Max Orenstein
2/27/2024

Moderating Hate Speech From the 2020 Presidential Election Using GPT-4

For this project I used data from a study entitled “Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection" in the Association for Computational Linguistics Journal by Laura Grimminger and Roman Kingler. The data included tweets from the 2020 Election cycle regarding 3 candidates: Donald Trump, Joe Biden, and Kanye West. The tweets were labeled across two dimensions: Stance Detection (Favor, Against, Neither, Mixed, Neutral) and Hateful (Hateful, Not Hateful.) For this project I wanted to see if I could reproduce or improve their results with a prompt I gave to OpenAI’s GPT-4 as well as expanding the study across three new dimensions: Discrimination Type (General, Sexist, Sexual Harassment, Homophobic, Racist, Transphobic, Ableist, Intellectual, Ageism), Severity (Mild, Moderate, Severe), and Directedness (Implicit, Explicit.) These new categories gave me new insights into the tweets as well as new ways to test the model.

I developed my taxonomy to understand what sort of hate speech was being dispersed about these candidates as well as which candidates got the most hate. I used the readme file attached to the data from the study as a starting point for the prompt. I used generative AI, readings from class, and my intuition to build up the prompt. I split it into clear sections for clarity and made the output structure very clear. Once I had something that I felt good about I began testing the prompt using a random selection from the dataset. To get the random selection I used the =RAND() function in Excel and ordered by largest to smallest, and to test the prompt I used OpenAI playground. These tests revealed some issues with the output structure and also some difficulties determining correct classifications. I made note of potential issues and went back into the prompt to make adjustments. For example, the initial tests revealed the model was ignoring hashtags so I adjusted the prompt to explicitly mention the use of hashtags. If I were to run the experiment again I would spend more time in the playground and make adjustments to improve accuracy and debug output errors.

Once I settled on a prompt and ran it through my dataset I began assessing the accuracy and effectiveness of my approach. I began by comparing how similarly the GPT did to the previously labeled data. Across all categories my prompt generated answers that were 80.71% similar to the previously labeled data. For individual categories the similarity looked like this: Trump Stance Detection: 69.54%, Biden Stance Detection: 74.25%, West Stance Detection: 98.54%, Hate Speech Detection: 80.54%. This distribution could suggest that the model was far more accurate in classifying stance as "Neither" than "Favor" or "Against" since most of the tweets weren't about Kanye. However, raw matching is not the best way to evaluate effectiveness since the initial taxonomy may be outdated, incorrect, or possibly mistake-ridden. Thus, I also looked at individual case examples to evaluate the effectiveness of this approach.

Text	Labels	Model	Notes
So biden will have scripted answers, will be drugged up, and will have an earpiece giving him notes. #LowEnergyJoe #DebateTuesday #debate #Trump2020 https://t.co/nBEy53ymuT	Favors Trump, Against Biden, Non-Hateful	Favors Biden, Against Trump, Hateful	In this example the model flip-flops who the tweet favors for no obvious reason. This sort of blatant error although egregious is relatively rare.
@dominick_massa @CBSNews @MajorPatriot Go say a pray for PRESIDENT-ELECT JOE BIDEN AND VICE PRESIDENT ELECT KAMALA HARRIS OH YEAH. F##k god	Favors Biden, Non-Hateful	Favors Biden, Hateful	This example shows a weird hiccup for the AI. It's hard to say what this tweet intends and whether or not it's targeted without adequate context.
Congratulations President-Elect Joe Biden &amp; Vice President-Elect Kamala Harris. #2020Election https://t.co/YaxDKskEGw	Biden Neutral, Non-Hateful	Favors Biden, Non-Hateful	This one the model classifies better than the labelers. This highlights the problem of overreliance on similarity.

There were three main challenges with my approach. First, the prompt asks a lot of the model in one output. I may have been better off splitting the various categories into individual prompts and running them on their own. Thus by focusing on a single task we could avoid some of the few but present structural mistakes the model made and zoom in on smaller areas for improvement. Second, the assessment of the model was a difficult task given the time frame. As noted, the similarity score while a helpful ballpark is unsatisfactory. I would've liked to go through each tweet individually and manually check the results myself. This also would've helped me understand the strengths and weaknesses of my prompt and know where it could use adjustment. Finally, there's the umbrella issue of AI's contextual difficulties. Adding more examples, fine-tuning the model for political discourse, and more time spent with the prompt could help alleviate some of these issues but likely not all of them.

Prompt:

1. Description

This file contains offensive and hateful content. We have collected tweets about the 2020 US presidential campaigns and election. Given the text of a tweet, we want to annotate the stance the tweet holds towards our predetermined targets and the presence or lack of hateful and offensive speech. Your output should consist of several different classifications of the text separated by commas. The classifications are **Stance Detection**, **Hate Speech Detection**, **Discrimination Type**, **Severity**, and **Directedness**. Your final output for each cell should look like this: [**Stance Detection**], [**Stance Detection**], [**Stance Detection**], [**Hate Speech Detection**], [**Discrimination Type**], [**Severity**], [**Directedness**]. No additional words should be added and you must classify each cell based on the categories I give. I'll now go through each classification one by one and how I want you to determine which category to use.

1.1 **Stance Detection**

This category identifies the stance of a tweet's text towards three specific targets: Donald Trump, Joe Biden, and Kanye West in that exact order. The stance is classified according to the tweet's support or opposition to these individuals, based on predetermined annotation labels:

- Favor: The tweet supports or advocates for the target.
  - Example: "Biden has the experience we need in a president."
- Against: The tweet opposes or criticizes the target.
  - Example: "Trump's policies have failed us."
- Neither: The target is not mentioned, either implicitly or explicitly.
  - Example: "Voting is our civic duty."
- Mixed: The tweet contains both positive and negative aspects about the target.
  - Example: "Kanye is innovative but unpredictable."
- Neutral mentions: The tweet states facts or quotes without taking a position towards the target.
  - Example: "Trump signed a new executive order today."

Guidelines:

- Tweets must explicitly or implicitly mention the targets to be labeled under any category other than "Neither."
- References to political parties or slogans (including hashtags) associated with the targets can indicate a tweet's stance.
- Explicit mentions include direct names or titles, while implicit mentions may involve slogans or references to the candidates' positions.
- The stance towards one target can be affected by mentions of associated parties or vice presidential candidates.
- Offensive or derogatory nicknames and hashtags that are critical of or supportive towards the targets or their parties should be considered in determining the tweet's stance.

- You should return one stance detection for each candidate separated by commas in the order Trump, Biden, West. This order must not change. Ex: Neither, Neither, Favor.

1.2 **Hate Speech Detection**

This category identifies tweets that contain hate speech or offensive language directed at individuals or groups, particularly in the context of the 2020 presidential election. Tweets are classified as either "Hateful" or "Non-Hateful."

- Hateful: Tweets that explicitly or implicitly demean or threaten a person or group based on aspects of their identity or political affiliations.
  - Example: "Candidate X's supporters are ruining our country."
- Non-Hateful: Tweets that do not contain language that demeans or threatens individuals or groups.
  - Example: "I disagree with Candidate Y's policies."

Guidelines:

- Hate speech can include abusive, degrading speech, violent threats, insults, and racial or sexist slurs.
- Name-calling or derogatory references to political figures or their supporters is considered hateful.
- The context and combination of words are crucial in determining whether speech is hateful or offensive.
- When determining if a tweet is hateful, consider whether the language used would be offensive to the average person.
- Slang and abbreviations (including hashtags) commonly understood to be derogatory should be flagged as hateful, though some words may have context-dependent meanings.

1.3 **Discrimination Type**

This category identifies the specific nature of discrimination present in the content. Classification is based on the targeted aspect of identity or characteristic. Each tweet should be evaluated for content that discriminates against individuals or groups based on the following types:

- N/A
- General: Discriminatory content that does not fit into the more specific categories listed below but still contains elements of discrimination.
  - Example: "All politicians are corrupt and cannot be trusted."
- Sexist: Content that discriminates based on gender, typically against women, including stereotypes or derogatory comments.
  - Example: "Women shouldn't be in politics; they're too emotional."
- Sexual Harassment: Content that includes unwelcome sexual advances, requests for sexual favors, and other verbal or physical harassment of a sexual nature.

    - Example: "She only got her position by sleeping her way to the top."
- Homophobic: Content that expresses fear, hatred, discomfort with, or mistrust of people who are lesbian, gay, or bisexual.
    - Example: "Being gay is not natural and should not be promoted by politicians."
- Racist: Content that discriminates against individuals based on their race or ethnicity.
    - Example: "This country doesn't need a president from that race."
- Transphobic: Content that expresses fear, hatred, discomfort with, or mistrust of transgender individuals or transsexuality.
    - Example: "Transgender people shouldn't be allowed in the military."
- Ableist: Content that discriminates against people with disabilities.
    - Example: "He's too mentally unstable to hold office."
- Intellectual: Content that discriminates against individuals based on their intellectual capabilities or expressions of intellectual elitism.
    - Example: "Only educated people should be allowed to vote."
- Ageism: Content that discriminates on people based on age, referencing senility or declining mental faculties.
    - Example: “He couldn't even remember his speech!”

Guidelines:

- Identify the Nature of Discrimination: Carefully assess each tweet to determine the specific type of discrimination it contains. Pay attention to the aspect of identity or characteristic that is being targeted.
- Use Examples for Clarity: Refer to provided examples to better understand the nuances of each discrimination type. This will aid in accurately classifying tweets that may not explicitly mention a category but imply discrimination.
- Consider Context: Sometimes, discrimination might be implicit or embedded in jokes, sarcasm, or cultural references. Consider the broader context of the tweet to accurately identify the type of discrimination.

1.4 **Severity**

The severity of the detected hate speech or discrimination can be rated on a scale from N/A to 3, where each level represents the intensity of harm or potential for harm:

- N/A
- 1: Mild - Content contains minimal discriminatory language or implications without intent to harm.
    - Example: "This candidate seems clueless about real-world problems."
- 2: Moderate - Content explicitly discriminates or uses hate speech but falls short of inciting violence or harm.
    - Example: "Politicians like her are what's wrong with this country, always lying."

- 3: Severe - Content contains clear, targeted hate speech or discrimination with potential to incite or endorse harm.
    - Example: "People from that place are destroying our society and need to be stopped."

Guidelines:

- Assess Harm Potential: Evaluate the content's potential to cause harm or incite violence against individuals or groups. This involves understanding the intensity and impact of the language used.
- Scale Appropriately: Use the provided scale from N/A to 3 to rate the severity of hate speech or discrimination. Ensure that the rating reflects both the explicitness and the potential harm of the content.
- Acknowledge Subjectivity: Recognize that assessing severity can be subjective. When in doubt, consider consulting with peers or supervisors to reach a consensus on the appropriate severity rating.

1.5 **Directedness**

This category assesses whether the hate speech or discriminatory language is directed towards a specific individual or group, or if it's implicit or general in nature.

- N/A
- Explicit: The content clearly identifies an individual or group as the target of hate speech or discrimination. This can include direct mentions, tags, or clear references to specific identities.
    - Example: "John Doe is a disgrace to his country."
- Implicit: The content suggests hate speech or discrimination without directly naming or identifying the target. This can include veiled references, stereotypes, or coded language that implies a target without explicit mention.
    - Example: "Some people just don't belong in positions of power, especially if they can't understand basic values."

Guidelines:

- Determine Targetedness: Identify whether the hate speech or discrimination is directed towards specific individuals or groups, or if it is more implicit or general in nature.
- Explicit vs. Implicit: Distinguish between content that explicitly names or identifies its target and content that implies targets through veiled references or coded language.
- Use Judgment: Evaluating directedness may require judgment calls, especially with implicit content. Consider the potential targets and the context within which the statement is made to accurately classify the directedness of the content.

1.6 Recap

Remember:

- Reason in order, step-by-step
- Only pick one category for each classification
- If something isn't hateful be consistent and respond with N/A to 1.3 1.4 and 1.5
- Your final output for each cell should look like this: [**Stance Detection**], [**Hate Speech Detection**], [**Discrimination Type**], [**Severity**], [**Directedness**]

Acknowledgment of AI Use:

For this project I used AI both to generate an output, but also to help me fill out the prompt. I did the latter for two primary reasons: 1) I've developed prompts for AI systems using AI systems in the past and found it not only to be effective but also a massive time saver, and 2) I wanted to see what a zero-shot attempt at defining various aspects of hate speech would look like from ChatGPT. That being said I was unsatisfied with various aspects of the prompt that the model produced including some milk-toast definition and various things it didn't account for. I used my knowledge from this class as well as several of the articles cited to manually improve the prompt and tested it several times for inaccuracies and inefficiencies. Overall it did a pretty good job generating basic examples and some of the guidelines and was a good starting point for the project. Going forwards I would use this method again in certain scenarios but would be highly skeptical of any approach that includes no human oversight.

References:

Citron, Danielle Keats. Hate Crimes in Cyberspace. Caimbridge, MA: Harvard University Press, 2014.

Grimminger, Laura, and Roman Kingler. "Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection." Association for Computational Linguistics, April 2021, 171-80.

Klonick, Kate. "THE NEW GOVERNORS: THE PEOPLE, RULES, AND PROCESSES GOVERNING ONLINE SPEECH." Harvard Law Review 131, no. 6 (2018): 1598-670.

Newton, Casey. "THE TRAUMA FLOOR: The secret lives of Facebook moderators in America." The Verge. Last modified February 25, 2019. Accessed February 27, 2024. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-condi

Wilner, Dave, and Samidh Chakrabarti. "Using LLMs for Policy-Driven Content Classification." TechPolicy.PRESS. Last modified January 29, 2024. Accessed February 27, 2024. https://www.techpolicy.press/using-llms-for-policy-driven-content-classification/.