Ethics in Natural Language Processing 2024 Homework 4

Due until Wednesday, 17.07.2024 at 11:59pm

Submission Guidelines for Homework

- This homework is worth 20 points
- Submit subjective answers using the given latex template.
- Extra credit shall be given to well-structured submissions.
- In case of questions or remarks, please contact:
  - Aishik Mandal, aishik.mandal@tu-darmstadt.de

Before you start, make sure you read the Submission Guidelines instructions associated with this homework for important setup and submission information.

1 Anthropomorphisation (8 points)

In the lecture, we discussed anthropomorphisation – the tendency to attribute human-like behaviours and properties to inanimate objects. We discussed AnthroScore – a metric developed by Stanford University researchers to measure the amount of anthropomorphisation in a given sentence. Your tasks are as follows:

- Write yourself or find online five distinct sentences that talk about AI models, systems and technologies as if they were human. Example: "ChatGPT is very polite and nice to talk to."
- Write yourself or find online five distinct sentences that talk about AI models, systems and technologies in a non-humanizing way. Example: "ChatGPT produces text that is perceived as polite and helpful by most users."
- Calculate the AnthroScore for each of the sentences using the online demo http://anthroscore.stanford.edu. You need to input your sentence as well as the entity that you want to check (e.g. ChatGPT). In your homework solution, report the score next to your sentence, like this:
  - "ChatGPT is very polite and nice to talk to." (6.38)
- A positive score means that the sentence is humanizing the entity. Do the scores align with your expectations, or did some sentences receive an undeserved high or low score?[1]

2 Technology stages (6 points)

In the lecture, we discussed that AI technologies can belong to different stages: (S1) Fundamental Theories, (S2) Building Blocks, (S3) Applicable Tools and (S4) Deployed Applications. Given the following eight technologies, which stage would you place them on?

[1] Note that this is just a research demo, and it has bugs, e.g. you can get a "No words found" error. In this case, paraphrase the sentence or find a new one.

a) Long-short Term Memory Networks (LSTMs)

b) Research prototype of a face detection system

c) A face search engine application

d) A benchmark for evaluating natural language understanding

e) Argumentation theory

f) The AnthroScore demo from the previous task

g) Grammarly[2]

h) The Flair library[3]

Compare Long-short Term Memory Networks to Grammarly in terms of their impact. Which one has a broader potential impact? Which one's impact is easier to measure?

3 Disagreement hierarchy (6 points)

Place the following online commentaries on Graham's disagreement hierarchy, from "DH0: Name calling" to "DH5: Refutation". If you are unsure, use a higher hierarchy level. For example, if you are unsure about DH2/DH3, then answer DH3.

a) (in Wikipedia discussion) "Could you PLEASE stop being a formatting warrior and wasting everyone's time"

b) "Nah, I disagree"

c) "I don't think you know what you are talking about, I bet you never lived in London for longer than a month."

d) "This is a common misconception about vaccines. It is based on the publication from several years ago, that has been since then retracted. The Nature journal made an editorial about this, here is a link:"

e) (in a peer review) "The language of the paper is very complex and the figures are poorly formatted, thus I recommend it to be rejected."

Were you unsure about any of the points? What would be the alternative hierarchy class you can think of?

[2] https://www.grammarly.com

[3] https://flairnlp.github.io