final_project

May 31, 2023

0.1 Content Moderation Bot - Group: CS-1

0.1.1 Project Description:

Our project is a content moderation bot that reads fifteen posts in a subreddit and calculates a sentiment analysis score from the submission's body. Then depending on if it has a positive or negative score it will comment on the post with "This post has a (positive or negative) sentiment analysis score of (score)." If the post is negative it will follow that sentence with "Do you want to take it down?"

0.1.2 Reddit PRAW setup

[ ]:
```
# Importing PRAW
import praw
```

[ ]:
```
# Running praw to get Reddit keys
%run reddit_keys.py
```

[ ]:
```
# Give the praw code your reddit account info so
# it can perform reddit actions
reddit = praw.Reddit(
    username=username, password=password,
    client_id=client_id, client_secret=client_secret,
    user_agent="a custom python script for user /" + str(username)
)
```

0.1.3 Sentiment analysis

Load sentiment analysis library and make analyzer

[ ]:
```
import nltk
nltk.download(["vader_lexicon"])
from nltk.sentiment import SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data…
```

Create String variables that respond to certain sentiment analysis scores of posts. Three responses should be considered: positive score, negative score, and whether the user would like to take down their post.

[ ]:
```
positive_sia = "This post has a positive sentiment analysis score of "
negative_sia = "This post has a negative sentiment analysis score of "
take_it_down = "Do you want to take it down?"
```

Looks up the subreddit info103_group_test and gets the latest fifteen submissions, for each submission it calculates a sentiment analysis score using the body of the submission, then replies to the submission with the prompts stated above. If the submission has a neutral score it prints neutral.

[12]:
```
# Look up the subreddit "_________", then find the "hot" list, getting up to 15␣
↪submission
submissions = reddit.subreddit("info103_group_test").new(limit=15)

for submission in submissions:
    submission_sentiment = sia.polarity_scores(submission.selftext)["compound"]

    if(submission_sentiment < 0):
        submission.reply(f'{negative_sia} {submission_sentiment}.␣
↪{take_it_down} {negative_img}')
    elif(submission_sentiment > 0):
        submission.reply(f'{positive_sia} {submission_sentiment}.␣
↪{positive_img}')
    else:
        print("Neutral")
```

```
Neutral
```

[12]:
```
Submission(id='13wehyc')
```

0.1.4 Reflection (NOTE: answer each reflection question with two ethical frameworks)

1. Is it ethical for the content moderation bot to prompt users to take down posts with negative sentiment analysis score?

A: The content moderation bot reads fifteen Reddit posts in a subreddit and generates sentiment analysis scores for each post. When the bot replies to posts with negative sentiment anaylsis scores and asks if the user would like to take it down interferes with the user's autonomy by asking them if they would like to delete their thoughts or opinions. This is a limitation of our content moderation bot. While it can be valuable to inform users about the sentiment of their posts, it is essential to approach this with sensitivity.

Directly suggesting the removal of a post solely based on sentiment analysis may suppress their natural rights. Examining this bot with the Natural Rights ethics framework, our bot does infringe on the user's right to use Reddit for free speech. The bot creates an environment where users are unable to express themselves and takes away their right to the pursuit of happiness given to them by the Natural Rights framework. The bot's limitations is that it cannot detect the meaning or context of a post. For example, if the post has a negative sentiment analysis score but the meaning of the post is not necessarily negative or harmful it will still reply to their post asking if they would like to take it down. Therefore, our bot may be unethical in the lens of the natural rights framework.

Using a consequentialist framework to analyze the content moderation bot, it may be ethical for the bot to prompt users to take down their posts with negative sentiment score when the post is actually hateful or harmful. The consequences of asking a user if they would like to take it down influences the user to reevaluate their post and how it might be harmful. If the user does believe it is harmful or not they still have freewill to decide whether they want to take it down or not.

2. Is it ethical to publicly display sentiment analysis scores for individual posts?

A: Using the virtue ethics framework to analyze the content moderation bot and how it replies to posts in a subreddit with the sentiment analysis score of the post, there are negatives and positives. By publicly displaying a sentiment score it encourages users in the subreddit to think responsibly about what they post, making users aware of the potential impact of their words. On the other hand, it may lead to other users to participate in public shaming of the post, which discourages people from participating in discussions. Analyzing the bot displaying sentiment analysis scores with virtue ethics promotes users to take responsibility for their own decisions and actions which creates accountability and transparency within the subreddit to improve its virtue.

Using the ethics of care framework, by publicly displaying the sentiment analysis scores for individual posts allows for others to provide feedback or guidance to the posts who receive negative scores. This allows for a cultivation of empathy, mutual care, and growth to foster in the subreddit, while still addressing any concerns or issues that arise. If the post is harmful and also receives a negative score, other Reddit users can create a discussion about why the content of their post is harmful, human content moderators can also use the context of the post and the negative score to ban users from the subreddit or mute them.

3. Does a negative sentiment analysis score indicate that a post is harmful or should be taken down?

There are multiple instances in which a negative sentiment analysis score does not imply that the post's intention or message is negative. For example, people may choose to make posts that give constructive criticism and choose to use negative words in their feedback. Although such a post would result in a negative sentiment analysis score, it is not considered morally wrong in most cases to give feedback to others. Analyzing this specific scenario in the deontological framework, there is more overall benefit if people are able to recieve constructive criticism and feedback for improvement.

Other cases may include general frustrations that people may choose to post online. Although some posts may use words that have a negative sentiment score, it does not imply that the post should be taken down. This would likely violate the user's right to freedom of speech in which they should have the ability to speak their own frustrations.

If we look at this under the lense of the Egoism ethical framework, negative sentiment analysis scores do not indicate a post is harmful. Egoism urges individuals to put their own interests first, so posts that gain a negative score but benefit the user in some way are not necessarily harmful nor should they be taken down. For example, a post where the user shares their negative opinion on a topic (earning them a negative score) does not deserve to be taken down because the user is only serving their self-interests by venting online. On the other hand, the Deontological framework argues that negative sentiment analysis scores *do* indicate a post is harmful. The Deontology framework asserts that it is each person's duty to follow absolute moral rules, such as being kind to others or being honest. Posts that have earned negative scores from the content moderation bot have violated these rules in some way, meaning they deserve to be taken down since the user has

failed to uphold their moral duty.