Data: legal and ethical issues

Sharon Goldwater

5 November 2019

School of informatics

Orientation

Last few lectures: distributional semantics (technical aspects). Your next assignment (out Friday) explores some of these ideas.

- Work with data extracted from Twitter (co-occurrence counts)
- Compare different ways to contruct context vectors and compute similarities
- Analyze and discuss differences between approaches, qualitatively and quantitatively.

Also an opportunity to consider many other issues...

Remainder of the course

- Only two more lectures on purely technical topics (sentence semantics)
- Mostly focusing on broader picture: NLP in practice (scientific, legal, and ethical issues)
  - Where does the data come from? Annotation, licensing, privacy
  - The messy world of data: user-generated text, biases
  - Issues in evaluation: reliability, human evaluation
- Your assignment ties in with several of these: a step closer to real research/practice.

Today's lecture

- What issues must you consider when using or collecting data?
  - Legal issues
  - Ethical issues and procedures
- What about social media in particular?

Data set for assignment 2

We provide word counts/cooccurrences from 100 million tweets. We do not provide original tweets. Why?

1. Working with that much data is very challenging!
   - We already did a lot of preprocessing for you.
   - Even then, very large files!
   - **Lab 8** walks you through what we did and how to use the files. Do it before you start the assignment!

Data set for assignment 2

We provide word counts/cooccurrences from 100 million tweets. We do not provide original tweets. Why?

2. We have to respect Twitter's licensing agreements.
   - Twitter data may be downloaded for research purposes (i.e., by this University).
   - Twitter data may not be redistributed (i.e., **do not copy to your personal machine or upload elsewhere**—use DICE, in person or remotely).
   - If storing Tweets, must respect users' deletion of them (i.e., remove tweets that user deleted.)

Data set for assignment 2

We provide word counts/cooccurrences from 100 million tweets. We do not provide original tweets. Why?

3. Perhaps other ethical considerations? But first let's talk about licensing.

NLP data, more generally...

- Most NLP systems are **supervised**
  - Training data is annotated with tags, trees, word senses, etc.
- Increasingly, systems are **unsupervised** or **semi-supervised**
  - Unannotated data is used alone, or along with annotated data
- All systems require data for **evaluation**
  - Could be just more annotated data, but could be judgements from human users: e.g., on fluency, accuracy, etc.

Where does the data come from?

- Annotated data: annotators usually paid by research grants (government or private) or by companies
- Unannotated data: often collected from the web
- Human evaluation data: collected in physical labs or online: again, usually paid by research grants or by companies

All of these raise legal and ethical issues which you need to be aware of when using or collecting data.

Intellectual property issues

Annotation is expensive and time-consuming, so annotated data is usually distributed under explicit user/licensing agreements.

- Paid licenses: e.g., the Linguistic Data Consortium (LDC) uses this model.
  - Researchers/institutions pay for individual corpora or buy an annual membership.
  - Edinburgh has had membership for many years, so you can use corpora like Penn Treebank (and treebanks in Arabic, Czech, Chinese, etc), Switchboard, CELEX, etc.
  - But you/we may not redistribute these outside the Uni (which is why we put them behind password-protected webpages).

Intellectual property issues

Annotation is expensive and time-consuming, so annotated data is usually distributed under explicit user/licensing agreements.

- Freely available corpora: e.g., Child Language Data Exchange (CHILDES) uses this model.
  - Anyone can download the data (corpora in many languages donated by researchers around the world).
  - If used in a publication, must cite the CHILDES database and the contributor of the particular corpus.
  - Redistribution/modification follows Creative Commons license.
- Other free corpora may have different requirements, e.g., register on website, specific restrictions, etc.

Privacy issues

To build NLP systems for spontaneous interactions, we need to collect spontaneous data. But...

- Are individuals identifiable in the data?
- Is personal information included (or inferrable) from the data?
- What type of consent has been obtained from the individuals involved?

The answers to these questions will determine who is permitted access to the data, and for what.

Example: CHILDES database

Many of the corpora are recordings of spontaneous interactions between parents and children in their own homes.

- Usually 1-2 hours at a time, at most once a week.
- Parents must sign consent agreement, including information about who will have access to the data.
- In some cases, only transcripts (no recordings) are available, often with personal names removed.

Example: Human Speechome Project

Deb Roy (MIT researcher) instrumented his own home to record all waking hours of his child from ages 0 to 3 (starting around 2006).

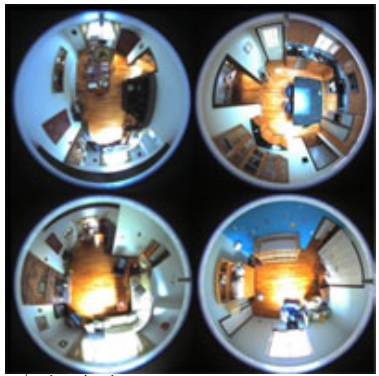


Image: http://www.media.mit.edu/cogmac/projects.html

Example: Human Speechome Project

Deb Roy (MIT researcher) instrumented his own home to record all waking hours of his child from ages 0 to 3 (starting around 2006).

- Huge project involving massive storage and annotation issues; incredible effort and expense.
- Huge potential to study language acquisition in incredible detail.
- But for privacy reasons, "there is no plan to distribute or publish the complete original recordings". Roy may consider "sharing appropriately coded and selected portions of the full corpus."

Use of existing data sets

Usually straightforward to follow legal and ethical guidelines.

- Don't redistribute data without checking license agreements
  - This includes modified versions of the data
- In most cases, you may store your own copy of data licensed by UofE to use **for University-related work only**; if not, we'll say.

Except: datasets from social media and others that reveal human behaviour may need ethical consideration for new types of use.

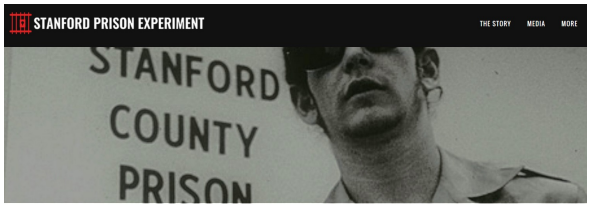
- If in doubt, check with your instructor or project supervisor.

New uses/ new collection of data

Creating a new corpus, getting human evaluations of a system, etc.

- Any work involving human participants, personal or confidential data requires ethical approval.
- Heightened approval requirements if participants include "vulnerable groups": children, people with disabilities, etc.
- Raw social media data almost always includes personal data! (data related to an identified **or identifiable** person).

Why ethical approval?



THE STANFORD PRISON EXPERIMENT:
A SIMULATION STUDY ON THE PSYCHOLOGY OF IMPRISONMENT

- Image from http://www.prisonexp.org/, where you can find details, movie, etc.

The Belmont Report and others

Following this and even worse cases (e.g., the Tuskegee Syphilis Study) the US govt commissioned a report laying out ethical principles for studies with human participants. Three core principles:

- **Respect for persons**: e.g., informed consent.
- **Benificence**: Maximize benefits while minimizing risks.
- **Justice**: Fair/non-exploitative for potential/actual participants.

Many professional and academic bodies now have ethics codes with similar principles.

- British Psychological Society, British Sociological Assoc, etc.

How is it enforced?

- Funding agencies and journals normally require universities to have ethics approval procedures, and researchers to follow them.
- Companies must follow privacy laws, also self-policing based on public relations. Some have their own ethics panels.
  - though sometimes data which purports to be "anonymized" can still be identifiable... see for example
    `https://www.wired.com/2010/03/netflix-cancels-contest/`
    `https://en.wikipedia.org/wiki/AOL_search_data_leak`

What about our School?

School ethics panel reviews applications for research, ensuring:

- Appropriate plans for acquiring and storing personal data (required by GDPR: General Data Protection Regulations).
- Informed consent from human participants (exceptions may be granted if compelling reasons).

If your research involves personal data or human participants, you or your supervisor needs to fill in an ethics approval form.

Example: Evaluating a system

You develop a machine translation system and want people to rate the output of the system for fluency and accuracy.

- If you bring people into your lab to do this, you will need to get ethical approval.
- If you use crowdsourcing on the Internet to do this, you will still need to get ethical approval.

Generally, approval should be straightforward using the School ethics approval form.

Example: anti-spambot

Real student project proposal: develop a system to automatically respond to spammers, trying to engage them in email conversation for as long as possible.

- Does it require ethics application/approval?

Example: anti-spambot

Real student project proposal: develop a system to automatically respond to spammers, trying to engage them in email conversation for as long as possible.

- Does it require ethics application/approval?
- Yes: the person on the other end of the spam is still a person!
- This project involves human participants, and ones who cannot give informed consent. So may be problematic even though spammers are not the object of study.

Example: user localization from audio

Real student project proposal: learn what individual's daily patterns are using always-on audio recording from mobile phone.

- Plans to avoid needing subjects' consent by running the data collection on own phone. (No ethical approval required for self-experimentation.)
- Does it require ethics application/approval?

Example: user localization from audio

Real student project proposal: learn what individual's daily patterns are using always-on audio recording from mobile phone.

- Plans to avoid needing subjects' consent by running the data collection on own phone. (No ethical approval required for self-experimentation.)
- Does it require ethics application/approval?
- Yes: always-on recording will still capture other people's speech, and this is personal data!

Back to Twitter

Legally, Twitter allow free downloads of 1% sample of Tweets...

- ...subject to restrictions noted above (e.g., no redistribution, must delete any Tweets as user deletes them, etc.)

But is it ethical to use this for any/all research we want to do?

One issue: public data and consent

- Historically, "human participants" are viewed as people the researcher interacts or intervenes with.
- Professional and academic codes have not required ethical scrutiny to collect, store, or study publicly available data.
- However, with social media many researchers are questioning these old assumptions, partly due to studies of users' own attitudes.

What do users think?

- Do Twitter users know their tweets might be used for research?
- How do they feel about potential research on their tweets? E.g.,
  - Is it ok at all?
  - Do they want to be asked permission?
  - Is it ok to publish their tweets in papers?

Recent papers surveyed Twitter users to find out:

- Williams et al. (2017) [WBS17]: Surveyed British Twitter users
- Fiesler and Proferes (2018) [FP18]: Surveyed Twitter users on Amazon Mechanical Turk (i.e., they have US bank accounts)

Findings from FP18

- Most (62%) of respondents did not know that tweets are sometimes used for research.
- Only 20% were uncomfortable with the idea in general, but this rose to 48% if the study includes their entire Twitter history.
- Other factors also matter, such as:
  - whether they give permission
  - what the study is about and who is doing it
  - whether their tweets are part of a much larger dataset
  - whether profile information is also used
  - whether tweets are analyzed by humans or by computers

Findings from WBS17

- **Most** respondents were at least slightly concerned about use of their tweets in research (less for University research, more for commercial research).
- 80% of respondents expected to be asked for consent.
- 90% of respondents expected their tweets would be anonymised if published in research papers.

Like FP18, results imply that despite the data being "public", many users expect some level of consent and privacy!

User expectations vs Twitter guidelines

- Users seem to want anonymity if tweets quoted in research publications.
- Twitter guidelines explicitly state that published tweets should be reproduced verbatim, including username and twitter handle.
- But Twitter also says users should be able to delete their tweets (effectively impossible if published non-anonymously).

What do you think?

If a researcher needs to discuss an example tweet, should they follow Twitter's guidelines (include user ID) or user expectations (make it anonymous)?

- I prefer not to say
- I am not sure/depends on the situation
- They should include the user ID
- They should make the tweet anonymous

So, what to do?

Ethics of research on social media data is complex and rapidly evolving, with legitimate disagreements. Still, there are some best practices:

1. Consult outside your research group (e.g., with ethics panel in our School or wherever you end up), as you would for any research involving humans.
   - You **don't** need to do this for Assignment 2 as long as you're mainly just working with our pre-processed data set, since it no longer has information that is identifiable to users or groups.

So, what to do?

Ethics of research on social media data is complex and rapidly evolving, with legitimate disagreements. Still, there are some best practices:

1. Consult outside your research group (e.g., with ethics panel in our School or wherever you end up), as you would for any research involving humans.
2. Think twice before publishing specific tweets, and avoid it if there is sensitive content.
   - You **should** consider this for Assignment 2 if you look at examples from Twitter in your analysis.

So, what to do?

Ethics of research on social media data is complex and rapidly evolving, with legitimate disagreements. Still, there are some best practices:

1. Consult outside your research group (e.g., with ethics panel in our School or wherever you end up), as you would for any research involving humans.
2. Think twice before publishing specific tweets, and avoid it if there is sensitive content.
3. Consider both benefits and risks of the research. (More on this in later lectures.)

Summary

NLP (and other AI) requires lots of data. You need to comply with:

- Legal issues: check licenses and don't collect or redistribute data unless explicitly permitted.
- Ethical procedures: such as the School's approval process, especially if using human participants or personal data.

But established laws and guidelines don't cover all possible ethical issues, and areas such as social media can be complex. Consider potential harms before proceeding!

References

Fiesler, C. and Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1):1–14.

Williams, M. L., Burnap, P., and Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, 51(6):1149–1168.