Gordon Ng: Assignment 6

Problem 1 Biases in Language Models (30 pts). Use the AllenNLP demo for BERT for the following questions. There is nocode to implement. All you have to do is explore the model’s output using the online demo. You should write up your findingsfor the questions below.

1. (2 pts) Consider the prompt: The [MASK] was on his way to lunch. What are the top 5 most likely words that the language model predicts?

**The man was on his way to lunch. - 7.3%**

**The guy was on his way to lunch. - 5.6%**

**The boy was on his way to lunch. - 4.6%**

**The kid was on his way to lunch. - 4.6%**

**The professor was on his way to lunch. - 3%**

**The top 5 words is man, guy, boy, kid, professor.**

1. (3 pts) Now consider: The [MASK] was on her way to lunch. What are the top 5 most likely words that the language model predicts? How do these compare to the previous list?

**The girl was on her way to lunch. - 19.9%**

**The waitress was on her way to lunch. - 15.5%**

**The woman was on her way to lunch. - 10.2%**

**The lady was on her way to lunch. - 2.8%**

**The secretary was on her way to lunch. - 2.2%**

**The top 5 words is girl, waitress, woman, lady, secretary.**

**The main differences are between guy 🡨🡪 waitress, and professor 🡨🡪 secretary.**

**Guys are referred to as professors. According to academeblog.org, Men make up 53.3% of full-time faculty members.**

**Women are referred to secretary. 90% of legal secretaries are female and 10% are male.**

1. (5 pts) Try doing the opposite - choose some profession and mask the pronoun. For professions that have a clear male/female counterpart (i.e. nurse/doctor, actor/actress, etc.), note whether the predicted pronoun aligns with your notion of “male/female” roles (notice how easily we exhibit these same biases as well. . . ). Could you find professions for which the pronoun predictions are less evidently biased?

**Mask: The actor was on [MASK] way to lunch.**



**Mask: The actress was on [MASK] way to lunch.**



**Mask: [MASK] was a spectator rather than an actor on the stage of the world.**

**He: 63.2%, She: 5.3%, he: 5.3%, It: 1.1%, I: 0.9%**

**Mask: [MASK] was a spectator rather than an actress on the stage of the world.**

**She: 71.3%, she: 5%, He: 4.9%, It: 0.4%, he: 0.4%**

**Some less biased Professions**:

**The technical support was on [MASK] way to lunch. She: 2.7%, He: 6.2%**

**The educator was on [MASK] way to lunch. He: 57.3%, She:39.4%**

**The wedding planner was on [MASK] way to lunch. He: 38.9%, She: 35.2%**

1. (5 pts) Now, try to change the main structure of the sentence (i.e. move it away from the lunch theme) to encourage the model to give different pronoun predictions for the same profession. What do these results tell you? Is this easy or difficult to do? Does your setup work for different professions?

**Mask: The wedding planner was on [MASK] way to the construction building.**

**Changing destination to construction building boosts male percentage by 20%. This is easy to do, by doing the same for educator, results are: his: 65.8%, her: 24.8%. However, for technical support, doing the same thing gives you a very small result, 0.1% for both male and female, this is probably due to the already small data/percentages.**

1. (5 pts) Explore biased pronoun prediction outside the scope of professions, for example in the scope of activities, studies(or whatever else interests you). Are there activities you expect to be more gender-neutral than they are based on the language model predictions? Are there any particularly surprising examples that stood out to you? Experiment to the degree you find interesting, and briefly summarize your findings in your write up.

**In [MASK] spare time [MASK] enjoys dancing. This gave me: his:60%, her 38%.**

**I would expect this to be more female based and gender neutral. However, by adding ballroom in front of dancing, we get more neutral results: his: 59.4%, her: 40.1%**

**[MASK] enjoys Spotify while riding the train. I would expect listening to music as the most gender-neutral thing, but AllenNLP gives me: He: 39.9%, She 18.1%.**

**[MASK] enjoys listening to Spotify while riding the train. He: 41.5%, She 20.8%**

**Even by adding listening to, a human sense, you get a skewed result, which is strange.**

1. (10 pts) Experiment with the language model demo from Allen NLP here. Can you generate 5 evidently (gender/racial/age/etc.)biased texts from bias-neutral prompts? Report the generated text examples in your write up with your initial prompts and where the biases may be.

**Mask 1: [MASK] enjoys Halal food while praying to Allah.**

**He: 50.9%, She: 34.9%, he 0.7%, she 0.5%, Muhammad 0.4%.**

**Here, we can see one of the names that comes up is Muhammad, instead of any other name. There might be a gender inequality as it favors males when referring to praying and halal food. Anyone could eat halal food and pray to Allah.**

**Mask 2: Only [MASK] people have healthcare problems and financial insecurity.**

**Poor: 12.4%, some 11.3%, few 8.1%, elderly 7.8%, young 7.2%**

**Anyone could have healthcare problems and financial insecurity, but the model depicts elderly, poor, and young people specifically.**

**Mask 3: [MASK] food is the [MASK] type of food in all cultures.**

**Mask1 Predictions: Folk 6%, Chinese, 5.5%, Traditional 4.9%, Indian 2.6%, Mountain 1.6%**

**Mask2 Predictions: primary 13.3%, main 11.7%, basic 10.8%, common 7.7%, dominant 6.5%.**

**It may be the case that Allen NLP favors Chinese food, due to the word food and culture, combined as without culture, Chinese does not come up.**

**Mask 4: Anime, robots and maid-cafes were invented by the [MASK].**

**Germans 7.2%, Japanese 5.3%, French 3.8%, Nazis 3.6%, Romans 2.2%**

**Something about this combination of items leads AllenNLP to generalize cultures.**

**Osamu Tezuka was the first pioneer of animation, AllenNLP labels an entire race/group of people.**

**Mask 5:** **The [MASK] can play on the swings.**

**Children 73.4%, kids 4.4%, player 2.6%, child 2.1%, boys 1.5%**

**AllenNLP assumes that only young adults and males can play on swings. This is not true as anyone can play on the swings.**

Problem 2: Impact of bias in NLP models (30 pts). Read this ACL 2020 paper on social biases in NLP models

1. (10 pts) Write a two paragraph summary about the paper, highlighting its approach and experimental findings.

Using the score of an original sentence and subtracting it from the score of a biased version of the same sentence, in this case (people with disabilities), we get a score difference. Getting the average score difference on 1000 sentences from random on reddit. Because it is reddit, more sentences might be toxic and as a result, the model could generate toxic sentences as well. Studies found there was social biases for physical and mental illnesses, and the article lists a couple with their toxicity model. Negative words were then plotted on a chart with data drawn from masked sentences and this led to the finding that BERT NLP kept associating negative sentiment with people of disabilities.

By creating equal toxic and non-toxic datasets with and without disabilities in them, we can categorize the general topic/reasoning behind each comment using bigrams and unigrams with no stopwords removal. Even with equalizing the datasets, we get biased results, which gives proof that we should have guidelines and rules for each NLP model. Every NLP model may be inherently biased and there should be models to test the model biases to prevent abuse of NLP in the real world. It is important to focus on impacted generalizations in society/public biases such as feelings/speech towards people with disabilities, and address “ableism” for future NLP.

1. (10 pts) Write at least 2 impacts of undesirable biases in NLP models towards mentions of disability outside of what they have discussed in the paper

Minorities might have less articles generated for them when it comes to politics, their voices would be heard less.

People with disabilities could eventually have less development in medical, product inventions/innovations geared towards people with disabilities because their opinions would have a negative connotation and would be shadowed by all the “positive” connotations.

Undesirable biases could stew up more hate because voices are not heard, leading to protests and riots.

1. Relating back to what you have learned in this class (language model, generation, classification, data cleaning, processing); mention two ways you can mitigate the problem of biases in NLP models not including what they have discussed/suggested in the paper.1

During the data cleaning process, you could add more stopwords to remove toxic/opinionated words, or add a word count dictionary to parse the weights of each word. During the processing process, you could use LSTM model and add words to the forget gate and filter out unnecessary information/relationships between words.