

Natural Language Understanding, Generation, and Machine Translation (2023–24)

School of Informatics, University of Edinburgh
Alexandra Birch

Tutorial 3: Ethics for NLP (Week 8)

1 Credit Risk Assessment

Question 1:

A startup wants you to help build an app that can help banks predict the creditworthiness of a person based on their social media activity. In simplified form: the input is the text of a user’s social media posts, and system should classify the user into one of three categories that the bank will use to determine whether or not they receive a loan: LOW RISK, MEDIUM RISK, or HIGH RISK. A bank provides you with a large amount of training data based on their loan records. Each example consists of a user’s twitter handle (from which the text of the user’s tweets can be recovered) and the bank’s classification of the user.

- a. Who are the stakeholders in the system? Your answer should include *direct stakeholders* who participated in system’s creation (funders, developers, speakers or writers, annotators and curators) or who are the system’s users. It should also include *indirect stakeholders* who do not use the system but are nonetheless impacted by it.
- b. What could go wrong?
- c. Who would be harmed? How?
- d. Who would benefit? How?
- e. What (if anything) might be a way mitigate such harm?

Solution 1:

- a. Direct stakeholders include you, your startup, and your customers in the banking industry: specifically the funders of the project and the loan officers who will use the system. Indirect stakeholders include loan applicants and even those they interact with on twitter—for instance those they retweet or reply to—since their data could be captured by the system.
- b. One possible problem is that a system like this can codify any discriminatory practices that were already present in the historical loan data, even if the system itself does not explicitly refer to demographic variables. We know that the banking industry has a long history of racial discrimination.¹ We also know that language use on social media is correlated with demographics.² Since bank decisions and language use are both correlated with race, they are also correlated with each other, and a machine learning system can and will learn this correlation even if race is not an explicit variable in the model.

A second possible problem is that decisions made by the system will not be explainable if they use the kinds of deep learning models that we

¹<https://en.wikipedia.org/wiki/Redlining>

²E.g. <https://www.cc.gatech.edu/~jeisenst/papers/nipsws2010.pdf>

have focused on in lecture. These models cannot be easily inspected to understand how they produce a decision.

These problems, and others, arise because the system design conflates two quite distinct concepts. The first is *creditworthiness*, which you might think of as a prediction about whether an applicant will pay back a loan, and is what a bank is primarily interested in. The second is *previous decisions* about creditworthiness, which are themselves predictions that are likely to contain human biases. We might be interested in knowing the *outcome* of previous loans (i.e. whether or not someone paid a loan back), but even with that information, we cannot know what would have happened with individuals who were denied loans. Some of them may have been creditworthy.

There are other problems around privacy and coercion. Your app may ask for consent to access someone's social media data—but if someone needs a loan, and this is the only way to get it, is that person really in a position to freely consent?

- c. Loan officers in banks may face increased pressure on their jobs or wages because of automation. Loan applicants from populations that have been systematically discriminated against could face continued (now automated) discrimination. Banks and their stakeholders may make worse decisions, if the machine learning system learns irrelevant correlations that happen to be predictive on the training data.
- d. Banks (and you) will profit if the cost of making loan decisions is lowered, and individuals without traditional credit history may benefit if they can be approved for loans that they previously could not obtain.
- e. This is fundamentally a social and legal problem, which careless use of machine learning can easily amplify: the historical data on which your system depends cannot easily be collected again, and it cannot easily be scrubbed of biases that it already contains. Continuing with the status quo quite likely also means continuing with existing biases. So, there are potential benefits in this application, but it will be critical to involve all of the stakeholders in the process of understanding *what* information in a social media profile is likely to be relevant, and automating around that, rather than blindly hoping for machine learning to discover it.

2 Automated Medical Coding

A company wants to hire you to develop an NLP system that fully automates **medical coding** from doctor's reports. One type of medical code identifies a diagnosis indicated in the report, while another identifies a prescribed treatment. Each code comes from a finite set. Here is a simplified example:

Input (doctor's report): *Patient is a 27-year-old white male. Height is 74 inches, weight 220 lbs. Patient states he is allergic to penicillin, but has no other outstanding medical history. Does not smoke, exercises moderately. Patient presents with chills, headache, cough, fever (101 degrees), difficulty breathing. Examination via stethoscope yields heavy rales. Percussion test on thorax suggests buildup in lungs. Streptococcal pneumoniae suspected. Prescribed patient two weeks of 500mg azithromycin (Zithromax), and scheduled follow-up for next week.*

Output 1 (medical diagnosis code): pneumonia

Output 2 (medical treatment code): azithromycin-500mg

Question 2:

Medical codes determine medical bills and authorize patients to receive certain treatments like prescription medicines, so they are a critical part of the modern healthcare system. They are also costly because they are produced by trained professionals. Many companies want to reduce this cost by using AI to automate medical coding.

- a. Who are the stakeholders in the system?
- b. What could go wrong?
- c. Who would be harmed? How?
- d. Who would benefit? How?
- e. What (if anything) might be a way mitigate such harm?

Solution 2:

- a. Direct stakeholders include you, your company, and customers in the medical industry: hospitals, doctors, medical coders, and other workers whose decisions are entered into or affected by the system. Indirect stakeholders include patients and their families.
- b. This type of system is different from the banking example, in that it has a known answer: the diagnosis and prescription are in the doctor's report, and the goal of the system is to correctly identify them, by classifying the document with the corresponding codes.

However, we know that the medical system is systematically biased in certain ways. For example, we know that doctors are less likely to prescribe painkillers for women than they are for men with similar levels of pain.³⁴ Since the language in the reports clearly varies with the patient's gender, a deep learning system could learn this correlation. This could further harm women by amplifying the effect. To see this, consider the same doctor's report, changing only the word "male" to "female" and the word "he" to "she". Given many examples differing only in these words, but with differing rates of painkiller prescription, the system can learn representations of gendered words that enable it to assign higher probabilities of prescription for words associated with men, and lower probabilities for words associated with women. So, when correct, the system will pass along an existing bias, and when incorrect, it is likely to amplify that bias.

Observe that many other types of harm to patients can result from incorrect classification. For example, in the case above, the system must correctly understand that penicillin should *not* be prescribed.

Side note: This question was on the 2019 NLU+ exam. A few months after the exam, a widely-publicized study demonstrated that racial bias had been learned by decision support systems used in medicine.⁵ These systems were trained on reports paired with future health costs of patients, which was used as a proxy for the severity of the patient's health

³<https://www.bbc.com/future/article/20180518-the-inequality-in-how-women-are-treated-for-pain>

⁴<https://www.ncbi.nlm.nih.gov/pubmed/18439195>

⁵<https://www.nature.com/articles/d41586-019-03228-6>

problems. In this historical data, black patients were less likely to be recommended for treatments and thus incur costs, even when a white patient with similar conditions were. Hence, the system learned not to recommend black patients for needed treatments, because the real variable of interest (whether the treatment would be effective) was replaced by a poor proxy (future health costs).

Further harms may accrue to medical coders, skilled workers whose livelihoods may be threatened by automation.

- c. Patients may be harmed by system error and the codification of bias. As a result, the reputation and finances of doctors, hospitals, and your company may also be harmed.
- d. Effective automation of medical coding may result in greater reliability of codes and greater efficiency in medicine, which can benefit medical workers and their patients.
- e. As in the banking case, this is fundamentally a legal and social problem, and as in that case, the primary approach must be to work with stakeholders in order to understand *what* aspects of medical coding can be reliably improved or automated, rather than relying on the routine application of machine learning.

3 Bonus Question

Question 3:

Find a recent news article about harms caused by a machine learning system, ideally in natural language processing.

- a. Who are the stakeholders?
- b. What went wrong?
- c. Who was harmed? How?
- d. Who benefited? How?
- e. What (if anything) might be a way mitigate such harm in the future?

This is an open-ended question. Keeping in mind that ethics is an ongoing conversation, we encourage you to discuss your articles and findings with other participants in the course, including on piazza.

Solution 3:

This is up to you. If you haven't yet found an example, you may want to review "Examples of Government Use Cases for Automated Decision Systems", which will give you some ideas of application areas to look at.⁶

⁶<https://ainowinstitute.org/nycadschart.pdf>