AP6: Project reflection

We had a thorough review of the annotation process in the class, and we just followed the process depicted by a flow chart from the readings. However, implementing everything just right was impossible --- our members were always busy on some days during the week, so we did not have regular meetings to change the guidelines, and this caused problems later down the line. I thought the annotation part of the process was easy, and that it could be done by anyone, but having dedicated people working around the clock together would be much better than people doing it as a part-time job.

At first, we thought rating the family-friendliness of comments would be easy, as we as humans, are able to read a text once, and discern whether it is family-friendly or not for ourselves. Our task's population scope was global initially, but due to the disparate idea of family-friendliness different cultures hold around the world, we had to restrict ourselves to Reddit's primary userbase, based in North America. But defining family-friendliness was still difficult since there are also numerous cultures in North America as well. Even amongst our teammates, we had different values, but we strived to rate the comments from the view of an average person in North America. For example, some families would be uncomfortable with controversial or misleading news, or even news from channels leaning to the other side of the spectrum, but we excluded this from the criteria, as it was too subjective. Another example is drinking alcohol: adults drinking alcohol around children is pretty much normal, but it is a 21+ activity. Is it family-friendly then?

I started to think that for tasks that depend heavily on human judgement, restricting the geographical, cultural, temporal scope of the dataset would yield better agreement among annotators.

The most important skill I think I learned was organizing team projects. Coordinating people who were not always 100% motivated to do the projects taught me a lot about setting goals, inspiring motivation, and patience. Secondly, it reinforced a lot of skills I learned in both inside and outside of 490A. The annotation project allowed me to use BERT and logistic regression in practice, and create a database for the annotators, reinforcing my knowledge of databases as well.

The most memorable memory for me is when we compared ordinal regression with custom features and fine-tuned BERT model's performance. Our accuracy with ordinal regression was quite tragic, and we did not have a sense of

accomplishment. But the fine-tuned model we created performed much better and we felt relieved.

One thing we should have done was to start much earlier. One of our members, Saakshaat, did all the annotations of documents in the evaluation set on the last day. That hurriedness probably hurt our IAA as well as dataset, and model performance. Lastly, we should have relied less on context we inferred from the comments, treating every comment as if it is in its own box. The reason it that there is no realistic way to incorporate outside context from images, links, and older replies in our dataset, and for a model to take all of that into account in its classification.