Opinion Mining in Social Media: Modeling, Simulating, and Visualizing Political Opinion Formation in the Web

Michael Kaschesky
Bern University of Applied Sciences
E-Government Unit
Bern, Switzerland
+41 31 848 3433
ksm1@bfh.ch

Pawel Sobkowicz
Bern University of Applied Sciences
E-Government Unit
Bern, Switzerland
+41 31 848 3440
pawelsobko@gmail.com

Guillaume Bouchard
Xerox Research Center Europe
Data Mining and Machine Learning
Grenoble, France
+41 31 848 3440
guillaume.bouchard@xrce.xerox.com

ABSTRACT

Affordable and ubiquitous online communications (social media) provide the means for flows of ideas and opinions and play an increasing role for the transformation and cohesion of society – yet little is understood about how online opinions emerge, diffuse, and gain momentum. To address this problem, an opinion formation framework based on content analysis of social media and sociophysical system modeling is proposed. Based on prior research and own projects, three building blocks of online opinion tracking and simulation are described: (1) automated topic and opinion detection in real-time, (2) topic and opinion modeling and agent-based simulation, and (3) visualizations of topic and opinion networks. Finally, two application scenarios are presented to illustrate the framework and motivate further research.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing
H.3.1 [**Content Analysis and Indexing**]: Linguistic processing
H.3.3 [**Information Search and Retrieval**]: Information filtering
H.3.4 [**Systems and Software**]: Information networks
H.3.5 [**On-line Information Services**]: Data sharing

General Terms

Management, Measurement, Design, Economics, Reliability, Experimentation, Human Factors, Standardization, Languages, Verification.

Keywords

Management, Measurement, Design, Economics, Reliability, Experimentation, Human Factors, Standardization, Languages.

1. INTRODUCTION

The goal of opinion research is to identify emerging societal trends based on views, dispositions, moods, attitudes and expectations of stakeholder groups or the general public. One major application of opinion research is the area of policymaking in order to better anticipate likely impacts of policy measures and better communicate expected benefits and consequences. Models of opinion formation based on real-world online communication enable the simulation and prediction of the evolution of communication patterns on a specific policy issue within a region or cross-regionally for global comparison. For example, using real-world observations of the communication patterns around the smoking ban in public places may show that resistance against such a directive is waning and support is increasing. The opinion formation model for real-world complex systems – based on early work of the economist and policy advisor Thomas C. Schelling and subsequently refined for online networks – would be able to anticipate the likely further evolution and tipping point, after which a large majority may support a policy initiative.

The democratization of web publishing has led to the explosion of the number of opinions expressed over the internet. At the same time, citizens are becoming more actively engaged in policy issues, more empowered, and more demanding in their relations with traditional institutions while political clubs, organizations, and editorials experience falling memberships [9].

Researchers identify a 'hunger' for and reliance upon peer advice and recommendations found online and this information hunger is strongly evident in the political sphere. For example, through a large-scale survey, researchers were able to infer the motivations of over 60 million U.S. citizens who gathered online information about the 2006 elections and exchanged their views [21]. For one third of these citizens, the motivation to engage online was to get perspectives from inside their community, while another third was motivated by getting perspectives from outside their communities. Another third was motivated by other citizens' endorsements or ratings. The ability to trace the evolution of online conversations has been employed in few longitudinal studies [15]. The political sphere appears particularly suited for investigating opinion-formation in the blogosphere, because "blogging as democratic practice" is inherently linked to the broader policy processes [7].

Affordable and ubiquitous information and communication technologies (ICT) promote the exchange of ideas and opinion across borders. Driving the structural transformation are information flows connecting individual ideas and opinions with others thereby creating the networked society [3]. Arguably, the ICT-enabled flows of ideas and opinions play a fundamental role for the transformation and cohesion of the information society –

Dg.o'11, June 12–15, 2011, College Park, MD, USA.

yet little is understood about how online opinions emerge, diffuse, and gain momentum. To address this problem, we follow an overarching research question:

In what ways can online content from various social networking resources be exploited to inform decision makers about constituent opinions, emerging trends, as well as feasibility and potential impacts of new initiatives?

In the following, we address this question by presenting three interlinked components for opinion mining, simulation, and visualization as well as briefly describing the underlying distributed processing architecture:

- *Social media content analysis*: a large set of online forums, blogs or other publicly available text streams are tracked and analyzed. Text understanding algorithms extract semantic information related to the topics targeted by the decision maker. In particular, the social network of individuals expressing their opinion online is reconstructed and for every analyzed text, the main subtopics are identified, as well as the associated sentiment (positive/negative opinions);
- *Opinion formation modeling, simulation and prediction*: An opinion diffusion model is estimated on the extracted data to recover the graph of influence and model current and future opinions trends. Every opinion is represented by a concept (or sub-topic) and a diffusion rate, and individuals are represented by interests, influence and disposition of being influenced;
- *Network visualization and interface design*: The results of opinion mining are presented to decision makers in an intuitive and customized dashboard. Implicit and explicit feedback is used to improve the accuracy of opinion mining and adapt to the user's topics of interest.
- *System architecture & run-time platform*: The underlying system architecture is based on semantic complex event processing in a cloud environment capturing different levels of information (such as event data, i.e., new content) as well as associations between them created during the opinion mining and sentient analysis process.

The next three sections describe these components in more details. Section Five then presents two application scenarios that naturally fit to the proposed opinion mining framework. Finally, we discuss the practical implementation choices and future research directions.

2. SOCIAL MEDIA CONTENT ANALYSIS

Topic and opinion detection in online content facilitates the identification of emerging societal trends and analysis of public reactions to policies. The next step beyond current web search is to rank information entities of varying type, complexity, and structure, rather than document-only (e.g. web pages). Being able to retrieve specific entities rather than whole documents allows building innovative applications for topic and opinion detection (e.g. extracting comments). These possibilities are made possible due to the proliferation of Semantic Web standards and methods, rise of machine learning methods in natural language processing, availability of datasets for machine learning algorithms to be trained on, and the spread of review-aggregation websites and user-rated content. Topic and opinion detection provides a fast and reliable way of transforming a set of unlabeled documents into a well-structured knowledgebase. There are two approaches, which currently develop rather unrelated to each other:

- *Natural language processing (NLP)*: implicit representation of meaning, based on a vector representation of texts and meaning, which enables the definition of similarities between texts and degrees of positive or negative opinions. The outcome of such models is accurate but difficult to interpret.
- *Semantic web approaches (SW)*: explicit representation of the domain based on semantic annotations that map a text to the domain ontology via keywords or tags. There are few large scale examples of efficient reasoning based on this approach.

Today, there are few hybrid systems combining the strengths of both approaches. We present an approach based on a robust method using the implicit representation of meaning (NLP) and extending it using domain ontologies (SW) to improve performance and allow more fine grained analysis of opinions.

2.1 Topic and Relationship Recognition

Topic recognition and classification concerns the process of parsing textual information (e.g. blog entries, comments, forum discussions) and deducing whether the parsed content belongs to an existing topic category or constitutes a new one. The general principle is mapping a given piece of text, such as a document, paragraph, or sentence, to one or more labels representing abstract concepts. Topics are constituted by a hierarchy of related events, that is, contributions regarding specific aspects of the (broader) topic or regarding the evolution of the topic, subtopics, or very specific concepts at the lowest level of the hierarchy.

Relationship recognition between topics becomes critical when the texts to be analyzed form part of a running discussion, such as in posts to online discussion boards and comments on blog posts. The context of the posts forms a rich information source based on references between the posts and such information can be exploited for better topic and relationship recognition of the texts. Recognition of relationships on the level of identified (anonymized) users should lead to reconstruction of social networks, which provide complementary information for constructing topic relationships.

2.2 Opinion Detection and Sentiment Analysis

The focus is on the automatic identification and extraction of opinions, emotions, and sentiments from text and multimedia [4]. Motivation for this component is based on providing support for decision makers to automatically track attitudes and moods in online media and user generated content [29]. For example, opinion detection and sentiment analysis has been proposed as a key enabling technology in eRulemaking, allowing the automatic analysis of the opinions that people submit about pending policy or government-regulation proposals [1][14][22].

The goal of opinion mining is to create a knowledgebase containing online opinions in a more structured and explicit form. The data is processed by a NLP engine based on a syntax analyzer and machine learning technique that detect which part of the sentence correspond to the expression of an opinion, and on which specific topic. For each text, the identified opinion is represented as a list of pairs (rhetorical concept, keyword) mentioned in the text. The rhetorical concept is defined a priori by linguists. To start with, the vocabulary will be simplified into four categories, such as 'positive opinion', 'neutral opinion', 'negative opinion' and 'information' (e.g. fact-like information such as quality news).

The representation of the domain knowledge will be centered around named entities, which are typically at the core of the online discussions. The knowledgebase for a specific domain would contain for each entity a list of relations between them (e.g. represented as RDF triplets), a list of attributes (e.g. name, age, location), as well as a list of entities that may be addressed in online discussions:

- major events (e.g. Deepwater Horizon oil spill);
- known people (e.g. Obama) or groups (e.g. fishermen);
- known organizations or concepts (e.g. BP, pollution, oil spill);
- locations (e.g. Houston, Keathley Canyon).

In addition, the knowledgebase for the domain contains the major aspects of domain knowledge, which is required to relate similar opinions. For example, the extracted pairs ("support", "government") and ("agree", "Obama") will both be counted as positive opinions when aggregating data to measure public opinion about the government. For example, the RDF triplet ("Obama" – "is part of" – "government") in the knowledgebase may be used to improve the accuracy of measuring opinions on the current U.S. government. The knowledgebase containing entities, relations and attributes can be implemented using a semantic web ontology language, but a traditional relational database could also be used in this context. The goal is not to do complex reasoning within the database, but to store information in a format that can be directly matched for the topic and opinion modeler and simulator (see next component).

The main challenges addressed are in the following areas [19]:

- Genre recognition (e.g. product reviews, political opinions);
- Content relevancy (e.g. determining which content is topically relevant to an opinion-oriented query);
- Sentiment identification, i.e. overall sentiment expressed and/or the specific opinions regarding particular features or aspects of the items or topics in question (the more pre-structured, the easier the identification);
- Sentiment aggregation and visualisation, i.e. aggregation of "votes" that may be registered on different scales (e.g., stars, numbers), selective highlighting of opinions, representation of points of disagreement and points of consensus, identification of communities of opinion holders.

3. OPINION FORMATION FRAMEWORK

Recent years have brought significant interest in interdisciplinary studies, combining tools and methods known from physics with social analyses. These studies are often referred to as sociophysics, and range from purely numerical studies of economic trends to descriptions of social activities. Among the latter, a significant role is played by computational models of opinion formation. Such models often combine results derived from statistical physics with agent based simulations. Within a simplified framework, focusing on a few selected aspects of social activities (such as communication network, susceptibility to influences, contrariness etc.), it is possible to derive general trends of behavior of large societal groups, starting from individual perspectives (similar to statistical, kinetic theory of matter).

One of the major problems with 'social physics' or sociophysics research on opinion modeling is the lack of connection to real-life examples and data. Recent works reiterate the need for a real-life connection by emphasizing real-life evidence over conceptual models and theory [16] as well as prediction and explanation based on real data for opinion modeling and observation [6],. The cost and difficulty of making real social analysis using large-scale observations and real data is enormous. The task in this component is a challenge and innovation, using real data for large-scale social observations, analyzing its past evolution, and simulating potential future developments. This work ties in observational data and requires a truly multidisciplinary way of conducting research to adapt the modeling approach better to realistic conditions. Opinion formation is a branch of sociophysics that would greatly benefit from a few modifications through integration of real-life data.

3.1 Opinion Formation Modeling

Models of opinion formation typically employ three stages, the initial state, the alert state, and the percolated state. Such models draw on the works of Watts [27][28] in sociology and of Payne, Dodds, and Eppstein [20] in physics and is here applied to the area of opinion cascading [11].

The model of opinion formation proceeds from an *initial state* in which the vast majority of participants ignores government action (or inaction) on a potentially important issue. These participants are inactive because they initially see no need to take action. But when confronted with prominent new information on government action, they make a approving or disapproving decision. A small minority of participants already believes that government action warrants debate, that is, they are active in either approving or disapproving government action. Such minority players may differ in their social influence and connectivity, which, in turn, may determine the effectiveness of their actions. Such differences may result from authority, communications skills and capacity to convince and gather followers as well as social connectivity.

The arrival of prominent new information changes the initial to an *alert state*. If the information is perceived as supporting the opinion that approves of government action, then a proportion of participants decides to approve and, at least some of them, to publicly promote their opinion while disparaging opposite opinions. With a higher proportion of participants now approving government action on an issue, the number of communication partners adopting a positive opinion rises and the interactions between them – due to increased information-seeking on the issue – become more intense.

The arrival of more new information changes the alert state either in the direction of the initial state or towards a *percolated state*. If the information is perceived as supporting an assessment opposite to the one in the alert state, then a proportion of participants decides to engage in the opposite activities resulting in an approval score opposite to the earlier direction. If, however, the new information is perceived as reinforcing the earlier assessment, then the corresponding activities further reinforce the earlier direction. In this case, the assessments of the earlier adopters are proven accurate whereas the vast majority of participants as well as the active participants promoting the opposite opinion are disproven. These reinforcements create a tipping point which set the system dynamics into motion: With an increasing number of communication partners now adopting the opinion and increasingly more intensive interactions between them, thresholds of other participants for adopting the opinion are more quickly reached and the opinion as well as the corresponding decisions and actions percolate.

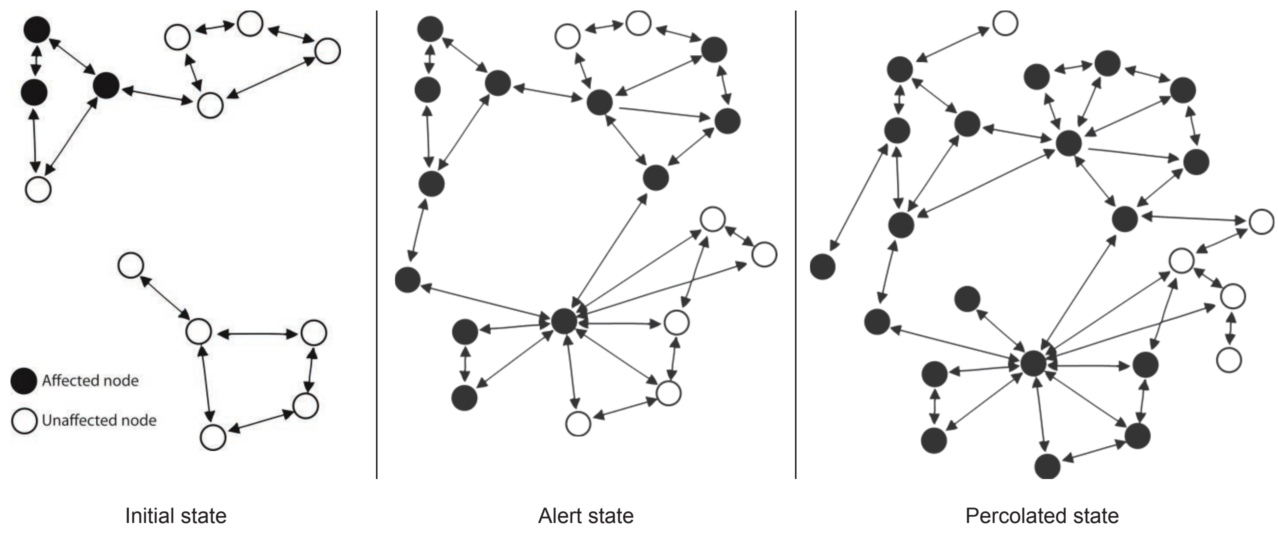

Figure 1: Simple model of opinion formation

The model takes full account of the reflexivity of agents, because participants adjust their behavior in cognizance and reflection of other agents' behavior and the arrival of new information. The reflexivity of agents who approve and disapprove government action can be modified and extended as follows: In addition to enhancing the links with those of the same opinions, participants may sever the connections with those of opposing ones. This dynamics methodology further intensifies the reinforcements in the communication network and changes the initial network characteristics, in extreme cases, towards a dominant mainstream and quasi isolation of minority opinion holders.

Figure 1 above exhibits the three states or stages of the model showing the proliferation of a new opinion (filled nodes represent affected nodes) to neighboring unaffected nodes. The double-headed arrows indicate dyadic linking behavior, that is, the adoption of new opinions constitutes an interactive process involving the active engagement of both participants. The adoption or rejection of opinions is a dynamic and reflexive process. Dynamic, because reinforcements lead to a tipping point after which dominant opinions become mainstream and may lead to quasi isolation of minorities (white nodes). The process involves reflexive agent behavior, because agents adopt or reject opinions in cognizance and reflection of other agents' behavior and the arrival of new information.

3.2 Opinion Formation Simulation

Opinion modeling based on sociophysics typically focuses on global properties of the modeled system [18][26][5][8][10]. However, to be able to provide practical insights and real-life tools for social monitoring and policy formulation, models must include factors absent from simplified viewpoints. An important way of extending such an approach is via agent-based simulations, which allow integration of agent descriptions that are more detailed on the micro-level of individual behavior and therefore enable the combination of observations across several levels, from the individual micro-levels to the aggregated macro-levels. Such non-classical socio-economic modeling goes beyond simplified economic models because it takes into account several and multi-faceted characteristics of individuals, rather than one monolithic characteristic (e.g. utility maximization).

The simulations presented here allow computer agents to cut the social links with those they disagree with and form a new links with agents sharing the same opinion. This changes the network structure. To take into account the fact that in real societies some links cannot be broken (e.g. family or work relationships) we have simulated situations where certain percentage of the links remain static, while the rest are free, allowing changes in topology of the social network.

The first novel aspect of the simulations is direct inclusion of agents with no preferred opinion (neutral agents). This allows significant change from models where only committed agents are present, changing both the social network and opinion change dynamics. Appropriate real life examples of application of the simulation include political preferences and highly controversial opinions on topics such as abortion. Within the model, the strength of influence between agents decreases with their social separation, reflecting the fact that our opinions are swayed less by remote acquaintances or strangers than by the closest associates.

Secondly, the opinion of a given agent may be changed in reaction to perceived cumulative social opinion of others, corresponding to a properly averaged 'peer pressure' rather than the individual encounters. Many of the classical models have stressed the importance of such individual contacts in opinion changes of agents, but the constant background of perceived opinions, resulting from numerous encounters and information on opinions held by other members is relevant. In a way, this can be described as each agent continuously measuring and responding to the 'discomfort' due to difference between own opinion and properly averaged opinions of other agents. It also becomes possible to simulate propagandist efforts via simple parameterized factor of external influence.

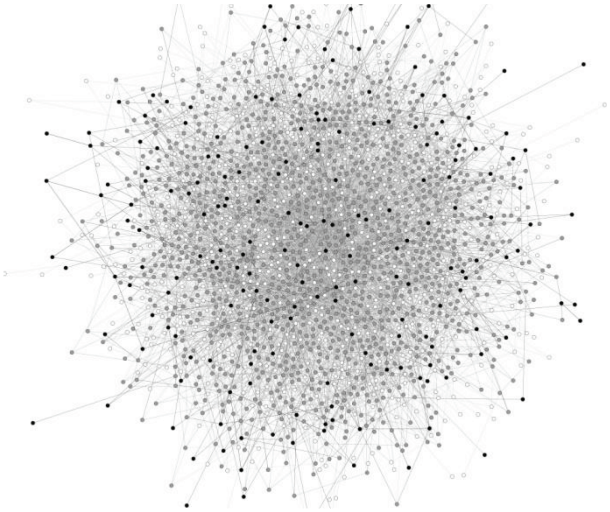

Figure 2: Initial state with large majority undecided (grey), proponents (light grey) and opponents (black)

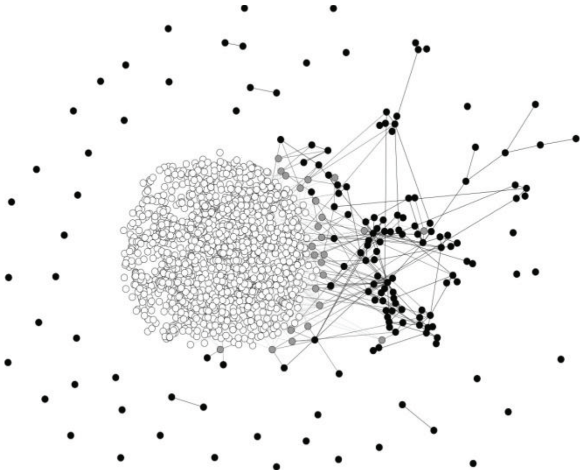

Figure 3: Final state of simulated opinion diffusion with majority having adopted the proponents' position (light grey)

Figure 2 and Figure 3 show examples of changes in simulated opinion distribution and social network topology leading to opinion diffusion and network rearrangement. While the predominant opinion reaches an increasing number of previously undecided participants, minorities may remain untouched, separated from the main part of society. Figure 2 shows the initial social state with a large majority of undecided participants without a specified opinion on the issue (grey), and smaller groups of proponents and opponents concerning the issue (respectively light grey and black). The social links enabling information flow and opinion influences exist between all groups. Figure 3 shows the final state of simulated opinion diffusion with most of the participants having adopted the proponents' position. However, small minorities persist at the outskirts of society, mainly because they have severed and cut most of the information links with the majority [24]. Their existence and vitality can be monitored for example, via internet media activity, especially when filtering for high negative emotions. This enables policymakers to adjust implementation to better integrate dissenters and communication to better target expected benefits and consequences, by monitoring the effects of activities of individual persons and leaders [23].

3.3 Opinion Formation Forecasting

Predictions within complex systems, such as communication networks, are possible thanks to statistical learning techniques through three distinct modeling steps. First, a model for the opinion diffusion network model is assumed. In a Bayesian framework, this corresponds to the definition of an a priori probability density function. The specific form of this prior distribution (i.e. the hyper-parameters) is taken from past observations, from experiences in Monte-Carlo simulations of the communication network and from the literature. Second, new observational data come with its own uncertainties (due to noise, partial information and error from the opinion mining system). There uncertainties can also be represented by a probability distribution. Third, the observations and the prior opinion diffusion distribution are coupled together to update the distribution and make a new network model consistent with the new data while following the prior assumption. This updated distribution is often called the *a posteriori* distribution in the Bayesian setting. The new network model is used to make predictions and is a key component for interpretation: it defines the way the latent opinion diffusion process is mapped to the observations. But since in practice no model is completely reliable, the uncertainty of the estimation has to be included in the analysis and the visualization of the results. Uncertainty measures are a natural byproduct of many inference algorithms and will be estimated and visualized in the dashboard component used to interpret the results (see next component).

4. NETWORK VISUALIZATIONS

Impact analysis of topics and opinions is provided by visualization and simulation techniques on texts (e.g. issues, topics, opinions) and on the network topology (e.g. groups/factions, relationships, evolution over time). We present a selection of text and network visualization techniques which are used for rapid data interpretation and analyses.

4.1 Text Visualization Techniques

Visualizing text. Tag clouds visualize text by laying out words from the text corpus in a visual space and encoding data about the relative frequency, popularity, and preference of each term using graphical attributes like color, size, and weight. However, standard tag clouds do not use spatial dimensions. The Wordle tag cloud technique enables tag clouds that make more efficient use of display space by packing terms together.

Visualizing textual relations. Arc diagrams show repetition in string data, used for text documents. For example, FeatureLens supports visual exploration of frequent text patterns in document collections. DocuBurst uses the existing Word-Net ontology to group similar words into a space-filling radial hierarchy. WordTree visualizes relations within a document on a per-word level, constructing an interactive hierarchy of the context of a word. Parallel Tag Clouds are a tag cloud technique supporting faceted browsing of text corpora, allowing comparison between documents and parts of documents.

The WordBridge is a novel graph-based visualization technique, based on GreenArrow for exposing not only entities in a large document collection, but also their relationships beyond mere co-occurrence information. It works by forming a bridge of words that connects one entity to the other using composite tag clouds in the shape of nodes and links [13].

4.2 Network Visualization Techniques

First, the task is to visualize the relevant issue-specific communication network for identifying key arguments and positions. This enables policymakers to take into account the impact that proposed policy measures may have on different groups who are affected by the policy. In this way, the tools support policymakers in understanding how the policy may actual impact various stakeholders and to see its implementation from the citizen's point of view, thereby minimizing the likelihood of unintended consequences and strengthen legitimacy of measures.

Second, the task is to visualize the likely future evolution of the communication network in order to detect dynamics leading to integration or radicalization of different opinions as shown in Figure 2 and Figure 3 above. This enables policymakers to predict extremism in minority opinions and to devise policy implementation and communication to better address minority opinions or specific political views of their constituents.

Third, dynamic analysis provides an interactive environment that allows end users to play (mode slider) with the value of each of the tracked criteria an indicators to monitor the possible result to be achieved. For example, a confusion matrix can be used to model the feasibility of new policy initiatives by introducing a free value in terms of cost effort (in a cost matrix) of implementation or communication measures necessary to target specific groups with specific messages to maximize understanding of a policy measure (e.g. better policy understanding in two different citizen clusters).

There are several possibilities to change modeling parameters (such as topic, region, time, social group) to incorporate relevant input data and to simulate for example:

- Views and ideas of people, filtered by supportive versus opposing groups;
- Emerging opinion trends (which social groups are supporting/opposing issues);
- Acceptance or rejection of policy initiatives (which social group will accept/reject the policy)
- How topics and policy proposals are connected to constituents' opinions
- How issues and initiatives are linked to other public sector data (e.g. statistics, environment)
- Communication measures Plan (e.g. target group specific focus on expected benefits)
- Impact assessment (e.g. how are target groups responding to policy implementation)

5. System Architecture & Run-Time Platform

In opinion mining massive amounts of content is gathered from various heterogeneous, distributed sources, such as online discussion forums, blogs, news sites or social network sites. Each of these sources provides continuously new content which needs to be processed in a reasonable time frame to enable opinion mining and sentiment analysis in near real-time. The vast amount of content that needs to be gathered has to be condensed to reduce the amount of separate information pieces that have to be processed to allow the overall system to perform extended analysis with a reasonable amount of resources. Effective pre-processing of gathered content is thus required to reduce the load on the following processing steps and to store only relevant pieces of information in databases and event histories. In addition, content gathering and processing must be able to scale in case of a general increase of published content, for example, before and after major events such as elections or natural disasters.

To provide the required level of abstraction form the underlying technologies and architectural principles, we introduce a runtime container that allows the access to the underlying systems in a unified manner. The container will provide the means necessary for the event based communication with other components and the service based access to central information sources (e.g. the domain specific model / the data repository for foreign data sources). Furthermore the runtime container provides a strong encapsulation of the components with clear interfaces to the other communication parties. Hence, the event processing components can be deployed dynamically within those containers where the containers themselves can are distributed across several locations, i.e. data centers. The results from this stage are then fed into the simulation and visualization engines of the system.

We propose the described container concept as a platform for a staged processing approach. This platform provides the required environment for components of opinion mining and sentiment analysis. It can further provide the means for the communication between these components and to access central model data to enable loosely coupled integration with other components.

6. Application Scenarios

The goal is to provide a decision environment that presents the main historical and current developments regarding topics and opinions as well as trends of constituents' opinion in temporal and spatial (i.e. regional) contexts and the likely future evolution of the relevant communication networks in an intuitive and easy exercisable way. Using geospatial distributions of analytical results, decision makers understand the topics and opinions of different local, regional or global stakeholders based on their past sentiments towards a policy issue. The ultimate goal is to learn from unexpected reactions and the evolution of general, minority or viral opinions to bring forward accurate decisions and maximize the likelihood of intended consequences. The following paragraphs introduce some illustrative examples to demonstrate possible application scenarios.

6.1 Governance of Java Standard

The role of online opinions in the governance of the Java software standard illustrates the link between online opinion diffusion and its impact on policy making (in this case decisions of the Java Governing Board on opensourcing Java). Decision-making on the governance of the Java standard Java thus serves as an application scenario [12]. Decision-making on Java governance used to be a closed-book exercise involving the largest players. Small software firms and individual developers had to accept what the Java Governing Board decided. But this community with high internet-affinity informed and communicated via online media and forums to address this issue and request changes leading to opensource the Java standard.

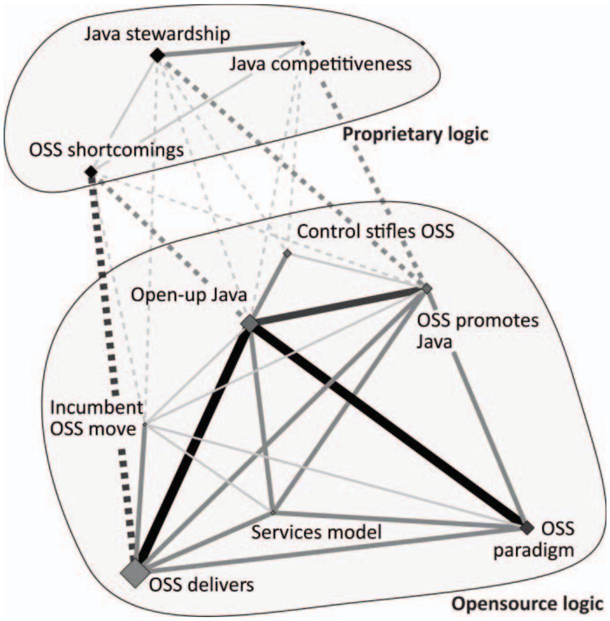

Figure 4: Topics, centrality, momentum and cross-references of important issues in Phase 1

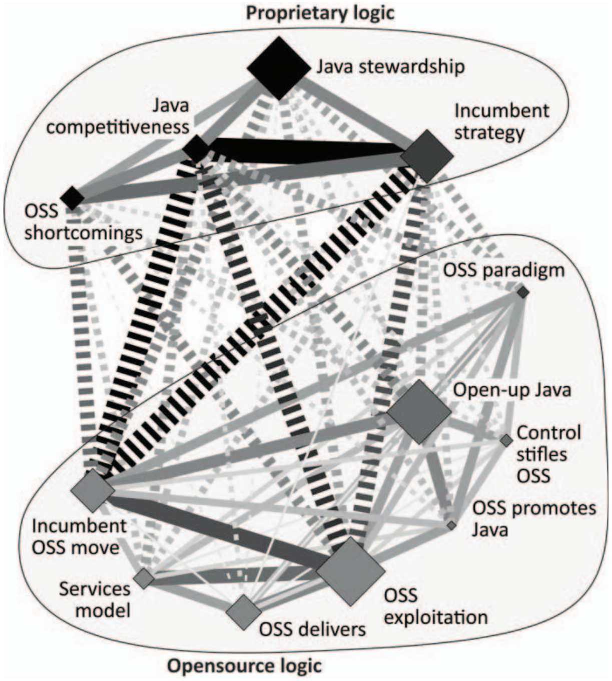

Figure 5: Topics, centrality, momentum and cross-references of important issues in Phase 2

Figure 4 and Figure 5 each present the two opposing communication networks on the issue at different points in time (Opensource logic vs. Proprietary logic). One communication pattern is around the 'proprietary logic' while the other is promoting the 'opensource logic'. On the left side, the communication network in 2002 is depicted, showing that Java opensource software was an issue that attracted some interest. On the right side, the same communication network is depicted in 2004, showing a massive increase in interest and engagement.

The proposed approach in this paper goes beyond these observations in three directions: 1) by tracking opinion formation in real-time, 2) by simulating the evolution of the communication networks (e.g. emergence of isolated minority opinions), and 3) by predicting its future evolution based on past observations and statistical learning techniques.

6.2 BP oil spill

Policymaking after the Deepwater Horizon oil spill serves as another example for an application scenario [11]. On 20 April 2010, an explosion on the oil rig caused by a blowout killed 11 crewmen and caused the second largest oil spill in history. Beside disaster relief operations, policymakers were reviewing the regulatory regime for oil exploration, the existing liability and compensation framework, the technological challenges involved with deepwater activities, and medium-term response activities (e.g. relief aid, use of chemical dispersants). At the same time, the general public accused BP and the government of inaction thereby asserting heavy pressure on policymakers to act swiftly. In addition, affected local citizens required medium-term help and support to cope with the consequences yet little was known to policymakers about the myriad of local problems that were caused by the oil spill.

Topic and opinion detection are illustrated using the example application scenario of the BP oil spill. Data collection retrieves online content related to the Deepwater Horizon oil spill. The field's boundaries are set so as to include all participants who exert some effect on opinion formation in the field.

Table 1: Top focal blogs related to U.S. politics

#	↑↓	Focal blog	Most recent post
1	→	Hot Air	Quotes of the day
2	→	CNN Political Ticker	Congressman involved in on-camera confrontation
3	→	Think Progress	Rep. Broun says CLEAN ENERGY LEGISLATION …
4	↓	Political Punch	BP Emails Show Disregard for 'NIGHTMARE WELL'
…		…	…
12	↓	RedState	TN State Rep: You have to lift …
13	↓	TPMMuckraker	Gov't GEOLOGIST Spoke Of Vast Economic …
14	↓	Power Line	Speaking of Gangster Government
…		…	…
18	↑	**Greenpeace Campaign Blog**	**Deepwater Horizon disaster and OIL SPILL will impact …**

For this example, the popular Technorati blog search engine is used for retrieving blog popularity rankings (arrows indicate popularity changes). The table to the left presents the 16 focal blogs representing the top 1 percent of all blogs related to U.S. Politics with an authority index above 1. Included in the sample are influentials or focal blogs, for example, the top 10 percent of blogs who maintain on average more connections to other blogs than do the remaining 90 percent. In addition, government news and publications and political news services such as Associated Press and Reuters are included for triangulation.

Topic detection may identify topics on the oil spill, such as 'Clean Energy Legislation', 'Nightmare Well', or 'oil spill' (uppercase words in Table 1). Opinion detection will then be able to analyses the content according to whether the topics are associated with primarily positive or negative opinions focusing on a specific region or the general public.

Let's take the post on the Greenpeace Campaign Blog is used to illustrate opinion detection and sentiment analysis. In this context, the accumulation of words such as Tragedy, Accident, Pay the price, Damage (in uppercase below) signify a negative sentiment.

Table 2: Illustration of opinion detection and sentiment analysis based on post from Greenpeace Campaign Blog

> The TRAGEDY we're witnessing right now is but the latest in a long line of OIL SPILLS, be they from pipelines, tankers, or exploratory drill rigs like the DEEPWATER HORIZON. Each ACCIDENT brings CONGRESSIONAL INQUIRIES, finger pointing, scathing editorials and PUBLIC OUTRAGE, yet we as a nation are no closer to weaning ourselves from oil than we were after any other big oil spill. So long as we remain dependent on oil we will continue to PAY THE PRICE IN HUMAN LIVES, as well as in ENVIRONMENTAL AND ECONOMIC DAMAGE.

7. Conclusion, Implications and Validation

7.1 Conclusion

Coinciding with the advent of the 'social web', citizens are becoming more actively engaged in policy issues, more empowered, and more demanding in their relations with traditional institutions while political clubs and organizations experience falling memberships. At the same time, citizens are becoming more interested in the political views that other citizens express online. However, little use is made of user-generated content concerning policy issues. Sourcing a wide range of views and concerns, which is being made possible by the proliferation of user-generated content across the web, enhances the effectiveness of policymaking by providing insights that are typically difficult to obtain, such as hidden costs and risks, likely winners and losers, or differing cultural perspectives.

The prominence of the 'social web' and of user-generated content online has created a new situation for the interaction between policymakers and citizens. Previously, there weren't many indicators of citizen opinions available except for sporadic surveys, making precise assessments of the policy impact on constituents' life almost impossible and, consequently, inhibiting the possibility to react swiftly to emerging societal challenges. What most people felt and thought about policy measures and how this influenced their opinions and subsequent decisions was inaccessible – a policymaking black box.

Rather, online or offline surveys and consultations are undertaken at great costs and expenditure of time while highly valuable qualitative information on potential benefits and consequences is often available online, particularly regarding controversial issues that attract wide interest. When implemented, the proposed opinion mining approach allows valuable ideas and discussions to be collected and analyzed. In this way, it enables citizens to voice their views and concerns in the format and the websites they prefer while ensuring that these inputs do not disappear without a trace. In this way, it supports political interest and engagement of citizens in two ways:

- By providing highly-usable, graphical web-based and mobile applications that inform citizens about political issues and perspectives inside and outside their regions and about other citizens' endorsements or ratings of political issues. The visualization tools provide such information based on intelligently extracted, classified, and privacy-preserving content from many internet sources.
- Because the multi-platform applications aggregate and summarize content from a wide range of internet sources, they provide long-term incentives for citizens to engage online on political issues thereby having their topics and opinions recognized in regional and cross-regional aggregations and summarizations. The components must address privacy issues and trustworthiness of sources in order to arrive at privacy-compliant and trustworthy aggregations and summarizations.

In addition, the proposed approach supports inclusiveness in policymaking by ensuring that policymakers take more comprehensively into account the impact that proposed policy measures may have on different groups who are affected by the policy, such as businesses, families, older people, ethnic minorities etc. In addition, it enables policymakers to extend their understanding of how the policy may actual impact various stakeholders and to see its implementation from the citizen's point of view, thereby minimizing the likelihood of unintended consequences and strengthen the legitimacy of policy measures.

When implemented, the proposed approach puts users – policymakers and citizens – in the position to track how policy topics and issues as well as institutions are viewed and judged by various publics. With such detailed knowledge about what stakeholders perceive and are interested in, users can inform the views and actions to take advantage of that perception. With opinion simulation, users can even assess the future evolution of communication networks on specific issues and investigate whether a dominant mainstream opinion will evolve around an issue and whether this leads to isolated minority opinions that require better integration.

7.2 Practical Implications

Policy impact assessment benefits greatly from an integrated and ICT-enabled approach taking into account the issues and concerns raised by citizens and businesses. At present, policymakers carry out separate policy impact assessments for affected areas, such as businesses or health as well as for particular groups, such as women, older people, or ethnic minorities. These assessments are typically done in isolation and are subjected to long and time-consuming processes. The proposed opinion mining approach supports more effective policy implementation and better identification of benefits and consequences in two ways: (1) Sourcing and integrating expert and lay stakeholder views and opinions regarding the impact of a policy measure and (2) predicting the evolution of constituents' opinions to better adjust policy implementation and communication.

Regarding the sourcing of expert and lay stakeholder views and opinions, the approach promotes more effective policy assessments using views and expectations expressed by citizens within a region. This may involve incorporating data derived from official online or offline (imported) consultations, from public online discussions, and/or from personal blogs and comments which are ranked and labeled according to their provenance. In addition, evidence-based expert opinions and recommendations derived from policy analysis and forecasting models can also be included and ranked by provenance. This creates an integrated platform and toolset for comprehensive and systematic assessment of impacts of policy measures which can take data ranging from specific consultations to user-generated content within the region or cross-regionally for comparison.

Regarding the prediction of the evolution of constituents' opinion, the proposed opinion mining approach supports more effective policy implementation and communication through early recognition of problems and opposition possibly leading to extreme positions. When implemented, the approach supports policymakers in assessing and anticipating potential policy impacts on public opinion throughout the policy cycle:

- *Agenda setting*: opinion tracking provides policymakers with issue-specific, policy-focused, on-topic perspectives and sentiments about a concrete problem that requires policy action. In this way, policymakers are better able to understand the pros and cons as well as the expected benefits and consequences voiced by citizens regarding the problem.
- *Policy formulation*: opinion forecasting enables policymakers to assess and anticipate the sentiment and likely impact of proposed policy measures on constituents' opinion within a region. The opinion simulation not only performs sentiment mapping for congruence between constituents' opinion on the specific issue and the corresponding policy action. It also predicts how constituents' opinion may evolve further taking into account past evolution of sentiment on the issue and the new policy action. In this way, it enables policymakers to better anticipate the impact of proposed policy measures on constituents' opinion and adapt policy implementation and communication.
- *Implementation and evaluation*: opinion tracking recognizes issue-specific and policy-focused arguments and sentiments of opponents and proponents about a concrete problem while opinion simulation analyses the actual impact of policy measures on constituents' opinion and predicts its further evolution. In this way, it enables policymakers to recognize and respond to the root cause and better communicate expected benefits and consequences of the policy.

7.3 Practical Validation

The more policymakers innovate, the less certainty they have about achieving intended results and the greater the need to assess policy impacts and be prepared to change tack.

In order to become useful, policymakers and interested citizens implementing the proposed approach must first trust that analytical results of opinion tracking, mining, and simulation are indeed valid. This requires the ability to compare and triangulate analytical results with results observed from other sources and 'facts'. Therefore the analytical results should be compared with surveys and polls and in user testing documenting the changes in social reactions to specific topics and policy implementations. Such comparisons extend in two directions:

- temporal, predicting the evolution of social trends; and
- representational, predicting behavior of large social groups based on data from smaller samples.

Feedback from such test cases allows the opinion tracking and simulation to become more reliable over time, and provide direct measurement as to the quality of the analysis, modeling and programming involved in the project.

If real change is to be achieved and sustained, it must impact on day-to-day policy work being done in agencies and government. The potential major stakeholders must be identified upfront and asked to provide feedback throughout the project. They include, for example, international, national and regional policymakers, national and regional governments, ministries, municipalities, environmental organizations, health institutions, and civic service organizations.

8. REFERENCES

[1] Cardie, C. et al. Using natural language processing to improve eRulemaking. Proceedings of dg.o, 2006.

[2] Castellano, C. Fortunato, S. and Loreto, V. . Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, 2009.

[3] Castells, M. The rise of the network society. Oxford: Blackwell Publishing, 1996.

[4] Chesley, P. et al. Using verbs and adjectives to automatically classify blog sentiment. AAAI-CAAW, 2006.

[5] Deffuant, G. Neau, D. Amblard, F. and Weisbuch, G. . Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3:87–98, 2000.

[6] Epstein, J.M. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008.

[7] Griffiths, M. E-citizens: Blogging as democratic practice. *Electronic Journal of E-Government*. 2(3), 2004: 155-166.

[8] Hegselmann, R. and Krause, U. . Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artifical Societies and Social Simulation (JASSS) vol*, 5(3), 2002.

[9] Inglehart, R. and C. Welzel. Modernization, Cultural Change and Democracy. Cambridge UK: Cambridge University Press, 2005.

[10] Kacperski, K. and Hołyst, J.A.. Opinion formation model with strong leader and external impact: a mean field approach. *Physica A*, 269:511–526, 1999.

[11] Kaschesky, M. and R. Riedl. Tracing opinion-formation on political issues on the internet. Proceedings of Hawaii International Conference on System Sciences, 2011.

[12] Kaschesky, M. and R. Riedl. Top-level decisions through public deliberation on the internet. Proceedings of dg.o, 2009.

[13] Kim, K. et al. WordBridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora. Proceedings of Hawaii International Conference on System Sciences, 2011.

[14] Kwon, N. et al. Multidimensional text analysis for eRulemaking. Proceedings of dg.o, 2006.

[15] Markham, A. Internet communication as a tool for qualitative research. In: Qualitative research – Theory, method and practice. Ed. David Silverman. Thousand Oaks CA: Sage Publishing, 2004.

[16] Moss, S. and B. Edmonds. Towards good social science. *Journal Artif. Societies and Social Simulation*, 8(4):13, 2005.

[17] Mullen , T. and R. Malouf. Taking sides: User classification for informal online political discourse. *Internet Research*, Vol. 18, 177–190, 2008.

[18] Nowak, A. Szamrej, J. and Latané, B. From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact. *Psychological Review*, 97(3):362–376, 1990.

[19] Pang, Bo and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, Iss. 1-2, 1-135, 2008.

[20] Payne, J.L, P. Sheridan Dodds, and M.J. Eppstein. Information cascades on degree-correlated random networks. *Physical Review E*. Vol 80 Issue 2, 2009.

[21] Rainie, L. and J. Horrigan. Election 2006 online. *Pew Internet & American Life Project Report*, January 2007.

[22] Shulman, S. et al. Language processing technologies for electronic rulemaking. Proceedings of dg.o, 2006.

[23] Sobkowicz, P. Modelling opinion formation with physics tools: call for closer link with reality. *Journal of Artificial Societies and Social Simulation*, 12(1):11, 2009.

[24] Sobkowicz, P. Studies of opinion stability for small dynamic networks with opportunistic agents. *International Journal of Modern Physics C (IJMPC)*, 20(10):1645–1662, 2009.

[25] Sobkowicz, P. Effect of leader's strategy on opinion formation in networked societies with local interactions. *International Journal of Modern Physics C (IJMPC)*, 21(6):839–852, 2010.

[26] Sznajd-Weron, K and Sznajd, J. Opinion Evolution in Closed Community. *Int. J. Mod. Phys. C*, 11:1157–1166, 2000.

[27] Watts, D.J. A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences of the United States of America. Vol 99 Issue 9: 5766-5771, 2002.

[28] Watts, D.J. and P. Sheridan Dodds. Influentials, networks, and public opinion formation. Journal of Consumer Research. Vol 34 Issue 4: 441-458, 2007.

[29] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe and Alexander Hauptmann. Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. Conference on Computational Natural Language Learning, 2006.