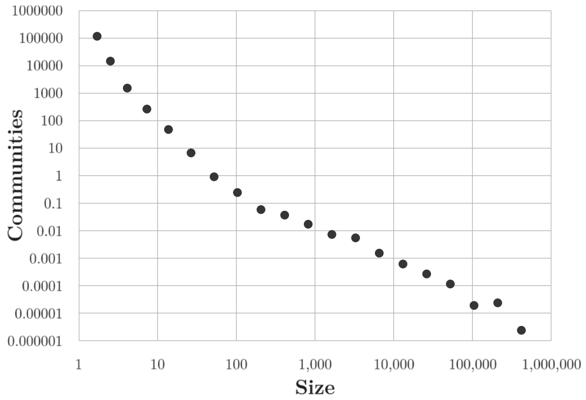

Fig. 2 Distribution of community sizes using logarithmic binning, as defined in [30]

In Fig. 3, we present the relation between the user influence, out-degree, and the number of unique tweets, for the top three most influential users of selected nine communities. The selection is explained in the subsequent section. The figure shows the magnitude of the top users in different communities and is consistent with the inequality measures by HHI. On the other hand, there is no obvious relation between the tweet volume and the influence of the users. It seems that higher out-degree is accompanied by higher influence, which can be seen also from Fig. 1.

Community content

A preliminary community categorization was performed by looking at the Twitter profiles of their most influential users and the contents of their tweets. We find that the communities could roughly be classified into six categories. Table 3 presents the community categories and examples of the most influential users in these categories.

The community categorization reveals that for our further investigations we can ignore certain categories of communities. First, in the “Humor” community, the presence of an actual leaning or sentiment towards a certain topic is for one questionable (every topic

Table 2 Structural properties of the 12 largest communities

k	Name	Users	Unique tweets	$I_{in}(C_k)$	$I_{out}(C_k)$	$\frac{I_{out}(C_k)}{I_{in}(C_k)}$	$HHI(C_k)$
1	Env 1	366,979	625,280	1,546,998	787,139	0.509	0.037
2	Env 2	324,518	561,659	2,189,373	796,861	0.364	0.034
3	News 1	275,172	325,867	1,160,347	385,355	0.332	0.035
4	Humor	272,780	12,971	330,897	150,148	0.454	0.065
5	News 2	254,159	44,587	363,539	307,039	0.845	0.036
6	Skeptic	160,257	236,618	983,672	132,509	0.135	0.029
7	India	96,158	32,981	311,754	37,849	0.121	0.045
8	Celebrity	92,434	13,480	174,105	36,414	0.209	0.158
9	News 3	91,446	95,415	274,704	91,323	0.332	0.032
10	Env 3	83,259	180,210	707,292	187,576	0.265	0.030
11	Other	65,363	13,697	115,709	41,309	0.357	0.031
12	Env 4	53,847	29,863	105,608	19,796	0.187	0.104

Community influence $I(C)$ is split into $I_{in}(C)$ and $I_{out}(C)$, intra- and inter-community influence, respectively. $HHI(C)$ is the Herfindahl-Hirshmann index of the intra-community influence

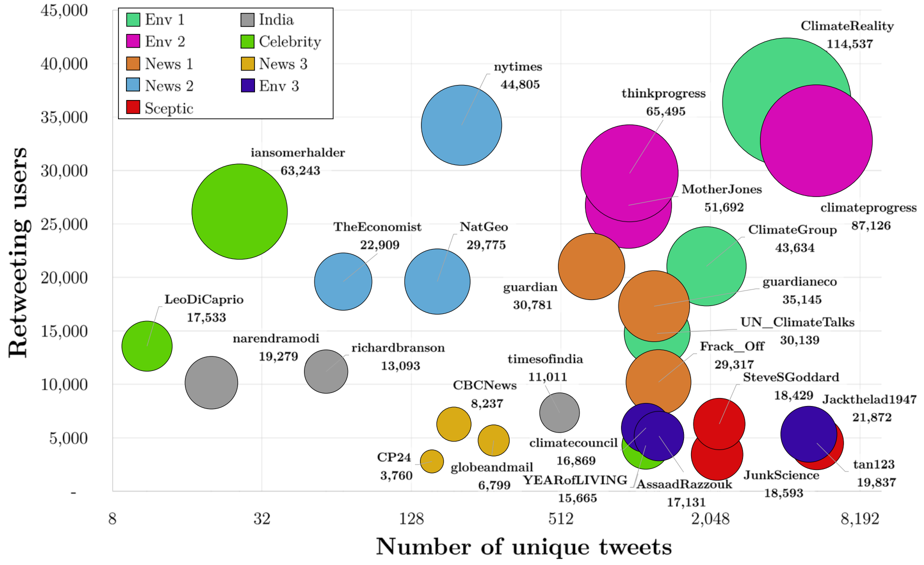

Fig. 3 Influence (bubble size), out-degree (number of retweeting users), and the number of unique tweets for the top three most influential users of the nine selected communities

can be made fun of using positive or negative words), and for two, it is hard to automatically identify the correct polarity due to frequent use of irony and sarcasm. Second, we also ignore a smaller community in the category "Other" that we are unable to strictly categorize.

One community from the "Environmental" category is also not included, because it contains numerous content duplicates as a result of marketing and spamming. The final selection includes three communities from the "Environmental" category (labeled as "Env 1", "Env 2", and "Env 3"), three from "News" ("News 1", "News 2", and "News 3"), the "Indian" community ("India"), one "Celebrities" community ("Celebrity"), and the "Skeptics" community ("Skeptic"). The network of these nine communities is outlined in Fig. 4.

Table 3 Community categories and their most influential users

Category	Count	Includes	Influential users
Environmental	4	Activists, organizations, green/eco news, and technology	ClimateReality, ClimateGroup, climateprogress, thinkprogress, Jackthelad1947, GreenrEnergy
News	3	News agencies, media	guardianeco, guardian, nytimes, NatGeo, TheEconomist, BBCWorld, CBCNews
Humor	1	Joke websites, commentators, comedians	emmkaff, StephenAtHome, TheTweetOfGod, neiltyson, michaelarria, pourmecoffee
Skeptics	1	Republicans, lobbyists	JunkScience, tan123, SteveSGoddard, hockeyschtick1, realDonaldTrump
Indian	1	Politics, news and business from India	narendramodi, richardbranson, timesofindia, MIB_India, EconomicTimes
Celebrities	1	Actors, musicians, athletes	iansomerhalder, LeoDiCaprio, YEARSofLIVING, JaredLeto
Other	1	Miscellaneous	-

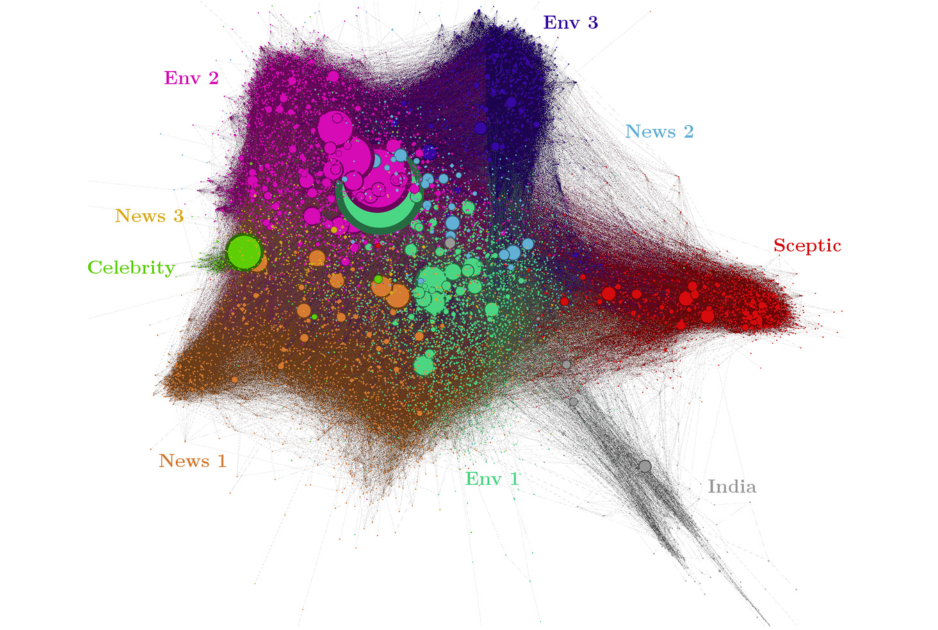

Fig. 4 Subgraph of the retweet network induced on the nine selected communities. Only users with influence larger than 100 retweets are displayed. The size of the nodes is proportional to the user influence and individual communities are distinguished by color

Each community is represented with its own color and the size of the nodes is proportional to the user's influence. The presented network layout shows a relatively clear segregation between the communities.

We analyze the content tweeted by a community in terms of (i) hashtags and (ii) plain text. Hashtags can represent entities in the tweet and/or user-inserted labels of a tweet, indicating the topic or broader context of the tweet. Content analysis in terms of hashtags, using the approach presented in section "Content identification and filtering", is therefore expected to show the characteristic entities and topics of interest in a selected community. On the other hand, plain text analysis is more appropriate for identification of actions, attitude, and phrases that are most distinctive for a particular community. The results of content analysis are presented in Table 4.

The most characteristic content of each community, as shown by the results in Table 4, reasonably distinguishes the communities of different categories. The hashtag content analysis supports the membership of the communities with the most influential users "ClimateReality" and "climateprogress" in the "Environmental" category, therefore from now on labeled by "Env 1" and "Env 2", respectively. Next two largest communities include topics present in the news in the United Kingdom and the United States of America, hence called "News 1" and "News 2", respectively. It reveals that the users retweeting "JunkScience" belong to the "Skeptic" community. Local topics from "India" are apparent from the hashtags of the next community. Similarly, the hashtags of the Ian Somerhalder Foundation (#isf) and their opinions point to the "Celebrity" community. Hashtag analysis of the last two communities shows interest in Canadian political and environmental issues, hence "News 3", and in environmental problems and political topics in Australia, therefore "Env 3".

Table 4 Characteristic content of the nine influential communities, selected on the basis of the largest number of unique tweets (in parenthesis are the most influential users)

Community	Users	Tweets	Content
Env 1 (ClimateReality)	366,979	625,280	#gridpowerstorage (0.49) #caribbeantech (0.20) #solars (0.19) #ag4dev (0.14) #jamaica (0.12) #idb (0.11) #energyefficiency (0.11) retw (0.18) global wind jobs: (0.14) global solar jobs: (0.11) green jobs: (0.09) lexinerus: retw (0.09) daily stories via (0.08) filed under: solar (0.07)
Env 2 (climateprogess)	324,518	561,659	#uniteblue (0.37) #p2 (0.28) #copolitics (0.26) #wiunion (0.16) #ofaction (0.14) #stoprush (0.14) #ctl (0.13) #libcrib (0.12) #coleg (0.11) without remorse please (0.12) dying plastic next. (0.11) next. watch share (0.11) companies poison water (0.09) stop now watch (0.09)
News 1 (guardianeco)	275,172	325,867	#olsx (0.34) #rhi (0.30) #bizitalk (0.26) #bartonmoss (0.24) #stopbrep (0.22) #udobiz (0.20) #besw14 (0.18) #gbhour (0.16) #ukair (0.15) 3 low (0.30) low 3 (0.21) average 2 low (0.16) pollution forecast tomorrow (0.14) moderate average (0.08) 40 % power 20 % (0.06)
News 2 (nytimes)	254,159	44,587	#la_chefs (0.44) #bos (0.43) #scistuchat (0.24) #lax (0.22) #ntrs (0.20) #washington (0.17) #sfo (0.15) #stockaction (0.15) #koreans (0.15) power personal branding (0.14) branding b2b lead (0.14) green chemistry pls (0.14) strange preferential treatment (0.11) japan privilege foreigners (0.11)
Skeptic (JunkScience)	160,257	236,618	#pjnet (0.84) #ccot (0.26) #tcot (0.19) #climatescam (0.15) #teaparty (0.15) #sgp (0.11) #tlot (0.11) #lnyhbt (0.10) #copolitics (0.09) man-made global-warming (0.14) conducts dangerous human (0.13) la dr. mengele (0.12) human experiments:a la (0.12) ibd obama's conducts (0.12)
India (narendramodi)	96,158	32,981	#invisiblekiller (0.37) #namo (0.26) #telangana (0.26) #mufflerman (0.23) #insubcontinent (0.20) #aap (0.18) #upa (0.15) web-app share ur (0.16) resources hands aam (0.16) plz join reduce (0.14) air pollution. web-app (0.14) shri (0.12) ganga (0.10) kejriwal (0.09)
Celebrity (iansomerhalder)	92,434	13,480	#coalsucks (0.66) #isf (0.61) #beyondcoal (0.34) #nofrackla (0.19) #isfcommcrew (0.11) #yearsproject (0.08) #yearssolutions (0.03) warm idea solar (0.23) help recycle (0.20) solar powered energy (0.13) coal get heated (0.12) fan wind power (0.10) coalsucks (0.10)
News 3 (CBCNews)	91,446	95,415	#cdnpoli (0.81) #nbpoli (0.23) #bcpoli (0.23) #hamont (0.13) #onpoli (0.12) #nspoli (0.12) #yeg (0.11) #yql (0.09) #nofrackns (0.08) big top thought (0.16) maritime electric (0.16) alberta (0.13) share resources stories (0.08) energy efficiency job: (0.08) ceea (0.07) hydro one (0.07)
Env 3 (Jackthelad1947)	83,259	180,210	#nswpol (0.48) #csg (0.47) #auspol (0.47) #springst (0.22) #qldpol (0.20) #ret (0.16) #qanda (0.15) #insiders (0.14) #vicvotes (0.12) business news (0.37) local banks (0.37) energy via full (0.19) can finance renewable (0.14) lnp (0.13) ret (0.05) agl (0.05) full story business (0.04)

Community contents is characterized in terms of hashtags and plain text (with the respective *TFiDF* values in parenthesis)

On the other hand, the results of the plain text analysis mostly show more specific topics that are shared in the observed communities. The top terms or phrases (*n*-grams) in the "Env 1", "Env 2", "News 1", "News 2" and "Celebrity" communities, reflect their interest in the promotion of alternative, renewable, and environmentally friendly energy sources, in contrast to the controversial energy supply solution provided by fracking, as well as raise awareness of global pollution. The two most distinctive topics that surface from the content of the "Skeptic" community are "man-made global-warming" and "conducts dangerous human experiments". The former is related to the community's skepticism regarding human-caused global warming, and the latter is about an article published by the "Investor's Business Daily" newspaper [31] that criticizes an allegedly harmful experiment by the Environmental Protection Agency (EPA). The plain text content results for the communities "India", "News 3", and "Env 3" show less specific topics, with the main focus on the local political situation, or environmental and energy policies.

Community sentiment

Finally, we investigate the sentiment leaning of the most content-rich communities. In our dataset of over 30 million environmental tweets, there are almost 3.2 million unique tweets. We label them by the SVM sentiment model, described in the "Sentiment analysis" section, as *positive* (1), *neutral* (0), or *negative* (-1). Only 31 % of the unique tweets are labeled as subjective, i.e., non-neutral. Furthermore, among the sentiment-carrying tweets, there are 52 % of tweets with positive sentiment and 48 % with negative sentiment.

We analyze the sentiment leanings towards selected topics related to the environmental issues. The selection is based on three major groups of topics that are of interest to environmental policy makers: energy sources and energy generation, environmental side effects, and actions or initiatives for solving the environmental issues. We separate the first group into four topics: renewable or green energy sources, nuclear energy, fossil fuels, and fracking, as a separate controversial topic. The second group is represented by the broader topic of global warming and climate change, more general pollution and contamination, and its more specific variant about emissions of greenhouse gases (CO_2 and methane). The last group is separated into recycling and waste management, and environmental policies and initiatives.

The nine communities selected for investigation produce over two thirds of the unique tweets in our dataset. We use the approach presented in the "Content identification and filtering" section to filter these 2.1 million tweets by the nine topics defined above. Table 5 presents the queries used in the filtering process to describe a particular topic. The number of tweets filtered by topic for each community is shown in Fig. 5.

The sentiment of a community towards a selected topic is computed from the tweets on that topic, tweeted by that particular community, as proposed in the "Sentiment analysis" section, Algorithm 1. The results of the community sentiment analysis on different environmental topics are presented in Fig. 6. Community leaning towards a specific topic is computed as the difference between the community sentiment on this topic and the community's average sentiment in our dataset. In Figs. 5 and 6, the topics of interest are in descending order from left to right by their average sentiment over all the communities.

Table 5 Selected environmental topics and the associated queries for tweet filtering

Topic	Query
Green energy	green renewable sustainable sustainability solar wind photovoltaic biomass biofuel biofuels #green #cleanenergy #renewable #renewableenergy #sustainable #sustainability #solar #wind #solarpower #windpower #photovoltaic #biomass #biofuel #biofuels
Recycling	recycling reuse re-use "waste management" waste-management "carbon capture" carbon-capture "carbon storage" "co2 capture" "co2 storage" sequestration decarbonization decarbonisation #reuse #recycling #wastemanagement #CCS #carboncapture
Emissions	emission emissions carbon co2 "carbon dioxide" carbon-dioxide greenhouse greenhouse-gas ghg ch4 methane #emission #emissions #carbon #co2 #carbondioxide #greenhouse #greenhousegas #greenhousegases #ghg #ch4 #methane
Nuclear	nuclear #nuclear #nuclearenergy #nuclearpower #nuclearmatters
Policies	ipcc cop19 cop20 cop21 kyoto 2030 #ipcc #cop19 #cop20 #cop21 #kyoto #2030 #2030now
Fossil fuels	oil gas coal fossil #oil #gas #fossilfuel #coal #oilgas natgas #natgas
Climate change	"climate change" climate-change climate warming "global warming" global-warming #climatechange #climate-change #climate_change #globalwarming #global-warming #global_warming
Pollution	pollution contamination pollute contaminate spill #pollution #polluted #contamination #contaminated #spill #spills #oilspill #oilspills
Fracking	fracking frack shale shalegas aquifer #fracking #frack #shale #shalegas #aquifer #aquifers

The first interesting finding is that the sentiment analysis is in accordance to the commonly accepted attitude towards different environmental topics. All communities show positive leaning towards "green energy" and "recycling", and negative towards "fossil fuels", "climate change", "pollution", and "fracking", except for two outlier communities that we examine separately. Regarding "emissions", "nuclear energy", and "policies", the sentiment leanings are less unanimous, which is to some extent also expected. These results indicate that the domain-specific sentiment model produces reasonable results.

Observing individual communities, we find that most of them follow the same trend; however, there are two notable exceptions: the "Skeptic" and the "Celebrity" communities. The "Skeptic" community is very segregated from the rest (see Fig. 4), and its sentiment leanings show greatest deviations from the leaning of other communities (see Fig. 6). It is the only community having a positive sentiment leaning about the topics "fossil fuels" and "fracking", which is considerably different from all other communities. These results

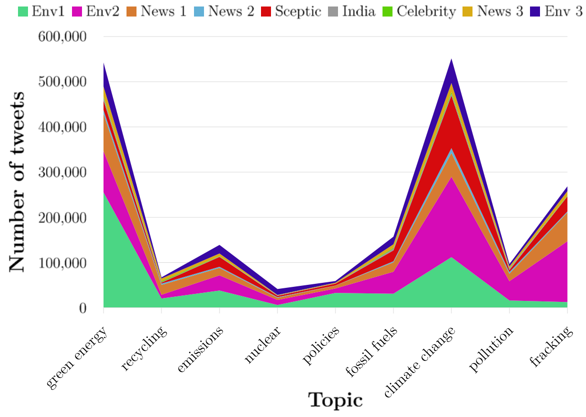

Fig. 5 The number of unique tweets for each selected topic published by each of the major communities

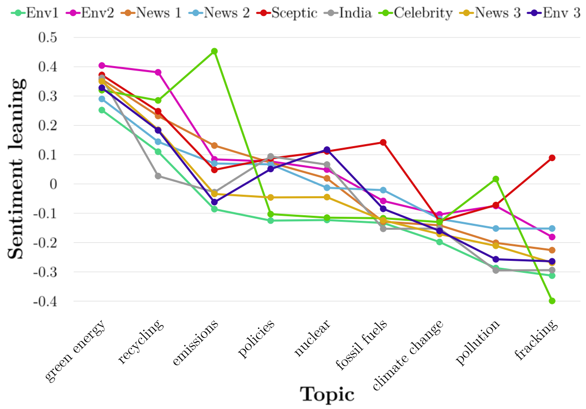

Fig. 6 Sentiment leaning of the nine communities towards different environmental topics

clearly indicate that the preferences of this community are diverging from the interests of the other communities.

The "Celebrity" community is dominated by "iansomerhalder", one of the most influential users overall (see Fig. 3). Despite the high influence, the community produces very low number of original tweets (less than 1 % of all the unique tweets, see Table 4). Its influence emerges from the large number of retweets, due to the large number of followers of "iansomerhalder". This hints at the possibility to engage high-profile celebrities, with the commitment to environmental issues, in promotion and spreading of influential contents.

This is exactly what can be observed for the topics "emissions" and "pollution". The extremely positive sentiment leaning towards these topics is predominantly (60 and 78 %, respectively) due to only three tweets by the two most influential users of the "Celebrity" community: "iansomerhalder" and "LeoDiCaprio". They are expressing their happiness and thankfulness regarding the "action to limit carbon pollution" and "cutting carbon pollution", which will "clean up our air and tackle climate disruption", as they put it. Hence, the distinctively positive leaning for the topics "emissions" and "pollution". On the other hand, the "Celebrity" community seems to be least in favor of "fracking".

Conclusions

The paper contributes to the research on complex networks in social media by combining a structural and content-based analysis of Twitter data. From structural properties of the retweet network, we identify influential users and communities. From the contents of their tweets, we characterize discussion topics and their sentiment. Sentiment of different communities shows perceivable differences in their leanings towards different topics. We have identified two communities that considerably diverge from the rest, "Skeptic" with the most different sentiment leanings on several topics, and "Celebrity" with a low number of original tweets, but highly influential, with the potential to spread interesting information.

Our previous research in sentiment analysis of Twitter data in politics and stock market suggests that different vocabularies are used in different domains and that high-quality expert labeling of domain-specific tweets yields better sentiment models. The comparison of the three sentiment models (smiley-based general, hand-labeled general, and hand-labeled domain specific) presented in this paper confirms our intuition: hand-labeled domain-specific model yields lower error rate and higher combination of precision and recall (F-score) than the other two models. However, more extensive evaluations are required to determine the amount of hand-labeled tweets needed to approach the "maximum" performance, e.g., the inter-annotator agreement.

Another line of future research is the construction of more sophisticated SVM classifiers. In the case of smiley-based training data, only positive and negative tweets were available, and a binary SVM classifier was extended with a neutral zone to allow for the three-class classification. However, in the case of hand-labeled tweets, there are three sets of training data available: positive, neutral, and negative tweets, so we are dealing with a multiclass problem. Further, we can assume that the classes are ordered (neutral is between the positive and negative), and therefore, we are faced with the problem of ordinal regression [32], instead of binary classification. In the future, we plan to exploit various extensions of an SVM to deal with the multiclass [33] and ordinal regression problems.

In this paper, we present a general methodology of combining a structural and content-based analysis of Twitter networks, and then apply it to 1 year of Twitter data about environmental topics. There are several plans for future work. On the one hand, we plan to study the temporal aspects of community formation and sentiment spreading. In addition to the retweet networks, we will also construct mention networks (which model mutual engagement of users in conversations). We will investigate various spreading models and study the differences in sentiment spreading at such multilayer (retweet and mention) networks.

We are also collecting Twitter data in several other interesting domains: stock market, EU commission and parliament members, and lobbying organizations. The application of the presented structural and content-based analysis to these new domains will result in complex 'Twitter' networks. On the other hand, networks between the same entities can also be constructed by other means, such as correlations between stock returns, national and party membership of politicians, vote similarity, and ownership between the companies. The research challenge for the future is the comparison between the Twitter induced and other types of networks, and the mutual interplay and property spreading between these multilayer networks.

Endnotes

[1] Cosine similarity is a measure of similarity between vectors **a** and **b**. It is calculated as the normalized dot product between vectors **a** and **b**: $\mathrm{sim}(\mathbf{a},\mathbf{b}) = \cos(\angle(\mathbf{a},\mathbf{b})) = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}$

[2] The dataset was obtained from "For Academics" section, at http://help.sentiment140.com/for-students.

[3] The approach to feature vector construction was implemented using the LATINO (Link Analysis and Text Mining Toolbox) software library, available at http://source.ijs.si/mgrcar/latino.

Appendix

Our dataset of over 30 million tweets on environmental topics was acquired using the Twitter Search API [34]. Table 6 shows the list of search queries used.

Table 6 Queries for the "Environmental dataset" acquisition from the Twitter Search API

("2030 framework") OR ("2c objective")	("energy transition")	((#ccs (climate OR eu))
("abatement cost")	("energy utilities")	(carbon capture storage))
("adaptation fund")	("environment friendly" OR "environmentally friendly")	(((cer OR cers) (kyoto OR co2 OR emission OR emissions)) OR "certified emission reduction")
("affordable energy")	("environmental footprint")	((cdm (climate OR co2 OR carbon))
("algal energy")	("environmental protection" OR "environment protection")	("carbon development mechanism"))
("alternative energy" OR "alternative fuel")	("environmental regulation")	((emission reduction (unit OR units))
("arctic meltdown")	("environmental savings")	((Kyoto OR CO2 OR warming) (ERU OR ERUs)))
("building stocks")	("ets reform")	((eu OR european OR unified) "power market" OR "electricity market")
("car sharing" OR "car share" OR carsharing OR carshare)	("expenditure of energy" OR "energy expenditure")	((ipcc (eu OR climate OR energy))
("carbon bubble")	("feed in tariff" OR "feed in tariffs")	(intergovernmental panel "climate change"))
("carbon cap" (emission OR trade OR climate))	("fossil fuel" OR "fossil fuels")	((ipcc OR climate) assessment (impact OR report))
("carbon credits")	("fuel cost" OR "fuel costs")	((quota OR quotas) (renewable OR renewables))
("carbon dioxide" (emission OR emissions))	("fuel efficient" (car OR cars OR vehicle OR vehicles))	((reduce OR reducing) ("greenhouse gas" OR GHG) (emission OR emissions))
("carbon footprint")	("geo engineering" or "geoengineering")	((smart OR smarter) energy infrastructure)
("carbon leakage")	("geothermal energy" OR "thermal energy")	((vehicle ("zero emissions" OR "zero emission"))
("carbon lock in")	("global warming" OR globalwarming)	(zev (car OR vehicle)))
("carbon price")	("green cars")	((vsc (carbon OR climate))
("carbon tax" OR "carbon taxes" OR "carbon taxation")	("green chemistry")	(verified carbon (standard OR standards)))
("clean energy")	("green economy")	(95g fleet)
("clean growth")	("green energy")	(actonclimate)
("clean tech")	("green growth")	(alternative energy sources)
("climate action" OR "action on climate" OR climateaction)	("green job" OR "green jobs" OR "greener jobs")	(anthropogenic "climate change")
("climate adaptation")	("green transportation" OR "green transport")	(biofuel OR biofuels)
("climate change")	("grid control" OR (control "power grid"))	(biomass)
("climate deal")	("heat insulation")	(carbon (credit OR credits) (trading OR auctioning))
("climate denier" OR "climate deniers")	("hydro power" OR hydropower)	(carbon energy intensity)
("climate finance")	("hydroelectric energy")	(chemtrail OR chemtrails)
("climate goal" OR "climate goals")	("icecap meltdown")	(cleantech (investment OR investments))
("climate mitigation")	("industry exemptions")	(climate energy (target OR targets))
("climate policy")	("intelligent networks")	(climate resilient economy)
("climate report" OR "report on climate")	("joint implementation")	(climate2015)
("climate sensitivity")	("kyoto protocol")	(climatechange)
("climate system")	("life cycle approach")	(co2)
("co2 neutral")	("light duty" vehicles)	(cop19 OR "cop 19" OR (cop warsaw))
("coal industry" OR "oil industry" OR "nuclear industry" OR "gas industry")	("low carbon")	(cop20 OR "cop 20" OR (cop peru))
("cohesion policy")	("low carbon" (tech OR technology OR technologies))	(cop21 OR "cop 21" OR (cop paris))
("district heating")	("low carbon" economy)	(cross border infrastructure)
("e bike" OR ebike)	("merit order")	(decarbonisation)
("e mobility" OR emobility)	("micro cogeneration")	(deforestation)
("eco design")	("natural resources")	(demand side management)
("eco entrepreneurship")	("non ets")	(desertec)
("eco technologies")	("oil spill")	(desertification)

Table 6 Queries for the “Environmental dataset” acquisition from the Twitter Search API

("effort sharing")	("permanent set aside")	(eco best invest)
("electric motors" OR "electric motor")	("polar meltdown")	(emission allocation)
("electricity costs" or "electricity costs")	("power blackout" OR "energy blackout" OR "electricity blackout")	(emission cap)
("electricity mix")	("power plant" OR "coal plant" OR "gas plant")	(energy (price OR prices) (peak OR peaks))
("electricity storage" OR "energy storage")	("renewable energy" OR renewables)	(energy climate policy framework)
("emission reduction" OR "emission reductions")	("resource efficiency")	(energy efficiency (improvement OR improvements))
("emission trading" OR "emission trade")	("sea level rise")	(energy efficiency policy)
("energy affordability")	("shale gas" OR "unconventional gas" OR "unconventional hydrocarbons")	(energy import dependency)
("energy company" OR "energy companies")	("smart grid" (energy OR electricity OR supply OR power))	(energy supply security)
("energy consumption")	("smarter city" OR "smart city" OR "smarter cities" OR "smart cities")	(energyaware OR "energy aware")
("energy cost" OR "cost of energy")	("solar panel" OR "solar panels")	(environmentalist OR environmentalists)
("energy crisis" OR "crisis of energy")	("solar power" OR "solar energy")	(eu energy legislation)
("energy demand" OR "demand for energy")	("stranded assets" OR strandedassets)	(forestfinance)
("energy efficiency")	("sustainable finance" OR "sustainable investment")	(fossilfuel OR fossilfuels)
("energy efficient" (building OR buildings OR car OR cars OR home OR homes OR vehicle OR vehicles))	("sustainable manufacturing")	(fracking OR fracked)
("energy efficient" (tech OR technology OR technologies))	("tar sand" OR "oil sand")	(fukushima)
("energy firm" OR "energy firms")	("temp rise")	(global carbon (trading OR market))
("energy future" OR "future of energy")	("transport sector")	(green climate (fund OR funds))
("energy generation" OR "electricity generation")	("unburnable carbon" OR "unburnable coal")	(greentech OR "green tech" OR "green technology" OR "green technologies")
("energy independent" OR "energy independence")	("warming mitigation")	(greenvc)
("energy intensity")	("waste management")	(model shift (climate OR CO2 OR environment OR carbon OR warming OR energy))
("energy intensive" (industry OR sector OR business))	("wind farm")	(pollution)
("energy market")	("wind power" OR "wind energy")	(primary energy consumption)
("energy mix")	("wind turbine" OR "wind turbines")	(recycling)
("energy performance")	("zero emissions" OR "zero emission")	(renewableenergy)
("energy policy")	(("combined heat power")	(second generation (biofuel OR biofuels))
("energy price" OR "energy prices")	(chp (climate OR energy OR electricity)))	(single energy market)
("energy production")	(("emissions trading system") OR "eu ets")	(stopfracking)
("energy productivity")	(("energy efficiency directive")	(sustainability)
("energy savings" OR "energy saving" OR "conserving energy" OR "energy conservation")	(energy eed))	(sustainable2050)
("energy sector")	(("greenhouse gas" OR ghg) (emission OR emissions))	(wholesale (energy OR electricity) (cost OR prices OR price))
("energy security")	(("zero emissions" OR "zero emission" OR "low energy") house)	
("nuclear power" OR "solar power" OR "geothermal power" OR "thermal power" OR "electrical power" OR "electric power")		
((industry OR sector OR bussines) specific (targets OR target) (energy OR climate OR EU OR emission OR emissions))		
((offshore OR onshore) (climate OR CO2 OR environment OR carbon OR warming OR energy OR oil OR gas OR fracking OR wind))		
(power (coal OR gas OR oil OR biomass OR diesel OR biogas OR photovoltaic OR thermoelectric OR hydrogen OR fuel OR climate OR emission OR emissions OR CO2 OR carbon OR electricity OR fusion OR fission OR generation OR turbine))		

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BS and IM conceived and designed the experiments. BS and JS performed the experiments. BS, JS, IM, and SB analyzed the data and results and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

This work was supported in part by the European Commission under the FP7 projects SIMPOL (Financial Systems SIMulation and POLicy Modelling, grant no. 610704) and MULTIPLEX (Foundational Research on MULTIlevel comPLEX networks and systems, grant no. 317532), by the H2020 project DOLFINS (Distributed Global Financial Systems for Society, grant no. 640772), and by the Slovenian Research Agency programme Knowledge Technologies (grant no. P2-103). We thank Matjaž Juršič for his help on the construction of retweet networks, Petra Kralj Novak, Miha Grčar and Martin Žnidaršič for their help on sentiment models, their evaluation, and tweet preprocessing.

Author details

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. [2]Department of Banking and Finance, University of Zurich, Zurich, Switzerland.

Received: 18 February 2015 Accepted: 22 May 2015
Published online: 14 July 2015

References

1. Dorsey, J, Williams, E, Stone, B, Glass, N, Twitter online social networking service. http://www.twitter.com/. Accessed: Feb 15, 2015
2. Cha, M, Haddadi, H, Benevenuto, F, Gummadi, PK: Measuring user influence in twitter: the million follower fallacy. ICWSM. **10**, 10–17 (2010)
3. Avnit, A: The million followers fallacy. Pravda Media Group, Tel Aviv, Israel (2009)
4. Suh, B, Hong, L, Pirolli, P, Chi, EH: Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE Second Intl. Conf. on Social Computing, pp. 177–184. IEEE, Piscataway, New Jersey, (2010)
5. Bakshy, E, Hofman, JM, Mason, WA, Watts, DJ: Everyone's an influencer: quantifying influence on twitter. In: Proc. Fourth ACM Intl. Conf. on Web Search and Data Mining, pp. 65–74. ACM, New York City, New York, (2011)
6. Conover, M, Ratkiewicz, J, Francisco, M, Gonçalves, B, Menczer, F, Flammini, A: Political polarization on twitter. In: Proc. Fifth Intl. Conf. on Weblogs and Social Media (ICWSM). AAAI, Palo Alto, California, (2011)
7. Conover, MD, Gonçalves, B, Ratkiewicz, J, Flammini, A, Menczer, F: Predicting the political alignment of twitter users. In: Privacy, Security, Risk and Trust, 2011 IEEE Third Intl. Conf. on Social Computing, pp. 192–199. IEEE, Piscataway, New Jersey, (2011)
8. Agarwal, A, Xie, B, Vovsha, I, Rambow, O, Passonneau, R: Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media, pp. 30–38. Association for Computational Linguistics, Stroudsburg, PA, USA, (2011)
9. Bollen, J, Mao, H, Zeng, X: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
10. Gayo-Avello, D: A meta-analysis of state-of-the-art electoral prediction from twitter data. Soc. Sci. Comput. Rev. **31**(6), 649–679 (2013)
11. Smailović, J: Sentiment Analysis in Streams of Microblogging Posts. PhD thesis, Jožef Stefan International Postgraduate School, Ljubljana, Slovenia (2014)
12. Vapnik, VN: The Nature of Statistical Learning Theory. Springer, New York, NY, USA (1995)
13. Go, A, Bhayani, R, Huang, L. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1–12 (2009)
14. Sluban, B, Smailović, J, Juršič, M, Mozetič, I, Battiston, S: Community sentiment on environmental topics in social networks. In: Proceeding of the Tenth International Conference on Signal-Image Technology & Internet-Based Systems, pp. 376–382. IEEE Computer Society, Washington, DC, USA, (2014)
15. Fortunato, S: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
16. Blondel, VD, Guillaume, J-L, Lambiotte, R, Lefebvre, E: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), 10008 (2008)
17. Newman, MEJ: Modularity and community structure in networks. Proc. Natl. Acad. Sci. U. S. A. **103**(23), 8577–8582 (2006)
18. Lancichinetti, A, Fortunato, S: Community detection algorithms: a comparative analysis. Phys. Rev. E. **80**(5), 056117 (2009)
19. Werden, GJ: Using the Herfindahl–Hirschman index. In: Phlips, L (ed.) Applied Industrial Economics, pp. 368–374. Cambridge University Press, Cambridge, UK, (1998)
20. Feldman, R, Sanger, J: Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY, USA (2006)
21. Joachims, T: A support vector method for multivariate performance measures. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 377–384. ACM, New York City, New York, (2005)
22. Joachims, T: Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226. ACM, New York City, New York, (2006)
23. Joachims, T, Yu, C-NJ: Sparse kernel SVMs via cutting-plane training. Mach. Learn. **76**(2-3), 179–193 (2009)
24. Smailović, J, Grčar, M, Lavrač, N, Žnidaršič, M: Stream-based active learning for sentiment analysis in the financial domain. Inf. Sci. **285**, 181–203 (2014)

25. Smailović, J, Grčar, M, Lavrač, N, Žnidaršič, M: Predictive sentiment analysis of tweets: A stock market application. In: Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science, pp. 77–88. Springer, Berlin Heidelberg, (2013)
26. Feldman, R, Sanger, J: Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY, USA (2006)
27. Baccianella, S, Esuli, A, Sebastiani, F: Evaluation measures for ordinal regression. In: Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference On, pp. 283–287. IEEE, Piscataway, New Jersey, (2009)
28. Kiritchenko, S, Zhu, X, Mohammad, SM: Sentiment analysis of short informal texts. J. Artif. Intell. Res. **50**, 723–762 (2014)
29. Zhang, W, Skiena, S: Trading strategies to exploit blog and news sentiment. In: Proc. Fourth Intl. AAAI Conf. on Weblogs and Social Media (ICWSM), pp. 375–378. AAAI, Palo Alto, California, (2010)
30. Newman, MEJ: Power laws, Pareto distributions and Zipf's law. Contemp. Phys. **46**(5), 323–351 (2005)
31. Obama's EPA Conducts Dangerous Human Experiments. Investors.com. http://news.investors.com/ibd-editorials/040414-696061-epa-conducts-pollution-experiments-on-humans.htm. Accessed: Sep 5, 2014
32. Cardoso, JS, Da Costa, JFP: Learning to classify ordinal data: the data replication method. J. Mach. Learn. Res. **8**, 1393–1429 (2007)
33. Crammer, K, Singer, Y: On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. **2**, 265–292 (2002)
34. Twitter search API. Twitter, Inc. https://dev.twitter.com/rest/public/search. Accessed: Jan 1, 2014