Should you keep the tweet?:
Balancing reproducibility, open data and participant privacy

Dr. Rachael Tatman
Computational Sociolinguistics Workshop @
#NWAV47, October 18 2018

kaggle

@rctatman

These are my personal recommendations based on my own experiences and ethical principles.

One size does **not** fit all.

- **Reproducibility:** We want to be able to repeat computational analyses done in the course of research
- **Open Data**: Datasets are part of our research contribution and we want to share data with other researchers and citizen scientists
- **Participant Privacy:** We need to abide by the developer's agreements & respect users' wishes about how their data should be used

Reproducibility

- Reproducibility is more important in computational fields
- For machine learning projects, sharing data is necessary for reproducibility (discussion in Tatman et al 2018)

> An article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.
>
> Claerbout & Karrenbach, 1992

Open Data: Not just “nice to have”

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

NSF DATA SHARING POLICY

Reproducibility & Open Data

Reproducibility & Open Data

Participant Privacy

“Participant” Privacy

- “Participant” isn’t quite right, since it implies informed consent (“Data producers”, maybe?)
- Usually social media is exempt from or expedited through IRB review
- We’re already legally constrained
  - Need to abide by the developer’s agreement (if using API)
  - GDPR (for personal data about individuals in the European Union)
  - US regulations around data in specific domains (HIPAA, FERPA)
- Beyond legality, how can we proactively respect the wishes of the communities we’re studying?
  - Here discussing Twitter (Fiesler & Proferes 2018) but it will (presumably) vary by community

Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. Social Media + Society, 4(1), 2056305118763366.

More acceptable:

- Large datasets
- Analyzed automatically
- Social media users informed about research
- Anonymized tweets being quoted (note that enough words should be changed that a reverse search isn't possible)

Less acceptable:

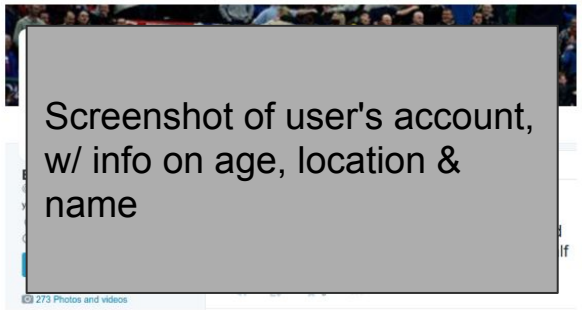
- Small datasets
- Analysis done by hand (presumably including analysis by Mechanical Turk workers)
- Tweets from protected accounts or deleted tweets analyzed
- Quoting with citation (very different from academic norms!)

Areas of concern:

- Small dataset, from a single account
- Analysis done by hand
- Tweets from protected accounts analyzed (account was briefly protected & then made public again)
- Directly quoting tweets in paper
- Sharing info shared on user profile

Conclusion: Style

- How does this interact with style?
- Case study: Twitter user [User name]
  - https://twitter.com/[User name]
  - 100 most recent tweets on April 23, 2015



Screenshot of [User name's] Twitter Profile taken April 23, 2015.

Areas of concern:

- Small dataset, from a single account
- Analysis done by hand
- Tweets from protected accounts analyzed (account was briefly protected & then made public again)
- Directly quoting tweets in paper
- Sharing info shared on user profile

Conclusion: Style

- How does this interact with style?
- Case study: Twitter user [User name]
  - https://twitter

This was my study! I presented this slide.

w/ info on age, location & name

Screenshot of [User name's] Twitter Profile taken April 23, 2015.

What could go wrong?

My main area of concern are cases where a user deletes a post from a platform but you retain it (which is against developer agreements but easy to fix) or it’s referenced in a publication (which is much harder to fix)

- Social media posts are admissible as evidence in court & research data (at least in the US) is not protected from subpoena
- Research data on members of sensitive groups (like minority language speakers) might be used to find\target them
- You quote a user saying something controversial and trolls reverse search them and engage in a harassment campaign

What could go wrong?

My main area of concern are cases where a user deletes a post from a platform but you retain it (which is against developer agreements but easy to fix) or it's referenced in a publication (which is much harder to fix)

- Social media posts are admissible ch data (at leas
- Rese guage spea
- You rse search them

We don't know what type of data will become sensitive in the future

What can we do?

1. Share data/code necessary to repeat an analysis but not the raw text
    a. Scripts used for querying/scraping (remember not to share your API key!)
    b. Tweet ID's (See Documenting the Now for examples & rehydrator)
    c. Preprocessing scripts to get from text to your features
    d. Features that can't be reverse engineered to get text back (word vectors yes, sentence vectors no)
2. Don't directly quote tweets in your papers
    a. My rule of thumb: Replace 20% of the content words with synonymous or equally frequency words
    b. Check to make sure you can't reverse search the tweet from the text
    c. Exception: viral or newsworthy tweets

Exceptions: Bots, Brands and Bigwigs

I think these measures are important for private individuals, but I'm less concerned with other categories of accounts:

- Bots: automated accounts
  - Some bot accounts will clearly disclose (e.g. @infinite_scream, @MagicRealismBot)
  - Others will attempt to mimic real users & require detection
  - I'd recommend automatic detection (e.g. Botometer by Davis et al) w/ a hand-tuned threshold
- Brands: accounts speaking on behalf of a corporation or other organizations
- Bigwigs: accounts for public individuals
  - Verified accounts (self-selecting as an account of public interest)
  - Politicians (ProPublica's politwoops archives deleted tweets by public officials)
  - Celebrities or very popular accounts, even unverified (e.g. @dril)

Thanks! Questions?

Slides:

@rctatman
rachael@kaggle.com