Allen Cheung, Kaiyuan Wang, Beibei Du
Professor Bergen
LIGN 167
February 8, 2022

Final Project Proposal

1. **What tasks are you trying to accomplish?**

   We are aiming to develop a deep learning model that can classify hate speech. In the world of rapid technological development, we are more connected than ever before. Consequently, access to the internet provides opportunities that allow people to freely speak whatever they want, resulting in racist, sexist, or homophobic speech. Thus our task is to create a language model that can classify the hate speech from various datasets compiled from the Internet, such that this information can be filtered out if needed, or the user can be warned.

2. **How are you planning to do it？**

   We are planning to replicate and potentially improve upon a project from the Stanford CS224n course, *AItBERT: Domain Specific Pre-training on Alternative Social Media to Improve Hate Speech Classification*, to perform the hate speech classification task. More specifically, we aim to investigate whether domain specific pre-training could produce an effective model to solve the classification problem. We will refer to the BERT model and the other language models potentially mentioned in this paper as a baseline for our experiments. We plan to extend the BERT model, which learns word embeddings, for the specific task of classifying hate speech. Pretraining the model on a large corpora of hate speech text, such as 4chan or Parler, may help to improve the accuracy of the model for the task at hand. Additional datasets such as the HuggingFace hate speech tweet dataset may also be used for training our model.

3. **How are you planning to evaluate whether you have succeeded?**

   We plan to evaluate the performance of our model based on the comparison with several baseline models. The baseline model could be SVM (support vector machine) or some other more advanced models without task-specific fine-tuning. Achieving a superior performance with our model would indicate a success.

4. **Potential challenges and backup plan**

   One pitfall is the task being too simple - detecting hate speech by looking for specific hateful words could be a simple yet effective solution. If it turns out that our baseline model performs exceptionally well, we plan on extending our project object beyond merely classifying. We could use generative models to neutralize the hateful component with minimal alteration to the overall meaning.

Resources for Reference:
StanfordCS224: https://web.stanford.edu/class/cs224n/project.html
Hugging Face (Learn SOA Models): Hugging Face

There are generally two routes to go for the final project: build something relatively simple (but beyond what we learned in class) from scratch, or **build something more complex by extending existing libraries.** Both are good options, depending on your interests.

For getting ideas about projects, take a look here: https://web.stanford.edu/class/cs224n/project.html The papers here will give a sense about what's currently possible, and also will include references to useful prior work.