# Content analysis of social media communications

Andrei Kirilenko
University of Florida

## Summary

1. Introduction to social media analysis
2. Computer assisted text analysis
   - Example: climate change discourse over time
3. Sentiment analysis
   - Example: Sochi Winter Olympics
4. Using social media data with environmental data
   - Example: Do people feel climate change?



## Social networks in research

- Twitter: One of the largest microblogging services (0.5 bil. Twitter accounts)
  - China: Sina Weibo (>0.2 bil. Users)
- Twitter is popular in research due to…
  - Information exchange on "What's happening now" vs. e.g. "What are you doing?" on Facebook
  - Twitter users can be treated as a distributed network of noisy sensors
  - Geotagged data
  - API interface allows easy access to the last 7 days of data
  - Drawbacks: archived data are expensive; fake accounts (noise); biases
- Facebook: unknown legal status; useful when data from few accounts are collected
- Instagram: API access rights changed in 2017. But data for purchase exist; high percentage is geotagged
- Flickr: falling popularity between general public; good for arched data
- Reddit - ?



| | |
|---|---|
| RT | Y |
| text | RT @AB_EmergAlert: Overland Flood Alert Updated Jun20 741AM Take necessary precautions. Black Diamond http://t.co/B6uO3E9qa7 #ABemerg #ABfl… |
| firstpost_date | 1371736404 |
| hashtags | abemerg#abflood |
| user_mentions | |
| in_reply_to_user_id_str | |
| retweet_count | 0 |
| favorite_count | |
| iso_language_code | en |
| user_name | Jason van Rassel |
| user_verified | Y |
| user_followers_count | 4323 |
| user_listed_count | 182 |
| user_statuses_count | 16337 |
| user_description | Calgary Herald justice/social issues reporter, Alberta Primetime panellist & access to info advocate. Also: runner, XC skier, Habs fan & craft beer evangelist. |
| user_utc_offset | -21600 |
| user_friends_count | 1076 |
| user_screen_name | JasonvanRassel |
| user_favourites_count | 1826 |
| user_created_at | Fri Feb 05 22:32:05 +0000 2010 |
| date time | 2013 – 6 – 20 13:53:30 |
| Domain | emergencyalert.alberta.ca |
| search | calgary%20alberta |
| placename | Calgary |
| countryCode | CA |
| adminCode | 01 |
| fcode | PPL |
| population | 1019942 |
| lat | 51.05011368 |
| lon | -114.085281 |
| geolat | 51.04551315 |
| geolon | -114.064598 |

**Message, hashtags: content analysis; sentiment analysis**

**Language**

**Author's influence and activity**

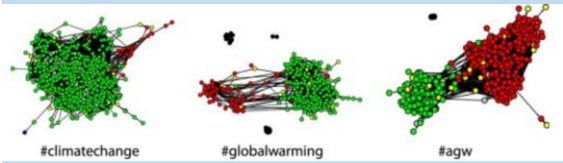
**Time**

**Geolocaton**



## More from Twitter

- The user on previous slide has 5860 followers and follows 1835 users. He published over 25K tweets.
- GET followers & friends -> social network
- Social Network Analysis (SNA): investigate the social structures using the graph theory. Example: Distribution of attitudes across the followers' network on Twitter. Dataset represents Twitter search results. Green: activist; red: sceptic. From: Williams et al., 2015






## Text Analytics and Sentiment Analysis

- Tweets contain unstructured text data
- Text analytics (content analysis): Extract meaning of the text or text parts. Main topics discussed.
- Sentiment analysis: extract expression of the entire text / parts / expression towards certain objects (positive, negative, neutral)

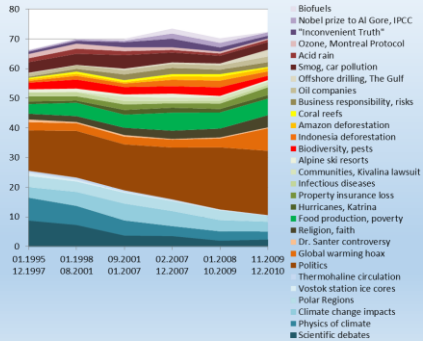
Text analysis

## Topic extraction basics

- We have a set of documents; each of these documents may discuss one or more topics. **Goal: extract these latent topics**
- Simple (and frequently used) bag-of-words approach (Harris, 1954):
  - Assumption: each topic is verbalized with specific words
  - In the medical papers, the topic "flu symptoms" will have words like 'headache' and 'fever' and the topic "flu treatment" will have words like 'aspirin' and 'rest'
- Create a global dictionary of "important" words from all processed documents
  - Important words are those that have high frequency word count but low frequency document count
- Make a matrix that show which words appear in each document
  - Simplest, a 0/1 matrix
  - Identify patterns in the matrix (that is, parts with high density of 1s)
  - Interpret these patterns as topics
- Data mining provides a wide set of tools to find the patterns
  - Clustering: based on a notion of distance; the documents with small distance are close one to another
  - Dimension reduction methods (Principal component analysis)
  - Latent Dirichlet Allocation: a very popular generative naïve Bayes approach

Text analysis

Example: changing discussion on climate change in The New York Times, 1995 – 2010 (Kirilenko et al., 2012)




Sentiment analysis

## Sentiment analysis

- Goal: extract a sentiment expressed in a document towards a certain aspect based on the subjectivity and the linguistic characteristics of the words within an unstructured text
- Basic task: identification of polarity: positive, neutral, or negative
- Lexical approach:
  - start with a set of words, for which a typical sentiment (positive, negative, or neutral) is defined. The sentiment of the entire textual unit is derived based on the balance of words with negative and positive sentiment and subject to linguistic rules
- Non-lexical approach:
  - based on machine learning, where an algorithm is trained on a thematically close text corpus
  - the sentiment is expressed differently in different types of content, e.g., in blogs and newspapers, which requires diverse algorithms

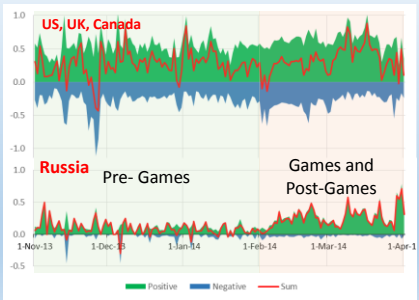
Sentiment analysis

## Example: sentiment towards Sochi Winter Olympic Games (Kirilenko et al, 2016)

- Research questions:
  - Compare hosts' and guests perspectives on the Games: Which issues are country-specific and influenced by political events?
  - Compare hosts' and guests sentiment prior and after the Games
- **Data:** 616,333 tweets spanning the period between November 1, 2013, and March 31, 2014.
- **Methods:**
  - **Content analysis**: main discussion topics were extracted with cluster analysis
  - **Sentiment analysis**: the emotions were extracted from the messages with SentiStrength

Sentiment analysis

## Pre-Games vs. Games/Post-Games sentiment in tweets




Mean normalized sentiment:

Pre-Games 1.12

Post-Games 1.29

No change:

Mean= -.92

Pre-Games 0.83

Post-Games 1.89

No change:

Mean= -.41

Social data with environmental data

## Twitter in natural sciences: People as sensors

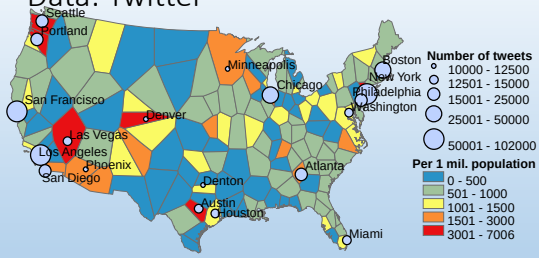
- Natural science: can we use observations of environment from the social network?
- Types of sensor networks :
  - Traditional: static, inert sensors designed to capture specific measurements of their local environments
  - Sensors carried by humans, vehicles, or animals
  - Humans themselves using their own senses
- 2010s: The spread of smartphones promotes unifying the human-sensors into networks exchanging georeferenced data
  - Volunteered geographic information (VGI) Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69 (4) 211–221
  - A concept of mining social media for geographical data

# Example: combining social media data and weather/climate data

- Hansen et al. (1998): objectively measured subjective climate change indicator, which can relate public feeling that the climate is changing to the observed meteorological parameters.
  - Never tested
- Research goal: do people living in the US connect their sensory experiences with local temperature to climate change?
- Research questions
  - Is the number of tweets on the topic of climate change/global warming positively associated with the changes in local weather conditions?
  - Is the number of tweets on the topic of climate change/global warming positively associated with the number of publications on the subject in mass media?
  - Could the observed variability be explained by media coverage alone?



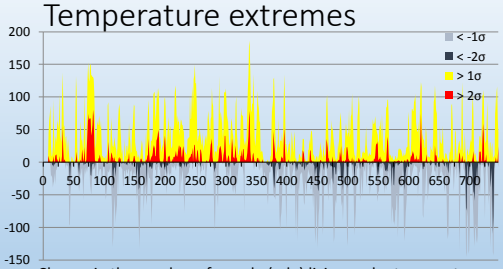
# Data: Twitter




Intensity of climate change tweeting. Shades show the annual number of tweets per 1 mln people; the areas are obtained as Voronoy tessellation with major US urban areas used as the seeds. The database contained 1,309,177 georeferenced tweets sent from the continental US during 105 weeks starting Monday, January 9, 2012.



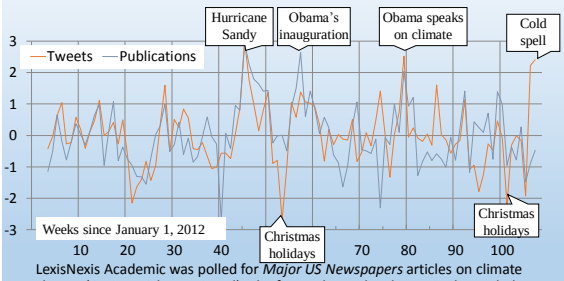
# Weather variable example: Temperature extremes




Change in the number of people (mln) living under temperature anomalies of one and two standard deviations below the climate norm temperature (below the abscissa) and above the norm temperature (above the abscissa).



# Data: Mass media




LexisNexis Academic was polled for *Major US Newspapers* articles on climate change (6,421 articles retrieved). The figure shows the change in detrended normalized weekly number of tweets and weekly number of publications on climate change. The explanatory text boxes are added to spikes exceeding 2σ.



# Study outcome

- At both national and local aggregation levels, high significance of the mass media and temperature variables in the majority of regression models confirmed
  - Both the weather and mass media coverage control public interest to the climate change
  - Substantial positive or negative local temperature anomalies do increase tweeting on climate change
- People are able to recognize local weather deviations from the climatic norm and attribute them to climate change



# Social Media Data Mining Quality Issues

The volume of data brings confidence in data mining results; however, the underlying assumptions in searching for patterns in data are not always valid

1. 1. Data quality: The data is not necessarily representative of the population we are interested in. Problems with bias, errors, noise
   - Dan Ness (MetaFacts): "A lot of big data today is biased and missing context, as it's based on convenience samples or subsets."
   - Fake social media

2. Data processing: the algorithms are not necessarily robust to give consistent results

3. Quality indicators. Sentiment analysis: accuracy, precision, sensitivity – comparable to human raters. Kohen's kappa: inferior to human raters. Knowledge of context is important.

4. The role of theory is not necessarily known: the theories in social science are formulated on snapshots of interactions of few dozen people; are they applicable to the observed interactions between millions of people?

## Summary

- The interest to text analytics (extraction of information from the text) is exploding
- Social network data is of a particular interest to practitioners since it allows to learn the opinion of large groups of people worldwide – unobtrusively, fast, and inexpensively.
- Text analytics is based on a variety of data methods originating in statistics, machine learning, and natural language processing. Practical significance of LDA and similar methods in topical analysis and machine learning in sentiment analysis.
- Social media data can augment environmental data
- Data quality control is a must

## Questions?