Computational Impact Assessment

Jana Diesner, PhD
Assistant Professor at the iSchool & Dept of Computer Science
University of Illinois Urbana-Champaign
Team

- From left to right in pic: Andrew Higgings, Phil Shubhanshu Mishra, GSLIS Sean Wilner, Informatics Kiumars Soltani, Informatics Jinseok Kim, GSLIS Liang Tao, Agricultural Eng. Amirhossein Aleyasen, CS
Measuring Impact of Social Justice Documentaries: Problem Statement

• **Goal of (Social Justice) Documentaries: Storytelling**
  – Create memories, imagination, sharing (Rose 2012)

• **Goal of funders and producers (Sundance Institute, Ford Foundation, BritDoc): Impact**
  – Evoke **change** in people’s knowledge and/or behavior (Barrett & Leddy 2008)

• **Common Approach and status quo:**
  – Big data: frequency counts of screenings and viewers
  – Thick data: small-scale, in-depth interviews with focus groups
  – Science: psychological effects of media on individuals
  – Strong need for **comprehensive, empirical, rigorous** impact assessment
Our Questions

• How can we know if a documentary or media product has what impact?
  – Computational, scalable, theoretically grounded
  – Generalized question: measure impact of information and media in terms of change

• At what point/ how early in the life cycle of a production can we answer this question?
  – Prediction models for impact trajectories

• Usefulness for filmmakers and producers
  – Strategic allocation of limited resources for outreach and campaign work
  – Leverage existing social capital and discourse
Approach: A story of microscopes and telescopes

- Assumption: documentaries produced, screened, watched as part of larger, dynamic ecosystems of stakeholders and information flow

- Method: identify, map, monitor, analyze social (stakeholders) and semantic (information) networks to study their structure, functioning and dynamics
<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>LEVEL</th>
<th>INDEX</th>
<th>ANALYTICS</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENT</td>
<td></td>
<td>Guiding</td>
<td>Description</td>
<td>Report by producers or funding agencies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factor</td>
<td>Ranking</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weighting</td>
<td></td>
</tr>
<tr>
<td>EXPECTED OUTCOME</td>
<td>OFFLINE</td>
<td>Outreach</td>
<td>Stats</td>
<td>Number of movies, CDs distributed</td>
</tr>
<tr>
<td></td>
<td>ONLINE</td>
<td></td>
<td></td>
<td>Number of theatrical, Internet release</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Duration of release; Sales of product</td>
</tr>
<tr>
<td>EVALUATION PRIORITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RESOURCE</td>
<td>MEDIUM</td>
<td>Mass Media</td>
<td>Web Mining</td>
<td>Frequency of news coverage weighted by influence (article, opinion/editorial)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attention</td>
<td></td>
<td>Domestic, international broadcast</td>
</tr>
<tr>
<td></td>
<td>USER MEDIA</td>
<td>User Media</td>
<td>Web Mining</td>
<td>Frequency of talking about, links included, user-created contents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Attention</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PROFESSIONAL</td>
<td>Prestige</td>
<td>Web Mining</td>
<td>Number of festival acceptance</td>
</tr>
<tr>
<td>MEDIUM</td>
<td></td>
<td></td>
<td>Survey, Interview</td>
<td>Number of reviews</td>
</tr>
<tr>
<td></td>
<td>RESPONSE MEDIUM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>INTERPERSONAL</td>
<td>Intimate</td>
<td>Web Mining</td>
<td>Number of advocacy communities, colleges, schools, or NGOs</td>
</tr>
<tr>
<td>INTERACTION</td>
<td>OFFLINE</td>
<td></td>
<td>Survey, Interview</td>
<td>Number of advocacy communities, colleges, schools, or NGOs</td>
</tr>
<tr>
<td></td>
<td>ONLINE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TARGET</td>
<td></td>
<td>Diversity</td>
<td>Web Mining</td>
<td>Number of viewers or visitors</td>
</tr>
<tr>
<td></td>
<td>SINKER</td>
<td>Passiveness</td>
<td>Web Mining</td>
<td>Number of inactive viewers</td>
</tr>
<tr>
<td>HOMOSEGENEITY</td>
<td></td>
<td>Diversity</td>
<td>Web Mining</td>
<td>Geography &amp; demography: location, age, gender, education, income</td>
</tr>
<tr>
<td></td>
<td>TRANSMITTER</td>
<td>Leadership</td>
<td>Web Mining</td>
<td>Number of opinion leaders</td>
</tr>
<tr>
<td>AUDIENCE SIZE</td>
<td></td>
<td>Reachability</td>
<td>Web Mining</td>
<td>Number of viewers or visitors</td>
</tr>
<tr>
<td>AUDIENCE TYPE</td>
<td></td>
<td></td>
<td>Survey, Interview</td>
<td></td>
</tr>
<tr>
<td></td>
<td>COLLECTIVE ENTITY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>INDIVIDUAL</td>
<td>Advocacy</td>
<td>Web Mining</td>
<td>Number of advocacy communities, colleges, schools, or NGOs</td>
</tr>
<tr>
<td>IMPACT</td>
<td>COMMUNAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEHAVIORAL</td>
<td>GLOBAL</td>
<td>Awareness</td>
<td>Web Mining</td>
<td>Frequency of positive, negative, neutral sentiments of comments</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Survey, Interview</td>
<td>Personal, critics, mass media, and organizational responses</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reaction to calls for action</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sentiment</td>
<td>Web Mining</td>
<td>How well connected</td>
</tr>
<tr>
<td>ATTITUADINAL</td>
<td></td>
<td></td>
<td>Survey, Interview</td>
<td>How much &amp; far disseminated</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>How centralized is the impact</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The route of diffusion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Number of action pledges</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Alliance and allied action of organization</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Discussion or decision by organizational, governmental, international</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>policy/legislation makers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sponsorship of bills, adoption, donation, funding, implementation,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Social movement or intervention</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEMPORAL</td>
<td></td>
<td>Impact</td>
<td>Web Mining</td>
<td>Comparison b/w multiple time points</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dynamics</td>
<td>Survey, Interview</td>
<td>Duration of impact</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increase vs. decrease</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Change vs. stability vs. reinforcement</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Introduction or shifts of topics</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detection of social norm change</td>
</tr>
</tbody>
</table>

This is no computational fishing expedition.
We have theory: CoMTI Framework

Scientific Logic

Baseline

Social Structure

Meta Data

Content

Reality/Change

Ground truth Transcript

Content

Social Structure

Meta Data

Content

Social Structure

Meta Data

Content

Social Structure

Meta Data

Content

Movie

Theme

Technology: ConText

http://context.lis.lis.illinois.edu
Technology: ConText

- Social Networks of agents (FB, Twitter, YouTube)
- Number, type and quality of social agent
- Semantic networks of content
- Disambiguate data
- Create curated dataset
- Create meta-data database
- Construct semantic networks

Text Mining & Natural Language Processing:
  - Summarization
    - Corpus Statistics
    - Topic Modeling
    - Sentiment Analysis
  - Pre-Processing
    - Stemming
    - Stop Word Removal
    - Parts of Speech Tagging
  - Codebook construction and Application
  - Entity and Relation Extraction

(http://context.lis.illinois.edu/)
Case studies & Lessons Learned

- **House I Live in** (Eugene Jarecki 2012)
  - Mandatory minimum sentencing
- **One Mile Away** (Penny Woolcock 2012)
  - Transforming inner city youth gangs
- **Pandora's Promise** (Robert Stone 2012)
  - Nuclear *and* sustainable energy
- **Solar Mamas** (Eldaief & Noujaim ‘12)
  - Technical education of women in developing world
  - Part of “Why Poverty Series”
From Raw Data to Actionable Knowledge
Media Data:
From Raw Data to Actionable Knowledge
Opportunity Space (Baseline Model): News Coverage of Theme: “education” + “women empowerment”

Core issues: “children”, “women”, “development”, “poverty and homelessness”

News Coverage of “Solar Mamas”

- Mainly: “documentary films”, “movie industry”, “festival”
- And a little bit of: “children”, “poverty and homelessness”, “women”
energy
year
industry
job
support
environmental
include
clean
company
subsidy
green
nuclear
investment
government
bill
electricity
target
policy
Britain
work
turbine
people
change
carbon
time
make
power
renewable
wind
climate
time
future
people
change
carbon
fuel
gas
Energy
year
industry
project
energy
clean
documentary
movie
write
film
director
star
minute
Michael
find
James

Issue before
after

Transcript
Press on Film
Social Media: Facebook: Social Network of Co-commentors

72% are male, 80% are from U.S.
Ian Dechert: CEO at Safe Energy Association
Social Media: Facebook:
Semantic Network of Posts (Stimulus)
Social Media: Twitter:
Semantic Network of Comments (Response)

Followees: 1,549
(Minus intersection)

Followers: 1,283
(Minus intersection)

22: NGO (9), media (8), energy companies (5)
Strongest:
International Justice Mission 82K,
ABCEnvironment 22K

Reciprocated Followers: 445
(Intersection)
Text Mining: What node classes to consider?

- **Who?** (people, groups)
- **Where?** (places)
- **Why?** (beliefs, sentiments, mental models)
- **What?** (tasks, events)
- **How?** (resources, knowledge)
- **When?** (time)

Conflicts RoE ICT Mission NATO Vessel
How to find and categorize nodes in text data?

- Sequences of $x$ (words) and $y$ (label)
  $P(x, y)$: generative models, e.g. Hidden Markov Model (HMM).
- $P(y|x)$: conditional models, e.g. Maximum Entropy Markov Models (MEMM) and Conditional Random Fields (CRF).

- CRF:
  - Consider **arbitrarily large bag of features**
  - Consider and **any property of $x$**, incl. long-range features
How to find and categorize nodes in text data?

- Model relationship among hidden states \( (y) \) as **Markov Random Field** (MRF) conditioned on observed data \( (x) \) (Lafferty et al. 2001)
- Compute **conditional distribution** of entity sequence \( y \) and observed sequence \( x \) as normalized product of potential functions \( M_i \):

\[
M_i(y_{i-1}, y_i | x) = \exp \left( \sum_{\alpha} \lambda_{\alpha} f_{\alpha}(y_{i-1}, y_i, x) + \sum_{\beta} \mu_{\beta} g_{\beta}(y_i, x) \right)
\]

- **Edge and transition features** plus **node and emission features**
- \( f, g \): boolean feature vectors with **learned** weights
- Tool: CRF project page, training data: BBN
Solution to node identification and classification

- Parallelized version of CRF base implementation
  - publicly available
- Currently training models, stay tuned...
Next steps

• Impact assessment of broader range of information
  – Feature stories
  – Books
  – Articles

• Additional sources
  – Legal data
  – Corporate reports

• After all those case studies
  – Prediction models of likely trajectory of impact
Acknowledgement and Publications for Impact Assessment

• Ford Foundation 0125-6162, 0145-0558
• Start-up allocation award from Extreme Science and Engineering Discovery Environment (XSEDE)
Thank you!

Q&A

- For questions, comments, feedback, follow-up:
  Jana Diesner
  Email: jdiesner@illinois.edu
  Phone: (412) 519 7576
  Web: http://people.lis.illinois.edu/~jdiesner