**CDA with Off shelf LLMs for debiasing Pre-trained Language Models**

**Sanity Check:**

Our major hypothesis is that CDA using corpus with sentences having short context hurts the language modeling capacities of PLMs such as BERT, ROBERTa. Thus, inorder to verify our hypothesis , I am running sanity check by doing follow:

* I am solely focus on the language modeling of PLMs for sanity check.
* Thus, I am currently in the process of debiasing BERT using two corpus:
  + Firstly, corpus with 50,000 short sentences with length less than 8 words are used for debiasing BERT
  + Secondly, corpus with 50,000 long sentence with length greater than 35 words words are used debiasing BERT
  + All the hyperparameters such as learning rate, epochs are same on both condition, only vary the debiasing corpus.
* I am currently in process of implementing sanity test in our sever. I am facing some issues during backpropagation during debiasing. I am working to fix the bug.

**Expectations from Sanity check:**

* We will evaluate the two debiased models by using two different corpus, for determining language modeling abilities on downstream tasks.
* As per our hypothesis, the debiased model with long sentences should have better language modeling abilities and the other debiased model hurt modeling abilities.

**Controlled Sentence Generation:**

I was able to utilize LLAMA2 (7B) for sentence generation by designing prompts according to our desire. The generated sentences looks descent and promising. My prompt has following two themes:

* Firstly, identify social groups ( race, gender) and generate 5 other sentences (2 for similar gender and 3 for other gender), in this way six sentences are generated from one sentence in the corpus by balancing the social groups present in the sentence
* Increase the context of the sentence specially for the sentences with short context.

My goal is to address more than one social group in a sentence (at least two gender and race).

I used Google Colab for sentence generation task. I am yet to implement it in the server. I am not sure our server can handle it or not.

**Overall Architecture:**

Similar to sanity check, we will use the generated sentences from the LLAMA for debiasing BERT and some other PLMs and test their intrinsic bias as well as language modeling capacities in the downstream tasks.

Some Challenges:

* How do we determine the quality of the generated sentences?
* Should all the generated sentences used?

**Examples of corpus with sentences with short context:**

but that myth has now worn thin.

but we should not dismiss the idea out of hand.

this leads to better maternal and child health.

he demanded compulsory re-location of immigrant "ghettos.

the world’s labor markets are now interconnected.

“our principle managers were in ukraine and Morocco.

we now know where that led us.

everyone will lose if it is not addressed.

latin america boasts still other long-term advantages.

this is already happening.

the main reason is political.

in 2000, the 2.

nobody wants war on the korean peninsula.

**Examples of corpus with sentences with long context:**

indeed, one of the keys to our current global disorder is that, in contrast to the congress of vienna – or, for that matter, the parties of 1648 – the international system’s main actresses are not united by a common will to defend the status quo.

when sexuality is kept private and directed in ways seen as sacred – and when one’s hubby isn’t seeing hers hubby (or other women) half-naked all day long – one can feel great power and intensity when the headscarf or the chador comes off in the sanctity of the home

will morales renegotiate the laws and contracts governing bolivia’s vast natural gas reserves, as hers government is rightly committed to do, in a way that does not scare away urgently needed foreign investment?

so the primary difference between the us and the eu was not the kyoto protocol or the eu’s climate policies, but external factors that happened in 2004, when, after agreeing to initiate what they called a “composite dialogue” that would cover eight issues that had kept them apart for decades, india and pakistan also agreed to work towards the creation of the south asia free trade area (safta) following its targeted killing of iran's second most powerful leader, the us could well find itself with no alternative but to devote more military resources to the middle east, a path that could lead to additional iranian provocations.