JeDEM 4(2) 198-221, 2012
ISSN 2075-9517
http://www.jedem.org

Determining Citizens' Opinions About Stories in the News Media

Analysing Google, Facebook and Twitter

Timo Wandhöfer*, Steve Taylor**, Paul Walland***, Ruxandra Geana****, Robert Weichselbaum*****, Miriam Fernandez******, Sergej Sizov*******

**GESIS – Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany, timo.wandhoefer@gesis.org, +49 (0) 221 47694 544*

***IT Innovation Centre, University of Southampton, Gamma House, Enterprise Road, Southampton SO16 7NS, UK, sjt@it-innovation.soton.ac.uk, +44 (0) 23 8059 8866*

****IT Innovation Centre, University of Southampton, Gamma House, Enterprise Road, Southampton SO16 7NS, UK, pww@it-innovation.soton.ac.uk, +44 (0) 23 8059 8866*

*****University of Southampton, UK, geana.ruxandra@gmail.com*

******GESIS – Leibniz-Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany, robert.weichselbaum@gesis.org, +49 (0) 221 47694 0*

*******Knowledge Media Institute, Open University. Walton Hall, Milton Keynes, MK7 6AA, UK, m.fernandez@open.ac.uk, +44 (0) 1908 659400*

********WeST – Institute for Web Science and Technologies, University of Koblenz, University Street 1, 56070 Koblenz, Germany, sizov@uni-koblenz.de, +49 (0) 261 287 2736*

Abstract: *We describe a method whereby a governmental policy maker can discover citizens' reaction to news stories. This is particularly relevant in the political world, where governments' policy statements are reported by the news media and discussed by citizens. The work here addresses two main questions: whereabouts are citizens discussing a news story, and what are they saying? Our strategy to answer the first question is to find news articles pertaining to the policy statements, and then perform internet searches for references to the news articles' headlines and URLs. We have created a software tool that schedules repeating Google searches for the news articles and collects the results in a database, enabling the user to aggregate and analyse them to produce ranked tables of sites that reference the news articles. Using data mining techniques we can analyse data so that resultant ranking reflects an overall aggregate score, taking into account multiple datasets, and this shows the most relevant places on the internet where the story is discussed. To answer the second question, we introduce the WeGov toolbox as a tool for analysing citizens' comments and behaviour pertaining to news stories. We first use the tool for identifying social network discussions, using different strategies for Facebook and Twitter. We apply different analysis components to analyse the data to distil the essence of the social network users' comments, to determine influential users and identify important comments.*

Keywords: e-participation, e-government, social networks, news articles, news stories, political engagement, citizens' opinions, dialogue, Facebook, Twitter

Acknowledgement: The WeGov project (no. 248512) is funded with support from the European Commission under the SEVENTH FRAMEWORK PROGRAMME THEME ICT 2009.7.3 ICT for Governance and Policy Modelling.

The work reported here has been done within the context of the WeGov IST FP7 project. The project's primary remit is to enable effective dialogue and engagement between governments and citizens, and a key feature of the project is that it uses social networking sites (SNS) as the primary communication channel. Before the project, there were a number of efforts to engage citizens with governmental policy, mainly using bespoke websites whose main drawback was that they were rarely used (see Hansard Society, 2009 for an example). WeGov is aiming to address this drawback by using tools the citizens already use: social networking sites, blogs, forums, etc.

The project supports its target audience of governmental policy makers with tools to enrich the dialogue with citizens on SNS. The project's philosophy has been to develop a set of tools enabling the user to find and analyse SNS postings, and to make responses into SNS; along with a dashboard-based environment where the tools can be used individually or together.

In this paper, we provide two related case studies to illustrate how the WeGov project can assist a policy maker, and both these case studies fall within the same scenario. During initial meetings with external end users, a particular need of WeGov's target users, governmental policy makers, was requested. This is the gathering citizens' opinions as feedback to a particular statement by a politician. The first WeGov prototype covered this scenario as a basic use case. Here, the policy-maker posts a statement into a social network, collects the citizens' feedback (where it is publicly available) and runs the analysis components on the feedback. The result is a summary of the key themes and opinions over the sum total of the citizens' comments REFERENCE TO: Wandhöfer et al, 2010.

The initial toolbox was presented to 29 office employees working for a parliamentarian of the German Bundestag with the aim of gathering feedback for the further development process REFERENCE TO: WeGov, 2011. During discussions with them, the consensus was that parliamentarians' posts are unlikely to solicit a large amount of feedback, unless the politician is high-profile: "ordinary" parliamentarians' posts typically generate below 100 comments. They confirmed that the requirement to test citizens' reactions to politicians' statements is important, but they need more comments to provide a statistically significant sample of opinions.

A modification of the original use case was proposed by the Bundestag employees, where politicians' statements are covered on the internet through news articles, which are in turn disseminated and discussed by citizens. Figure 1 outlines "The Newspaper Story" which capitalises on the effect of "indirect injections" REFERENCE TO: Joshi et al, 2011 – this means the politician's statement is disseminated by citizens rather than the politician. For example, a news article is written around the statement, and this is discussed over many different locations by citizens.

In the example in Figure 1, www.bbc.co.uk published a news article with the headline "State multiculturalism has failed, says David Cameron"[1]. Internet news sites provide the opportunity for readers to share and discuss news articles over diverse internet locations, and thus the story may be propagated and discussed in many places on the internet by citizens. This news article was shared 31,309 times on Facebook and 1,922 times on Twitter.

[1] BBC news website. News article "State multiculturalism has failed, says David Cameron", 2011. URL: http://www.bbc.co.uk/news/uk-politics-12371994 (Retrieved 20 November 2011).

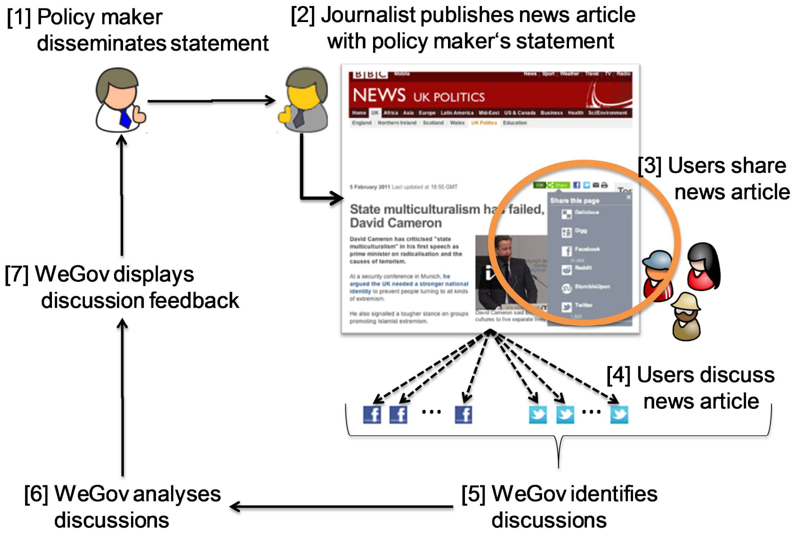

[1] Policy maker disseminates statement
[2] Journalist publishes news article with policy maker's statement
NEWS UK POLITICS
State multiculturalism has failed, David Cameron
[3] Users share news article
[4] Users discuss news article
[5] WeGov identifies discussions
[6] WeGov analyses discussions
[7] WeGov displays discussion feedback

Figure 1: The Newspaper Story

In addition to the question of finding where discussions are taking place, there is also the question of understanding what is being said. In many cases there will be a large number of postings made by citizens and the policy-makers need a way of summarising the discussions to determine the overriding sentiments and themes of the discussion. However, given the potentially large numbers of postings, manual analysis ranges from very difficult to impossible.

The scenario above gives rise to two major challenges related to finding and understanding the discussions:

- *Where are citizens discussing the news stories related to their policies?*
- *What is the essence of the citizens' discussions?*

Should the policy maker wish to engage with the citizens discussing the policies, an additional subsidiary question arises:

- *Who are the influential users, and which are the influential comments or posts?*

We thus had two major challenges to address in our work, and the results are reported in this paper. This paper has two major sections, each addressing one of the challenges above. Section 1 addresses the question of locating where discussions are taking place and Section 2 describes how we can take a discussion on the popular social networks and summarise the many comments and postings that make up a discussion of a news story. Section 2 also discusses the approach to addressing the subsidiary question of influential users and posts. We then wrap up with a brief conclusion.

1. Where Discussions are Found

Addressing the challenge of finding discussions related to a news story (step 5 in Figure 1) is the core of the work described in this section: to identify the websites where a news article is disseminated and discussed.

Once we had a good idea of the core problem to be addressed, and began to consider its implications, a number of related challenges presented themselves. These are discussed next, and determine our requirements.

- Once people start sharing the news article, discussion spreads out over different sites. This adds a new dimension to the requirements – it would be most helpful to the policy-maker to track where the discussion on an article occurs over time from initial publication of the article.

- News events are not usually covered with a single news article on one website – it is more likely that there are multiple articles written from different perspectives in different newspapers and on different websites, and each generates their own set of comments from the citizens that read them. In addition, news events develop and multiple articles are written, adding new developments and analysis. This adds a further requirement: to be able to track reactions to multiple articles, and to group them into sets, so they can be presented to analysis in logical groups.
- It is also probable that the policy maker does not know the exact news articles they wish to track, or they only know of a subset of articles, so a related requirement is to enable searching for news articles to "bootstrap" the tracking of reactions to them.
- Finally, policy makers are often specialists in, or are responsible for, a certain discipline or topic area, and it would be most helpful to them to determine key sites that are worth monitoring for general discussion and ideas around this topic.

Given the problem statement and the requirements above, a number of research questions arose:

1. How can we find out where a news article is being discussed?
2. How do the discussions' locations change over time?
3. How can we track a news story containing many news articles?
4. How can we find news articles related to a press release or an MP's statement?
5. Which are the important places a policy maker needs to monitor for discussion of items relevant to them?

1.1. Strategy

Once we had identified the research questions, we determined methods and plans to provide answers to these questions:

- To address research question 1 (how to find where a news article was being discussed on the internet), we proposed a strategy whereby we perform internet searches for references to the news article and store the results in a database. The assumption underlying this strategy is that if the news story is referenced in a web page (i.e. it is returned as a hit in an internet search for the news article), then that web site has at least some relation to the news story.
- To track how the discussions' locations changed over time (research question 2), we proposed to repeatedly (automatically) execute the same search on a regular basis, and store these results along with the original results in the database.
- Tracking a news story containing multiple articles (research question 3) is a matter of grouping searches and results together into a set for the story. This is simply a question of management of the searches and results; and maintaining links between sets of results, searches, news articles and news stories. We proposed to utilise relational database patterns to maintain these links - databases are well suited for this task.
- If we do not know the URL of the news articles we want to track, we will need to find the articles themselves (research question 4). For this we have proposed that we search selected newspapers' websites for keywords from the press release or MP's statement. The result set should be links to articles about the press release etc. Once we have this set of articles, they can be used as input to the strategies above.
- Finally, in order to determine useful sites for general monitoring (research question 5), we have proposed that we analyse groups of search results, to determine which web domains are most frequently featured, and how they are ranked.
- We also determined that we should be able to select arbitrary sets of search results for this analysis, so that we can determine the best sites given any set of results, from one single set to all sets. This enables maximum flexibility to "data mine" the search results. We should be able to determine their own groups – for example multiple stories may be related because they are

about a similar subject area. Grouping the results from these and analysing them produces a set of web domains pertinent for that subject area.

1.1.1. Searches

There are two main types of search. The first type is a search for comments and references to news articles. This is the major form of search (addressing research question 1, to find where news articles are being discussed), and returns result sets containing ordered hits for references to the news article. The second search type is a search for news articles given the text or keywords of an MP's statement. This is used when the news articles are not known, and the results can feed into the search for comments and references to news articles. This addresses research question 4.

Searches for Comments and References to News Articles

The basic strategy we chose for this search was to automate data collection based on Google searches for places where the news article occurs, and to store the search results (hits) in a database. The same searches can be repeated periodically over time, and differences in where the news item is discussed can be highlighted.

The first thing we did was determine how to search the internet, and the solution was straightforward: we chose to use Google because it is the most popular search engine in the western world. It has by far the largest market share of the search engines[2], and appears to be the de facto choice for most users of the internet. As such, it is very likely to be used by many people who may want to comment on a news article – if someone wants to comment on a news article, they are likely to either:

- comment directly on the news site itself;
- comment on a website / SNS / forum / blog they already know about; or
- Google search for the news article to find places where it is being discussed.

Therefore using Google, we will find pages with comments from people who do any of these actions. We chose to perform two Google searches for each news article: firstly the article's headline and secondly the news article's web page URL. This enabled us to capture references to the news article (when the URL is quoted in a referring text), and the more general search for the headline, which uses natural language processing such as stemming, removal of stop words and fuzzy matching.

Google searching returns references to web pages (hits), ranked by Google's proprietary algorithm. The Google ranking gives us a measure of how useful a hit will be. Google's ranking of a search hit is important to us, as it determines the "popularity" of a site in response to the search, and we wish to find the most likely sites where people will go to make comments. Therefore our goal is in line with Google's: to find the "best" sites for the search. We do not know the exact details of Google's search algorithm as it is proprietary, but we do know it is founded upon citations – links to a web page behave as "votes" to increase its rank. What the algorithm contains is not important to us, but that it is used and relied upon by a vast user base, is. Google has a vested interest in returning useful hits to its users, and its market share indicates that it is doing just this.

To perform the actual searching, we used a Google Custom Search Engine (CSE) with automated control to repeatedly execute searches. The frequency of searching is configurable, and our initial configuration is that we search at the same time every other day. We created a relational database to hold the search results. For each hit in a search result set, we record the URL of a hit, its domain and the Google rank, as well as the search information such as the search query and

[2] See for example Netmarketshare. URL: http://www.netmarketshare.com/ (Retrieved 20 November 2011).

the date of searching. This enables us to perform analyses of how the ranking can change over time as well as aggregate analyses to determine the best sites given arbitrary sets of search results.

An example of search results for a news article, collected over a time period, is shown in Figure 2. The figure illustrates the relationship between the news article and repeated searches for its headline and URL, and also illustrates how we can address research question 2 (tracking the changes of discussion location for the news article over time).

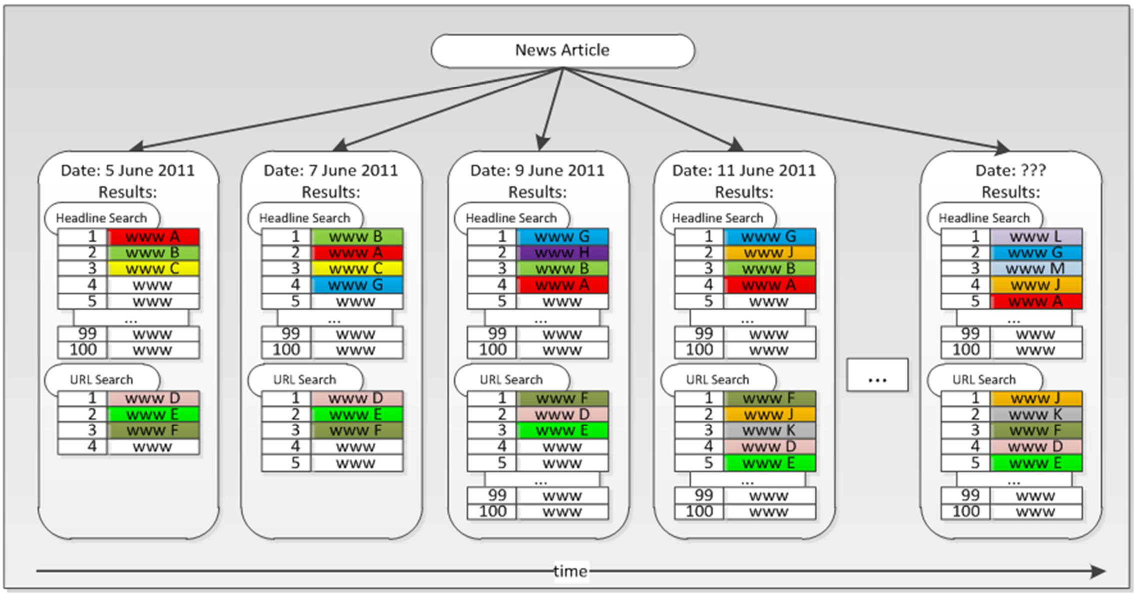


Figure 2: Example News Article, Associated Searches & Results

The sets of hits in the search results above are analogous to the music charts denoting which records are the most popular at a given time. Our “hit parade” is of the top 100 web sites where the news article is discussed. The charts are updated periodically and web sites move up and down the chart according to their popularity. We also get new entrants to the chart and other web sites drop out of the top 100. Figure 2 illustrates this by showing the progress of some example websites — we can see the progress of domain A over time — it starts off at the top, and then drops down the list. Domain G is another example: it does not feature until the second search, but then rises to the top before tailing off. Figure 2 also illustrates that we may not get a full set of hits early on in the search, especially for a fresh news article. This is particularly true for the URL search, as it is highly specific and the search engine cannot use any natural language or fuzzy matching techniques to widen the result set. This is a desirable property for our purposes, as it means we are getting exact matches for the URL, and we can see its propagation.

Searching for News Articles

The second search type, where news articles themselves are found (research question 4), is only required when the news articles are not known, or the user wants to see newspaper reports of a particular policy statement. This search also uses Google, but requires that a limited section of the Internet is searched – we only want news websites to be searched here. We created a second Custom Search Engine (named here the “Newspaper CSE”), that only searched a sample of UK news and newspaper sites (for example www.bbc.co.uk/news, www.telegraph.co.uk, etc.).

The Newspaper CSE can be searched using keywords from a government press release or an MP’s statement, and because it is configured only to search newspaper and news sites, the results will be news articles. These news articles can then be used as inputs to the other search type (references and comments) to see where the articles are discussed on the wider internet.

The choice of news sites is customisable; the Newspaper CSE can be altered at any time, so additional news sites can be included. The Newspaper CSE could also be targeted to a specific purpose, for example a subject area, or geographical location (the Newspaper CSE searches could be local newspapers rather than national ones).

1.2. Data Analysis

By utilising different criteria to group the search results, we can answer research questions 3 (tracking the discussion locations for multiple news articles related to a news story) and 5 (a general aggregated analysis showing useful sites given multiple different data sets).

The grouping process may be thought of as a form of data mining known as an OLAP cube[3]. This allows data to be grouped and analysed along different dimensions. The dimensions we can utilise are: story names / keywords / subject areas, web domains, dates and article titles.

An example OLAP cube for a complete story is shown in Figure 3, alongside an analogous set of searches for news articles. Here we are comparing news articles against dates, and showing the ranked set of domains for each article and date.

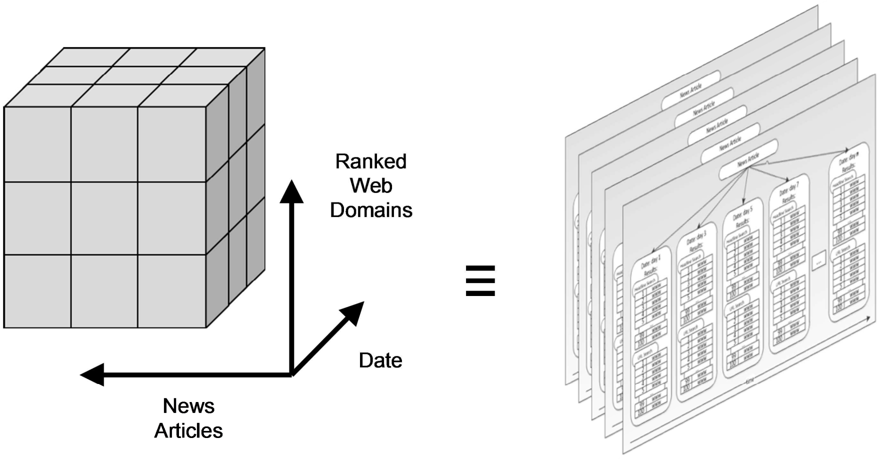


Figure 3: Example OLAP Cube

We can use slices from the OLAP cube as well – this means we are only interested in one value in a particular dimension. For example, if we are only considering one story keyword, we only have a single value on that dimension.

In OLAP-style analysis we often want to aggregate values in a particular axis so we can examine the effect of other axes on the overall result. For example we may wish to investigate the aggregated ranking of all news articles in our story and how it changes over time.

In the right hand side of Figure 3, this is collapsing the five layers into one and finding aggregated rankings in all the hit parades, taking into account all layers. This provides us with an overview of the top sites where the discussions on all our news articles are taking place over time, regardless of the articles.

[3] URL: http://en.wikipedia.org/wiki/OLAP_cube (Retrieved 29 November 2011).

We thus needed a means of aggregating the rankings, and we attempted different methods of determining the aggregated ranking. We originally attempted simple averaging, e.g. we took all the ranks for a particular web domain and computed their average. The major problem with this method was that it was highly sensitive to the number of records for that domain. If, for example, domain A had a single record at position 1, this would have an average value of 1, because there is only a single record. If domain B had ten records, and nine were at position 1 and the remaining record was at position 2, the average value would be 1.1. Given that the lower the average the better in this example, this means that the consistent high performer, domain B, with 9 top positions, was apparently outperformed by domain A, who only appeared once.

Next, we looked at different weighting algorithms, so as to give more importance to the higher positions, but these suffered similar problems to the straightforward average. It was therefore decided that we needed to take account of the number of occurrences a domain has, as well as its position in each occurrence. What we wanted to find was consistent good performers (e.g. domain B above), rather than ones with few high positions but no other records (who could be considered "lucky" without further evidence).

The Bayesian Average method[4],[5] is purpose-built for this task. It reduces the effect of anomalous values by considering the average number of occurrences for each domain as well as the average value per occurrence. It does this by calculating a corrected ranking that takes the number of occurrences a domain has into account using the two following principles: the more occurrences a domain has, the closer its corrected ranking value is to its uncorrected value; and the fewer occurrences a domain has, the closer its corrected ranking is to the average ranking value of all domains. Thus, the more times a domain appears in search results, the more "believable" its scores are.

After the data slicing, aggregation and Bayesian averaging, we have ranked tables of "chart positions" for each domain that take into account the way we have sliced the data, the chart positions and the number of votes for each chart position. Using this aggregation and the OLAP cube technique, we can show the aggregated ranking of multiple search results, for example:

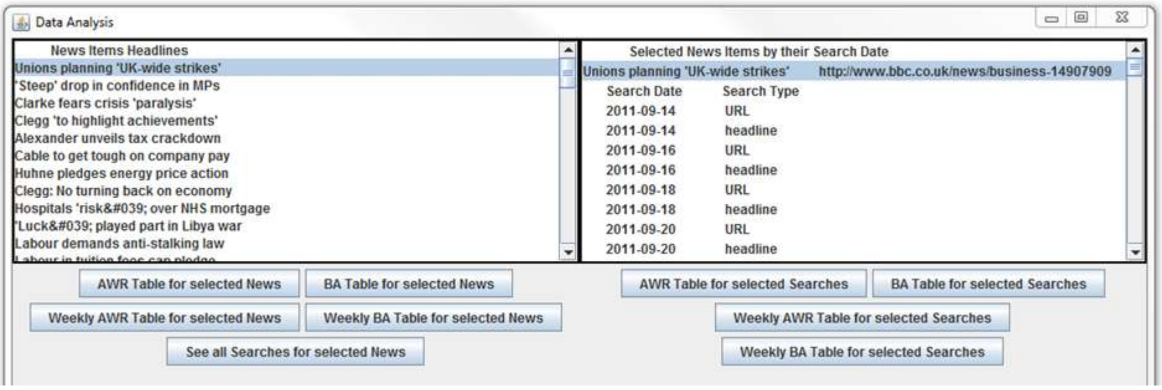
- A single story (e.g. all the searches over time to date for the story)
- The time development of a single story (e.g. where discussions for all news articles in the story change over time)
- Multiple stories (e.g. the user can select related stories to find out where people are talking about them)
- One day (e.g. all searches on one single day independent of story)
- Everything (e.g. all results for all stories to date).

1.3. Results & Initial Evaluation

We implemented a software tool to perform the searches and analyses described above. We show here an example of its output for a case study of a news article. The main UI of the software tool with the article and its searches displayed is shown in Figure 4.

[4] URL: http://en.wikipedia.org/wiki/Bayesian_average (Retrieved 30 November 2011).

[5] URL: http://leedumond.com/blog/the-wisdom-of-crowds-implementing-a-smart-ranking-algorithm/ (Retrieved 30 November 2011).


Data Analysis
AWR Table for selected News
BA Table for selected News
Weekly AWR Table for selected News
Weekly BA Table for selected News
See all Searches for selected News
AWR Table for selected Searches
BA Table for selected Searches
Weekly AWR Table for selected Searches
Weekly BA Table for selected Searches

News Items Headlines

Unions planning 'UK-wide strikes'
'Steep' drop in confidence in MPs
Clarke fears crisis 'paralysis'
Clegg 'to highlight achievements'
Alexander unveils tax crackdown
Cable to get tough on company pay
Huhne pledges energy price action
Clegg: No turning back on economy
Hospitals 'risk&#039; over NHS mortgage
'Luck&#039; played part in Libya war
Labour demands anti-stalking law

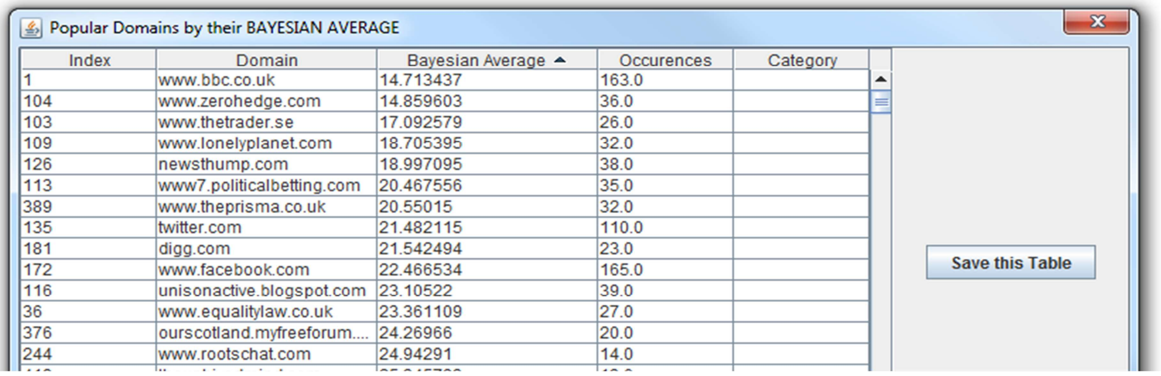
Selected News Items by their Search Date

Unions planning 'UK-wide strikes' http://www.bbc.co.uk/news/business-14907909

Search Date	Search Type
2011-09-14	URL
2011-09-14	headline
2011-09-16	URL
2011-09-16	headline
2011-09-18	URL
2011-09-18	headline
2011-09-20	URL
2011-09-20	headline

Figure 4: News Story Tracking

The example is based on a single news story, from the BBC news website. This concerned a story about plans for a national public-sector strike, published on 14 September 2011[6]. Figure 5 shows a results page, showing the Bayesian Average ranking, position, and number of occurrences for different domains over all times the news article is searched for (14 September 2011 to 30 November 2011).



Index	Domain	Bayesian Average ▲	Occurences	Category
1	www.bbc.co.uk	14.713437	163.0	
104	www.zerohedge.com	14.859603	36.0	
103	www.thetrader.se	17.092579	26.0	
109	www.lonelyplanet.com	18.705395	32.0	
126	newsthump.com	18.997095	38.0	
113	www7.politicalbetting.com	20.467556	35.0	
389	www.theprisma.co.uk	20.55015	32.0	
135	twitter.com	21.482115	110.0	
181	digg.com	21.542494	23.0	
172	www.facebook.com	22.466534	165.0	
116	unisonactive.blogspot.com	23.10522	39.0	
36	www.equalitylaw.co.uk	23.361109	27.0	
376	ourscotland.myfreeforum....	24.26966	20.0	
244	www.rootschat.com	24.94291	14.0	

Figure 5: Bayesian Average Overall Rankings

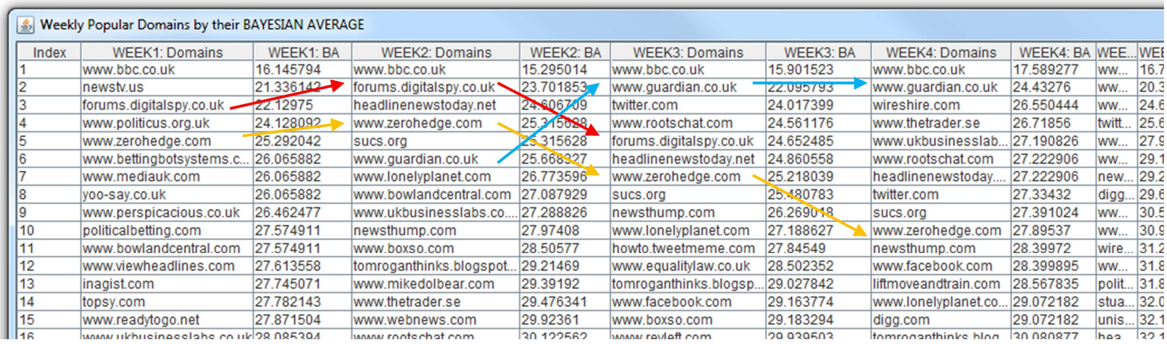
Because they reference the news article, the sites can be useful places for the policy-maker to watch for opinions on their policy statements. In the results table, the domain is only shown, and we use this as a gateway to the actual results: we click on a domain and we can see the hits associated with that domain. A hit contains the URL of the page, Google's "snippet" and the rank. The table is sorted by the Bayesian Average of the Google rankings. The smallest is first – this means the most interesting sites as determined by Google are at the top. The number of occurrences of each site indicates how many data points were used to compute the overall Bayesian Average ranking, and we can see that there are reasonable numbers of samples (i.e. search results) for each site. We can also see that www.bbc.co.uk is the highest position – this is reasonable and to be expected, since the BBC needs to index its own pages, but is rather obvious. This is no problem; it is a known good site, so we just look further down the list.

The "category" column allows the user to mark some sites as known good sites, or sites that are not useful. The policy maker can mark a site as "uninteresting" in the "category" column, and it is

[6] Available at: http://www.bbc.co.uk/news/business-14907909 (Retrieved on 1 December 2011).

ignored in this and all other result tables. The data from ignored sites is still collected and analysed, and the blacklist of uninteresting sites can be edited at any time to return the blacklisted sites to the analysis results.

The overall rankings may be broken down by time, and Figure 6 shows rankings per week. The results are grouped per week and are for the single news article. Here we see a more varied set of sites, and their positions change week by week. The arrows in the figure illustrate the movement of some particular domains, and we discuss one of them in more detail next.



Weekly Popular Domains by their BAYESIAN AVERAGE

Index	WEEK1: Domains	WEEK1: BA	WEEK2: Domains	WEEK2: BA	WEEK3: Domains	WEEK3: BA	WEEK4: Domains	WEEK4: BA	WEE...	WEE
1	www.bbc.co.uk	16.145794	www.bbc.co.uk	15.295014	www.bbc.co.uk	15.901523	www.bbc.co.uk	17.589277	ww...	16.7
2	newstv.us	21.336142	forums.digitalspy.co.uk	23.701853	www.guardian.co.uk	22.095793	www.guardian.co.uk	24.43276	ww...	20.3
3	forums.digitalspy.co.uk	22.12975	headlinenewstoday.net	24.606709	twitter.com	24.017399	wireshire.com	26.550444	ww...	24.6
4	www.politicus.org.uk	24.128092	www.zerohedge.com	25.315628	www.rootschat.com	24.561176	www.thetrader.se	26.71856	twitt...	25.6
5	www.zerohedge.com	25.292042	sucs.org	25.315628	forums.digitalspy.co.uk	24.652485	www.ukbusinesslab...	27.190826	ww...	27.9
6	www.bettingbotsystems.c...	26.065882	www.guardian.co.uk	25.668327	headlinenewstoday.net	24.860558	www.rootschat.com	27.222906	ww...	29.1
7	www.mediauk.com	26.065882	www.lonelyplanet.com	26.773596	www.zerohedge.com	25.218039	headlinenewstoday....	27.222906	new...	29.2
8	yoo-say.co.uk	26.065882	www.bowlandcentral.com	27.087929	sucs.org	25.480783	twitter.com	27.33432	digg...	29.6
9	www.perspicacious.co.uk	26.462477	www.ukbusinesslabs.co....	27.288826	newsthump.com	26.269018	sucs.org	27.391024	ww...	30.5
10	politicalbetting.com	27.574911	newsthump.com	27.97408	www.lonelyplanet.com	27.188627	www.zerohedge.com	27.89537	ww...	30.9
11	www.bowlandcentral.com	27.574911	www.boxso.com	28.50577	howto.tweetmeme.com	27.84549	newsthump.com	28.39972	wire...	31.2
12	www.viewheadlines.com	27.613558	tomroganthinks.blogspot...	29.21469	www.equalitylaw.co.uk	28.502352	www.facebook.com	28.399895	ww...	31.8
13	inagist.com	27.745071	www.mikedolbear.com	29.39192	tomroganthinks.blogsp...	29.027842	liftmoveandtrain.com	28.567835	polit...	31.8
14	topsy.com	27.782143	www.thetrader.se	29.476341	www.facebook.com	29.163774	www.lonelyplanet.co...	29.072182	stua...	32.0
15	www.readytogo.net	27.871504	www.webnews.com	29.92361	www.boxso.com	29.183294	digg.com	29.072182	unis...	32.1
16	www.ukbusinesslabs.co.uk	28.085394	www.rootschat.com	30.122562	www.revleft.com	29.939503	tomroganthinks.blog	30.080877	hea	32.1

Figure 6: Search Results - Bayesian Average Rankings per Week

The Digital Spy forums contained two threads that discussed the BBC news article, and was therefore a hit in the search. Digital Spy started off in position 3, peaked in the second week at position 2, and then dropped to position 5 before dropping out of the visible data. The first thread[7] was simply discussing the article, and the first post included a link to the BBC news story. This thread had 120 posts – the first was made on 14 September 2011, and the last was made on 7 October 2011. The bulk of the activity was in the first week. The second thread[8] was a poll about whether people supported the strike, together with opinions given in the thread. The thread accompanying the poll contained 522 posts, the first was made on 14 September 2011, and the last was made on 19 September 2011. The bulk of the posts were made over 15-18 September.

The pattern is common – an event occurs, and there is a flurry of activity concerning it, which peaks and then tails off. Given that there may be a lag in Google's indexing of the Digital Spy forums, the activity on the forum and the rankings in the table give a reasonable match. The important point is there is genuine and useful debate in this forum. Other forums with debate on this topic were also highlighted by the results table.

This initial evaluation has demonstrated that the approach works, but it is worthwhile assessing the positive and negative aspects of the approach.

- *Positive aspects.* Firstly, our approach uses Google, the most popular search engine in the western world. Its performance is attested by the fact that millions of people use it daily. Secondly, the Bayesian average approach to aggregating search scores has the advantage that it reduces the effect of infrequent anomalous values, resulting in a score supported by the bulk of the data points. Finally, the approach allows the user to see the changes over time in the rankings of the sites discussing the article.
- *Negative aspects.* The major drawback with our approach is that the ranking of the results are determined by a proprietary and unknown algorithm. However, we can easily adopt another search engine without adjustment of our technique. Another drawback is

[7] Public Sector Workers Balloted On Strikes. URL: http://forums.digitalspy.co.uk/showthread.php?t=1534213 (Retrieved 8 December 2011).

[8] Do you support the union strikes? URL: http://forums.digitalspy.co.uk/showthread.php?t=1534436 (Retrieved 8 December 2011).

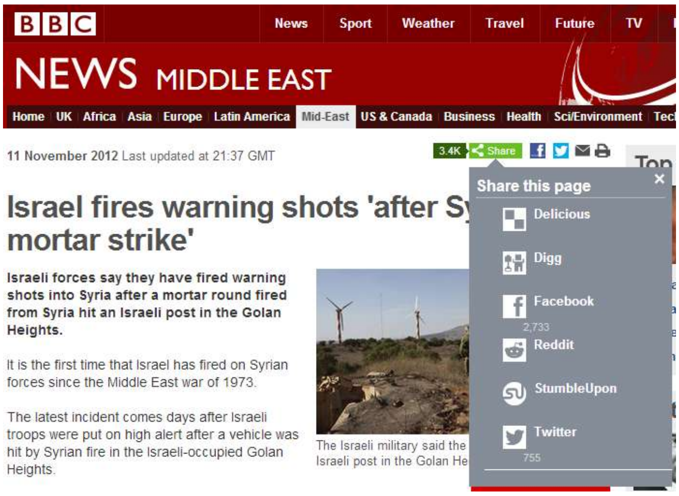
that in order to see where discussion of a story changes over time, the story has to be tracked from its beginning. Automation can assist here – the user can specify queries to find news stories as they happen, and these can be tracked automatically. A further drawback is that our approach is based on a numerical aggregation of rankings to determine the relevance or popularity of a site pertaining to a particular query – no account is taken of the actual postings on the sites. The lexical analysis of postings is addressed in other aspects of the WeGov project (see for example Sizov, 2010), and this work provides starting points for searches that can provide input to these analyses.

2. Summarisation of Discussions on Social Networks

Once we have determined where people are talking about a news item, we next want to understand what they are saying. This work addresses the challenge of sifting through a potential deluge of comments and posts from social network to find the key themes, sentiments and postings. Once we have understood what is being said, in order to fully interact with citizens on the social networks, we will need to know the key users and posts we should respond to, retweet or follow. In this section we aim to address these questions through introduction of the use of the WeGov toolkit, and illustrations of its use with an example of discussions about a news article on the social networking sites Facebook and Twitter.

2.1. Starting with BBC News Story

On 11 November 2012 (Last update at 21:37 GMT) the BBC published the news story “Israel fires warning shots ‘after Syria mortar strike’” on its web page. Figure 7 shows the screenshot with the news story, and readers can share and discuss the stories. Here the grey box displays that readers shared the story 2,733 times on Facebook and 755 times on Twitter. We know in general the sites where discussions were taking place, but there is no functionality to retrace how the story was disseminated within the social networks and what comments users made.



11 November 2012 Last updated at 21:37 GMT

Israel fires warning shots 'after Sy
mortar strike'

Israeli forces say they have fired warning shots into Syria after a mortar round fired from Syria hit an Israeli post in the Golan Heights.

It is the first time that Israel has fired on Syrian forces since the Middle East war of 1973.

The latest incident comes days after Israeli troops were put on high alert after a vehicle was hit by Syrian fire in the Israeli-occupied Golan Heights.

The Israeli military said the
Israeli post in the Golan He

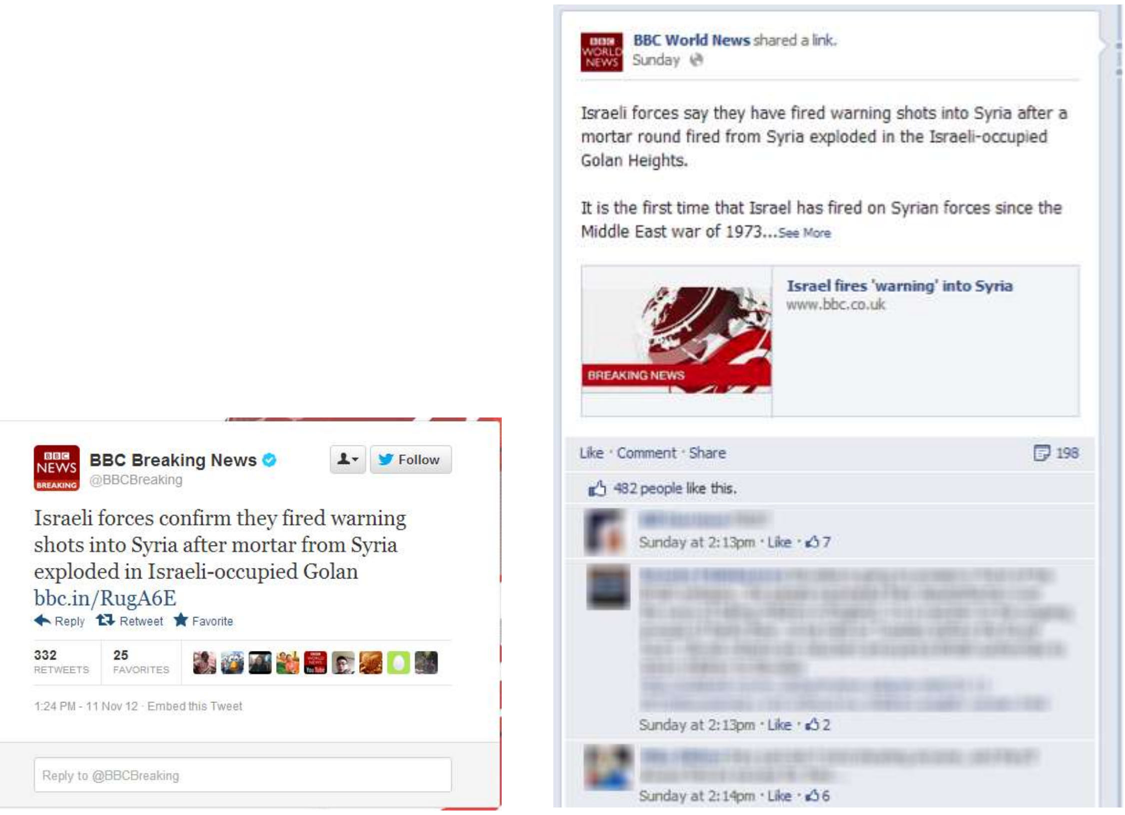
Figure 7: BBC News Story[9]

[9] Original news story published on the BBC web page. URL: http://www.bbc.co.uk/news/world-middle-east-20288263 (Retrieved 18 November 2012).

2.1.1. Seed Posts on Twitter and Facebook

On 11 November at 13:24 GMT the news story was published on Twitter via the BBC Breaking News[10] profile. Figure 8 (Cp. left) shows the screenshot of the tweet and additional information: In total 332 users retweeted the message and 25 users marked the tweet as a favourite tweet. The functionalities for replying to the user, for retweeting the message to the own followers and for highlighting the tweet as a favourite are directly available under the message.

After the story was published on Twitter, at 14:12 GMT on the same day, the BBC shared the news story on its Facebook page BBC World News[11]. Figure 8 (Cp. right) shows the screenshot of the BBC post. This created 482 comments and 199 shares on Facebook, at the time of writing. The shared link refers again to the web page with the news story.



BBC Breaking News
@BBCBreaking

Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan bbc.in/RugA6E

332 RETWEETS 25 FAVORITES

1:24 PM - 11 Nov 12 · Embed this Tweet

Reply to @BBCBreaking

BBC World News shared a link.
Sunday

Israeli forces say they have fired warning shots into Syria after a mortar round fired from Syria exploded in the Israeli-occupied Golan Heights.

It is the first time that Israel has fired on Syrian forces since the Middle East war of 1973...See More

Israel fires 'warning' into Syria
www.bbc.co.uk

198

482 people like this.

[redacted]
Sunday at 2:13pm · Like · 7

[redacted]
Sunday at 2:13pm · Like · 2

[redacted]
Sunday at 2:14pm · Like · 6

Figure 8: BBC News Story Published on Twitter[12] (Cp. left) and Facebook[13] (Cp. right)

From the number of comments and shares of the story on Facebook and Twitter, determining the reaction to it manually would be very difficult. In addition, at the time of writing, the story was quite new, so there is a strong chance that more comments will follow, making the problem more challenging.

[10] Official Twitter Account BBC Breaking News. URL: https://twitter.com/BBCBreaking (Retrieved 18 November 2012).

[11] Official Facebook Page BBC World News. URL: https://www.facebook.com/bbcworldnews/ (Retrieved 18 November 2012).

[12] News story Twitter account BBC Breaking News. URL: https://twitter.com/BBCBreaking/status/267603641077219328 (Retrieved 18 November 2012).

[13] BBC news story published on the BBC World News Facebook page. URL: https://www.facebook.com/bbcworldnews/posts/297088283728675 (Retrieved 18 November 2012).

Our approach of providing automated tools to summarise the themes and opinions in the discussions is presented in the remainder of this section. Our approach is implemented as a kit of search and analysis tools for discussions on social networking sites, integrated into the so-called “WeGov Toolbox”, and this is described next.

2.2. Introducing the WeGov Toolbox

WeGov is a web-based system that enables the user to collect and analyse social network postings and users, and to inject posts into social networks. The system is deployed and hosted at a server, and the user connects to this using their web browser. The key features of WeGov are as follows.

- The user can specify and run searches on social networks and feed the search results into WeGov’s two analysis components to provide summaries and automated insights into the (sometimes very large) data set returned from the social networks.
- The user can search on the two social networks currently supported: Facebook and Twitter. On Facebook, the user can monitor public groups and pages: the user can instruct WeGov to collect posts and comments on those posts from a Facebook group or page by specifying the URL of the page. On Twitter, the user can search for keywords or hashtags. Searches can be scheduled, so that they repeat automatically. This is useful for collecting data over an extended period, which is particularly suitable for monitoring a news story. The system is designed so that when a search is executed multiple times by a schedule, it will not collect any duplicate posts, as duplicates can skew analysis results.
- The first of the analysis components of WeGov is behaviour analysis, developed by the Open University, Knowledge Media Institute (KMi), which monitors the discussion activity, categorises users into behaviour types and highlights posts and users to watch.
- The second analysis component is topic-opinion analysis developed by the University of Koblenz, which determines themes of the documents (posts, comments, etc) in the discussion by identifying sets of terms that frequently occur together in multiple posts and grouping them together into topic groups. In addition, opinions are determined by sentiment analysis, and the topic groups can be measured in terms of whether they express positive or negative opinion.
- WeGov presents data in two formats: “widgets” and “advanced”. Both formats provide search and analysis functions: widgets provide simple and quick functionality whilst the advanced search and analysis provides more control and flexibility in the way results can be viewed.
- We have adopted a methodical approach for the development process of the software with frequent and iterative end user engagement REFERENCE TO: Wandhöfer et al, 2012 so as to get requirements and feedback on development progress REFERENCE TO: Joshi et al, 2012. As part of user engagement, a number of use cases were designed showing how the WeGov analysis tools could provide a two-way dialogue with citizens REFERENCE TO: Addis et al, 2010, and the work reported here develops one of these use cases.
- An important aspect of the work in WeGov is to protect the rights and privacy of citizens and policy makers. To address this, a legal and ethical analysis was conducted to provide us with an understanding of data protection issues and give an insight into transparency. This work has influenced the design and use of all parts of the toolbox, and has been reported elsewhere (Wilson & Fletcher, 2010). The impact is has on the work here is that we only collect posts from publicly-accessible sources.

2.2.1. Analysis Tools

WeGov's analysis tools provide the functionality to address the challenge of what people are saying, and because of their relevance to addressing the challenge, they are discussed in more detail next.

Discussion Activity Analysis

With billions of users generating information in online communities, it is becoming increasingly important to distinguish those users who are most likely to generate more activity than others. This knowledge will help policy makers focus their attention on popular discussions and leading users. To this end, the WeGov analysis tools study the characteristics of those posts and users that generate lengthy discussions within the network. This study is based on the extraction of the following user and content features: REFERENCES LIST Rowe et al, (2011a) & (2011b)

- User features describe the author, 'U', of a post by capturing his standing and engagement in the system. These features are: in-degree (number of users following U), out-degree (number of users U follows), post count (number of posts U has made), user age (the length of time U has been a member of the community), and post rate (number of posts made by U per day)
- Content features define quality measures of a post 'P' such as novelty of language, sentiment and time of posting. These features are: post length (number of words), complexity (cumulative entropy of P's terms to gauge the concentration and dispersion of language), readability (Gunning fog index, gauging how hard the post is to parse by humans), referral count (number of hyperlinks within the post), time in day (number of minutes through the day), informativeness (the novelty of the post's terms with respect to other posts), and polarity (average polarity of the post using Sentiwordnet[14]).

The objective of this WeGov analysis is to distinguish the key user and content features that spike activity in an online community. Identifying important features and predicting high-attention posts offer two benefits to the policy maker. Firstly, it assists the policy maker in focusing his attention where the largest participation occurs therefore maximising his own involvement to the community. Secondly, it provides the policy maker with recommendations on where and when to make their own posts (content placement strategies) for provoking high activity around his own posts.

Predicting the discussion activity a given post is likely to generate is carried out in two steps:

- Identifying seed posts: A seed post, 'P', is a post that generates at least one reply. The goal of this step is to understand which of the User and Content features render P as a seed and to provide classification models that are able to identify seed posts from non-seed posts. For this purpose, three different classification models (Naive Bayes, Maximum Entropy and J48 decision tree) are assessed using user features, content features, and the combination of both. We use F-measure, precision, recall and the area under the ROC curve to measure the accuracy of these models so that we select the best performing one. Once the best-performing seed classification model is selected we analyse which features are the most important in identifying seeds. For this purpose, we remove one feature at a time from the best performing model and measure the reduction in accuracy. The outcome of this step is the ranking of the features that helps us identify seed posts from non-seed posts.
- Predicting Activity levels: The goal of this step is to rank the previously identified seed posts, sorting in the higher positions those ones that are predicted to generated higher levels of attention (number of replies). The ranking of seeds post is done by means of a regression model. In order to choose the most accurate model three different regression models (Linear Isotonic and Support Vector Regression) are assessed considering user features, content features and the combination of both. To evaluate these models we compare their output with

[14] URL: http://sentiwordnet.isti.cnr.it/ (Retrieved 22 November 2012).

the actual rank based on activity volume (number of replies) by using the Normalised Discounted Cumulative Gain (nDCG) evaluation metric. Once the best model has been selected we assess the relevance of the features by looking at the model coefficients and how they are associated with activity volume.

We performed this analysis on different datasets collected from online communities. In the scope of the WeGov Project we analysed a large (1.5M posts) randomly collected dataset from Twitter. The results of our analysis indicate that in order to generate attention the content of the post is more important than the reputation of the user within the SNS. In particular, those posts that generate high levels of attention generally fit the following characteristics: are NOT written in the afternoon (time of the day), are written in a familiar language to the users of the networks (i.e., high readability and low Informativeness), are written by users who follows a lot of people, i.e., listen what others say (high out-degree), and tend to be negative (negative polarity of the post).

User Behaviour Analysis

The User Behaviour Analysis component aims to monitor and capture citizen's behaviour over time. This analysis draws the attention of the policy maker to a smaller, more manageable, set of users, with whom he may want to engage more closely (read their contributions, monitor their opinion, answer their questions, invite to participate in further discussions, surveys etc.). This analysis is particularly useful when there is a large number of participants that the policy maker cannot possibly pay equal attention to. The behaviour analysis categorizes users in online communities with the roles they hold in the context of these communities and in a specified timeframe. To perform such labelling we first need to capture the behaviour of users in online communities, define what sort of behaviour is associated with particular roles and classify user's in different roles according to their exhibited behaviour. REFERENCES LIST Angeletou et al, (2011) & Rowe et al, (2012)

The behaviour and role analysis literature showed a large divergence in roles. The identified roles are generally subjective (i.e., no empirical basis drives their derivation) and their assignment is down to the interpreter. There role identification approaches are divisible into two main types: (i) interpretive analysis (ethnography, surveys, and interviews) and, (ii) structural analysis (formal computational methods). For the purpose of this work, we analyse a dataset of more than 1.5M posts (randomly collected from Twitter) following one of the more formal structural analysis methods found in the literature at the time of writing REFERENCE TO: Chan et al, 2010 and empirically derive the following set of roles:

For representing users in Twitter, we selected the roles of:

- Broadcaster (users who post a lot and are followed a lot but rarely follow anyone),
- Information Source (users who post a lot, are followed by many people but they also follow many people themselves),
- Information Seeker (users who follow many users but do not post frequently themselves),
- Rare Poster (users who post very rarely) and
- Daily User (users who follow and are followed by an average number of users, and that posts with a medium level frequency).

- For associating users with roles in a particular community and time frame the behaviour features of each users (post-rate, in-degree, out-degree, etc.) are compared against the features that characterize each role. For example, in order for a user to be classed as an Information Source he should have high values of post-rate and out-degree. The outcome of this step is the classification of a given set of users into roles that best represent their behaviour.

Analysis of Topics and Opinions

In many cases, discussion tracks in social media become quite long and complex. Stakeholders of WeGov technology (such as politicians, political researchers, active users) are often interested in gaining a quick overview of such a discussion, including understanding its thematic aspects, identifying key pro and contra arguments and finding the most influential users. However, completely reading hundreds (or even thousands) of posts is a time-consuming enterprise. The Topic-Opinion Analysis toolbox of WeGov aims to provide appropriate summarization techniques by identifying latent themes of discussion (topics), most relevant contributions and arguments for each topic, as well as identifying the most active users that influenced a certain aspect of discussion. REFERENCE TO: Sizov, 2010. The topic-opinion tool employs state of the art methods of Bayesian learning and opinion mining for finding the most relevant pieces of information that should be presented to the user, and these are briefly described next.

Modelling topics: Probabilistic Bayesian models are used for mining the latent semantic structure of the online discussion. The WeGov approach can be seen as an extension to the state-of-the-art method coined "Latent Dirichlet Allocation" (LDA). The collection of postings is represented by means of probabilistic distributions over terms (words) that appear in particular discussion postings with different frequencies. The Bayesian learning process provides estimates of multinomial distributions over terms for a limited number of topics (themes). In other words, each topic can be characterized by its most relevant terms. Consequently, postings are represented by means of distributions over topics. Postings that belong to a certain topic with high probability are considered as most characteristic examples for the certain aspect of online discussion.

Modelling opinions: The WeGov toolbox employs state of the art techniques for mining user opinions and affect states. Conceptually, they are based on structured vocabularies of affect-specific terms (including ANEW, LIWC, ADU, WordNet-Affect) that indicate a certain emotional state of the posting writer (e.g. scepticism, positive or negative emotions, anger, etc.). Consequently, postings with strong, characteristic opinion/emotion expressions are selected for presentation to the user.

Topic-opinion summarization: Results of topic and opinion analysis are combined for achieving suitable diversification of content that will be presented to the user. First, candidate postings are chosen with respect to their high relevance regarding particular discussion aspects (i.e. topics). Second, for each pre-selected posting, the opinion/emotion analysis is performed. The output is constructed in such a way that a) all topics identified in the dataset are appropriately reflected, and b) postings chosen for each topic reflect different opinions and emotions. As a result, the output contains a limited number of "must-see-first" contributions from the online discussions, covering a broad spectrum of its contextual and emotional facets. Furthermore, the toolbox output contains most characteristic terms for each topic that can be presented to the user as an explanation of the latent discussion structure.

The topic-opinion tool has been evaluated in various realistic settings, including summarization of Twitter tracks of postings, comments to editorial articles on Yahoo News, and commented online blogs of political parties. In all cases, the diversified summaries of discussion tracks have been positively evaluated by test users as a helpful tool for gaining a quick and systematic overview over long and fragmented discussion tracks. Quantitative evaluations have shown that the use of the topic-opinion tool allows for a statistically significant reduction of the time necessary for reading and analysing online discussions.

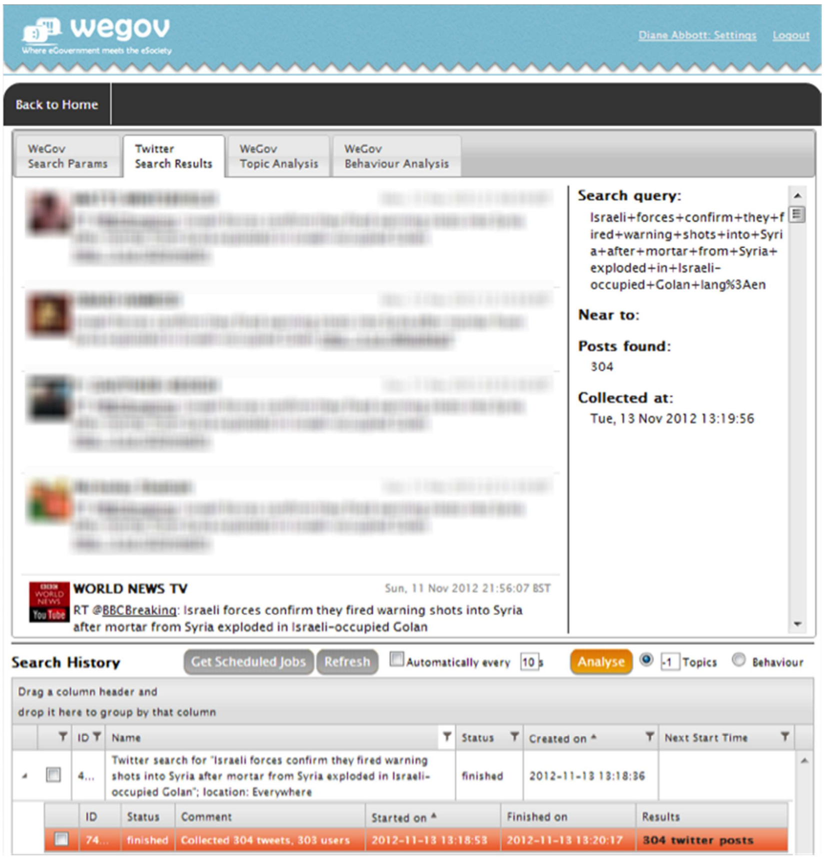
2.3. Strategy to Make Use of the WeGov Toolbox

For this paper we are using the WeGov toolbox to identify the posts that are shared by users on Facebook and Twitter. Afterwards the comments will be stored within a database and analysed by the analysis components that are provided by the WeGov toolbox. Facebook and Twitter work different concerning the dissemination of posts. Hence we are using two diverse strategies.

2.3.1. Twitter

For demonstrating the WeGov analysis we started with the seed post of the BBC Breaking News account on Twitter. The BBC tweet includes the URL that refers to the news story on the BBC web page and the text message: *"Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan*". We used this text message as input for identifying tweets with similar content on Twitter:

- Two days after the seed post – 13 November 2012, at 13:20:17 GMT, we queried[15] the Twitter web page to have a comparison against the WeGov search. The result page included 22 users that shared the similar content.
- Figure 9 shows the screenshot of tweets we queried with the WeGov toolbox nearly the same time we ran the Twitter search – 13 November 2012, at 13:19:56 GMT. Here we used the same search query and got a collection with 304 instead of 22 tweets via the Twitter web page.



WORLD NEWS TV Sun, 11 Nov 2012 21:56:07 BST
RT @BBCBreaking: Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan

Search query:
Israeli+forces+confirm+they+fired+warning+shots+into+Syria+after+mortar+from+Syria+exploded+in+Israeli-occupied+Golan+lang%3Aen

Near to:

Posts found:
304

Collected at:
Tue, 13 Nov 2012 13:19:56

Search History

□ Automatically every 10 s ● -1 Topics ○ Behaviour

Drag a column header and drop it here to group by that column

		ID	Name	Status	Created on	Next Start Time
	□	4...	Twitter search for "Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan"; location: Everywhere	finished	2012-11-13 13:18:36	

	ID	Status	Comment	Started on	Finished on	Results
□	74...	finished	Collected 304 tweets, 303 users	2012-11-13 13:18:53	2012-11-13 13:20:17	304 twitter posts

Figure 9: Twitter Search via WeGov Toolbox

Searching Twitter via the WeGov toolbox provides end users a richer result set and further functionality to work with the data collection. Figure 9 shows the WeGov advanced search that we used to query similar posts to the seed post. Here the user got a list with 304 tweets that is sorted chronologically and can be reviewed by the user from the first to the last post. Each WeGov user gets a unique account that allows the storing of data. The lower third of Figure 9 shows the "Search History" that currently displays the last query with a bar highlighted in orange. To analyse the posts, we select this last search by activating the checkbox at the left hand side of the search summary

[15] Twitter query. URL: http://tiny.cc/rr60nw (Retrieved 18 November 2012).

and confirm the analysis process by clicking the orange coloured “Analyse” button. The WeGov toolbox provides the following analyses on the collection of 304 tweets.

Frequency of Tweets and User Roles Analysis

The first analysis provides a graphical overview of the frequency of tweets when they were published. Figure 10 shows a curve that consists of 32 checkpoints over a 2 ½-day period after the story was published on the BBC’s website. Each checkpoint was plotted nearly every hour. The curve shows a common pattern: most activity about a news story is typically close to its publication time, when it is still fresh and interesting. The analysis identified the first checkpoint as the maximum peak (11 November 2012, at 12:24 GMT) with 226 tweets. Therefore this was the time when most of the tweets where published. Already one hour later the second checkpoint (11 November 2012, at 13:25 GMT) shows only 32 tweets. The next five checkpoints show that the users’ activity decreased continuously. After the seventh checkpoint each of them shows between zero and two published tweets.

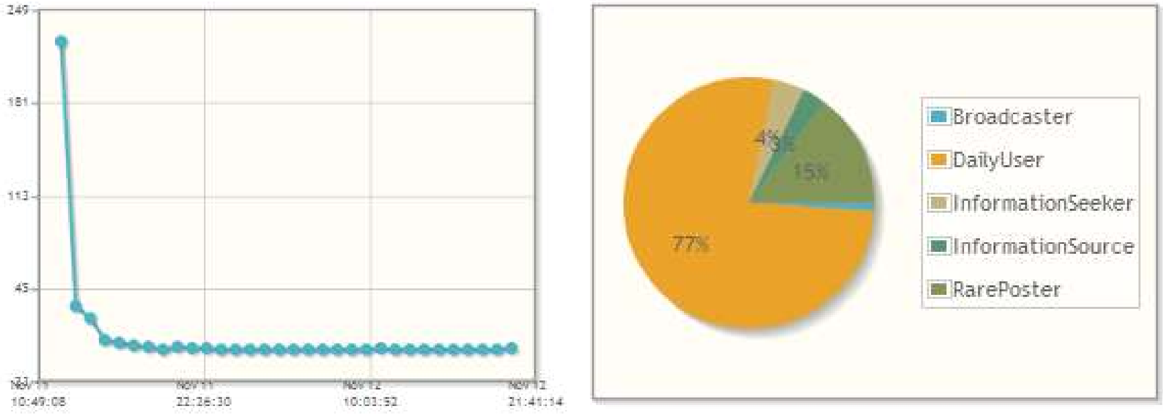


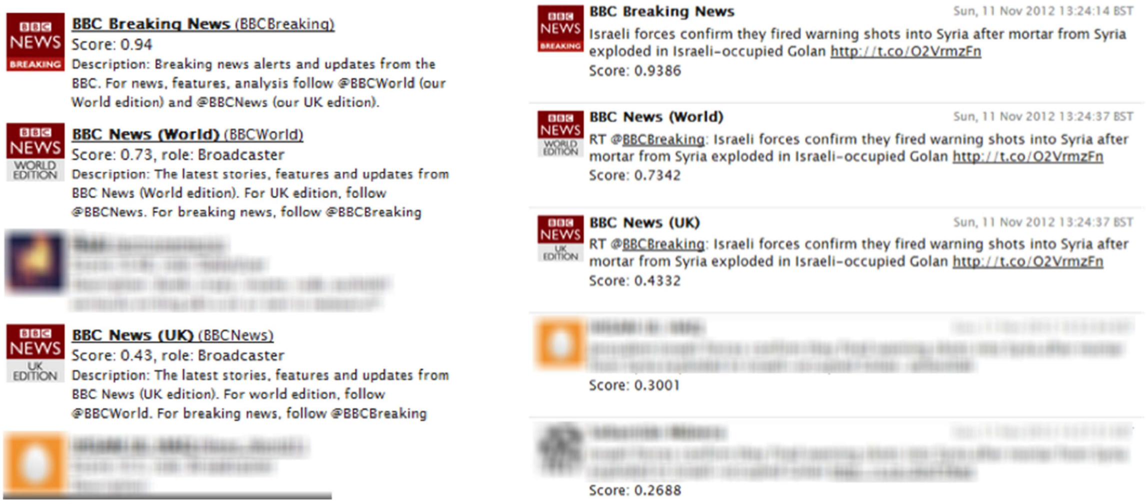
Figure 10: WeGov Analysis Components – Frequency of Posts (Cp. left) and User Roles (Cp. right)

The second analysis provides a classification of the authors of the 304 tweets into the roles described previously in Section 2.2.1. Figure 10 shows how the user roles’ analysis is translated in the user interface of the WeGov toolbox. The pie chart on the right hand side consists of five parts with different sizes. The orange coloured part shows that 77% of users are “Daily User” — this means they are publishing daily tweets and following the activities on Twitter frequently. The second biggest part is the “Rare Poster” — this means 15% of the total amount of users who shared the message are normally publishing on Twitter very rarely. Hence this information seems to be very important to them. The smallest parts are the roles “Information Seeker” (4%), “Information Source” (3%) and “Broadcaster” (1%). Concerning the “Information Seeker”, which characteristic is rather getting than publishing information, the tweet seems to be very important for them. The “Information Source” is well connected on Twitter and very active within discussion. They might be interesting to identify third party statements that their followers have replied. The “Broadcaster” is with 1% the smallest group. This can be explained with the background that these users — typically the press — publish their own seed posts rather than re-tweeting messages from third party news providers. The user roles are available via the WeGov toolbox. The total numbers are 152 “Daily User”, 29 “Rare Poster”, eight “Information Seeker”, six “Information Source”, and two “Broadcasters”.

Top five Users and Tweets to Watch

Figure 11 shows the output of the WeGov discussion activity analysis. The analysis shows the top five users to watch (see left hand side) and the top five tweets to watch (see right hand side).

While the scores for the tweets show its impact on Twitter the score for each user shows the impact of the user in general. For both lists the BBC Breaking News account is the top scorer with 0.94 for the user score and 0.9386 for the single tweet score. This ranking seems realistic because BBC Breaking News has a big standing on Twitter (over four million followers) and users re-tweeted the BBC's message. The second rang is hold by the BBC News (World)[16] profile. The score is 0.73 for the impact of the user and 0.7342 for the impact of the tweet. From the third position the score is quite smaller and the ranking of users is different in both lists.



BBC Breaking News (BBCBreaking)
Score: 0.94
Description: Breaking news alerts and updates from the BBC. For news, features, analysis follow @BBCWorld (our World edition) and @BBCNews (our UK edition).

BBC News (World) (BBCWorld)
Score: 0.73, role: Broadcaster
Description: The latest stories, features and updates from BBC News (World edition). For UK edition, follow @BBCNews. For breaking news, follow @BBCBreaking

BBC News (UK) (BBCNews)
Score: 0.43, role: Broadcaster
Description: The latest stories, features and updates from BBC News (UK edition). For world edition, follow @BBCWorld. For breaking news, follow @BBCBreaking

BBC Breaking News Sun, 11 Nov 2012 13:24:14 BST
Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan http://t.co/O2VrmzFn
Score: 0.9386

BBC News (World) Sun, 11 Nov 2012 13:24:37 BST
RT @BBCBreaking: Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan http://t.co/O2VrmzFn
Score: 0.7342

BBC News (UK) Sun, 11 Nov 2012 13:24:37 BST
RT @BBCBreaking: Israeli forces confirm they fired warning shots into Syria after mortar from Syria exploded in Israeli-occupied Golan http://t.co/O2VrmzFn
Score: 0.4332

Score: 0.3001

Score: 0.2688

Figure 11: WeGov Analysis Components – Top Users to Watch (Cp. left) and Top Posts to Watch (Cp. right)

The WeGov toolbox provides the opportunity to engage with all users that are shown in the top lists. In the WeGov User Interface, the user name is interlinked with Twitter and provides the standard Twitter functionality within a new frame, meaning the WeGov user can directly follow the users, write a tweet to the user, reply to a user's tweet, re-tweet a user's tweet or highlight it as a favourite. The URLs that are embedded within the tweets are also clickable and can be seen within a new frame.

2.3.2. Facebook

We used a different strategy for Facebook than we did for Twitter. Facebook itself provides a good starting point for identifying shared posts and comments for an initial seed post, so is a good place to get reactions to news stories. If one post is shared by a user Facebook provides an icon to show the total number of shares. When the user clicks on the icon, the application shows a list with all the users that shared the seed post with their own network. In addition to the user names the list includes all comments from third party users, who commented on the shared posts.

Identifying Shared Posts

As one concrete example we started with the BBC seed post "*Israel fires 'warning' into Syria*". Figure 8 shows that the seed post was shared 198 times by Facebook users. Hence we manually analysed this to identify all shared posts that generated user comments. Table 1 show that nine of the 198 shared posts' were also commented by friends of the nine users (Cp. *lines 1-9*). Line 0 shows even the seed post. The table shows the user names in the first column and the IDs to identify the shared posts in the second column. The discussions can be retrieved by extending the

[16] Official Twitter account BBC News (World). URL: https://twitter.com/BBCWorld (Retrieved 18 November 2012)

Facebook domain with the post ID. The example for the seed post is https://www.facebook.com/59145437587_297088283728675.

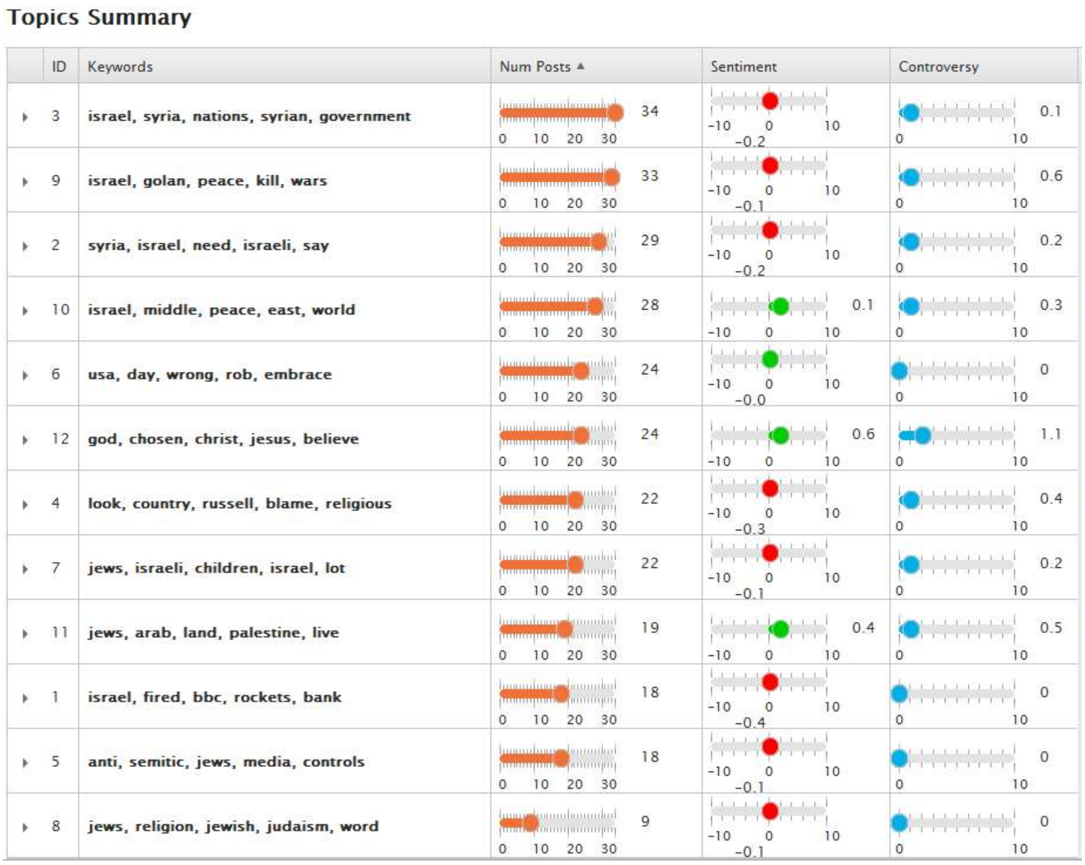
	Facebook User	Post ID	Characteristic of post	Amount of comments
0	bbcworldnews	59145437587_297088283728675	Seed post	361
1	anonymized user 1	100002176611146_405767042828091	Shared post	8
2	anonymized user 2	100002438768018_442412382483272	Shared post	1
3	anonymized user 3	1060137943_130459803772792	Shared post	2
4	anonymized user 4	175698555881824_296466607129144	Shared post	3
5	anonymized user 5	100002367606370_437094809673191	Shared post	1
6	anonymized user 6	515110789_170151173124236	Shared post	5
7	anonymized user 7	100001589968872_508650732493113	Shared post	2
8	anonymized user 8	100001441730180_165574513585108	Shared post	3
9	anonymized user 9	616624506_366871953403288	Shared post	2

Table 1: Facebook Users who shared a seed post and aggregated comments on their shares

Within the next step we used the nine identified seed post IDs (e.g. 100002176611146_405767042828091) as input for the WeGov advanced search. After the toolbox successfully collected the tweets we selected all the searches as input to start the topic analysis.

Topic Opinion Analysis

Topic-opinion analysis is intended to provide quick summaries of the themes in a debate and the opinions expressed by the citizens on the social networks. As an example of this, Figure 12 shows the topic analysis results when the input was multiple sets of responses on Facebook to the BBC news story, sorted by the number of posts.



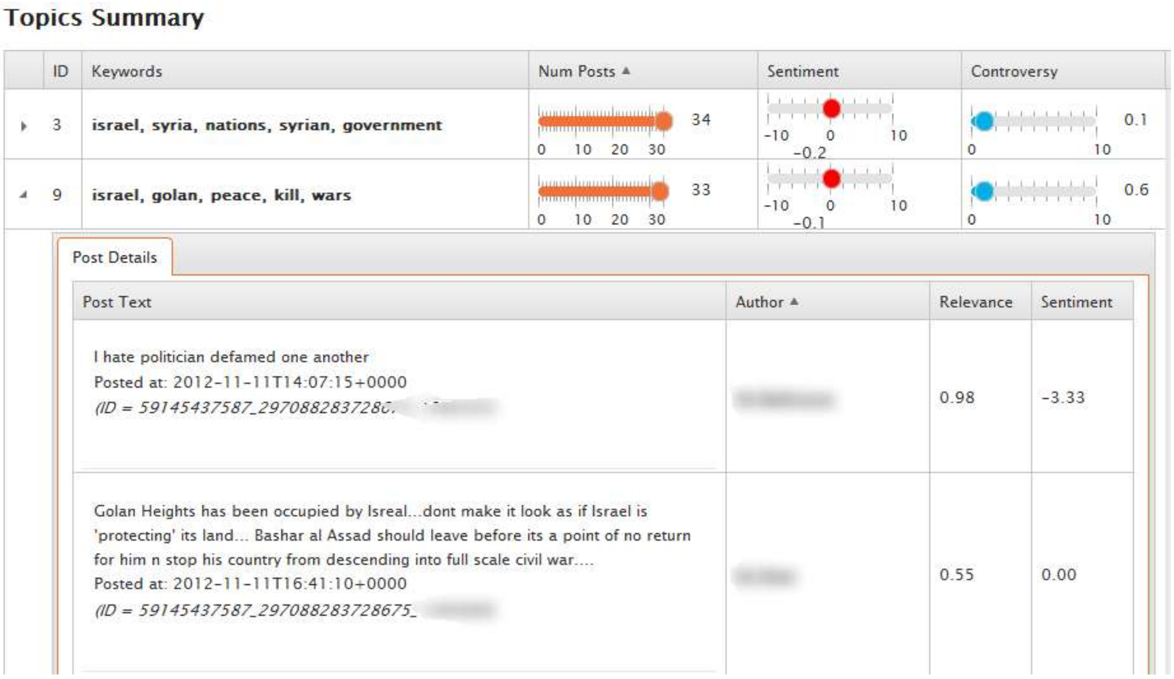
Topics Summary

	ID	Keywords	Num Posts ▲	Sentiment	Controversy
▸	3	israel, syria, nations, syrian, government	34	-0.2	0.1
▸	9	israel, golan, peace, kill, wars	33	-0.1	0.6
▸	2	syria, israel, need, israeli, say	29	-0.2	0.2
▸	10	israel, middle, peace, east, world	28	0.1	0.3
▸	6	usa, day, wrong, rob, embrace	24	-0.0	0
▸	12	god, chosen, christ, jesus, believe	24	0.6	1.1
▸	4	look, country, russell, blame, religious	22	-0.3	0.4
▸	7	jews, israeli, children, israel, lot	22	-0.1	0.2
▸	11	jews, arab, land, palestine, live	19	0.4	0.5
▸	1	israel, fired, bbc, rockets, bank	18	-0.4	0
▸	5	anti, semitic, jews, media, controls	18	-0.1	0
▸	8	jews, religion, jewish, judaism, word	9	-0.1	0

Figure 12: Topic Analysis for Posts and Comments in Table 1

Each line includes a list of five keywords that build the topic (e.g. “israel”, “Syria”, “golan”, “trying”, “rebels”). The next column shows the number of tweets that are sorted to each topic (e.g. 34 tweets for the first topic). The last two columns show the sentiment and controversy of tweets that are measured for each of the twelve topics. The indication of sentiment shows if the tweets that are related to one topic are rather positive, neutral or negative. The indication of controversy shows the ratio of positive and negative posts.

The purpose of the analysis and showing the keywords is to give the user an idea of the posts in each topic group. For example the second post from the top (topic ID 9) contains the words “Israel”, “Golan”, “Peace”, “Kill”, “Wars”, so it is a reasonable assumption that the posts in this topic will concern aggressive activity in the Golan Heights. The user can click on the arrow to the left of the topic ID, and this will expand the topic group to show the posts inside if required, and an example is shown in Figure 13.



Topics Summary

	ID	Keywords	Num Posts ▲	Sentiment	Controversy
▸	3	**israel, syria, nations, syrian, government**	34	-0.2	0.1
◂	9	**israel, golan, peace, kill, wars**	33	-0.1	0.6

Post Details

Post Text	Author ▲	Relevance	Sentiment
I hate politician defamed one another Posted at: 2012-11-11T14:07:15+0000 *(ID = 59145437587_297088283728*[redacted]	[redacted]	0.98	-3.33
Golan Heights has been occupied by Isreal...dont make it look as if Israel is 'protecting' its land... Bashar al Assad should leave before its a point of no return for him n stop his country from descending into full scale civil war.... Posted at: 2012-11-11T16:41:10+0000 *(ID = 59145437587_297088283728675_*[redacted]	[redacted]	0.55	0.00

Figure 13: Post Details

By using the toolkit to summarise the debates, the user can quickly get a feel for the content of the comments people are making, and this helps them to navigate quickly through the deluge of information facing them when they are searching for reactions to the news stories. Once they have the overview of the themes and sentiments, the user can dig into the detail further if they so wish.

3. Conclusions

First this paper has described how we addressed two major questions a governmental policy maker has when they want to determine the reaction to policies or proposals:

- where are citizens discussing the policy or proposal?; and
- what are they saying?

To answer the first question, our strategy was to assume that news articles are written about the policy statements, and these are discussed over the internet. To enable us to find these discussions, we automatically scheduled and repeated Google searches for references to news articles' headlines and URLs. We collected the results in a database, enabling us to aggregate and analyse them to produce ranked tables of sites that reference each news article. Using data mining techniques such as the OLAP cube, we can group data so that the result reflects an overall aggregate score, taking into account multiple datasets, averaging out individual differences. We can also examine the differences between datasets, for example how the sites where the article is discussed change over time.

To answer the second question about comprehending large discussion threads from the internet, this paper has described how the WeGov toolbox, as an analysis environment for social networks, can be used to identify and analyse discussions that refer to one news story. We looked at seed posts on Facebook and Twitter posted by the BBC that refer to a particular BBC news story, and showed different strategies to find user comments in different areas of social networks. After the data collecting process we showed different analysis components that are available throughout the WeGov toolbox to support the policy maker's everyday work. To provide summarisation of the discussion threads, we have used "topic-opinion" analysis to provide appropriate summarization techniques by identifying latent themes of discussion (topics), most relevant contributions and arguments for each topic, as well as identifying the most active users that influenced a certain aspect of discussion. Here we showed a table with ten topics and scales to visualize the number of posts, the value for the sentiment and as well the value for the controversy.

The scenario also provided additional questions, should the policy maker wish to interact with the population discussing their policy. In order to maximise the engagement, the policy maker needs to find out who are the most influential users and comments in the discussion. For this we applied further analyses. Firstly, we used “discussion activity”. This distinguishes those users who are most likely to generate more activity than others. Secondly, we used behaviour analysis. This categorizes users in online communities with the roles they hold in the context of these communities and in a specified timeframe.

Combined or separately, these tools and techniques provide a novel way for governmental policy makers to gauge reaction to their policies, and interact with key members of the population.

References

Addis, M.; Taylor, S.; Fletcher, R.; Wilson, C.; Fallon, F.; Alani, H.; Mutschke, P. & Wandhöfer, T. (2010): “New ways for policy makers to interact with citizens through open social network sites - a report on initial results”. In: H. Margetts & S. Gonzalez-Bailon & S. Ward & D. Sutcliffe (Eds.), Internet, Politics, Policy 2010: an impact assessment, 16-17 Sep 2010, Oxford, UK. Retrieved 20 November 2011, from http://eprints.soton.ac.uk/271073/1/AddisIPP2010PaperWeGov.pdf.

Angeletou, S.; Rowe, M. & Alani, H. (2011): “Modelling and Analysis of User Behaviour”. In: Online Communities, 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany, October 2011. Retrieved 20 November 2011, from http://people.kmi.open.ac.uk/rowe/files/mrowe-iswc2011.pdf.

Hansard Society (2009): “Digital Dialogues Phase Three”. Retrieved 20 November 2011, from http://digitaldialogues.org.uk/reports/digital-dialogues-phase-three/

Chan, Jeffrey; Hayes, Conor & Daly, Elisabeth (2010): “Decomposing discussion forums using common user roles”. In Proc. Web Science Conf. (WebSci10), Raleigh, NC: US, 2010. Retrieved 20 November 2011, from http://www.deri.ie/fileadmin/documents/uimr/jkcchan_websci10.pdf.

Joshi, Somya; Wandhöfer, Timo; Koulolias, Vasilis; Van Eeckhaute, Catherine; Allen, Beccy & Taylor, Steve (2012): “Paradox of Proximity – Trust & Provenance within the context of Social Networks & Policy”. In: Proceedings of The 4th International Conference on Social Informatics, 5–7 December 2012, p. 14. Retrieved 20 November 2011, from http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-642-35385-7.

Joshi, S.; Wandhöfer, T.; Thamm, M.; Mathiak, B. & Van Eeckhaute, C. (2011): “Rethinking Governance via Social Networking: The case of direct vs. indirect stakeholder injection”. In E. Estevez & M. Janssen (Eds.), Proceedings of 5th International Conference on Theory and Practice of electronic Governance. New York New York 10121-0701: ACM PRESS, S. 429. Retrieved 20 November 2011, from http://dl.acm.org/citation.cfm?id=2072117.

Rowe, M.; Fernandez, M.; Alani, H.; Ronen, I.; Hayes, C. & Karnstedt, M. (2012): “Behaviour analysis across different types of Enterprise Online Communities”, Web Science Conference, Evanston, US. Retrieved 20 November 2011, from http://oro.open.ac.uk/33432/1/.

Rowe, M.; Angeletou, S. & Alani, H. (2011 b): “Predicting Discussions on the Social Semantic Web”. In: Proceedings of Extended Semantic Web Conference 2011. Heraklion, Crete. Retrieved 20 November 2011, from http://www.lancs.ac.uk/staff/rowem/files/mrowe-eswc2011.pdf

Rowe, M.; Angeletou, S. & Alani, H. (2011 a): “Anticipating Discussion Activity on Community Forums”. In: The Third IEEE International Conference on Social Computing. Boston, USA. Retrieved 20 November 2011, from http://www.lancs.ac.uk/staff/rowem/files/mrowe-socialcom2011.pdf .

Sizov, S. (2010): “GeoFolk: latent spatial semantics in web 2.0 social media”. In B. Davison & T. Suel & N. Craswell & B. Liu (Eds.), Proceedings of the third ACM international conference on Web search and data mining, February 04-06, 2010, (pp. 281-290). ACM. Retrieved 20 November 2011, from http://www.wsdm-conference.org/2010/proceedings/docs/p281.pdf.

Wandhöfer, Timo; Taylor, Steve; Alani, Harith; Joshi, Somya; Sizov, Sergej; Walland, Paul; Thamm, Mark & Bleier, Arnim; Mutschke, Peter (2012): “Engaging politicians with citizens on social networking sites: the WeGov Toolbox”. In:

International Journal of Electronic Government Research (IJEGR), Vol. 8/ No. 3, p. 22-43. Retrieved 20 November 2011, from http://www.igi-global.com/article/engaging-politicians-citizens-social-networking/70074

Wandhöfer, T.; Thamm, M. & Mutschke, P. (2011): “Extracting a basic use case to let policy makers interact with citizens on Social Networking Sites: a report on initial results”. In P. Parycek & M. Kripp & N. Edelmann (Eds.), Proceedings of the international conference on e-democracy and open government; 5-6 May 2011, Danube University Krems, Austria, Krems: Ed. Donau-Univ. Krems, S. 355-358. Retrieved 20 November 2011, from http://works.bepress.com/cgi/viewcontent.cgi?article=1006&context=timo_wandhoefer.

WeGov (2011): Presentation at the German Parliament. Retrieved 20 November 2011, from URL: http://goo.gl/oqgZu.

Wilson, C. & Fletcher, R. (2010): “Appendix A, Legal Analysis of Issues Surrounding Social Networking Sites”. In: WeGov Deliverable 5.1. Retrieved 20 November 2011, from URL: http://goo.gl/Eu94j.

About the Authors

Timo Wandhöfer

The computer scientist Timo Wandhöfer is affiliated to GESIS – Leibniz Institute for the Social Sciences. His fields of research contain e-participation, Social Web and Information Retrieval. He has now working for GESIS for six months. Up to this point, he was responsible for the work package stakeholder engagement and evaluation of the EU project WeGov. Timo Wandhöfer is in the process of finishing his PhD where he surveys the effects of social network analysis tools, as WeGov, to enrich the dialogue of policy makers with citizens.

Steve Taylor

Steve Taylor is a research engineer at the University of Southampton IT Innovation Centre. He is the technical manager of the IST FP7 WeGov project.

Paul Walland

Paul Walland joined the management team of IT Innovation in 2006 where he is Principal Investigator on several EC and UK collaborative projects including WeGov. His experience ranges from pure and applied research to product development and innovation management in a wide range of UK based commercial organisations. Prior to joining IT Innovation Paul was Projects Group Manager responsible for collaborative project research at Snell & Wilcox Ltd, UK, where he initiated and led a wide range of UK and European research projects in the field of digital video and multimedia. Paul is a member of the Steering Board of the Networked Electronic Media Technology Platform of the EC. He has chaired workshops, conferences and working groups as well as publishing widely at international conferences and in technical journals. Paul graduated from the University of Manchester (UMIST) with a BSc in Pure and Applied Physics.

Ruxandra Geana

Ruxandra Geana is a third year University of Southampton undergraduate studying BSc Computer Science.

Robert Weichselbaum

Robert Weichselbaum M.A. is a student assistant at GESIS – Leibniz Institute for the Social Sciences. He is currently working on his Diploma thesis in economics.

Miriam Fernandez

Dr Miriam Fernandez is currently a research associate at KMi. She received her MSc and PhD from the Universidad Autonoma de Madrid, Spain. Her research is focused on the synergy of Information Retrieval, Semantic Web and Social Web technologies. She has participated in several European projects (aceMedia, Mesh, X-Media, SmartProducts, WeGov, Robust), published in top-level conference proceedings (ECIR, SIGIR. ESWC, ISWC) and journals (TKDE, TCSVT, JWS), and worked for one of the main companies in the search engine and Web development market (Google Zurich).