**Trial of GPT-3.5 on Anti-Racism Dataset**

**NB:** **The results shown below was by analysing each tweet individually based on the context of parent tweet only without considering any reply tweets above it.**

**System prompt used:** Whenever given a thread of tweets (The first tweet being parent tweet and following which are only 1st level replies to the same), you have to analyze and tell whether a tweet is hateful, counter speech or neutral statement labelling as 1, 0 or 2 respectively. Give the response in tabular format with reasons. Labelling should be in context with the parent tweet and your general awareness of the topic being discussed in the conversation.

1. **Response for some initial tweets of Thread-1:**

| **Tweet** | **Label** | **Reason** |
| --- | --- | --- |
| **Parent Tweet**: "The people of Germany are turning against their leadership as migration is rocking the already tenuous Berlin coalition. Crime in Germany is way up. Big mistake made all over Europe in allowing millions of people in who have so strongly and violently changed their culture!" | 1 | The tweet portrays immigrants negatively and suggests they have violently changed European culture, which can incite hostility. |
| **Reply 1**: "Trump does not love you, He loves himself. Don’t be like him, Don’t teach you children to be like him. Stand up for humanity, Oppose Trump." | 0 | This reply opposes Trump and encourages standing up for humanity. |
| **Reply 2**: "We don't believe you." | 2 | This reply expresses disbelief in Trump's statement. |
| **Reply 3**: [Empty reply] | 2 | There is no content to analyze. |
| **Reply 4**: "#WeDontBelieveYou" | 2 | This hashtag expresses disbelief in a neutral manner. |
| **Reply 5**: "You don’t know what you’re talking about." | 2 | This reply dismisses Trump's statement. |
| **Reply 6**: [Empty reply] | 2 | There is no content to analyze. |
| **Reply 7**: "Trump is more interested in the crime rates in other countries than he is in our own. That's because he was installed by Putin to divide the USA and to divide the West. He's doing Putin's work" | 0 | This reply criticizes Trump's focus on other countries' crime rates and suggests ulterior motives. |
| **Reply 8**: "You’re the guy who locks little kids in cages. I don’t think you get to tell anyone anything any more Donald." | 0 | This reply criticizes Trump's actions of separating children from their families at the border. |
| **Reply 9**: "Germany: apologies for our @POTUS." | 0 | This reply apologizes on behalf of Trump, indicating disapproval of his message. |
| **Reply 10**: "At least Germany is still a democracy with good leadership. We used to be. #WeDontBelieveYou" | 0 | This reply contrasts Germany's democracy with criticism of Trump and expresses disbelief. |
| **Reply 11**: "#WeDontBelieveYou #WeDontBelieveYou #WeDontBelieveYou #WeDontBelieveYou #WeDontBelieveYou | 1 | This reply contains a hateful hashtag calling for Trump's resignation using derogatory language. |

1. **Response for some initial tweets of Thread-14:**

| **Tweet** | **Label** | **Reason** |
| --- | --- | --- |
| **Parent Tweet**: "Congress has finally passed a targeted COVID bill to give much-needed relief to Americans who've shouldered the weight of this pandemic & provide more vaccine resources. This deal comes at a critical time as vaccines are being distributed. There's light at the end of the tunnel." | 2 | The parent tweet expresses relief and positivity about the passage of a COVID relief bill, indicating a neutral sentiment. |
| **Reply 1**: "That'll be fine. Earmarks simply aren't the problem these days. Getting literally anything done is the problem. If someone has to roll logs and make sausage to get something done, then again, that'll be fine." | 0 | This reply counters the notion that earmarks are problematic and highlights the difficulty of achieving progress in Congress, aligning with the neutral sentiment of the parent tweet. |
| **Reply 2**: "No question $600 per person is a joke, but since when is culture disposable? Are artists not Americans as entitled to survive as corporate employees?" | 0 | This reply critiques the relief amount but advocates for support for artists, reflecting a nuanced view in line with the topic of relief during the pandemic discussed in the parent tweet. |
| **Reply 3**: "Won't make a difference as long as people vote Republican or Democrat. No changes will happen." | 1 | The tweet expresses frustration with the political system, diverging from the positive tone of the parent tweet and reflecting a negative sentiment towards partisan politics. |
| **Reply 4**: "This is a Big FUCK YOU to Americans people who's suffering. They know we can't live off of $600" | 1 | This tweet expresses anger and frustration about the perceived inadequacy of the relief amount, aligning with the discontent expressed in the parent tweet about the need for relief during the pandemic. |
| **Reply 5**: "Mush of this money for the arts could have gone a long way to help the restaurant industry for the next 3 months." | 0 | This reply criticizes the allocation of funds while suggesting a different allocation, contributing to the discussion on relief distribution without expressing overt negativity, fitting with the overall topic of relief in the parent tweet. |
| **Reply 6**: "Um. Israel is spelled wrong. Maybe spell check before posting made up things, next time?" | 2 | This tweet points out a spelling error without contributing to the discussion on relief or expressing sentiment towards the topic, warranting a neutral label. |
| **Reply 7**: "I’m not fine. I got a 10% wage decrease, my company matching everyone’s 401k also no bonuses this yr. They received government & suffered only a 3% decrease in total numbers for the year. The following yr was a record yr for them w/less taxes. My 3 adult kids moved home" | 2 | The tweet shares personal hardships related to the pandemic without directly engaging with the relief bill, warranting a neutral label. |
| **Reply 8**: "It’s time we March on dc and get rid of these politicians" | 1 | This tweet suggests drastic action to address political dissatisfaction, which may not directly relate to the relief bill but reflects broader discontent with governance, warranting a negative label. |
| **Reply 9**: "Israel 500,000,000 ???? Is this true?" | 2 | This tweet questions the allocation of funds to Israel without expressing clear sentiment towards the relief bill, warranting a neutral label. |
| **Reply 10**: "Ole Nancy needs to go! She has gotten rich off of us." | 1 | This tweet criticizes Nancy Pelosi and suggests personal enrichment, diverging from the discussion on relief and reflecting a negative sentiment, warranting a negative label. |
| **Reply 11**: "Why tf does the smithsonian need $1B" | 1 | This tweet questions the allocation of funds without engaging directly with the relief bill, reflecting skepticism about fund allocation and warranting a neutral or slightly negative label. |
| **Reply 12**: "At least replucan have normal rules and anti cancel culture." | 0 | This tweet expresses a political opinion related to cancel culture, which is not directly relevant to the relief bill but aligns with broader political discourse, warranting a neutral label. |
| **Reply 13**: "How could they read the bill? they were handed it hours before the actual vote in the house and then an hour (give or take) after that vote they voted in the senate. it was almost 6k pages." | 0 | This tweet questions the process of reading and voting on bills without expressing sentiment towards the relief bill, warranting a neutral label. |
| **Reply 14**: "Wtf is a replucan?" | 2 | This tweet questions a term without contributing directly to the discussion on relief, warranting a neutral label. |
| **Reply 15**: "The Smithsonian is getting 1 billion dollars??" | 1 | This tweet questions the allocation of funds without engaging directly with the relief bill, reflecting skepticism about fund allocation and warranting a neutral or slightly negative label. |
| **Reply 16**: "Nancy did not get rich off “us” or “you”. Her husband is much like Donald Trump, a wealthy developer. She didn’t get rich off of being in Congress or being a politician. That’s a total farce that you morons throw around every time she is mentioned." | 0 | This tweet defends Nancy Pelosi against accusations of personal enrichment, which is not directly related to the relief bill but reflects broader political discourse, warranting a neutral label. |
| **Reply 17**: "that's for the annual year-end funding package, not the relief bill. nice piece of propaganda you have there, though" | 0 | This tweet corrects misinformation without expressing sentiment towards the relief bill, warranting a neutral label. |

1. **Response for some initial tweets of Thread-66:**

| **Tweet** | **Label** | **Reason** |
| --- | --- | --- |
| **Parent Tweet**: "If this doctor, who so recklessly flew into New York from West Africa, has Ebola, then Obama should apologize to the American people & resign!" | 1 | The parent tweet suggests blaming Obama for the potential spread of Ebola, expressing a negative sentiment. |
| **Reply 1**: "omg stfu with these lies and this nonsense" | 1 | This reply expresses frustration and dismisses the parent tweet as lies and nonsense, reflecting a negative sentiment. |
| **Reply 2**: "you are right" | 2 | This reply agrees with the parent tweet without expressing any strong sentiment. |
| **Reply 3**: "agreed, but the Dr should be held accountable as well. He should know better." | 2 | This reply agrees with the parent tweet but suggests holding the doctor accountable as well, expressing a neutral sentiment. |
| **Reply 4**: ".@realDonaldTrump You should apologize to that American doctor for wanting to keep him out of his own country! What's wrong with you?" | 1 | This reply criticizes Trump and suggests he should apologize, expressing a negative sentiment. |
| **Reply 5**: ". @realDonaldTrump I'll give Â£1 to any charity you name if you let him lick you." | 1 | This reply makes a sarcastic and mocking comment directed at Trump, expressing a negative sentiment. |
| **Reply 6**: "This dude here â€œ@realDonaldTrump If this doctor, who so recklessly flew into New York from West ... <http://tmi.me/1eP4Hy>" | 2 | This reply shares a link but does not express a clear sentiment towards the parent tweet, warranting a neutral label. |
| **Reply 7**: "@realDonaldTrump you have smart words I love your bulbous brain tell me your thoughts on ritual foot binding" | 1 | This reply sarcastically praises Trump and asks for his opinion on a random topic, expressing a negative sentiment. |
| **Reply 8**: "@realDonaldTrump Is this a parody account or are you an idiot?" | 1 | This reply questions Trump's credibility or intelligence, expressing a negative sentiment. |
| **Reply 9**: "is there anything anyone anywhere does that SHOULDN'T cause Obama to resign? No, right? @realDonaldTrump" | 1 | This reply sarcastically questions the constant calls for Obama to resign, expressing a negative sentiment. |
| **Reply 10**: "@realDonaldTrump Sure sucks when it hits home! I would know, I'm in Texas !!" | 2 | This reply sympathizes with the potential spread of Ebola, but without expressing a strong sentiment towards the parent tweet, warranting a neutral label. |
| **Reply 11**: "@realDonaldTrump We could not ever be THAT damned lucky, Mr. Trump!" | 1 | This reply sarcastically suggests luck in the context of Trump's statement, expressing a negative sentiment. |
| **Reply 12**: "@realDonaldTrump Would you be surprised?!" | 2 | This reply questions whether Trump would be surprised by the potential spread of Ebola, expressing a neutral sentiment. |
| **Reply 13**: "@realDonaldTrump Can you explain what Obama has to do with this man who monitored himself for symptoms and poses no harm to anyone else?" | 2 | This reply questions the connection between Obama and the doctor, expressing a neutral sentiment. |

**FINE TUNING MODEL**

* Fine-tuning is currently available for the following models: gpt-3.5-turbo-0125 (recommended), gpt-3.5-turbo-1106, gpt-3.5-turbo-0613, babbage-002, davinci-002, and gpt-4-0613 (experimental).
* Fine-tuning OpenAI text generation models can make them better for specific applications, **but it requires a careful investment of time and effort**. We recommend first attempting to get good results with prompt engineering, prompt chaining (breaking complex tasks into multiple prompts), and function calling, with the key reasons being:
  + There are many tasks at which our models may not initially appear to perform well, but results can be improved with the right prompts - thus fine-tuning may not be necessary
  + Iterating over prompts and other tactics has a much faster feedback loop than iterating with fine-tuning, which requires creating datasets and running training jobs
  + In cases where fine-tuning is still necessary, initial prompt engineering work is not wasted - we typically see best results when using a good prompt in the fine-tuning data (or combining prompt chaining / tool use with fine-tuning)
* But, if you can achieve good results with GPT-4, you can often reach similar quality with a fine-tuned gpt-3.5-turbo model by fine-tuning on the GPT-4 completions, possibly with a shortened instruction prompt.
* Cost associated with fine tuning **gpt-3.5-turbo**:
  + **Training: $8.00 / 1M tokens**
  + **Input Usage: $3.00 / 1M tokens**
  + **Output usage: $6.00 / 1M tokens**
* **Rough estimates for the dataset we have will require less than $20 for fine tuning.**

**DETAILS OF GPT-4o**

* GPT-4o is our most advanced multimodal model that’s faster and is 50% cheaper than GPT- 4 Turbo with stronger vision capabilities and it generates text 2x times faster. The model has 128K context and an October 2023 knowledge cutoff.
* **Cost of gpt-4o:**
  + **Input: $5.00 / 1M tokens**
  + **Output: $15.00 / 1M tokens**
* **Rate limits:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TIER | MODEL | RPM | TPM | BATCH QUEUE LIMIT | QUALIFICATION | USAGE LIMITS |
| 1 | gpt-4o | 500 | 30,000 | 90,000 | $5 paid | $100 / month |
| 2 | gpt-4o | 5000 | 450,000 | 1,350,000 | $50 paid and 7+ days since first successful payment | $500 / month |

**NB: Higher tiers are not required for us.**