Bringing Representativeness into Social Media Monitoring and Analysis

Michael Kaschesky
Bern Univ. of Applied Sciences
ksm1@bfh.ch

Pawel Sobkowicz
Independent
pawelsobko@gmail.com

José Miguel Hernández Lobato
Cambridge University
jmh233@cam.ac.uk

Guillaume Bouchard
Xerox Research Centre Europe
guillaume.bouchard@xrce.xerox.com

Cedric Archambeau
Xerox Research Centre Europe
cedric.archambeau@xrce.xerox.com

Nicolas Scharioth
Gallup Europe
Nicolas_Scharioth@gallup-europe.be

Robert Manchin
Gallup Europe
robert_manchin@gallup-europe.be

Adrian Gschwend
Bern Univ. of Applied Sciences
adrian.gschwend@bfh.ch

Reinhard Riedl
Bern Univ. of Applied Sciences
reinhard.riedl@bfh.ch

Abstract

The opinions, expectations and behavior of citizens are increasingly reflected online – therefore, mining the internet for such data can enhance decision-making in public policy, communications, marketing, finance and other fields. However, to come closer to the representativeness of classic opinion surveys there is a lack of knowledge about the socio-demographic characteristics of those voicing opinions on the internet. This paper proposes to calibrate online opinions aggregated from multiple and heterogeneous data sources with traditional surveys enhanced with rich socio-demographic information to enable insights into which opinions are expressed on the internet by specific segments of society. The goal of this research is to provide professionals in citizen- and consumer-centered domains with more concise near real-time intelligence on online opinions. To become effective, the methodologies presented in this paper must be integrated into a coherent decision support system.

1. Introduction

The opinions, expectations and behaviors of citizens are increasingly reflected online – therefore, mining the internet for such data can enhance decision-making in public policy, communications, marketing, finance and other fields. However, despite the abundance of user-generated content online, few decision makers feel comfortable basing their judgments on opinions expressed on the web. In fact, there is a lack of knowledge about the socio-demographics characteristics of those voicing opinions online. Hence, for stakeholders depending on detailed socio-demographic information of opinion holders, the harvesting of online opinions so far remains of little use.

This paper outlines the conceptual approach for a broader research project that addresses the shortcomings of existing and commercially available opinion mining solutions regarding the contextualization and representativeness of online opinion mining. To become effective, the methodologies presented in this paper must be integrated into a coherent decision support system.

The approach not only consists of calibrating online opinions with traditional surveys that gather rich socio-demographic information in order to provide insights into which opinions are being expressed on the internet by a specific segment of society. It also includes the use of agent-based simulations of online diffusion models to forecast the development of sentiment and other important indicators.

The impetus to extend this type of research stems from previous collaborations of the authors and discussions with policy makers and government agencies [7]. Based on this work, several use scenarios and use cases were defined in order to address to the following research questions:

- How to guarantee that analyzed online opinions are representative of the general population or pre-defined subgroups in order to base real-world decisions on these data?
- How can user-generated online data be used to gauge sentiment and forecast the development of leading indicators?

1530-1605/12 $26.00 © 2012 IEEE
DOI 10.1109/HICSS.2013.120

The research tackles a number of real-world challenges by applying and extending:

- The current research on big data integration approaches;
- Sentiment analysis, Linked Open Data (LOD) interlinking, and opinion learning;
- Online opinion representativeness and calibration;
- Opinion diffusion simulations and opinion prediction methods.

The research outcomes aim at supporting a large number of stakeholders both from public and private sectors to improve decision-making processes (in terms of accuracy and responsiveness) as well as the quality of their decisions (in terms of related and context knowledge).

Some of the use scenarios and use cases are briefly described in the following callouts.

Consumer 'buzz' on reputation and products or services

Decision problem: While needs and expectations expressed online are increasingly exploited, existing social media analysis technologies lack the representativeness and reliability required as basis for business decisions.
Decision question: How can online opinions become a more accurate and more reliable source of information for real-world business decisions?
Example: Much can be learned about products, features and brands by tracing and aggregating opinions about them. But few decision makers feel comfortable basing their judgements on results from online opinion mining, because it is unclear to what extent the online population represents the actual or potential share of buyers of the product or service.
Background: Many actors in the private sector and in public institutions are strongly interested in measuring the buzz around products, ideas or protagonists. A potential end user expressed interest in developing the instrument while benefiting from the research outcomes for their clients' strategic marketing and product development strategy.
Solution: Understanding the needs and preferences of consumer groups during a product launch phase by mapping the frequency and sentiment of mentions and thereby creating a 'web footprint' in order for marketing strategists to adapt their product's features and optimise their communications efforts. A detailed analysis supports marketing intelligence and helps anticipating opinion trends and dynamics necessary to foster innovations and develop products addressing tomorrow's constituencies' needs.

Public opinions on policy making

Decision problem: Approval or disapproval score of policy actors and policies are used to guide future actions but the lack of representativeness of online opinions constraints their use in real-world scenarios.
Decision question: How make online opinions representative and accessible as basis for real-world policy decisions?
Example: Due to the tremendous growth in user-generated content, most of which contain opinions, policy actors are unable to make sense of this 'opinion overload' not only in terms of aggregating the data but also – and more importantly – in terms of ensuring that online results are representative and thus reliable for decision making.
Background: The aim of an United Nations agency is to complement traditional polls by properly mapping the topics and prejudices associated to migration and migrant groups in online discussions as well as analysing the discourse and measuring the change of relative importance of concepts and subtopics within a given topic. This information helps them understand public fears, negative attitudes and hostility, address citizens' concerns, and redress misinformation.
Solution: By developing a trusted and calibrated way of analysing online opinions on migration issues, the research supports the work of policy actors and analysts to place rationality above fear at low costs for opinion research.

Sentiment analysis for economic forecasting

Decision problem: When setting interest rates, central banks study closely the sentiment towards the economy. For this, central banks typically rely on surveys and economic statistics which are available with significant time lags.
Decision question: How can user-generated online data used to gauge and forecast economic sentiment?
Example: The Financial Times (13 Jun 2011) reported that the Bank of England (BoF) uses Google searches for specific terms in order to track economic conditions in real time. BoF found that searches for "unemployment benefits" are at least as reliable for estimating current unemployment rates as are the actual number of benefit claimants.
Background: The institute is interested in the research as an additional and, more importantly, timely source of information for its own regular forecasts.
Solution: One of the key outcomes concerns the timeliness of reports given that results are available immediately and each day in contrast to official statistical data: "Monitoring current economic activity closely is an important aspect of policymaking, but official economic statistics are generally published with a lag" [17].

2. Opinion mining and data interlinking

This section presents the steps of opinion mining and data interlinking to identify in the extracted content the opinions towards relevant issues and their connections to each other as well as to other topics and content available as structured Linked Open Data in the internet. Based on available metadata, feature interlinking is used to identify relationships and correlations between content items (opinions) based on shared metadata.

Opinion mining and interlinking is mainly concerned with large-scale opinion harvesting. In its most basic form, opinion mining corresponds to sentiment analysis, i.e. the identification of a positive, neutral or negative tone about the topic discussed by the author of the document, post or tweet. Interlinking concerns the contextualization of extracted opinions to add more detail for both the human analyst as well as for the machine learning methods. While the amount of harvested opinions is huge, the data is still sparse as individuals only express their opinions about a limited number of topics. Hence, the challenge is to capture correlations between topics, so that they can be better understood and monitored over time by human analysts.

2.1 Sentiment analysis

Sentiment analysis focuses on the automatic identification and extraction of opinions, emotions, and sentiments from text and multimedia [2]. Motivation for this task comes from the desire to automatically recognize stances (opinions) in online debates and user generated content to be used by aggregators (see the subsection below on Opinion learning) [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][13][27].

The approach is based on a combination of machine learning methods with dedicated background information, such as dictionaries and labeled examples and sentiment words as features. After an initial training phase based on a supervised classification of regression technique, the polarity of the opinion expressed in free texts can be automatically estimated, enabling large-scale analyses of opinions [25]. With the rapid growth of online media and user-generated content, it has been demonstrated that relatively simple methods can be used to estimate the political orientation of people [22] or legislative speeches [15]. Contextual knowledge embedded in online social networks can significantly improve accuracy of opinion mining systems. For example, the social relation graphs increases classification accuracy in an opinion detection algorithm [9].

The goal of content extraction and analysis is to create a knowledgebase containing online opinions in a more explicit form. This is accomplished through Natural Language Processing (NLP) based on a syntax analyzer and machine learning techniques that detect which part of the sentence corresponds 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, such as (“dislike”, “BP”), (“agree”, “Obama”), (“support”, “government”), etc. The pairs corresponding to someone else’s opinion are discarded (e.g. a text such as “My father does not like BP“). 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 weather).

The representation of the domain knowledge is centered on named entities, as they are mostly at the heart of online discussions. The knowledgebase for a specific domain contains for each entity a list of relations between them (represented as triples, see interlinking below), a list of attributes (e.g. name, age, location), as well as a list of entities that may be addressed in online discussions (e.g. known people or organizations, locations, or major events). In addition, the knowledgebase contains the domain knowledge data which is required to relate similar opinions. For example, the extracted pairs (“support”, “government”) and (“agree”, “Obama”) are both counted as positive opinions when aggregating data to measure the public opinion about the current U.S. government: The presence of the triple (“Obama” – “is part of” – “government”) in the knowledgebase will be used to improve the measure of positive opinions.

2.2 Linked Open Data (LOD) interlinking

The next step concerns transforming the extracted content – enriched, as described above, with rich metadata on topics, entities, and stance (e.g. pro/positive, contra/negative) – into an openly available and reusable Linked Open Data (LOD) format following Open Government Data Principles [24]. This step also extensively reuses existing LOD sources because extracted content and metadata must put in context to increase value and provide unambiguous results. In the past years, billions of triples were published creating a cloud of semantic information often called LOD Cloud, mainly through applied research projects (e.g. LOD2, LACT,

PlanetData). Taken together, the web-harvested extracted data provides the terms and metadata to be interlinked with same and similar terms/knowledge thereby contributing to as well as exploiting the ever growing LOD cloud by:

- enriching the extracted content with existing information available in the internet;
- interlinking as much information as possible to increase the value of knowledge extraction;
- exploiting available public sector resources in Semantic Web and LOD format.

Many of the above mentioned and similar projects use RDF (Resource Description Framework) as its data model and started to bootstrap the web of data. To combine text mining and entity extraction with public data sources, semantic web technologies such as RDF are used.

The power of the semantic web lies in two things: A simple, scalable data model combined with links (Uniform Resource Identifiers, URIs) to uniquely describe and identify digital data. The data model is using a simple subject-predicate-object combination to form so-called triples. RDF is the preferred data model for triples in the semantic web and makes it possible to store this knowledge in both human and machine usable form. The fact that machines can access and 'understand' RDF makes it possible to automate certain tasks on huge amounts of data which otherwise would have to be carried out by human experts. Even though data sets may specialize on a certain set or domain of information, interlinking different data sets alone enables creating additional knowledge. This interlinked knowledge can again be consumed by the provider of the data itself.

The real value of interlinking data from semi-structured sources (e.g. online comments, Twitter messages) and structured sources (e.g. LOD) lies in the representation of the relationships between topics, opinions and individuals. This requires a data model which is able to represent this complex graph-based information and maintain and express the semantics of the relationship between entities in an unambiguous and machine-readable form. It is also important to be able represent the time-aspect of information in such a system. Topics, opinions and individuals evolve over time and semantics capable of tracking this change can help to understand trends in the future.

For opinion mining to be meaningful and contextualized, the extracted and enriched data is therefore linked with topics, opinions and individuals referenced by the LOD cloud. This makes it possible to uniquely identify and contextualize the extracted data. This enables cross-referencing topics, opinions and individuals worldwide to the correct identifier, which is also independent of a specific language. RDF thus enables storing data independent of a specific use case or application, which facilitates reuse and archiving of data. The strength of RDF lies in the used vocabularies for the relationship between two entities based on existing light-weight ontologies to solve specific problems of storing our data:

- Dublin Core provides basic provenance and metadata attributes like creator, subject, and publisher.
- Marl (An Ontology for Opinion Mining) is a standardized data schema designed to annotate and describe opinions expressed on the web or information systems. It is used for storing the opinion itself in RDF. For example, existing work demonstrates that storing opinions in RDF can help to link opinions with other concepts on the web and lead to better search capabilities and improved exposure of data [31].
- SIOC (Semantically-Interlinked Online Communities) is for linking online community sites. It enables mapping blog post, wiki entries, message boards etc. into machine readable RDF statements. It is used to represent the original structure of existing topics, opinions and individuals in RDF.
- PROV Data Model and PROV Ontology enables storing the exact provenance of data in RDF. This is a crucial point because the provenance might influence the interpretation of opinions. For example, readers of a conservative blog might have a different perception on a topic than a left-leaning reader.
- Simple Event Model (SEM) models events in various domains, without making assumptions about the domain-specific vocabularies used. Event-centered modeling captures the dynamic aspects of a domain. In addition, events provide a natural way to explicate complicated relations between people, places, actions and objects. SEM is used to put topics, opinions and individuals into context.

These ontologies and standards enable the interlinking and contextualization of extracted opinions in a very powerful, machine readable way.

The use of RDF allows us to store it independent of a specific application or use case, which makes it also possible for external parties to consume, link and grind with/on our data. Getting links back from external entities will also improve our data set as we might find new relationships between topics, opinions and individuals. The more we know about the relationship between things, the more additional data we can crunch on top of the data set. This can radically improve the value of the outcome of opinion mining.

2.3 Opinion learning

Opinion learning goes beyond sentiment analysis by discovering emerging topics with their associated opinions, trends in opinions, and, importantly, their semantic organization. Opinion learning will use the output of sentiment analysis and data interlinking to aggregate opinions and enable opinion monitoring. It is a prerequisite for opinion prediction (described in the next section), where the goal is to infer opinions that are not expressed (data imputation) and to predict opinions about emerging topics (forecasting).

Large-scale probabilistic models of opinions are used because they enable dealing with uncertainty in a principled way. Uncertainty does not only occur because the opinions are not always expressed clearly or because they may vary across documents, but also because of ambiguity in the network and the interlinking process. While the amount of harvested opinions is huge, the data is still sparse as individuals only express their opinions about a limited number of topics. Hence, the challenge is to capture correlations between topics, links and anonymized individuals. This requires the use of large-scale latent variable models and probabilistic graphical models to aggregate opinions and capture these correlations [11].

The advantage of graphical models is that they are a natural tool for dealing with complexity. They can be built in a modular way by combining simpler parts. Probability theory connects the different parts, ensuring that the whole system is consistent. This modular approach is useful for fusing and combining different data sources and domain knowledge in a principled way. The problem of making wrong model assumptions is avoided by using non-parametric Bayesian methods [25]. These techniques can capture complex patterns with arbitrary precision as long as enough data are available. Prediction is then implemented in two steps. First, a posterior distribution is computed for the model parameters by conditioning to the data. This posterior distribution represents our current beliefs about the correct value of the model parameters. Second, the value of interest is computed by averaging with respect to the posterior distribution. Computation of the posterior is performed using algorithms for large scale approximate Bayesian inference. However, in practice no model is fully reliable and the uncertainty of the predictions has to be incorporated in the analysis and the visualization of the results. Uncertainty measures are a natural output of Bayesian learning methods. These uncertainty measures are visualized in the diagrams, which will be used to interpret the results.

3. Online opinion representativeness

Besides the interlinking, contextualization, and learning of opinions, this paper identifies another key challenge that existing and commercially available opinion mining solutions fail to address. In order to reach the representativeness of classic opinion surveys, socio-demographic characteristics of those voicing opinions on the internet must be obtained. This section describes the approach to calibrate online opinions against standard opinion polls and to use machine learning methods to compute representativeness. The problem of representativeness has not been addressed adequately if at all [16]. In the context of online opinion mining, it sometimes also refers to the representativeness of summaries obtained from individual reviews (i.e. how far the summary resembles the original meaning) [12].

3.1 Opinion calibration

When opinions are harvested, they are blindly extracted from the web. In general, the extraction and the models do not take into account the fact that these opinions are originating from specific population categories. For example, young people are overrepresented in social media. Similarly, minorities, political activists or religious groups tend be overrepresented on forums and blogs [23]. Hence, there is a need for calibrating opinions collected online so that they reflect reality. In fact, their distribution can be far from the opinion distribution of the global population. The goal here is to confront opinions harvested off the web with traditional opinion polls to reduce discrepancies and – in turn – to help design targeted questions for surveys.

Opinion calibration intends to create a feedback loop between large-scale online opinion mining, which is expected to exhibit a small variance, but a potential large bias, and small- to medium-scale opinion polls, which are expected to have a large sampling variance, but a small bias. The most important difference between online opinion mining and standard opinion polls is that with the latter the demographics of respondents are known. In other words, the strength of probability samples is that sampling theory can be used to estimate the accuracy or representativeness of the sample. Hence, opinion polls can be used to de de-bias large-scale opinion mining form online resources.

Figure 1 illustrates how a concrete application of this process would look: if the example of economic outlook expressed in a representative Gallup survey is chosen for a case study, then the probabilistic model filters out the online opinions dealing with economic topics from the country where the survey was

conducted. Opinion learning would allow assessing whether the opinions held on the country's economy are more positive or more negative. In order to assess the representativeness of online opinion mining, the subsequent and – from an analytical point of view – most complex step of the modeling process investigates whether the results of the opinion learning can be shaped to resemble the survey results representative of a given population, i.e. perform advanced matching of characteristics underlying online and offline results.

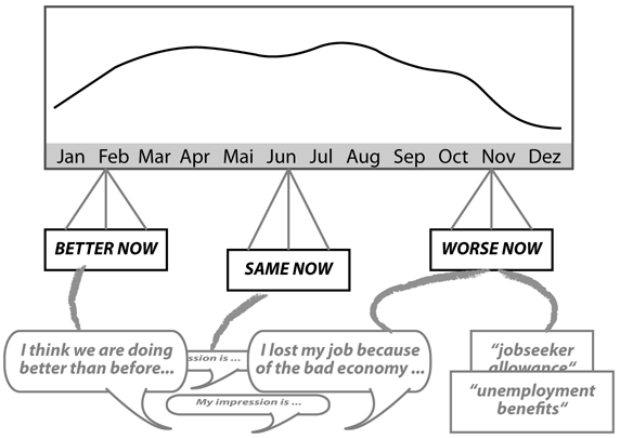

Figure 1: Extraction and clustering of online opinions as input to questions about personal economic conditions

The model may be further developed to account for 'rational ignorance' of people who choose not to be informed about an issue because the 'costs' of doing so (e.g. time, energy) outweigh the perceived benefits of being informed. Professor Fishkin of Stanford University argued that George Gallup, the pioneer of survey sampling, developed public opinion polling based on the inherently democratic assumption that citizens regularly gather together, make and hear political arguments, and vote for or against policies directly [17]. But in populations with high levels of rational ignorance, public opinion polling may be skewed by 'non-attitudes' or 'phantom opinions' on issues [4][5]. This can be tested by asking people the same questions before becoming informed and after providing the opinions and arguments around an issue.

The Gallup surveys to be used include, but might not be limited to following:

- Gallup Daily, a daily telephone survey among 1,000 US adults measuring their sentiment towards matters of economy and wellbeing, because this is the biggest daily survey with such a frequency;
- Gallup-Healthways Well-being Index, a monthly survey undertaken in the United Kingdom and Germany investigating respondents' wellbeing;
- Gallup World Poll a series of global surveys asking representative samples in more than 150 countries worldwide at least annually about health and wellbeing issues as well as opinions on societal matters such as politics, values and material satisfaction.

In addition, specific questions can be added to these ongoing surveys.

3.2 Survey optimization

The information gathered from the web can also be exploited to guide additional surveys. Indeed, *Gallup Daily* is repeated at a relatively high frequency. Repeated small-scale well-targeted opinion polls of this kind can be used to confirm the opinion trends observed on the web and to verify sudden opinion changes. They can also be used as a sanity check when there is a relatively large uncertainty on the opinion of certain subpopulations and/or about specific topics. It is expected that a balanced combination between traditional opinion polls and online opinion harvesting will lead to opinion models of unprecedented representativeness and reliability.

The *Gallup Daily* tracking methodology relies on live interviewers, dual-frame random-digit-dial sampling (which includes landline as well as cellular telephone phone sampling to reach those in cell phone-only households), and uses a multi-call design to reach respondents not contacted on the initial attempt. Gallup interviewers employ a "most recent birthday" selection method for choosing adult respondents within a household. Gallup Daily tracking includes Spanish-language interviews for Spanish-speaking respondents and interviews in Alaska and Hawaii. The data are weighted daily by number of adults in the household, number of phone lines in the household, and the respondents' reliance on cell phones, to adjust for any disproportion in selection probabilities. The data are then weighted to compensate for nonrandom nonresponse, using targets from the U.S. Census Bureau for age, sex, region, gender, education, ethnicity, and race. The resulting sample represents an estimated 95% of all U.S. households. Data that are summarized at the state, congressional district, and Metropolitan Statistical Area (MSA) level are weighted at each of these levels twice per year (for states) or once per year (for congressional districts and MSAs) to ensure that samples are representative of these areas.

One of the most salient features of the described approach is that we do not assume that data are provided by a source that cannot be controlled (i.e. not random data sources). This means, that data collection can generate predictions about what data points to collect in the next step so that the system can make the most of the new data inputs (e.g. in terms of learning from data and alignment with the opinion polls). This is especially important if collecting some of the new data points is very expensive, as is often the case with opinion polls. The framework of probabilistic graphical models enables the system to evaluate the expected information gain obtained by each candidate measurement and then select the most informative one.

4. Opinion diffusion and prediction

In order to expand the lead time and enable proactive decision making, simulation and prediction of the diffusion of opinions within selected communication networks are performed (such as distribution of opinions, changes of opinions over time, ratio of active vs. passive participants). This is based on rules observed from communication patterns in complex systems in order to model the further diffusion of observed parameters. For example, based on the rate of activity burst (short-term decision scenario) or based on the diffusion of opinions (medium-term decision scenario), it could trigger an alarm if a clear trend is detected that observed parameters will reach the specified threshold.

4.1 Opinion diffusion simulation

The goal of agent-based simulations in analyses of public opinion is to trace and forecast trend changes. The opinions harvested off the web are fed into the simulations thereby providing an understanding of the key factors driving opinion trend changes. Simulations are not aimed at perfect modeling of details, but at predicting general characteristics. As such, the agent-based simulations complement statistical analyses and machine learning methods by providing a different perspective and deeper understanding. The model combines three key areas that determine opinion diffusion. The first concerns nonlinear descriptions of individual reactions to influences, including aspects such as commitment to previously held opinions, persuasion and emotions. The second concerns communication modes (one-to-one messages, one-to-many communication, mass media influence) with their different characteristics. The third concerns descriptions of social communication networks, including factors such as time evolution and role of strongly linked sub-communities. Bringing the three aspects together in a single model enables creating simulation environments corresponding to different types of real-life situations and use cases.

Within a simplified framework focusing on selected aspects (such as interpersonal communication network, susceptibility to influences, contrariness etc.), it is possible to derive general trends of communication behavior. However, one of the major problems with opinion modeling by computer simulation is the lack of access to real-life data to test the simulations [29]. The steps outlined above to harvest and analyze continuous and real-life data streams make it possible to test the opinion diffusion simulation and prediction so as to learn and adapt simulations based on real-life data.

There are, however, examples of systems where the abstract models enabled discovery of unexpected universality of opinion change processes. An example of such discovery is the existence in the voting process of a general microscopic dynamics that does not depend on the historical, political, and/or economical context where voters operate [6].

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

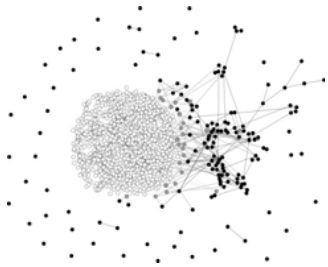

Figure 3: Later state of opinion diffusion with majority having adopted the proponent position (light grey)

The figures above show how the diffusion and prominence of opinions changes over time is outlind by agent-based simulations. Figure 2 shows the initial state with a large majority of undecided participants without a specified opinion on an issue in question (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 (light grey). However, small minorities persist, mainly because they have severed and cut most of the information flow links with the majority [29] (see also [21] for an example).

4.2 Opinion prediction

A characteristic of the opinion diffusion process described above is that some of the social links enabling information flow may not be directly observable or are corrupted by noise, for example, some users may have already adopted a position before starting the observation and running the simulation or the diffusion process may require knowledge about the probability with which two specific users that hold a social connection and opposite views may end up adopting the same opinion. Machine learning methods are used to estimate these quantities whenever they are not directly observable. Additionally, they can also provide reliable estimates when the required quantities are observable but are very noisy. The opinion prediction aims to develop reliable estimates of required quantities by using i) a probabilistic model for the data and ii) Bayesian machine learning methods.

The relationships between topics, opinions and individuals are represented using a probabilistic model in the form of a graph [18][20]. In this graph, the different nodes correspond to topics or individuals while opinions are encoded by edges or connections between individuals and topics. In addition to the edges (opinions) that link individuals and topics, the proposed model also allows for topic to topic connections and individual to individual connections.

The probabilistic model is dynamic and evolves with time [13][26]. As new topics emerge they are added to the graph. The same occurs with new individuals that start to produce relevant content. The model can also capture time dependencies in the network connections or opinions. This probabilistic model is combined with statistical machine learning methods [1] to implement the prediction functionality. The task here is to make use of information obtained from different heterogeneous data sources in order to infer the state of variables which are not directly observed or which are corrupted by noise. For the learning process, the framework of probabilistic graphical models is used in which all the variables, including observed data, are expressed as a graph as described in Section 2.

5. Discussion and next steps

5.1 Challenges of social media monitoring

The recent growth of web-based social media has resulted in a myriad of social media monitoring (SMM) solutions. They offer businesses or other actors the ability to monitor public opinion about their brand, products or services. At the end of 2011 approximately 300 SMM tools were available, among them free tools that belong to the social media platforms themselves (e.g. Google alerts, Facebook Insights or Twitter analytics), applications that use free APIs that are provided by platforms (e.g. Klout, Tweetreach) and more complex software tools. These tools categorize the content and visualize the data as charts and graphs. According to industry experts, applications build for the social media analytics market reached a value of $10.5 billion in 2010, expecting strong, steady future growth [10]. Yet, as Grimes and other industry experts from social media agency FreshNetworks point out, the industry is still in its infancy and in a permanent state of flux and change [7].

Some of the biggest challenges all SMM solutions still face concern sentiment analysis and the attempt to identify socio-demographics of discussants together with their different locations, not all of which may be relevant for the analyst. Automated sentiment analysis tools that aim to mark text as negative, positive or neutral would in theory be incredibly valuable, for instance if a manager wanted to know which topics are encouraging good or bad online conversations about the company. As FreshNetworks puts it: “we don’t believe that the tools on the market have nailed sentiment analysis yet”. For example, users of SMM tools are usually very interested in the location of the conversation they are trying to follow. But an online user has multiple locations attributed to him or her (e.g. current GPS data, location of IP provider, home address etc.).

For example, a manager interested in how a product is discussed in Brazilian social media will not be interested how the product is evaluated by social media users in Portugal. Sharing the same language, Portuguese and Brazilian internet users might use the same social media platforms to discuss this product. While social media sites have begun providing more and more geo-coding information through their API for analysts, more sophisticated solutions are needed.

The analysis of unstructured online data is more and more common in both science and technology and novel non-parametric methods are necessary for its analysis. However, the task involves many error prone steps. Therefore, the propagation and control of uncertainty is critical for success in the analysis of large-scale data streams. The Bayesian methods proposed in this paper are especially useful for reasoning and making optimal predictions in the presence of uncertainty. However, the operations required by these methods to update the current beliefs as new data becomes available are usually very costly. Some of the authors have already worked on parallel and distributed implementations of non-parametric Bayesian methods. However, these techniques are

often not suited to the problems addressed in this paper or are not yet able to scale to the amount of computations needed by the system. Therefore, the implementation will require of new modeling and algorithmic developments. These contributions are expected to have a strong impact in the field of large scale Bayesian machine learning.

5.2 Privacy concerns

Since the described approach is based on the use of data related to identifiable persons, privacy concerns have to be considered. This is a subtle topic. One may look at it from a political and from a juridical perspective. Since we rely on the processing of data which were published in order to be seen by others, there is little political concern about this. In Europe, the situation is more complicated from the juridical perspective of the European Data Protection Directive (Directive 95/46/EC) and the national laws of European Member States.

There are two key arguments why the described approach is lawful although it concerns data subject to data protection laws (that is, although some form of structured storage of data which can be traced back to individual persons takes place). One is that the results obtained are on an aggregated macro-level without reference to individuals. The other is that the data used are available in public and that the concerned persons have accepted that they are accessible in public. Unfortunately, the second argument partially breaks down in situations where data are only made accessible for "friends" or was not intended to be interlinked.

Hence, there is some room for juridical debate, which goes beyond the scope of this paper. From a practical perspective, this creates no real hindrance as long as individual tracking is impossible or obfuscated. Great care is mandatory, that data created for simulations cannot be used in a context other than predictions about social phenomena. Relinking these predictions with data referring to an identifiable person may not be compliant with laws in some countries.

5.3 Next steps

Interviews with the involved stakeholders from the public and private sectors (across various departments such as marketing, communications, public affairs, strategy, planning) must further identify contexts of use.

The shift of public sector agencies to make their data available as Open Government Data (OGD) constitutes a new and growing area that can benefit this research. The research outcomes must showcase applications based on OGD and LOD (Linked Open Data).

The authors working on the optimization of machine learning methods aim at extending graphical models (a.k.a. Bayesian networks) that encode conditional independencies between random variables. These would provide an efficient "language" to express complex probability distributions. Many real world systems can be modeled using this paradigm. Many machine learning tasks can be expressed in terms of inference in a probabilistic model. In many situations (e.g. in graphical models), this problem is intractable but can be efficiently addressed using approximate solutions. Another focus is on automatic classification. In Machine Learning, generative and discriminative methods have both advantages and drawbacks. Hence, a definition of a hybrid generative-discriminative estimation technique (called Generative-Discriminative Tradeoff) is favored as well as the proof of its optimality under weak conditions.

Finally, there are at least two salient application scenarios in the context of e-government, services for policy makers and services for e-government service marketing.

The first scenario concerning support for policy makers results from the lack of acceptance and adoption of dedicated e-participation websites. Based on a review of about a dozen European e-participation projects conducted for the European Commission, we identified severe problems with creating participation and attracting the attention of politicians. At the same time, sometimes online discussions outside of dedicated e-participation websites create such an impact to successfully overthrow government decisions, as in the case of ACTA legislation. This indicates that a prediction of emerging online opinions can be much more relevant for politicians than nurturing political discourse on dedicated e-participation websites.

The second scenario concerning e-government service marketing results from the huge acceptance problems faced by technology infrastructure initiatives such as electronic identity (eID). Although eID would provide tangible benefits if broadly used, its take-up is very slow in countries where it was introduced. For example, there is low adoption of the eID interoperability infrastructure established by the European STORK project, which enables the use of national eIDs across national borders. The key challenge here is to engage in convincing marketing activities in order to create a broader take-up. This requires a proper management of public opinions on these technologies.

References

[1] Bishop, C.M.. pattern recognition and machine learning. New York: Springer, 2007.

[2] Cardie, C., C. Farina and T. Bruce. Using natural language processing to improve eRulemaking. In: Proceedings of dg.o 2006, pp. 177-178, 2006.

[3] Chesley, P., B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In: AAAI-CAAW 2006, pp. 27-29, 2006.

[4] Fishkin, J.S. the case for a national caucus: taking democracy seriously. Atlantic Monthly (Aug.), 16-18, 1988.

[5] Fishkin, J.S. Realizing deliberative democracy. In: Leib and He (eds.), The search for deliberative democracy in China. New York: Palgrave MacMillan. pp. 44-45, 2006.

[6] Fortunato, S. and C. Castellano. Scaling and universality in proportional elections. Physical Review Letters, 99, 2007.

[7] FreshNetworks, Social media monitoring report 2011. [http://www.freshnetworks.com/files/freshnetworks/FINAL%20FreshNetworks%20version_0.pdf]

[8] Gallup Learn@Lunch. Social media in politics and policy-making. March 29, 2012. Brussels, BE. [http://eu.gallup.com/brussels/153680/gallup-learn-lunch-brussels-march.aspx]

[9] Goldberg, A. B., J. Zhu and S. Wright. Dissimilarity in graph-based semi-supervised classification. AISTATS 2007.

[10] Grimes, S. How I estimate (social/sentiment/text analytics) market size. socialmediatoday, 3 Jan 2012. [http://socialmediatoday.com/sethgrimes/421244/how-i-estimate-socialsentimenttext-analytics-market-size]

[11] Jordan, M.I. Graphical Models. Statistical Science, 19, pp. 140-155, 2004.

[12] Kim, HD and C. Zhai. Generating comparative summaries of contradictory opinions in text. In: Proceedings of 18th ACM CKIM, New York, NY, pp. 385–394, 2009.

[13] Kolar M., L. Song, A. Ahmed and E.P. Xing. Estimating time-varying networks. The Annals of Applied Statistics, 4(1), pp. 94-123, 2010.

[14] Kwon, N, S.W. Shulman, E. Hovy. Multidimensional text analysis for eRulemaking. In: Proceedings of dg.o 2006.

[15] Laver, M. Extracting policy positions from political texts using words as data. Am. Political Science Review, 97(2), 2003.

[16] Lu, Heng. Public opinion between blogsphere and real world. Annual conference of the World Association for Public Opinion Research. Hong Kong, June 14-16, 2012.

[17] Luskin, RC., J.S. Fishkin and R. Jowel. Considered opinions: deliberative polling in Britain. Brit. J. Pol. Sci., 32, pp. 457-460, 2002.

[18] McLaren, N. and R. Shanbhogue. Using internet search data as economic indicators. Bank of England Quarterly Bulletin, Q2 2011.

[19] Menon, A.K. and C Elkan. Link prediction via matrix factorization. ECML PKDD'11, 2011.

[20] Miller, Griffiths & Jordan. Nonparametric latent feature models for link prediction. Advances in Neural Information Processing Systems 22, MIT Press, pp. 1276-1284, 2009.

[21] Moral-Toranzo, F., J. Canto-Ortiz and L. Gómez-Jacinto. Anonymity effects in computer-mediated communication in the case of minority influence. Computers in human behavior, 23, pp. 1660-1674, 2007.

[22] Mullen, T. A Preliminary investigation into sentiment analysis of informal political discourse. AAAI-CAAW, 2006.

[23] Munson, S.A. and P. Resnick. The prevalence of political discourse in non-political blogs. International AAAI Conference on Weblogs and Social Media, 2011.

[24] Open Government Working Group. Sebastopol, USA, 2007. [http://www.opengovdata.org/home/8principles]

[25] Orbanz, P. amd Y.W. Teh. Bayesian nonparametric models. Encyclopedia of Machine Learning, Springer, 2010.

[26] Pang, B, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10, pp. 79-86, 2002.

[27] Sarkar, P. and M. Jordan. Non-parametric Link Prediction. CoRR abs/1109.1077, 2011.

[28] Shulman, S, E. Hovy, J. Callan and S. Zavestoski. Language processing technologies for electronic rulemaking. In: Proceedings of dg.o 2005, pp. 87-88, 2005.

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

[30] Somasundaran, S. and J. Wiebe. Recognizing stances in online debate. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116-124, 2010.

[31] Westerski, I. and T. Rico. Linked Opinions: Describing Sentiments on the Structured Web of Data. 4th workshop on Social Data on the Web (SDoW2011), co-located with (ISWC2011), Bonn, Germany, 23 October, 2011