CS410 – Explorations in Data Science Project Midpoint Report Kareem Aboueich

Updates to Project Objective

When I started the project at the beginning of the term, the idea was, initially, to get some sort of understanding as to how an AI processes hate speech. The data I obtained was a massive, 25,000+ line compilation of tweets that were taken down on Twitter. The data also included a scale as to how the tweets were measured – hate speech, offensive speech, both, neither, etc. Once I got the data I was using, which was entirely judged by humans working at an AI cloud computing company, my approach changed a little bit to "what to humans see as hate speech and what to they perhaps oversee?". There were some very unique takeaways from this, such as the fact that racism and homophobia are almost universally seen by humans as hate speech, whereas sexism often flies under the radar and is overlooked unless certain actions are mentioned. These were all interesting takeaways I plan to explore in more depth for the second half of this project.

Initially as well, I wanted to obtain data from Facebook as well, seeing as they are the dominant player in the social media world today. Facebook data was more obscure and harder to find – not exclusively because it came with a price tag, or at least the data that would be the easiest to use for this sort of project did. One of my objectives as well was to observe how different social media companies tackle hate speech, how their content moderators perform, what sorts of differences can we tell between them, and so on. This objective changed and ultimately was eliminated from the project as a whole because it was taking a lot of valuable time, and the data that was available costed a lot of money and/or was incomplete, so that part of my project was scrapped.

One of the things I was on the fence about earlier on was comparing how a robot would deal with hate speech compared to a human, and initially, I thought of creating a simple filter bot that would help me find tweets in a subsect of the file (~250 – 1,000 tweets or so) that would contain certain language, and then observing how the humans decided on that data. Comparing how humans and machines can respond and choose to judge the data will be an important step forward in understanding the future of content moderation. This aspect of the project was something I decided to take on. If I am unable to find a tool that can decide for itself what qualifies as hate speech or not, building a bot that will just flag tweets with certain words as hate speech will be interesting to compare to what the humans thought of a sect of the posts. This was also one of the suggestions Professor Tufte recommended I take my project in, considering I was unable to obtain Facebook data and also due to the fact that it was the question I was initially trying to answer when I thought of this topic.

These are the three primary things that changed in regards to my project objective.

Updates to Project Approach & Research Planned

The two things I wanted to do for my research were to:

1. Research the findings of existing research on this topic; what conclusions did they jump to, and how could that research help me in the questions I wanted to answer? If the researched the exact same topic, what questions did they leave out or what additional research would be needed?
2. Answer 4 specific questions regarding tweets that were exclusively flagged as hate speech, primarily if context plays a role, if they violated other Twitter terms of service, and whether or

not there would be any significant advantages to qualifying certain tweets as hate speech compared to offensive speech.

These aspects of my research have not changed. I have spent a lot of time reading existing research on the subject matter, and I plan to read a few more papers, specifically ones that pertain to the questions that I am trying to answer. I may come up with additional questions in addition to the top 4 I mentioned earlier.

If there was some new kind of research I would conduct for this project, it would be to find papers or articles comparing how an AI judges hate speech compared to humans, as that was a significant part of my project plan that has changed since I decided on it in Week 2. Some of the articles that I have read so far hint at this concept, but none discuss it as their sole goal of their research. I will try to find articles I can read for this to help me get some insight; if not, I will focus primarily on comparing how programs' and peoples' approaches differ on hate speech.

Updates to Team Structure

None; it is still just me, though I am getting advice from people who have done similar work on the subject, which has proven to be incredibly helpful.

Updates to Project Milestones

I set 5 very basic things to have done by this point in time in this project, and I have and continue to make significant progress on them:

1. Obtain research data set – Complete.
2. Begin search for additional data, primarily from Facebook – Decided to not dedicate more time to it, as Facebook data is much harder to access and is significantly more expensive compared to Twitter data.
3. Found webpages for scholarly articles – Complete, and also found a few additional ones as well that were incredibly helpful.
4. Begin data analysis and taking notes on articles – In progress. I read and taken notes on many articles, and I still plan to find a few more that cater to some of the specific questions I want to answer, so in regards to humans vs. robots in terms of detecting hate speech. If I cannot find anything additional that will be helpful, I will stick to the articles I have, as they have been incredibly informative.
5. Reaching out to people to discuss my topic – Also in progress. I have gotten a lot of insight and knowledge from them about the topic, certain factoids that were interesting, and some interesting questions to try and answer.

Overall, most of the milestones are either complete or have a lot of progress completed. The finding additional data one was scrapped, but I can say that, at this point, a few more were added:

1. Trying to find existing research comparing computers to robots at judging hate speech – what disparities exist, and why?
2. If that research is not available, try doing as much of it as I can, with a basic search / filter bot, and perhaps another feature to flag all tweets that use certain language.

Date & Time of midpoint meeting with Professor: 07/21, 3:20 PM