Small Data to Big Data - Coding and Culture with Social Media Data

Dhiraj Murthy
The University of Texas at Austin
@dhirajmurthy
Objectives

• There are unique challenges associated with data collection and analysis on social media platforms

• How do we integrate and weigh Big Data questions with more in-depth contextualized analysis of social media content?

• How do we categorize textual and visual content, addressing issues of ontology?

• How can we scale small data to big data?
Starting points

• Big data methods successfully applied to social media data (indeed 16% of research on Twitter employed sentiment analysis (Zimmer and Proferes 2014)

• There are also challenges associated with data collection and analysis on social media (boyd & Crawford, 2012).

• Closed coding systems are thought to be the best for studying social media data

• However, social media data involves very ‘messy’ elements and mixed approaches can have high utility
Beyond induction and deduction…

- **Abductive methods**: a form of reasoning ‘for finding the best explanations among a set of possible ones’ (Paul, 1993) are alternative approach.

- **Retroduction**: a type of abductive method that emphasizes “asking why” (Olsen, 2012: 215), researchers are able to probe the data regularly and to “avoid overgeneralisation but searching for reasons and causes” (p. 216) instead.

- Or put another way, “the retroductive researcher, unlike the inductive researcher, has something to look for” (Blaikie, 2004).
Methods

Emergent methods can be implemented operationally in a systematic fashion to build critical, reflective, conceptual knowledge of Social media-derived data.

Theory building, Adapted from Goulding, C. (2002), Grounded Theory: Sage, p. 115
The Case of ISIS

• Using ‘fuzzy’ methods

• I go from large scale data mining
  – to small scale human coding
  – to generalizable characteristics
  – with potential applications to big data methods
Do you have an Instagram account?

Login to Instagram to like or comment on this media
• ISIS uses social media as a key part of its recruiting strategy according to journalistic accounts

• For example, a Western ISIS defector named Ibrahim stated in an interview, “A lot of people when they come, they have a lot of enthusiasm about what they’ve seen online or what they’ve seen on YouTube” (Freytas-Tamura, 2015). *

• YouTube’s search and recommender algorithms may be inadvertently directing users to radical content whether the user’s search is actively seeking radical content. We can mine data but we can not effectively generalize attributes.

• **Research questions:** Given that Islamic state content is regularly deleted, but also reposted, what types of videos recommend radical videos as part of the recommended video algorithm? What linguistic or other markers are present? What methods can effectively be used for classification?
Case study methods

• News articles used to obtain the titles for 15 videos attributed to ISIS’ media wing, Al Hayat
• 11 videos were found through searches on YT and used as ‘seeds’ to crawl the YT network
• Resulting network represents videos recommended as well as videos recommended by the recommended videos
• Qualitative Content Analysis used to build a framework for recommending videos
<table>
<thead>
<tr>
<th>Video ID</th>
<th>Title</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQ9b6xmW1ZE</td>
<td>Mujatweets 6</td>
<td></td>
</tr>
<tr>
<td>KySHORzZu90</td>
<td>Mujatweets 2</td>
<td></td>
</tr>
<tr>
<td>9mKmnJotOTM</td>
<td>Islamic Caliphate Mujatweets Raqqah Market</td>
<td></td>
</tr>
<tr>
<td>toZ6m9lftKI</td>
<td>Mujatweets 3</td>
<td></td>
</tr>
<tr>
<td>b1WWFw3zin0</td>
<td>Mujatweets4</td>
<td></td>
</tr>
<tr>
<td>k2MupZqJBu4</td>
<td>The Religion of Kufr Is One - AlHayat Media Center</td>
<td></td>
</tr>
<tr>
<td>T2VjhJOQwnc</td>
<td>ISIS VIDEO Flames of War Trailer 720p</td>
<td></td>
</tr>
<tr>
<td>MSNCPj7s6vl</td>
<td>ISIS - The End of the Sykes-Picot Agreement</td>
<td></td>
</tr>
<tr>
<td>5tYiqWCTKY</td>
<td>No Respite FR MP4 360</td>
<td>French Language</td>
</tr>
<tr>
<td>Ta75NNb6MQ8</td>
<td>ISIS発布最新影片 驚見中華民国國旗!</td>
<td>Uncut English language ‘No Respite’ Video posted by China Television Corporation – returned by a search for ‘No Respite’</td>
</tr>
<tr>
<td>OPl9gLlAQdY</td>
<td>ISIS داعش تحدي العالم -</td>
<td>Arabic language 'No Respite'</td>
</tr>
<tr>
<td></td>
<td>Deterring the Hirelings</td>
<td>Not found on YouTube</td>
</tr>
<tr>
<td></td>
<td>Deterring the Hirelings 2</td>
<td>Not found on YouTube</td>
</tr>
<tr>
<td></td>
<td>Clanging of the swords</td>
<td>Not found on YouTube</td>
</tr>
<tr>
<td></td>
<td>Upon the prophetic methodology</td>
<td>Not found on YouTube</td>
</tr>
</tbody>
</table>

@dhirajmurthy
Case study methods

• This two degree network was stored as a directional network and is composed of 15,021 nodes and 190,087 edges

• Data was collected via YouTube Data Tools (Rieder)* in October 2016

• Gephi was used to visualize the network and observe what videos were recommending radical ISIS videos

YouTube Algorithms

• YouTube’s selection of what content is displayed in search results and as ‘recommended’ content is part of a black box

• We can only try to reverse engineer as we do not have access to YouTube’s proprietary search data

• Therefore, we do not know what percentage of users explicitly search for ISIS content vs. ‘accidental’ exposure
Results
ISIS’s embedding into YouTube

How can we combine big data, with ‘small data’ and then big data again?

@dhirajmurthy
• Zoom of ‘No Respite’ ISIS video; black nodes represent official Al Hayat ISIS videos and blue nodes do not; size of nodes is views; Arabic title names are translated using Google translate.
What linguistic, cultural, or other attributes do recommending videos have?

I used Qualitative Comparative Analysis (QCA) to code:

• All official ISIS videos collected that recommend ISIS content along with random samples of (1) videos collected that do not recommended radical content and (2) videos from YouTube

• Coded these for 11 attributes
A generalizable set of characteristics derived

• QCA requires many iterations using qualitatively coded attributes to see possible explanations. It is not just a 1-click process.

• The human in the loop needs to find where the fs/QCA algorithm (Ragin & Davey, 2009) makes sense
A generalizable set of characteristics derived

raw unique

<table>
<thead>
<tr>
<th></th>
<th>coverage</th>
<th>coverage</th>
<th>consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>radical_keyword</td>
<td>0.346939</td>
<td>0.244898</td>
<td>1</td>
</tr>
<tr>
<td>newscast<del>organization</del>arabic~recent</td>
<td>0.0408163</td>
<td>0.0204082</td>
<td>1</td>
</tr>
<tr>
<td>newscast<em>organization</em>english*recent</td>
<td>0.285714</td>
<td>0.244898</td>
<td>0.82352</td>
</tr>
<tr>
<td>explicit_reference<del>organization</del>arabic~recent</td>
<td>0.0612245</td>
<td>0.0408164</td>
<td>1</td>
</tr>
<tr>
<td><del>newscast*explicit_reference</del>organization*english</td>
<td>0.0612245</td>
<td>0.0408164</td>
<td>1</td>
</tr>
</tbody>
</table>

solution coverage: 0.693878
solution consistency: 0.918919
Why my approach?

- There is a lack of accessible examples of mixed methods to collect and interpret complex social media data

- We tend to either have big data or small data approaches and the two are not connecting synergistically

- Small data methods have further utility if they can scale up