Accelerated Natural Language Processing 2018

Lecture 30: Ethics of NL data collection, annotation, publication

Henry S. Thompson
Some material from slides by Sharon Goldwater
22 November 2018

1. Privacy: Ethical *and* legal issues

Non-fictional language data is not only copyright

- It's also potentially private

Both older (Protection of personal data, 1995) and very recent (General Data Protection Regulation (GDPR), 2018) EU law regulate the collection and storage of personal information

- And the definition of what is covered is quite broad
- For example, the fact that you are enrolled in this course

Any data that you collect may be covered

- Including data you collect for an employer

2. Privacy, cont'd

Publishing data you collect is a good thing

- See discussion of Open Data/Open Science earlier

But only if the privacy of the contributors is not at risk

- Are individuals identifiable in the data (authors/speakers *or* others)?
- 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

- That is, the terms under which the data is licensed

Just because you don't *intend* to publish

- Doesn't mean your responsbility for privacy ends
- You have to ensure your data doesn't leak
- For example, the University requires staff to encrypt any laptop which contains student-related information

- What if you have doubts about your employer's data security practices?

3. Anonymity

Is it enough if you don't identify your contributors?

- Necessary, but often not sufficient

Anonymising data is harder than you might think

- Simply removing names (which isn't necessarily all that easy) isn't enough

In many cases, the value of the data depends in part on the **metadata**:

- Age, gender, social class, postcode, ... of the contributor
- Context of production (if spoken)
- Task-relevant independent variables (illness/education/political party/...)

And it doesn't always take much in the way of metadata to allow identity to be discovered

4. Example failures of anonymisation

In the mid-1990s, a US health insurance company, intending to help medical research, released "records summarizing every state employee's hospital visits at no cost to any researcher who requested them" [1]

- "William Weld, then−Governor of Massachusetts, assured the public that [the insurance company] had protected patient privacy by deleting identifiers" [*ibid*]
- Each record contained around 100 medically-relevant items of metadata
  - Including gender, postal code and date of birth
- But by spending 20USD for the publicly-available voter roles for his home city, a researcher was able to identify Weld's *own* record!
- The voter role included full name as well as ...
  - gender, postal code and date of birth
- And those three provided a unique match in the medical dataset

A number of similar cases of successfully 'de-anonymising' by cross-matching across multiple datasets have been published

- See Iain Murray's MLPR lecture notes for an (in)famous case involving Netflix reviews

[1] Ohm, Paul, *Broken Promises of Privacy* 2009

Statistical approaches to anonymisation of very large datasets has become a hot topic for research

- Known as **Differential Privacy**

5. Ethics review

Many institutions require any research activity which involves human beings or animals to undergo some form of ethics review

- Funding agencies and journals normally require this
- The University of Edinburgh calls this the **Research ethics procedure**

The first step in the procedure is informal:

Before any research project can start, the following question[s] need[] to be answered:

- Does this research involve human participants or animals, confidential or personal data, developing countries, military applications? Does the research concern terrorist or extremist groups?

This is known as level 0 ethical screening, and takes the form of self-assessment by the researcher conducting the research. If the answer to this question in "yes", then screening at higher levels (levels 1-4) will follow, as appropriate.

- You *must* carry out this procedure in conjunction with your MSc dissertation project
- If you have any questions about the procedure, please discuss them with your project supervisor

Examples follow from Sharon Goldwater

Data and Ethics in NLP

Sharon Goldwater

14 November 2016

School of informatics

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 people on the Internet to do this, you will still need to get ethical approval.

Generally, cases like this only require a signed self-assessment confirming no further issues.

Example: language use on Twitter

Real paper: case study of one Twitter user's use of spelling to indicate regional pronunciation.

- The relevant data from the user is already public.
- But that isn't the same as giving informed consent to participate in a research study.
- Username, profile information, example tweets, and results of study are all described in the paper (i.e., personally identifying information).
- Requires further ethical consideration: presumably the researcher contacted the individual for approval (I hope!).

Example: anti-spambot

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

- 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.
- Requires further ethical consideration.

Example: user localization from audio

Real student 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.)
- Only plans to use non-speech audio data.
- However, always-on recording will still capture other people's speech.
- Requires further ethical consideration.

6. Kinds of participants

Any project with participants from 'vulnerable' groups requires at least level 2 screening

- And children and old people are considered 'vulnerable' in this context
- As are people with any physical or learning disabilities or serious or chronic illness

Also note that for ethical review it may be necessary to distinguish between

Annotators
Recruited (and/or trained) for their expert knowledge, and are not subjects of the study

Participants
Recruited as non-experts, and may themselves be subjects of study

Although annotators' privacy is not usually an issue

- They may well have quite challenging, stressful and even disturbing tasks, in which case ethical issues *do* arise
- As for example we heard from Jennifer last week

7. Transparency

The popular press has been getting very excited about 'algorithms' taking control of important decisions

- And there are real ethical and technical problems behind this concern

Machine learning systems whose deployed state consists of complex networks with thousands, hundreds of thousands or even more nodes and weighted connnections

- Are effectively black boxes
- There is no way to get an answer to a "why" question when they "make a decision"

This is not actually a new problem

- It was an issue as long ago as the Expert Systems boom of the 1970s/80s

One of the most successful (in scientific terms) Expert Systems of the early 1970s was **MYCIN**

- Developed at Stanford University in collaboration with the Stanford Medical Center
- To diagnose and suggest treatment for bacterial infections

But although its measured performance in trials was equal to or better than that of human doctors

- Early efforts to deploy it for actual use in hospitals totally failed
- Because it could not explain *why* its proposed treatment was the right one

Work on explanations of network-based Machine Learning systems is at a very early stage.

- So although there's a lot of pressure on government to introduce regulation requiring transparency
- It's not clear how to go about achieving it

8. Responsibility

Responsibility is an ethical issue that's closely tied to transparency

- When a computer-controlled x-ray machine overdoses a cancer patient and they die
- Who (or what) is responsible?
- Who is **culpable**?

When a launch-on-warning system mistakenly 'detects' a hostile missile launch and launchs its own missiles in retaliation

- Whose fault is it that World War III begins?

When an NLP system mistakenly 'detects' a terrorist tweet

- And the police break into an innocent person's flat
- Who pays for the damage?

(That's one real example, one partly real on two occasions (a human being overrode the system in time) and one made up)

Before we empower *human beings* to make decisions with an impact on other peoples' well-being

- We require training, and we test them
- Think doctors, teachers, even bus drivers

There's talk in goverments about the need to do the same for computer programs

- But nothing like a clear idea about how to go about it

9. Bias

NLP systems are particular vulnerable to undetected/unexpected bias

- Introduced by the material they are trained on

We need to update that old saying "Garbage in, garbage out"

- Train on biased data, you get a biased system

Bias may be political

- That is, minority opinions may get lost

Or cultural

- Reinforcing dominant cultural types
- At the expense of minorities
- A recent example:
    - A year ago the sentiment analysis component of Google's publicly available Cloud Natural Language API was discovered to be reflecting negative cultural biases
    - Treating descriptions as Jewish or homosexual as negative

Or linguistic

- Exaggerating the status of some dialects at the expense of others

We've come full circle: As NLP researchers, we have a responsibility for the impact on our users of the systems we build:

> *Data isn't a neutral tool. We embed our values in who benefits and who we protect*

Jeni Tennison, CEO, Open Data Institute

10. Over to you

As I said on Tuesday, these two lectures are inevitably grounded in my own particular (Western, liberal) context

What do *you* think?

- Talk to a neighbour about what *you* think about privacy, or responsibility, or Open Data, or whatever else has caught your interest
- Particularly if you disagree with the views I've expressed