Content Analysis and Natural Language Processing in the Social Media Era

Xinran (Joyce) Wang, Ph.D.

University of Missouri
Why is it important to learn content analysis of social media data (e.g., tweets)?

Traditional media data vs. social media data
Low
Massive, noisy, fragmented, exaggerated
High
Quick! You’re in a room with no key, a chair, two paper clips, and a lightbulb. How do you defraud investors? | #AskJPM @jpmorgan #Anonymous
6 steps
Step 1: Identify social media sources based on research interests

- social network (e.g., Facebook),
- video-sharing (e.g., YouTube),
- photo-sharing (e.g., Flickr),
- product and service review (e.g., Yelp),
- Emotions (e.g., Twitter).
Essay 2 of my dissertation

How do **social media coverage**, **national animosity**, and **nationalism** influence the diffusion of social disapproval from a home country to a host country?
Step 2: Identify key search terms

Brand name, firm name, activities, events, and emotions related words.

My study: The screennames of MNEs (e.g., 3m).
Here, we innovate with purpose & use #science every day to create real impact
Step 3: Write quarries to mine data

@3m OR to:3m OR from:3m OR #:3m
Step 4: Reorganize raw data

ISO_language_code: en

created_at: "Fri May 31 06:16:35 +0000 2013",
"id":340351078228971521,
"id_str": "340351078228971521",
"text": "#missyou #rip #grandpa #untilwemeetagain #guardianangel http:\/\/t.co\/ZGV4N6qqwV",
"source": "\u003ca href="http:\/\/www.apple.com" rel="nofollow\u003ePhotos on iOS\u003c/a\u003e",
"truncated":false,"in_reply_to_status_id":null,
"in_reply_to_status_id_str":null,
"in_reply_to_user_id":null,
"in_reply_to_user_id_str":null,
"in_reply_to_screen_name":null,
"user": {
"id":212277942,
"id_str": "212277942",
"name": "Jennara \u2741 Grandis",
"screen_name": "jgrandis",
"location": "New York",
"description": "",
"url":null,
"entities":{"description":{"urls":[]}}
,"protected":false,
"followers_count":35,
"friends_count":156,
### Step 4: Reorganize raw data

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>Stakeholder_name</th>
<th>Geo_location</th>
<th>Firm</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/3/11 1:43 AM</td>
<td>maartjemutsaers</td>
<td>Tilburg</td>
<td>3M</td>
<td>@3m Destine bij #3fm, o wat heb ik zin in 16 oktober @0</td>
</tr>
<tr>
<td>8/3/11 6:06 AM</td>
<td>Mohab11</td>
<td></td>
<td>3M</td>
<td>@3m 7azmbol : mubarak el mothag te7eb te2ol eh le mo</td>
</tr>
<tr>
<td>8/4/11 5:16 AM</td>
<td>kobusvanniekerk</td>
<td>Johannesburg, South Africa</td>
<td>3M</td>
<td>@3M - where ideas multiply...have a good idea. Duck tape</td>
</tr>
<tr>
<td>8/4/11 8:46 AM</td>
<td>bulldawgmktginc</td>
<td>Moorsville, NC</td>
<td>3M</td>
<td>Be sure to stop by the @3M display @Iowa Speedway this weekend</td>
</tr>
<tr>
<td>8/4/11 1:57 PM</td>
<td>MamaRiceCake</td>
<td>Lower Alabama (the OTHE)</td>
<td>3M</td>
<td>Back to School ain't happenin' at my house without Post-Iowa!</td>
</tr>
<tr>
<td>8/4/11 10:07 PM</td>
<td>themommyfiles</td>
<td>Pismo Beach, California</td>
<td>3M</td>
<td>Hoping Advil kicks in because I have a killer headache right now</td>
</tr>
<tr>
<td>8/5/11 1:36 PM</td>
<td>mikecook49</td>
<td>Minnesota</td>
<td>3M</td>
<td>Jay Haas aces fourth hole in opening round of @3M Championship schools</td>
</tr>
<tr>
<td>8/5/11 3:42 PM</td>
<td>psujewels</td>
<td>Lehigh Valley PA</td>
<td>3M</td>
<td>Thank u @3M for my back to school kit I heart post it</td>
</tr>
</tbody>
</table>
Step 5: Process data

Cleaning

1. Generate the plain English text excluding hashtags, screennames and URLs.

2. Clean the plain English texts by reducing them to the lower case, removing numbers and punctuations.

3. **Stem** each word to its root form.
Step 5: Process data

Match conversation

1. Whether a firm is the author (i.e., who posted the tweet) or recipient (i.e., being asked)?

2. Classify each tweet as “in” (i.e., incoming from a stakeholder to a firm) or “out” (i.e., outgoing from a firm to a stakeholder).

3. In the tweets classified as “in,” identify they are responses or initial posts.
Step 6: Generate variables

**Event:** Topic modeling

**Speed:** Timestamps

**Emotionality:** “afinn” indices. Finn Arup Nielsen (2011) "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs"

**Communality:** Communication network built by tweets and retweets.

**Country:** Identify countries of twitter users’ self-reported locations (“NYU,” “New York” or “New York University” → USA)
Challenges and opportunities

Overwhelming amount of data; Multiple languages; Lack context

The spread of fake tweets, celebrity tweets

and political tweets.
Thank you!

Any questions please contact wangxinr@missouri.edu
#McDStories Take a McDonalds fry, let it sit for 6 months. It will not deteriorate or spoil like a normal potato. It will remain how it was.
DATA & SAMPLES

MNE major overseas ownership (SDC M&A and JV databases) (Li et al., SMJ, 2017; Xia, SMJ, 2011; Makino & Beamish, JIBS, 1998).

Negative events (RavenPack) (Dang, JFE, 2015; Dai et al., JAR, 2015)

Blogs and Twitter posts (RavenPack & Twitter) (Hewett et al., J of Marketing, 2016; Ma et al., MS, 2015)

SAMPLE SIZE

Unit of analysis: national dyadic observations of a negative MNE event.
32,007 firm-event-national dyadic observations.
482 US-based MNEs, and 48 host countries during 2007 to 2014.
9,699,177 tweets and 186,937 blog posts.
One time I walked into McDonalds and I could smell Type 2 diabetes floating in the air and I threw up. #McDStories