Other Social Media Event Data Topics

Zachary Steinert-Threlkeld

Luskin School of Public Affairs, UCLA

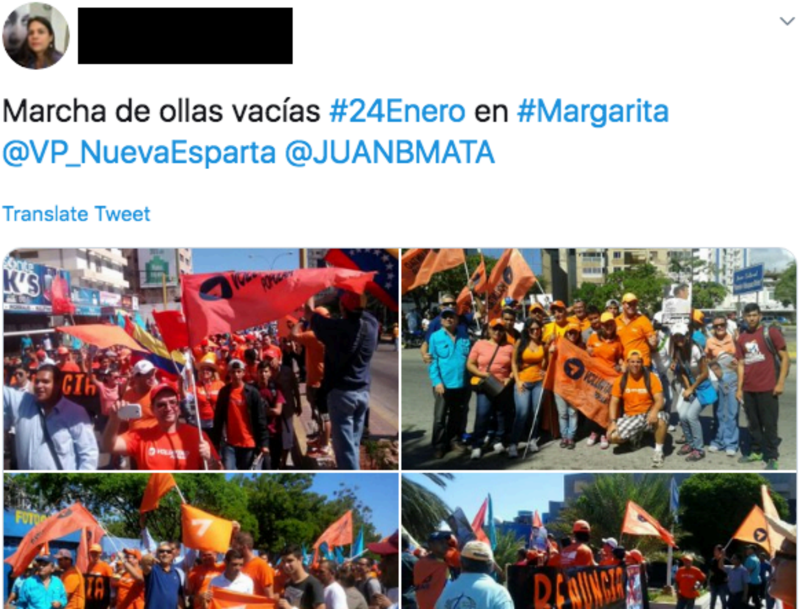
09.08.2020

1. INTRODUCTION
2. IMAGES
3. DATA INTEGRITY
4. OTHER SOURCES
5. MISCELLANY

INTRODUCTION

IMAGES

MOTIVATION



[redacted]

Marcha de ollas vacías #24Enero en #Margarita @VP_NuevaEsparta @JUANBMATA

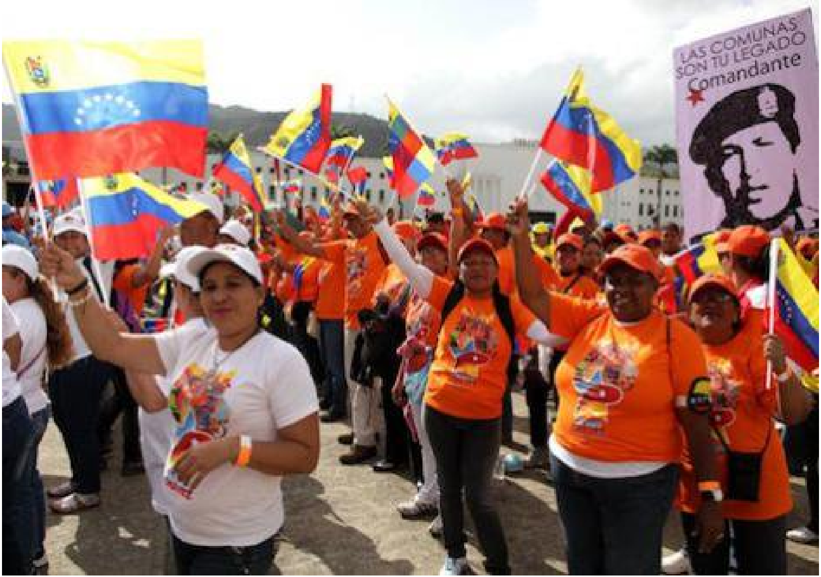
MOTIVATION

Improved or New Measures

1. Magnitude
    - Size (count faces in photographs)
    - Violence (protester, state)
2. Protester diversity (age, gender, race)
3. Emotions

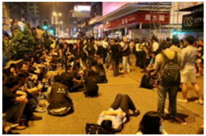
EXAMPLE

New event: Merida, 02.04.2015

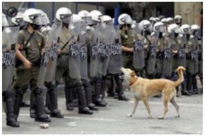


EXAMPLE

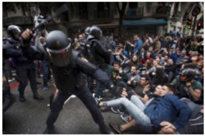
State violence continuously valued



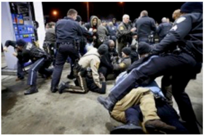
Hong Kong .107



Seoul .264



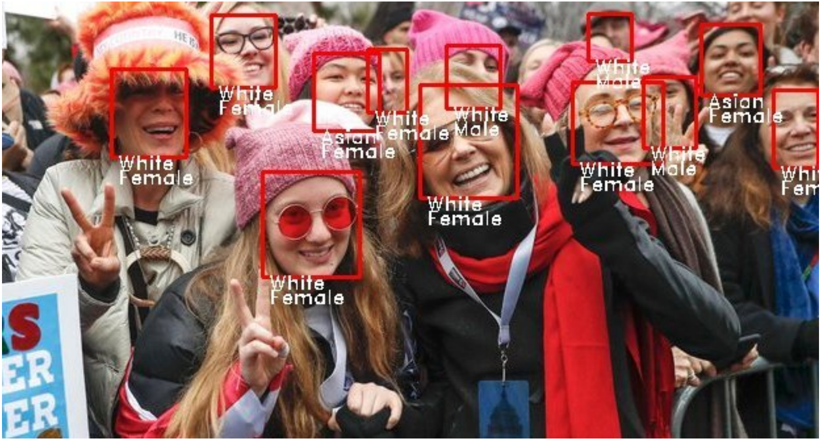
Barcelona .625



Caracas .849

EXAMPLE

Demographics



MULTIMODAL

Combining text and images into one dataset is a particularly promising approach.

1. Social media text particularly suited to measuring emotion.
2. Use other datasets, such as ACLED or MMAD, to identify events.

IMAGE ANALYSIS READING

Williams, Nora Webb, Andreu Casas, and John D. Wilkerson. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification.* Elements in Quantitative and Computational Methods for the Social Sciences (2020).

Joo, Jungseock, and Zachary C. Steinert-Threlkeld. "Image as data: Automated visual content analysis for political science." arXiv preprint arXiv:1810.01544 (2018).

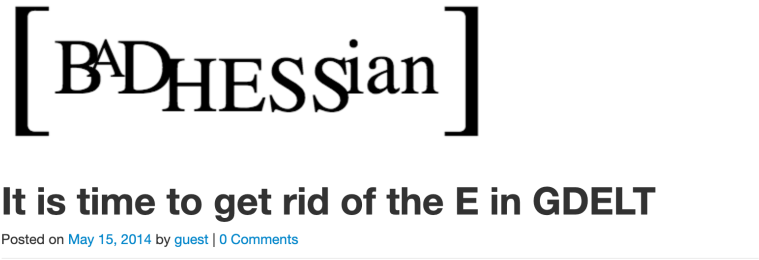
DATA INTEGRITY

DEDUPLICATION

Ensuring that event data do not record the same event multiple times is one of the most difficult issues in creating event data.

Will otherwise exacerbate newspapers' tendency to swarm to event.

DEDUPLICATION

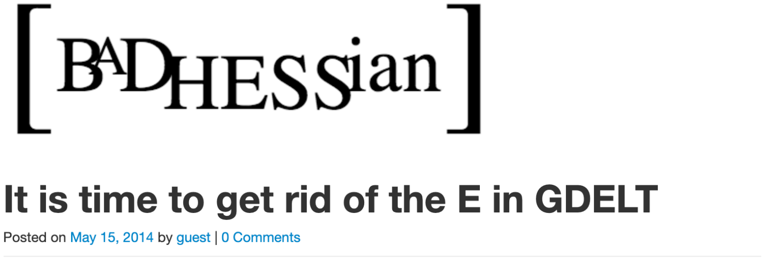


[BADHESSian]

It is time to get rid of the E in GDELT

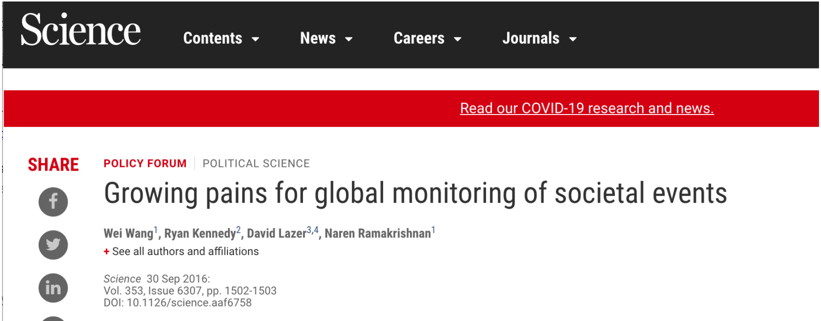
Posted on May 15, 2014 by guest | 0 Comments

DEDUPLICATION



It is time to get rid of the E in GDELT

Posted on May 15, 2014 by guest | 0 Comments



POLICY FORUM | POLITICAL SCIENCE

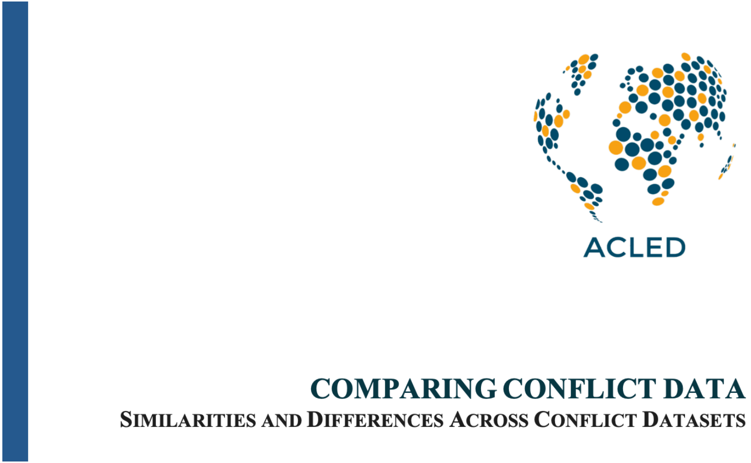
Growing pains for global monitoring of societal events

Wei Wang[1], Ryan Kennedy[2], David Lazer[3,4], Naren Ramakrishnan[1]

+ See all authors and affiliations

Science 30 Sep 2016:
Vol. 353, Issue 6307, pp. 1502-1503
DOI: 10.1126/science.aaf6758

DEDUPLICATION



ACLED

COMPARING CONFLICT DATA

SIMILARITIES AND DIFFERENCES ACROSS CONFLICT DATASETS

DEDUPLICATION - NOT SOCIAL MEDIA

Addressed

1. Use one source
2. Manual or hybrid methodology
3. Very conservative automatic coding in ICEWS, TERRIER

TWITTER DEDUPLICATION

1. Remove retweets (not an issue with geolocated tweets)
2. Remove duplicate images
3. One event type per city day
4. Humans in the loop

BOTS

Maybe bots rule social media.

Or maybe they do not.

BOTS - SOLUTIONS

1. Botornot

BOTS - SOLUTIONS

1. Botornot
2. Discard accounts tweeting the most

BOTS - SOLUTIONS

1. Botornot
2. Discard accounts tweeting the most
3. Join date

BOTS - SOLUTIONS

1. Botornot
2. Discard accounts tweeting the most
3. Join date
4. Following:follower ratio

BOTS - SOLUTIONS

1. Botornot
2. Discard accounts tweeting the most
3. Join date
4. Following:follower ratio
5. **Bots do not appear to geotag their tweets often.**

BOTS - MY EXPERIENCE

1. Appears to be much less prevalent in geotagged data.

BOTS - MY EXPERIENCE

1. Appears to be much less prevalent in geotagged data.

BOTS - MY EXPERIENCE

1. Appears to be much less prevalent in geotagged data. Including with hand coding.
2. Good reasons to exclude "special" (# followers, most common) account anyway.

BOTS - MY EXPERIENCE

1. Appears to be much less prevalent in geotagged data. Including with hand coding.
2. Good reasons to exclude "special" (# followers, most common) account anyway.
3. Unclear how serious is for many of topics.

BUT WHO KNOWS

AGGREGATION - PLACE

1. Very difficult to get intracity variation automatically from Twitter.
2. 1% sample puts floor on types of cities, countries observable.

AGGREGATION - TIME

1. Linking events across days.
2. Poorer, less populous the area, the coarser the temporal aggregation will need to be.

OTHER SOURCES

FACEBOOK

- Facebook used to make its data pretty available, especially if you worked with their teams.

Best

- I believe you can still get interesting insights from their Ads API.

FACEBOOK

- Facebook used to make its data pretty available, especially if you worked with their teams.
- Then some study scandals and really the 2016 Presidential Election changed things.

Best

- I believe you can still get interesting insights from their Ads API.

FACEBOOK

- Facebook used to make its data pretty available, especially if you worked with their teams.
- Then some study scandals and really the 2016 Presidential Election changed things.
- Now, I have stopped paying attention to their API.

Best

- I believe you can still get interesting insights from their Ads API.

FACEBOOK

- Facebook used to make its data pretty available, especially if you worked with their teams.
- Then some study scandals and really the 2016 Presidential Election changed things.
- Now, I have stopped paying attention to their API.

Best

- I believe you can still get interesting insights from their Ads API.

FACEBOOK

- Facebook used to make its data pretty available, especially if you worked with their teams.
- Then some study scandals and really the 2016 Presidential Election changed things.
- Now, I have stopped paying attention to their API.

Best

- I believe you can still get interesting insights from their Ads API.
- Crowd Tangle slash Social Science One.

INSTAGRAM

- See the previous slide.

INSTAGRAM

- See the previous slide.
- Some computer scientists scrape (Python library)

INSTAGRAM

- See the previous slide.
- Some computer scientists scrape (Python library)
- **So great for images.**

SINA WEIBO

Zhang, Han, and Jennifer Pan. "Casm: A deep-learning approach for identifying collective action events with text and image data from social media." Sociological Methodology 49.1 (2019): 1-57.

Goebel, Christian, and H. Christoph Steinhardt. Better Coverage, Less Bias: Using Social Media to Measure Protests in Authoritarian Regimes?. Working paper. University of Vienna, 2019.

CHAT APPS

Need to be a member of a group chat to see its data.

YOUTUBE

- Talk to Kevin Munger

YOUTUBE

- Talk to Kevin Munger
- Can gets lots of data, but I believe it is not as rich as Twitter.

YOUTUBE

- Talk to Kevin Munger
- Can gets lots of data, but I believe it is not as rich as Twitter.
- Very good to study communication, probably not so good for automatic or hybrid event data.

MISCELLANY

PROGRAMMING BACKGROUND

I do not consider myself a programmer.

PROGRAMMING BACKGROUND

- 1 semester of R undergrad.
- Started R in methods classes.
- Started Python summer before 3rd year.
- I do not consider myself a programmer.

PROGRAMMING

- YOU CAN DO IMAGE ANALYSIS EASILY WITH PYTHON.

PROGRAMMING

- YOU CAN DO IMAGE ANALYSIS EASILY WITH PYTHON.
- Use the `pytorch` library: do not rush to collaborate with a computer scientist.

GOOD IDEAS

1. Save tweets in raw form.
2. Consider using Python to process raw tweets.
3. Maintain data backup.
4. Use structured directories, ordered file names.

MY SET-UP

Graduate School

- 2009 Macbook Air with 8gb RAM. Eventually got an external hard drive.
- Manage AWS EC2 server for advisor's group, with local programming on the MBA.

PROGRAMMING IN CONSTRAINED ENVIRONMENTS IS GOOD PRACTICE.

MY SET-UP

Now

- Amazon Web Services t2.small to collect tweets
- Pipe to server in my office purchased with start-up money.
  - 24tb, running out of space
- Jungseock (collaborator) has his own servers with GPUs for image processing.

CONSIDERATIONS

CONSIDERATIONS