27 October 2013

Can text mining help handle the data deluge in public policy analysis?

Filed under: Civic Engagement,Journal,Methods,Politics & Government

By Aude Bicquelet

Governments are lagging behind when it comes to exploiting the advantages of text mining to handle and analyze the large quantities of text that result from large-scale e-consultations. In their paper "Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis?" Aude Bicquelet (LSE) and Albert Weale (UCL) analyze a public consultation on end-of-life medicines to evaluate the benefits of text mining for the analysis of online public consultations, weighing the benefits of increased automation against the potential risks.

Citizen participation poses certain challenges for the design and analysis of public policy; in particular, governments must demonstrate that all opinions expressed through participatory exercises have been duly considered and carefully weighted before decisions are reached. Image by brent granby.

Policy makers today must contend with two inescapable phenomena. On the one hand, there has been a major shift in the policies of governments concerning participatory governance – that is, engaged, collaborative, and community-focused public policy. At the same time, a significant proportion of government activities have now moved online, bringing about "a change to the whole information environment within which government operates" (Margetts 2009, 6).

Indeed, the Internet has become the main medium of interaction between government and citizens, and numerous websites offer opportunities for online democratic participation. The Hansard Society, for instance, regularly runs e-consultations on behalf of UK parliamentary select committees. For examples, e-consultations have been run on the Climate Change Bill (2007), the Human Tissue and Embryo Bill (2007), and on domestic violence and forced marriage (2008). Councils and boroughs also regularly invite citizens to take part in online consultations on issues affecting their area. The London Borough of Hammersmith and Fulham, for example, recently asked its residents for thier views on Sex Entertainment Venues and Sex Establishment Licensing policy.

ABOUT THIS BLOG

This blog investigates the relationship between the Internet and public policy. It covers work by the Oxford Internet Institute, and work published in its journal Policy & Internet (Wiley-Blackwell).

Policy & Internet calls for papers that present genuinely new approaches to policy questions or problems relating to the Internet and related ICTs. Ed-in-Chief: Helen Margetts. Submit your paper.

Submit your paper >

Top-Read Posts


Uber and Airbnb make the rules now — but to whose benefit?


How big data is breathing new life into the smart cities concept


Why are citizens migrating to Uber and Airbnb, and what should governments do about it?


Online collective action and policy change: new special issue from Policy and Internet


Digital Disconnect: Parties, Pollsters and Political Analysis in #GE2015


The promises and threats of big data for public policy-making


Time for debate about the societal impact of the Internet of Things


Five recommendations for maximising the relevance of social science research for public policy-making in the big data era


Young people are the most likely to take action to protect their privacy on social networking sites


Will digital innovation disintermediate banking -- and can regulatory frameworks keep up?

View all posts

Elections and the Internet

However, citizen participation poses certain challenges for the design and analysis of public policy. In particular, governments and organizations must demonstrate that all opinions expressed through participatory exercises have been duly considered and carefully weighted before decisions are reached. One method for partly automating the interpretation of large quantities of online content typically produced by public consultations is text mining. Software products currently available range from those primarily used in qualitative research (integrating functions like tagging, indexing, and classification), to those integrating more quantitative and statistical tools, such as word frequency and cluster analysis (more information on text mining tools can be found at the National Centre for Text Mining).

While these methods have certainly attracted criticism and skepticism in terms of the interpretability of the output, they offer four important advantages for the analyst: namely categorization, data reduction, visualization, and speed.

1. Categorization. When analyzing the results of consultation exercises, analysts and policymakers must make sense of the high volume of disparate responses they receive; text mining supports the structuring of large amounts of this qualitative, discursive data into predefined or naturally occurring categories by storage and retrieval of sentence segments, indexing, and cross-referencing. Analysis of sentence segments from respondents with similar demographics (eg age) or opinions can itself be valuable, for example in the construction of descriptive typologies of respondents.

2. Data Reduction. Data reduction techniques include stemming (reduction of a word to its root form), combining of synonyms, and removal of non-informative “tool” or stop words. Hierarchical classifications, cluster analysis, and correspondence analysis methods allow the further reduction of texts to their structural components, highlighting the distinctive points of view associated with particular groups of respondents.

3. Visualization. Important points and interrelationships are easy to miss when read by eye, and rapid generation of visual overviews of responses (eg dendrograms, 3D scatter plots, heat maps, etc.) make large and complex datasets easier to comprehend in terms of identifying the main points of view and dimensions of a public debate.

4. Speed. Speed depends on whether a special dictionary or vocabulary needs to be compiled for the analysis, and on the amount of coding required. Coding is usually relatively fast and straightforward, and the succinct overview of responses provided by these methods can reduce the time for consultation responses.

Despite the above advantages of automated approaches to consultation analysis, text mining methods present several limitations. Automatic classification of responses runs the risk of missing or miscategorising distinctive or marginal points of view if sentence segments are too short, or if they rely on a rare vocabulary. Stemming can also generate problems if important semantic variations are overlooked (eg lumping together ‘ill+ness’, ‘ill+defined’, and ‘ill+ustration’). Other issues applicable to public e-consultation analysis include the danger that analysts distance themselves from the data, especially when converting words to numbers. This is quite apart from the issues of inter-coder reliability and data preparation, missing data, and insensitivity to figurative language, meaning and context, which can also result in misclassification when not human-verified.

However, when responding to criticisms of specific tools, we need to remember that different text mining methods are complementary, not mutually exclusive. A single solution to the analysis of qualitative or quantitative data would be very unlikely; and at the very least, exploratory techniques provide a useful first step that could be followed by a theory-testing model, or by triangulation exercises to confirm results obtained by other methods.

Apart from these technical issues, policy makers and analysts employing text mining methods for e-consultation analysis must also consider certain ethical issues in addition to those of informed consent, privacy, and confidentiality. First (of relevance to academics), respondents may not expect to end up as research subjects. They may simply be expecting to participate in a general consultation exercise, interacting exclusively with public officials and not indirectly with an analyst post hoc; much less ending up as a specific, traceable data point.

This has been a particularly delicate issue for healthcare professionals. Sharf (1999, 247) describes various negative experiences of following up online postings: one woman, on being contacted by a researcher seeking consent to gain insights from breast cancer patients about their personal experiences, accused the researcher of behaving voyeuristically and “taking advantage of people in distress.” Statistical interpretation of responses also presents its own issues, particularly if analyses are to be returned or made accessible to respondents.

Respondents might also be confused about or disagree with text mining as a method applied to their answers; indeed, it could be perceived as dehumanizing – reducing personal opinions and arguments to statistical data points. In a public consultation, respondents might feel somewhat betrayed that their views and opinions eventually result in just a dot on a correspondence analysis with no immediate, apparent meaning or import, at least in lay terms. Obviously the consultation organizer needs to outline clearly and precisely how qualitative responses can be collated into a quantifiable account of a sample population’s views.

This is an important point; in order to reduce both technical and ethical risks, researchers should ensure that their methodology combines both qualitative and quantitative analyses. While many text mining techniques provide useful statistical output, the UK Government’s prescribed Code of

Visit the OII's Elections and the Internet website: exploring the extent to which data from the social web can be used to predict interesting social and political phenomena, especially elections.

Special series: The Internet in China

with posts from the ICA2013 preconference "China and the New Internet World" (Oxford, 14-15 June).

Practice on public consultation is quite explicit on the topic: "The focus should be on the evidence given by consultees to back up their arguments. Analyzing consultation responses is primarily a qualitative rather than a quantitative exercise" (2008, 12). This suggests that the perennial debate between quantitative and qualitative methodologists needs to be updated and better resolved.

References

Margetts, H. 2009. "The Internet and Public Policy." Policy & Internet 1 (1).

Sharf, B. 1999. "Beyond Netiquette: The Ethics of Doing Naturalistic Discourse Research on the Internet." In Doing Internet Research, ed. S. Jones, London: Sage.

Read the full paper: Bicquelet, A., and Weale, A. (2011) Coping with the Cornucopia: Can Text Mining Help Handle the Data Deluge in Public Policy Analysis? Policy & Internet 3 (4).

Dr Aude Bicquelet is a Fellow in LSE's Department of Methodology. Her main research interests include computer-assisted analysis, Text Mining methods, comparative politics and public policy. She has published a number of journal articles in these areas and is the author of a forthcoming book, "Textual Analysis" (Sage Benchmarks in Social Research Methods, in press).

Note: This article gives the views of the authors, and not the position of the Policy and Internet Blog, nor of the Oxford Internet Institute.

Related posts

Internet, Politics, Policy 2010: Political Participation and Petitioning
This panel was one of three in the first round of panels and has been focusing on ePetitions. Two contributions from Germany and two
In "ipp2010"

The scramble for Africa's data
In "Development"

Papers on Policy, Activism, Government and Representation: New Issue of Policy and Internet
We are pleased to present the combined third and fourth issue of Volume 4 of Policy and Internet. It contains eleven articles, each of which
In "activism"

Latest OII News

- Recruiting: Researcher (data science)
- Recruiting: Researcher
- Recruiting: Developer
- Now Available Part-time: MSc in Social Science of Internet
- Outstanding Paper Award for the OII's Darja Groselj
- Facial Recognition
- Political Communication
- Career Achievement Award for the OII's Grant Blank

Forthcoming OII Events

- VIRTUAL Student Open Day
- Student Open Day
- Workshop "Understanding the Responsibilities of Online Service Providers in Information Societies"
- CCS'15 Satellite Workshop on Computational Social Science
- Student Open Day
- Global Conference on Economic Geography 2015

Blog Tags

activism agenda-setting Arab Spring big data broadband campaigning censorship child protection China collective action crisis crowd sourcing data-sharing data protection data science democracy development digital divides digital goods economics elections governance government ICT4D ipp2010 ipp2012 media MENA MENA-Wikipedia mobilisation networks news media OII open data participation personal data politics privacy regulation representation social media trust Twitter voting Wikipedia