NLP: Value Sensitive Design Activity and Assignment

Names: Aaryan Sinha, Arnav Mahale, Kaan Tural, Jack Oehling, Arinjay Singh

Consider the following scenario: You are on a team developing an NLP application that will “read” letters of recommendation for Northeastern undergraduate applicants. For each letter, the application will generate a score 1-10 that reflects the strength of the applicant. The technology could be exported beyond the Northeastern network, if successful.

I Framing the Technology

(a) **Answer.** Why are letters of recommendation used in the application process?

The letters of recommendation are used because this shows how the candidate acts outside of their own wording, adding a new dimension, showing how others view the candidate and how easy they are to work with.

(b) **Answer each question.** What is the purpose of this technology? Who will use this technology?

The purpose of the NLP technology is to streamline the process of reviewing applications, the ones who will be using this technology would be the admissions team.

(c) **Argue for it.** Why (might some think this technology) is a good idea?

It gives different insights as to at has been successful before
It saves money and time.
It gets rid of an emotional bias that may be changing based on the human reviewer's emotions.

(d) **Give three relevant features of the social context in bullet point format** (this might include societal values regarding higher education and professional training, historical circumstance, surrounding cultural norms and expectations, economic circumstances, relevant features of Northeastern as an educational institution, laws or policies governing admissions, current admissions practices at

Northeastern, etc.). **Do some research here. Briefly explain why each seems possibly relevant.**

- Northeastern is test-optional, this acts as another datapoint
- Background of the person is an unknown, these letters can show maybe home issues affected their education
- Adaptation to changing educational needs, Northeastern or just schools in general adapt their education to the expectations of students which allows touching on topics that may not be traditionally taught within higher education.

II Empirical Investigations

(a) **Answer each question:** What training data will be used? How will the success cases be determined (to train the model)? Anticipate what problematic biases or stereotypes might be present in this data or the success cases?

- Needs data on different types of writers and maybe what types of schools the letters are coming from to determine the prestige of how the letters were created
- Training data will include old letters of recommendation ranging from good letters that should be rated a 10 to bad letters that should be rated a 1.
- Sampling at letters from different countries to get different styles of writing
- A way of classifying where it seems letters are coming from to decide whether wording should be taken at face value.
- Associate a letter to the graduating Northeastern GPA of a student to determine what words tend to be associated with high success students in certain majors.
- Bias could arise if there were situations outside of a student's control that impacted their GPA (financial, etc.)

(b) **List three empirical features about the world in bullet point form** (human psychology, culture, social practices, biases, etc.) that might be relevant for anticipating or assessing problems? **Do research here. For each of the three: state the feature, explain why it's a problem and list an article or source.**

- More conservative explanations outside of the U.S. for these letters of recommendation, as they may put some at disadvantage due to less excitable wording.
- Letters for men tend to be longer than letters for women, and due to this there may be more information on the men to benefit them.

- There are different words used based on which gender the letter is being written about, compassion for women vs problem solving for men, these associations can be damaging to the results of students whose trajectory could be determined by this difference in wording.

Stakeholder Analysis

(a) **Identify 4+ stakeholders** (i.e. whose interests stand to be impacted by this technology?)
(b) **For each of the stakeholders, identify 3+ interests or values that they each have at stake.**

(i) Students
    (1) Interests
        a) Future financial stability
        b) Get in
        c) Quality of education
        d) reputation
(ii) Families of Students
    (1) Interests
        a) Financial well-being
        b) Reputation of the family/Pride
        c) Legacy Admission
(iii) University/College
    (1) Interests
        a) Reputation for success
        b) Donations
        c) Diversity
        d) Best students
(iv) High School/Student's previous school
    (1) Interests
        a) Prestige/rank
        b) Donations
        c) Increased demand

(c) **Identify 3+ conflicts of interests or values. Explain how the conflict gets generated.**

- Getting in and getting the best students are conflicts between students and University
- Parents and Students opinions may differ on future career interests, including college major
- There is a conflict between financial well-being of the family and the profit incentives of the university

III Value Investigations

(a) Look at the conflicts of interests and consider what trade-offs should be made. **For each of the 3 conflicts, explain how we should approach the trade-off.** Consider: Are there illegitimate interests? Values that cannot be overridden? Priorities?

- For the first conflict, some could argue the idea of getting in is a legitimate interest but it may not be important to hold weight in this argument.
- Between parents and students the idea of having different interests in future career and college major, a good way of approaching the tradeoff is saying the student will be an adult, they should be allowed to make their own decision, the student can take on the financial burden if need be.
- The profit incentives of the university should be put on the backburner in certain cases when compared to the financial well-being of the family, because this could include different people of different socioeconomic statuses which contributes to diversity. This can also be alleviated by the government in certain cases.

(b) Of the stereotypes and biases you identified above, if these are left unmitigated, what sort of harms might result? **List and explain at least 2 for each of the categories below: representational harms, allocative harms, and other harms.**

Representational Harms (at least 2):

- Not all the different kinds of families of socio-economic status may be fully represented

- Conservative representations of students can prevent them from getting into the school, which is further damaging and may be cyclically bad for the NLP software, with less and less diversity reaching the school.

  Allocative Harms (at least 2):

- Trained on mostly american LORs, other cultures that use different language are not given that high of a score and are not accepted
- This may also create an allocative harm/bias in terms of extracurricular activities, it would harm students who do activities less recognized or valued by the AI, for example, and it might not be true, but community service in local neighborhoods for example. Putting a student at a disadvantage despite their meaningful impact.

  Other harms. Consider: stereotype threat, reinforcing implicit bias, double binds…. (at least 2).
- It could make the phenomenon of stereotype threat even worse, where students who are aware of negative stereotypes about their group (based on race, gender, socioeconomic status, etc.) may underperform in their applications due to the stress and anxiety these stereotypes cause. Knowing that an AI, which may not be capable of contextual understanding or empathy may not be aware or take into consideration these issues and may reinforce them.
- These AIs can inadvertently reinforce societal biases. For instance, women and minority students might face a double bind when expressing ambition or leadership in their cover letters. If they conform to stereotypically expected behaviors, they might be undervalued for not showcasing leadership or ambition. Conversely, if they do showcase these qualities, they might be penalized for not adhering to societal expectations for their gender or race.

IV Technical Investigations

(a) **Answer.** How might this technology be misused or abused?

Jack:
It might reinforce biases present in the training data, disadvantaging certain groups of applicants. It could also be manipulated by individuals who understand the system's criteria, leading them to craft letters that unfairly inflate an applicant's perceived qualifications.

(b) **Propose a technical intervention to mitigate an anticipated harm or reduce conflict of interests between stakeholders. Explain. Justify.**

One idea would be a holistic review support system that is transparently designed to weigh varied aspects of a student's application, including not just academic achievements and test scores, but also personal essays, recommendations, and evidence of personal and socioeconomic challenges. This system would explicitly account for and adjust biases related to the interests of students, families, universities, and high schools. This allows students who have skills that aren't as black and white to get a chance at getting in, allowing universities to get a diverse population and the best students, benefitting highschools and families indirectly as well.

(c) **Propose regulation or a policy to mitigate anticipated harms or reduce conflict of interests between stakeholders. Explain. Justify.**

I think law-wise, there should be a comprehensive policy requiring all educational institutions using AI for admissions to use a "multi-stakeholder feedback mechanism." This would essentially be regular consultations with students, families, and educational professionals from both the applying students' high schools and the universities themselves to review and adjust the criteria and algorithms used in the AI admissions system, this would allow a more equal chance for everyone, while it wouldn't be ideal for universities, as it is an extra expense it even benefits them as well to make sure their system is accurate. Though the benefit for this more so targets students, families, and highschools.

(d) Let's say you find that the dataset of successful applicants (those who have done well in programs and on which you intend to train the model) is 80% male. Someone on your team who is concerned about gender bias suggests upsampling to create a more representative dataset. **Analyze this. Answer questions.** Will this mitigate certain biases? If so, which ones? Exacerbate other biases? If so, which ones?

Jack:
Upsampling to balance the gender representation in the dataset can mitigate gender bias by ensuring the model is not unduly influenced by the overrepresented male examples. However, this approach might inadvertently exacerbate other biases, such as age, ethnicity, or socio-economic status if these factors are not equally balanced in the process.

V Reframing and Reflecting

(a) **Answer these questions.** What is the point of including letters of recommendation in applications? What values does this promote? Does this technology change the practice? Does it help achieve that aim? Does it promote or hinder the values?

Letters of recommendation provide a valuable perspective on applicants beyond grades and test scores. They offer insights into an applicant's character, work ethic, and skills from someone who has observed them in a professional or academic setting. This helps admissions committees gain a more holistic understanding of the applicant and assess specific qualities that are difficult to quantify through other application materials. Additionally, recommenders can often speak to an applicant's potential for success in a particular program or field. Using AI to review letters of recommendation has the potential to increase efficiency and objectivity in the application process. AI could automate the review of large volumes of letters, saving time and resources for admissions committees. It could also potentially offer a more objective assessment, reducing the influence of implicit bias and human error. But the issue comes about yet again, that this AI may have a bias from training and overlook some qualities that are important but can’t easily be found by some algorithm, so yes it can be beneficial for many reasons, but it should be recognized that it also can just as easily hinder values.

(b) **All things considered, what do you think:** Should we develop and implement this technology or not? If so, what regulation or technical fixes are needed? If not, why? **Provide your reasoning!**

This is a difficult question. The decision of whether or not to use AI to review letters of recommendation is a pretty complex issue with benefits on many sides. For me personally, this being Kaan, I don't think so, we have very well developed models, but capital greed would get in the way of truly caring if the model functions well or not, and for schools which treat themselves like a business, the odds of proper use and design are low until laws come into play. While there are potential benefits in terms of efficiency and objectivity, there are also significant risks of bias and undermining the values that letters of recommendation are meant to promote. If such technology is to be developed and implemented, laws that require rigorous testing and auditing for bias as I was kind of mentioning, transparency in decision-making, and human oversight are crucial to ensure fairness and protect the integrity of the application process. If these changes aren't made, I don't believe this technology should be implemented, though I can't speak for the group when I say this.