Ethical, Epistemological, Methodological, Social and Other Issues in Web/Social Media Mining

Marko M. Skoric

Department of Communication PhD Student Workshop
Web Mining for Communication Research
April 22-25, 2014

The Promise of Computational Social Science

* Using all available digital traces to create comprehensive explanations of individual, group and societal behavior
  * Longitudinal, networked, behavioral, cross-linked data
  * Increasing share of social interactions are mediated by technology
* Leveraging on exponential increase in both data availability and computational power ("Big Data")
  * Terabytes vs. megabytes of data
  * Physics, biology, and chemistry have been transformed by such approaches
  * Why not social science?
* Expected to bring benefits to industries, governments, communities and citizens
  * Increasing productivity and competitiveness, improving functioning of the public sector

Traditional Social Science

* Typically uses data from
  * Small-scale, qualitative designs (interviews, focus groups, ethnography)
  * Small-scale, laboratory based experiments
  * Larger-scale snapshots of mainly self-reported data (surveys) or larger-scale content analyses
* Each of the approaches has inherent limitations, which we have learned to accept and address
  * Absence of large-scale behavioral or self-reported data (with some exceptions, e.g. elections and censuses)
  * Many issues of validity, generalizability, representativeness, social desirability, artificiality, and invasiveness

Computational Social Science/Web Mining/Social Media Analytics

* Low(er) cost
* (Near) real-time analysis
* Greater variety of topics and contexts
* Inobtrusive measures
* Continuous, longitudinal, panel type data
* Cross-national/comparative data
* Captures the structure, not just the content
  * Behavioral logs
  * Social conversations (text, audio, & video)
  * Social networks and relationships
* Universe vs. sample

Surveys vs. Web/Social Media Mining

* Survey as a structured, systematic method of collecting fairly large number of solicited responses
  * Survey data frequently criticized for being based on "self-reports"
  * (Preferably) utilizing probability sampling
    * Ability to generalize based on sound statistical/mathematical principles
* Social media messages are more similar to short conversations and are typically not solicited (at least not by researchers), nor structured
  * Word-of-mouth (WoM) messages
  * Diaries
  * Ethnography

Challenges in Web/Social Media Mining

- Data collection
  - Know-how
  - Open vs. closed/limited API
    - Constantly changing APIs/lack of published specs
  - Scale
- Sampling and representativeness
  - Social media users are different from average citizens
    - Yes, but how and for how long more?
  - Self-selection issues
  - Spam and astroturfing
- Analysis
  - Techniques and approaches (problems with black-boxing)
  - Statistical assumptions are often violated (which ones?)
  - Scalability/computational issues
- Lack of good theories – data-driven research is dominant

Data Collection and Sampling:
Lack of Established Reporting Procedures and Standards

Sampling procedure description from a survey study (from *Public Opinion Quarterly*)

RESPONDENTS

Data for our first study were collected via the Face-to-Face Recruited Internet Survey Platform (FFRISP), which involved a national area-probability sample of American adults who completed monthly surveys via the Internet between October, 2008, and September, 2009. Interviewers from Abt/SRBI visited a set of randomly selected homes across the country to invite one randomly selected adult in each household to join the panel and complete one 30-minute questionnaire per month in exchange for a free laptop computer (or the cash equivalent of its value), free high-speed internet access at home (if the household did not have that already), and small cash payments each month. The present experiment was included in the questionnaire for the 11th wave of data collection, which was launched in September, 2009; 90.6% of the panelists completed that survey (N = 906). The AAPOR RR4 for recruitment of the panel was 43%, yielding a Cumulative Response Rate 1 of 39% for Wave 11 (Callegaro and DiSogra 2008). All analyses were conducted using survey weights that adjusted for features of the area-probability sample design and that included post-stratification adjustments so that the proportions of respondents in various demographic groups closely matched the true proportions in the population of American adults.

Twitter data collection procedure description (from *ICWSM*)

Data set and methodology

We examined 104,003 political tweets, which were published on Twitter's public message board between August 13th and September 19th, 2009, prior to the German national election, with volume increasing as the election drew nearer. We collected all tweets that contained the names of either the 6 parties represented in the German parliament (CDU/CSU, SPD, FDP, B90/Die Grünen, and Die Linke) or selected prominent politicians of these parties who are regularly included in a weekly survey on the popularity of politicians conducted by the research institute "Forschungsgruppe Wahlen". CDU and CSU, often referred to as the "Union", are sister parties which form one faction in the German parliament.

Our query resulted in roughly 70,000 tweets mentioning one of the 6 major parties and 35,000 tweets referring to their politicians.

Challenges in Web/Social Media Mining- continued

* Social science PhDs are (typically) not sufficiently trained
  * Inherent need for interdisciplinary collaboration (e.g. social scientists + physicists + computational scientists)
* Issues of consent, anonymity and privacy
  * Even anonymized data can be *deanonymized*
    * 87% of US population can be identified with 3 variables—gender, date of birth and ZIP code
    * Physicists don't need IRBs or do they?
* Lack of suitable theoretical models
  * "The end of theory"
* Inertia and lack of understanding in the social science community
  * Challenges with peer-review and professional evaluation
    * Where to publish? Journals or conferences?

Challenges in Web/Social Media Mining- continued

* Broader issues of epistemology and ethics
  * What is knowledge? What is research? What's reality?
  * "*Numbers (don't) speak for themselves*"
  * Quantification vs. objectivity
* "Tools'R'Us"
* Is "bigger data" better data?
  * Self-explanatory?
  * Methodologically sound?
    * Are messages on social media platforms genuine and authentic?
      * Or curated, managed, edited?
      * Long-tail of participation and content creation (80/20 rule)
    * Is Twitter data collected via APIs representative? If so, of what?
      * "Firehose", "gardenhose" & "spritzer" types of access
      * API characteristics may shift over time (without warning)

Challenges in Web/Social Media Mining- continued

* (Really) big data is mostly proprietary (Google, Facebook, Weibo) or owned by governments (national security agencies)
  * Lack of proper data-sharing norms, protocols and procedures
  * Only big players (countries, companies, universities) have the privilege of full access
  * Difficulty in replicating findings
* Supply of web/social media data varies highly across different societies and contexts
  * Dependent on level of technological/infrastructural development
    * "Big data" divide?
* Global shortage of computational talent

Some Future Directions

* Triangulation of traditional and web/social media mining methods
  * Establishing validity of measures
    * Surveys combined with web/social media mining
  * Combining human coders and machine-learning algorithms
    * Amazon.com's Mechanical Turk
* Developing standardized sets of methods and procedures for data collection, processing and analysis
  * Preserving comparability and allowing for replicability
    * Data sharing
* Doing RQ or theory-driven research
* Social science as scholarship aimed at understanding and/or improving lives of human beings

Scientists and Conferences to Follow

* Albert-Laszlo Barabasi
* Alex “Sandy” Pentland
* Lada Adamic
* David Lazer
* Gary King
* Jure Leskovec
* Michael Macy
* Noshir Contractor

* ICWSM
* iConference
* WSDM
* WebSci
* WWW

Useful References

* boyd, d. & Crawford, K. 2012. Critical Questions for Big Data, *Information, Communication and Society*, Volume 15, no 5, pp 662-679. pdf
* King, G. 2012. Ensuring the Data Rich Future of the Social Sciences. *Science* 331, no. 11 February: 719-721. pdf
* Manovich, L. 2012. Trending: The Promises and the Challenges of Big Social Data. *Debates in the Digital Humanities*, edited by Matthew K. Gold. The University of Minnesota Press. pdf

Thank you!

Questions? Any comments?