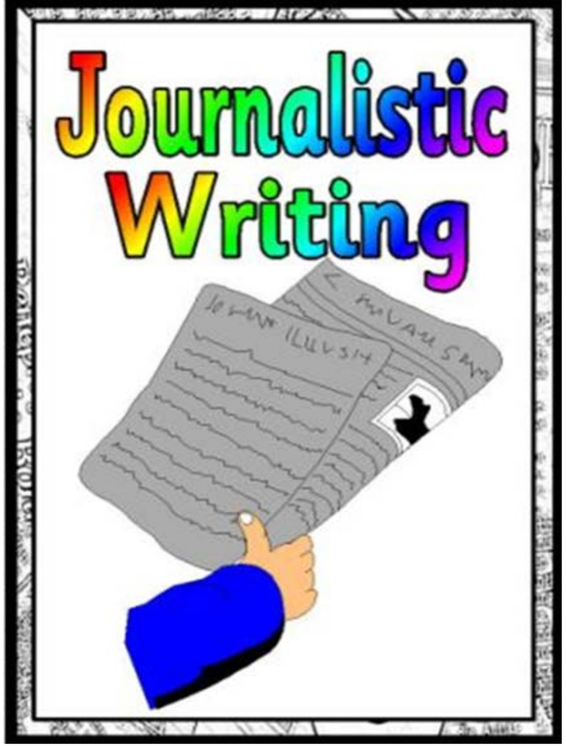
Content Analysis and Natural Language Processing in the Social Media Era

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

High

Massive, noisy, fragmented, exaggerated

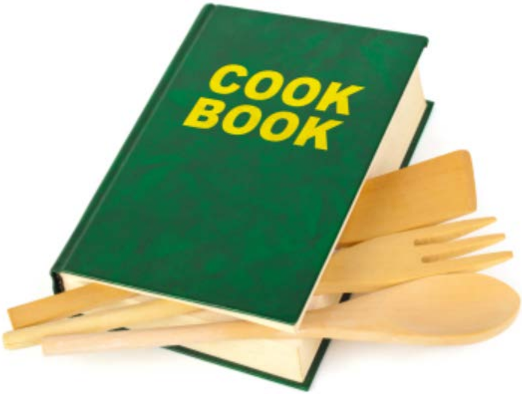
V
@TheAnon0ne

Follow

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

Reply Retweet Favorite More

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).**

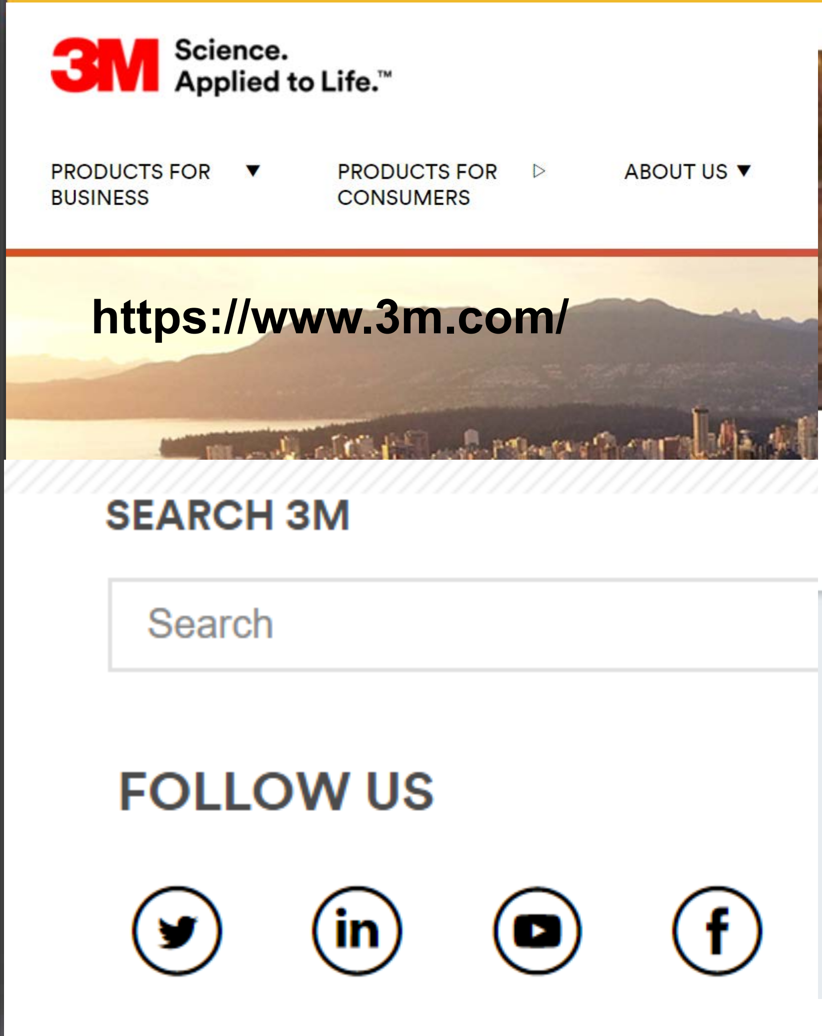
Essay 2 of my dissertation

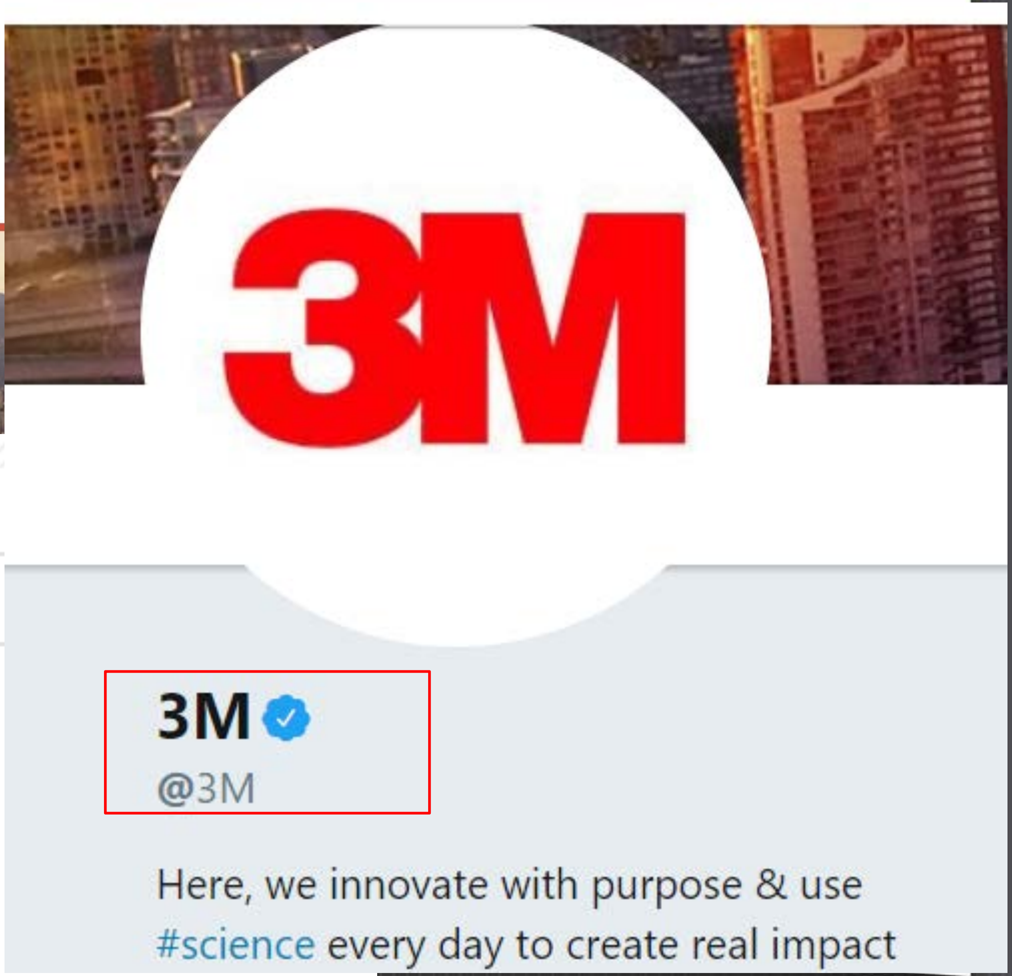
How do ***social media coverage***, ***national animosity***, and ***nationalism*** influence the diffusion of social disapproval from a home country to a host country?

Step 2: Identify key search terms

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

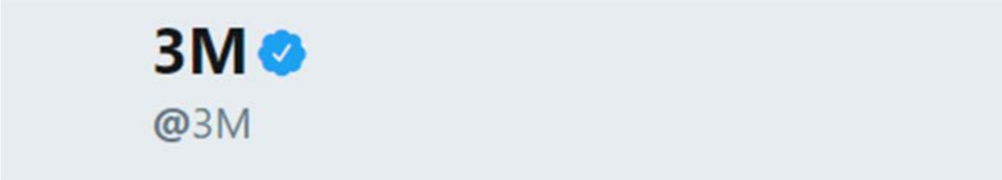
My study: The screennames of MNEs (e.g., 3m).


Science.
Applied to Life.™
PRODUCTS FOR BUSINESS
PRODUCTS FOR CONSUMERS
ABOUT US
https://www.3m.com/
SEARCH 3M
Search
FOLLOW US


3M
@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



From: Jun 08, 2013 00:00

To: Aug 06, 2013 23:59

Filters: to:3dsystems OR to:3m OR to:aarcorp OR to:abbottglobal OR to:aci_worldwide OR to:adi_news OR to:ADMupdates OR to:adtran OR to:advenergy OR to:AdventSoftware OR to:aecom OR to:agcocorp OR to:agilent OR to:aiginsurance OR to:airproducts OR to:akamai OR to:alere OR to:Allscripts OR to:alphatecspine OR to:altramotion OR to:amazon OR to:amd OR to:americanaxle OR to:americanexpress OR to:amerigas OR to:ametekinc OR to:ansys OR to:Aon_plc OR to:ap OR to:apachecorp OR to:aplncdsgn OR to:apogeeglass OR to:applied_ind OR to:aptar OR to:arrow_dot_com OR to:ashlandinc OR to:AskCiti OR to:askmastercard OR to:aspentech OR to:autodesk OR to:avid OR to:avnet OR to:azzincorporated OR to:badger_meter OR to:baldwin_corp OR to:ballcorphq OR to:baxter_intl OR to:bdandco OR to:BeaconCareers OR to:BeldenInc OR to:BemisCompanyInc OR to:bettykdevita OR to:bhinc OR to:BigLots OR to:biogen OR to:biomedrealty OR to:blackbaud OR to:blackboard OR to:blackrock OR to:blackstone OR to:BlountCareers OR to:bmcsoftwar OR to:bmsnews OR to:borgwarner OR to:bostonsci OR to:BriggsStratton OR to:brighthorizons OR to:brinks OR to:broadridge OR to:bruker OR to:cabotcorp OR to:Cainc OR to:calgoncarbonccc OR to:calliduscloud OR to:cameron_intl OR to:CapellaCareers OR to:carters OR

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\/ZGV4N6qq
"source":"\u003ca href=\"http:\/\/www.apple.com\" rel=\"nofollow\"\u003ePhotos on iOS\u
"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,
```

Step 4: Reorganize raw data

Timestamp	Stakeholder_name	Geo_location	Firm	Text
8/3/11 1:43 AM	maartjemutsaers	Tilburg	3M	@3m Destine bij #3fm , o wat heb ik zin in 16 oktober @0
8/3/11 6:06 AM	Mohab11		3M	@3m 7azmbol : mubarak el motham te7eb te2ol eh le mo
8/4/11 5:16 AM	kobusvanniekerk	johannesburg, South afric	3M	@3M - where ideas multiply...have a good idea. Duck tap
8/4/11 8:46 AM	bulldawgmktginc	Moorseville, NC	3M	Be sure to stop by the @3M display @Iowa Speedway thi
8/4/11 1:57 PM	MamaRiceCake	Lower Alabama (the OTHE	3M	Back to School ain't happenin' at my house without Post-I
8/4/11 10:07 PM	themommyfiles	Pismo Beach, California	3M	Hoping Advil kicks in because I have a killer headache righ
8/5/11 1:36 PM	mikecook49	Minnesota	3M	Jay Haas aces fourth hole in opening round of @3M Cham
8/5/11 3:42 PM	psujewels	lehigh valley pa	3M	Thank u @3M for my back to school kit I heart post it http

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 twitters 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

CATE STORM

@Cate_Storm

Follow

#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

3:22 PM - 18 Jan 2012

21 RETWEETS 1 FAVORITE

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 .

Skip Sullivan
@SkipSullivan

Follow

One time I walked into McDonalds and I could smell Type 2 diabetes floating in the air and I threw up. #McDStories

50+ RETWEETS | **12** FAVORITES

3:30 PM - 18 Jan 12 via Twitter for iPhone · Embed this Tweet

Reply Retweet Favorite