CS6830 Final Project Proposal

Laurel Bingham, Megh KC, Matt Rau

As generative AI continues to grow more competent at replicating human speech and art, the internet will be flooded with generic generated content. There are already AI models designed specifically to create marketing pages, social media posts, blog posts and emails. This continues to expand month by month to more complicated mediums. AI made YouTube videos, AI created podcasts or AI written news articles. As every corner of the internet becomes filled to the brim with generated content, humanness will become more and more important. People will strive to find the real person in a sea of fake, but it will get increasingly difficult.

For this reason, models which can detect AI generated content will be in high demand. Our goal with this project is to build multiple models which can sort out human written news articles from articles written by OpenAI's ChatGPT. These AI generated articles will be created using a real headline scraped from various news sites. An example of a prompt we would use is "Act as a journalist from the New York Times. Write a 600 word news article with the headline 'E.P.A. Is Said to Propose Rules Meant to Drive Up Electric Car Sales Tenfold'".

When it comes to getting data and preprocessing, Laurel will focus on scraping the web for articles and headlines to use in our dataset, while Matt will focus on using OpenAi's API to automatically generate articles based on said scraped headlines. Then Megh will focus on preprocessing the data, removing punctuation and non-words, converting everything to lowercase, and creating some bag of words or tf-idf representation of word frequency, as well as any other feature extraction deemed relevant. Once the dataset is complete and ready for analysis, we will each take on one different machine learning model type, such as recurrent neural networks, naive bayes, or random forests, and optimize those as much as possible, along with creating any relevant visualizations and graphs.

Our goal is to have all the data collection and some preprocessing done by 4/12, the day of the proposal presentation. We will finish up preprocessing and feature extraction, then start on our own individual models the rest of that week, up 4/16, and then have three days before the due date to focus on the presentation.

There could be a potential issue of finding features that are significantly correlated to our target. Word frequency will likely not be enough for an effective model. Some potential other features that we could look at would be paragraph/sentence length, use of/frequency of punctuation. We could also use potentially harder to quantify

features like vocabulary richness or use of named entities, such as specific people or organizations.

It could also be possible that we are unable to build a model with significant results. OpenAI's own AI classifier only correctly identifies AI-written text about 26% of the time. There are a few successful models, like zeroGPT, which claims a 98% accuracy, but there's a decent chance we won't be able to make a very successful model.

Thus far, we have fully been focused on collecting data to create the dataset for our model. Each news website will have articles and headlines scraped from many different news subjects, from politics to sports to finance. There have been some holdups on the web scraping side that have slowed things down. The major hurdle for gathering the news articles has been both that many APIs for news article scraping have been deprecated, and that current APIs do not return the full body of text for the article. Though it has taken some experimenting with various tools, a solution has finally been found in a more generalized web scraping tool. It's been proven to work on articles off of CNN (The news site, not the neural net architecture), and should be trivially easy to do the same for Fox, and other online news sources. Now that this hurdle is cleared, we are ready to take those scraped headlines and generate the fake articles to finish up making the dataset. Then we will be able to continue on with the data preprocessing and the creation of our models.

After preprocessing and data cleaning steps, each one of us will use different machine learning models to identify the fake news generated by chat gpt. Each of these models will be trained on both the real articles and headlines scraped from the web, and the fake articles made with the same headlines. The different models tested will be Naive Bayes, Random forests and a Recurrent Neural Network. Of course, dataset visualization with different plots and descriptive statistics of the dataset will also be presented. The model evaluation results, prediction accuracy and visual effect would be added in our model implementation.

Overall, our AI generated content detection project should help media platforms be certain that they're hosting human created content. While we don't yet know which model will prove to be the most accurate at identifying generated writing, discriminating between real and 'fake' news or writing of any kind has become a genuine issue for social media platforms around the world.It is our hope that we can find some identifying features typical of chat-gpt generated text, even if we do not achieve high accuracy in our classification.