Research Challenge on Opinion Mining and Sentiment Analysis*

David Osimo[1] and Francesco Mureddu[2]

Draft

Background

The aim of this paper is to present an outline for discussion upon a new Research Challenge on Opinion Mining and Sentiment Analysis. This research challenge has been developed in the scope of project CROSSOVER "Bridging Communities for Next Generation Policy-Making" in the view of the definition of a new Research Roadmap on ICT Tools for Governance and Policy Making, building on the model and the research roadmap developed within the scope of the CROSSROAD project[3], but with a stronger focus on governance and policy modeling. To this aim CROSSOVER focuses on amending two Grand Challenges, already part of the CROSSROAD roadmap: GC1 - Model-based Collaborative Governance and GC2 - Data-powered Collective Intelligence and Action. Each Grand Challenge consists in a number of research challenges. In particular the Grand Challenge 2 embeds the research challenge "Peer to peer public opinion mining", which we aim to amend, update, improve and validate during the workshop.

Introduction and definition

The explosion of social media has created unprecedented opportunities for citizens to publicly voice their opinions, but has created serious bottlenecks when it comes to making sense of these opinions. At the same time, the urgency to gain a real-time understanding of citizens concerns has grown: because of the viral nature of social media (where attention is very unevenly and fastly distributed) some issues rapidly and unpredicably become important through word-of-mouth.

Policy-makers and citizens don't yet have an effective way to make sense of this mass conversation and interact meaningfully with thousands of others.

As a result of this paradox, the public debate in social media is characterized by short-termism and auto-referentiality. Many experts consider social media as a missed opportunity for better policy debate.

At the same time, the sheer amount of raw data is also an opportunity to better make sense of opinions. The key asset that Google exploited to reach dominance in the search market is not a better algorythm, but the power of more data (quote).

* The research activity leading to this paper has been funded by the European Commission under the activity ICT-7-5.6 – "ICT Solutions for governance and policy modeling" within the Coordination and Support Action (FP7-ICT-2011-7, No. 288828) CROSSOVER project "Bridging Communities for Next Generation Policy-Making". The views expressed in this paper are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission

[1] Tech4i2 ltd., UK - david.osimo@tech4i2.com

[2] Tech4i2 ltd., UK - francesco.mureddu@tech4i2.com

[3] Osimo, D. et al., 2010. The CROSSROAD Roadmap on ICT for Governance and Policy Modeling.

We are therefore at a crucial underpinning where the challenge of information overload can become not a problem, but an opportunity for making sense of a thousand voices and identify problems as soon as they arise.

Opinion mining can be defined as a sub-discipline of computational linguistics that focuses on extracting people's opinion from the web. The recent expansion of the web encourages users to contribute and express themselves via blogs, videos, social networking sites, etc. All these platforms provide a huge amount of valuable information that we are interested to analyse. Given a piece of text, opinion-mining systems analyse:

- Which part is opinion expressing;
- Who wrote the opinion;
- What is being commented.

Sentiment analysis, on the other hand, is about determining the subjectivity, polarity (positive or negative) and polarity strength (weakly positive, mildly positive, strongly positive, etc.) of a piece of text – in other words:

- What is the opinion of the writer

Opinion Mining and Open Data

In recent times we witnessed an explosion of data availability, the so-called data deluge[4], determined by an increased amount of electronic action performed (such as using social networks online) and the progressive pervasive reach of IT in all devices. The first of these trends is the so called "open data" movement, characterized by the fact that all across Europe and the US, governments are increasingly publishing their data repositories for other people to access and use it. Another trend concerns the vast amount of data is made available by citizens through "participatory sensing": ordinary take a proactive role in publishing comments and complaint on line, and increasingly use technology to record additional information, such as photos or audio recordings, typically through smartphones. Furthermore sensors are becoming embedded in everyday non-ICT, such as cars or the urban landscape, so that usable data are automatically collected at a very fast pace. Finally these data, as well as government data, are not only collected by citizens but now made available to citizens so that new information becomes available. At analytical level there are several technological innovations that help making sense of the large amount of data availability. In this short paper we will take into account opinion mining, The limits of human attention, combined with the existing simple interfaces available for browsing discussion and comments, often leads to low levels of engagement and flaming wars, driving to polarisation of arguments and enhanced risks of conflicts. To address this challenge, opinion mining differs from pure data and text mining insofar it deals with subjective statement. In this sense, it is a specific development of a discipline dealing with unstructured information extraction (IE) that was previously mainly working with objective data such as natural disasters or bibliographic information. The explosion of user-generated content widens the application scope of public opinion mining tools, which are becoming more pervasive and available to the majority of citizens.

Opinion Mining Applications

Opinion mining and sentiment analysis cover a wide range of applications.

---

[4] Anderson, C. (2008). *Wired Magazine*, *16*(7), 16–07. Retrieved from http://www2.research.att.com/~volinsky/DataMining/Papers/AndersonWired.pdf

1. Argument mapping software helps organising in a logical way these policy statements, by explicitating the logical links between them. Under the research field of Online Deliberation, tools like Compendium, Debatepedia, Cohere, Debategraph have been developed to give a logical structure to a number of policy statement, and to link arguments with the evidence to back it up.
2. Voting Advise Applications help voters understanding which political party (or other voters) have closer positions to theirs. For instance, SmartVote.ch asks the voter to declare its degree of agreement with a number of policy statements, then matches its position with the political parties.
3. Automated content analysis helps processing large amount of qualitative data. There are today on the market many tools that combine statistical algorythm with semantics and ontologies, as well as machine learning with human supervision. These solutions are able to identify relevant comments and assign positive or negative connotations to it (the so-called sentiment).

The first two point reflect mature application areas, while the third area is emerging and with relevant research issues. We will therefore mainly focus on this area for the research issues.

Why it matters in governance

Opinion mining applications are the basic infrastructure of large scale collaborative policy-making. They help making sense of thousands of interventions. They help to detect early warning system of possible disruption in a timely manner, by detecting early feedback from citizens. Traditionally, ad hoc surveys are used to collect feedback in a structured manner. However, this kind of data collection is expensive, as it deserves an investment in design and data collection; it is difficult, as people are not interested in answering surveys; and ultimately it is not very valuable, as it detects “known problems” through pre-defined questions and interviewees, but fails to detect the most important problems, the famous “unknown unknown”. Opinion mining is helpful to identify problems by listening, rather than by asking, thereby ensuring a more accurate reflection of reality.
Argument mapping software is then useful to ensure that policy debates are logical and evidence-based, and do not repeat the same arguments again and again.
These tools would finally be helpful not only for policy-makers, but also for citizens who could more easily understand the key points of a discussion and participate to the policy-making process.

Recent trends

Opinion mining is not in itself a new research theme. Automated methods for content analysis have been increasingly used, and have increased at least 6 folds from 1980 to 2002 (Neuendorf, K. A. 2002. The Content Analysis Guidebook. Sage). The research theme is based in long established computer science disciplines, such as Natural Language Processing, Text Mining, Machine Learning and Artificial Intelligence, Automated Content Analysis, and Voting Advise Applications.
However, according to Pang and Lee (2008), since 2001 we see a growing awareness of the problems and opportunities, and “subsequently there have been literally hundreds of papers published on the subject.”

What is new today is the sheer increase in the quantity of *unstructured data*, mainly due to the adoption of social media, that are available for machine learning algorythm to be trained on. Social media content by nature reflects opinions and sentiments, while traditional content analysis tended to focus on identifying topics ((Pang, Lee, and Vaithyanathan 2002). As such, it deals with more complex natural language problems. Because of the combination of increase in the volume of data available and more complex concepts to analyse, in recent years there has been a decrease in interest on semantic-based application, and a move towards greater use of statistics and visualisation. Just as any other scientific discipline, also automated content analysis is becoming a data-intensive science.

Inspiring cases

Usage of DiscoverText in government
OpinionSpace

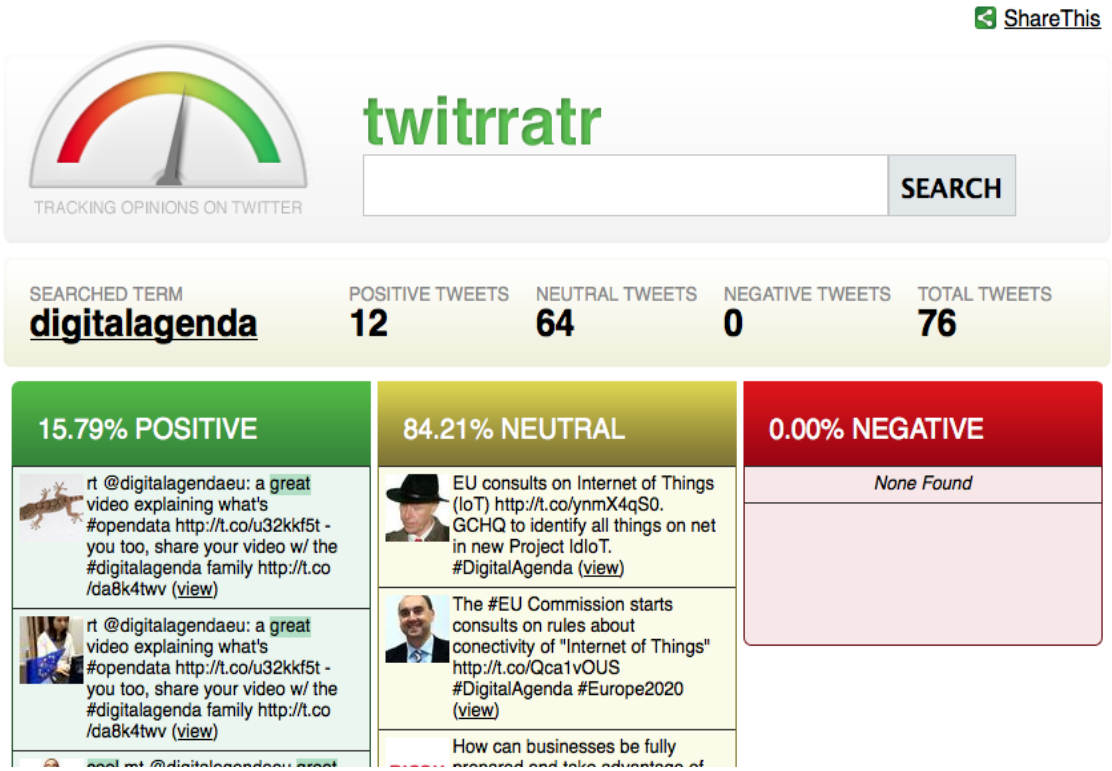
Tools on the market

The market of opinion mining tools is crowded with solution providers. Most of these applications are geared towards analyzing customers feedback about products and services, and therefore skewed towards sentiment analysis that detects positive/negative feelings by interpreting natural language.

Freely available tools

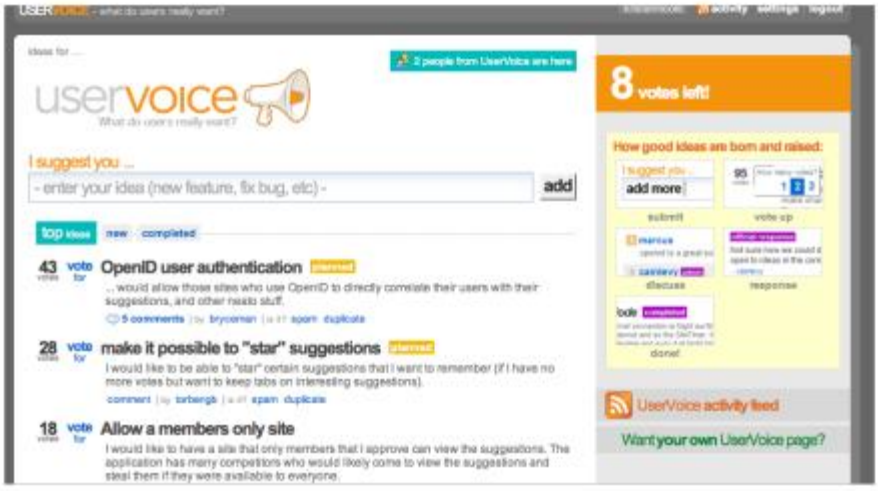
Most of the state-of-the-art argument mapping and voter advise applications are freely available, because they derive largely from academic community or NGOs. A comprehensive list of such tools is available in
http://groups.diigo.com/group/crossoverproject/content/tag/argumentmapping and
http://groups.diigo.com/group/crossoverproject/content/tag/VAA

There are currently freely available applications that simply analyse terms based on a pre-defined glossary, and giver highly simplified and unreliable results. One example is
http://twitrratr.com/



Another stream of simple, free and popular solutions is the word visualisation. Wordclouds are becoming more and more used to make sense of large quantities of information in a snapshot. Obviously, such tools are also extremely siplified and only offer a visualisation of the most common used terms, which is helpful to have an idea of what the document is about, but little more. Tools such as wordle.com provide an appealing design solution that can serve as an entry level in the opinion mining market. They are therefore important to involve a much wider public in this kind of activities.

Finally, another way of making sense of large amount of information is by relying on human effort, by crowdsourcing and collective intelligence: people are not only submitting their opinions, but actually filtering them by signaling the most important ones. Tools such as uservoice.com allow customers to submit feedback and to rank other people ideas, thereby allowing the emergence of the most popular ideas. These tools are available at very low cost, but research shows that they are effective in gathering feedback but not in identifying good ideas, as voting tends to focus on easier and most popular issues.



Enterprise-level software

Beside these simple and free applications, there is then a flourishing market of enterprise-level software for opinion mining which much more advanced features. These tools are largely in use by companies to monitor their reputation and the feedback about products on social media. In the government context, opinion mining has long been in use as an intelligence tool, to detect

hostile or negative communications (Abbasi 2007). More recently, politics has become a key area of applications, as politicians monitor public opinion on social media to understand public reaction to their position.
Technically, these tools rely on machine learning with regard to identifying and classify relevant comments, through a combination of latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation.This process requires significant human effort aided by machines: all the tools on the market rely on a combination of machine and human analysis, typically using machines to augment human capacity to classify , code and label comments. Automated analysis is based on a combination of semantic and statistical analysis. Recently, because of the sheer increase in the quantity of datasets available, statistical analysis is becoming more important.

Key challenges and gaps

Current solutions for opinion mining and sentiment analysis are fastly evolving, typically by reducing the amount of human effort needed to classify comments.
Among the challenges identified we can select:
- the detection of spam and fake reviews, mainly through the identification of duplicates, the comparison of qualitative with summary reviews, the detection of outliers, and the reputation of the reviewer (Liu 2008)
- the limits of collaborative filtering, which tends to identify most popular concepts and to overlook most innovative / out of the box thinking
- the risk of a filter bubble (Pariser 2011), where automated content analysis combined with behavioural analysis leads to a very effective but ultimately deviating selection of relevant opinions and content, so that the user is not aware of content which is somehow different from his expectations
- the asymmetry in availability of opinion mining software, which can currently be afforded only by organisations and government, but not by citizens. In other words, government have the means today to monitor public opinion in ways that are not available to the average citizens. While content production and publication has democratized, content analysis has not.
- the integration of opinion with behaviour and implicit data, in order to validate and provide further analysis into the data beyond opinion expressed
- the continuous need for better usability and user-friendliness of the tools, which are currently usable mainly by data analysts

Current research

Current research is focussing on:

- improving the accuracy of algorithm for opinion detection
- reduction of human effort needed to analyze content
- Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classification
- Identification of policy opinionated material to be analysed
- Computer-generated reference corpuses in political/governance field
- Visual mapping of bipolar opinion
- Identification of highly rated experts

Future research: long term and short term issues

Short-term:

- Enhanced discoverability of content through Linked Data
- Visual representation
- Audiovisual opinion mining
- Real-time opinion mining
- Machine learning algorithms
- SNA applied to opinion and expertise
- Bipolar assessment of opinions
- Multilingual reference corpora
- Comment and opinion recommendation algorithm
- Cross-platform opinion mining
- Collaborative sharing of annotating/labelling resources

Long-term

- Autonomous machine learning and artificial intelligence
- Usable, peer-to-peer opinion mining tools for citizens
- Non-bipolar assessment of opinion
- Automatic irony detection

Summary overview

Current free tools	Top market tools	Current research	Short term future research	Long term future research
filtering opinion based on rating; assessing sentiments based on keywords; visual word counting Argument mapping and VAA	Machine learning + human analysis	Statistical + Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classification Identification of policy opinionated material to be analysed Computer-generated reference corpuses in political/governance field Visual mapping of bipolar opinion Identification of highly rated experts	Visual representation Audiovisual opinion mining Real-time opinion mining Machine learning algorithms Natural language interfaces SNA applied to opinion and expertise Bipolar assessment of opinions Multilingual reference corpora Recommendation algorythms	Multilingual audiovisual opinion mining Usable, peer-to-peer opinion mining tools for citizens Non-bipolar assessment of opinion Automatic irony detection