16/07/2024

CHATGPT REDDIT POSTS

WEB AND SOCIAL NETWORKS SEARCH AND ANALYSIS

By: Marie Elyse bassil 503962

PREPARED FOR :

Prof. Marco Viviani
Prof. Marco Braga
Prof. Georgios Peikos

PROJECT OVERVIEW

1 OBJECTIVE

ChatGPT is an artificial intelligence chatbot developed by OpenAI and launched in November 2022. It is built on top of OpenAI's GPT-3 family of large language models and has been fine-tuned using both supervised and reinforcement learning techniques. Due to its advantages over traditional chatbots, ChatGPT attracted more than 1 million users within 5 days and reached 100 million users in just 2 months after its launch, surpassing popular online platforms such as Netflix, Facebook, and Instagram in terms of adoption rates.

The primary objective of this project is to analyze the sentiment of Reddit posts related to ChatGPT to gain a deeper understanding of public opinion. By examining user-generated content, this analysis aims to uncover how users perceive ChatGPT, identify recurring themes in discussions, and detect significant trends over time. This information is crucial for understanding the strengths and weaknesses of ChatGPT from a user perspective, enabling stakeholders to make informed decisions for future improvements.

2 DATA COLLECTION

Method:

- Scraping Reddit Posts: I utilized the Reddit API to scrape posts from various subreddits. This involved querying Reddit for posts that mentioned specific keywords in their title or content.
- The subredits were the following: OpenAI, ChatGPT, artificial, machinelearning
- The keywords: ChatGPT, OpenAI GPT-3, GPT-3, GPT-4, OpenAI, AI chatbot, language model, natural language processing, AI assistant, conversational AI

Tools:

- PRAW (Python Reddit API Wrapper): PRAW is a Python library that provides a simple and efficient way to interact with the Reddit API. It allows for easy access to Reddit data, enabling us to scrape large volumes of posts and comments systematically.

Description:

On the right, the entire dataset can be seen.
However, since most of the columns were not useful for my analysis, I decided to retain only the important ones.

My new dataset contains the following columns:
'author', 'comments', 'date', 'id', 'num_comments', 'score', 'subreddit', 'upvote_ratio', 'text', ‘title],

Since some reddit posts don't have a selftext (simply just posted as a title), I will be combining the title and the text together into one column and considering that column as my target.
--> combined_text

The focus period for the analysis was chosen based on when ChatGPT started gaining popularity, as it allows capturing the most relevant and significant data. Therefore, the data spans from 2023-01-01 to 2024-07-06

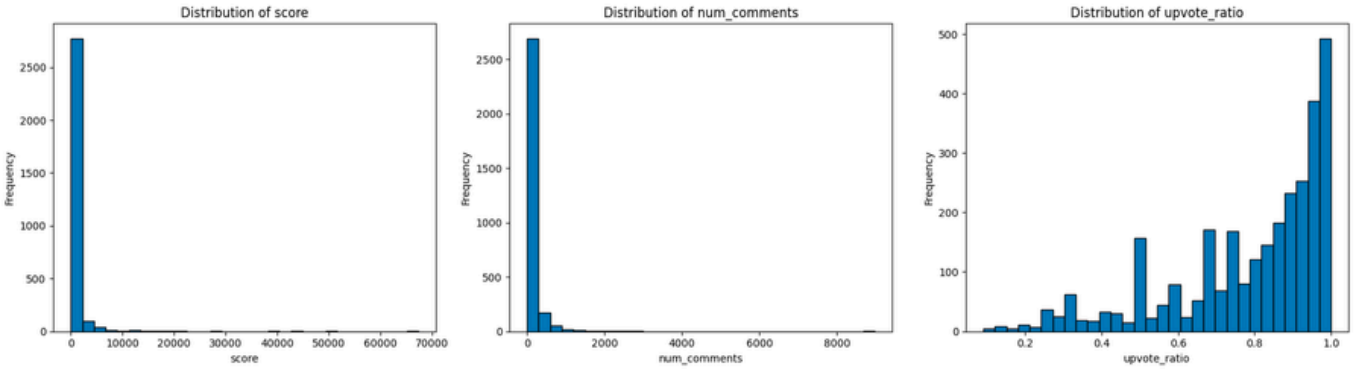
Column Name	Description
author	Username of the Reddit user who created the post.
author_flair_text	Flair text associated with the author's username, if any.
clicked	Whether the post has been clicked.
comments	List of comments associated with the post.
created_utc	Timestamp of when the post was created in UTC.
distinguished	Indicates if the post is distinguished (e.g., by a moderator).
edited	Indicates if the post has been edited.
id	Unique identifier for the post.
is_original_content	Indicates if the post is marked as original content.
is_self	Indicates if the post is a self-post (text post).
link_flair_template_id	Template ID of the link flair, if any.
link_flair_text	Text of the link flair, if any.
locked	Indicates if the post is locked (no further comments allowed).
name	Full name of the post (e.g., t3_18hgbrt).
num_comments	Number of comments on the post.
over_18	Indicates if the post is marked as NSFW (not safe for work).
permalink	Permanent URL link to the post.
poll_data	Data related to any poll associated with the post, if any.
saved	Indicates if the post has been saved by the user.
score	Score (upvotes - downvotes) of the post.
selftext	The text content of the post, if it is a self-post.
spoiler	Indicates if the post is marked as a spoiler.
stickied	Indicates if the post is stickied (pinned to the top) by moderators.
subreddit	Name of the subreddit where the post was made.
title	Title of the post.
upvote_ratio	Ratio of upvotes to total votes.
url	URL of the post or linked content.

3 EXPLORATORY DATA ANALYSIS (EDA)

1- Summary Statistics on my numerical features

Summary Statistics:

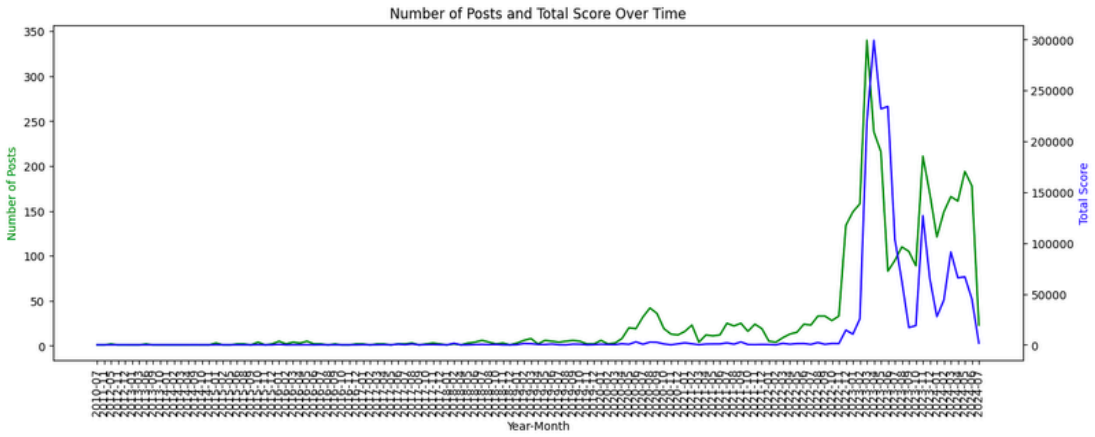
	score	num_comments	upvote_ratio
count	2955.000000	2955.000000	2955.000000
mean	594.893063	100.446701	0.785465
std	2614.193805	274.698796	0.202859
min	0.000000	0.000000	0.090000
25%	2.000000	2.000000	0.670000
50%	20.000000	15.000000	0.860000
75%	280.000000	99.500000	0.950000
max	67407.000000	8979.000000	1.000000



Observations:

- The distributions for score and num_comments are highly skewed, indicating the presence of outliers or highly popular posts.
- The upvote_ratio shows a higher concentration around the upper end (0.8-1.0), suggesting most posts receive a positive reception.

2- Time series analysis



Observations:

- There are clear peaks in the number of posts and total scores, suggesting periods of heightened interest or activity related to ChatGPT and AI topics.
- Certain months show a significant spike in both the number of posts and the total score, indicating viral or highly engaging discussions.
- The significant increase in activity around late 2022 aligns with the release of ChatGPT, indicating that its introduction has had a substantial impact on community engagement. Despite the decline from the peak, the sustained higher levels of posts and scores suggest ongoing interest and discussions about ChatGPT and related topics.

3 - Author analysis

```
author
Excellent-Target-847    53
Maxie445                43
NuseAI                  39
Singularian2501         23
lostlifon               20
Successful-Western27    17
ShotgunProxy            17
jaketocake              15
MysteryInc152           13
Altruistic_Gibbon907    13
Name: count, dtype: int64
```

- Highly Active Authors: The dataset shows that a small group of authors are highly active in posting content related to ChatGPT. The top three authors alone contribute a significant number of posts (Excellent-Target-847: 53 posts, Maxie445: 43 posts, NuseAI: 39 posts).
- Engagement Leaders: These authors likely play a key role in driving discussions and shaping the community's perception of ChatGPT.
- Potential Influencers: These authors could be considered influential within the community due to their high activity levels.

	author	avg_score	avg_comments
668	NeedsAPromotion	51315.0	2258.0
717	Outrageous_Bee4464	44499.0	662.0
1324	dtutubalin	40054.0	1519.0
1784	rich_awo	34259.0	619.5
380	ForceTypical	28559.0	2056.0
1619	meth_addicted_lama	27418.0	325.0
1851	sniperxp21	21279.0	544.0
296	DrDejavu	20160.0	1018.0
1922	throwaway9au	18856.0	451.0
1816	savatrebein	18463.0	780.0

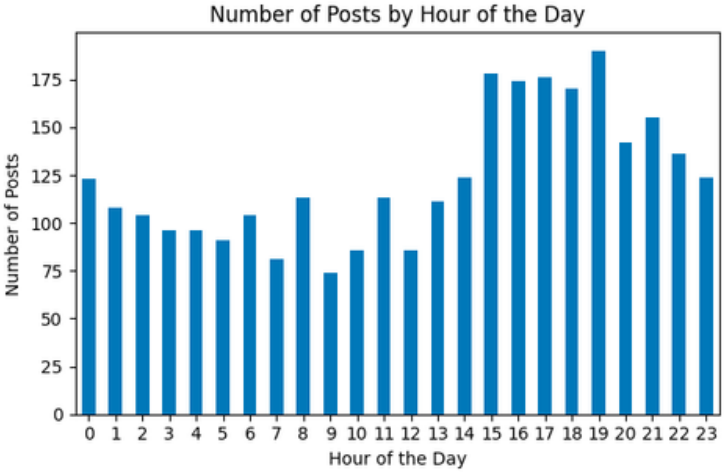
- High Engagement Posts: The authors listed here have posts with exceptionally high average scores and comments. For example, NeedsAPromotion has an average score of 51,315 and 2,258 comments, indicating highly engaging content.
- Quality vs. Quantity: Unlike the most frequent posters, these authors may not post as often but their posts generate significant interaction, highlighting the quality and impact of their contributions.

4 - Engagement by subreddit

	subreddit	avg_score	avg_comments
0	ChatGPT	1517.327886	198.452503
2	OpenAI	181.219277	74.827711
1	MachineLearning	145.943262	42.411348
3	artificial	68.249141	28.364261

The ChatGPT subreddit shows the highest average score (1517.33) and the highest average number of comments (198.45) per post. This indicates that discussions in this subreddit are highly engaging and resonate well with the community.

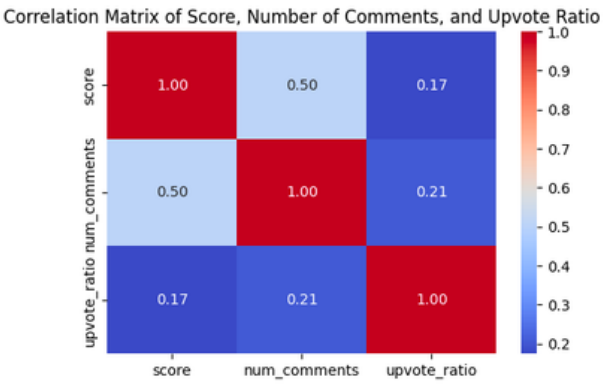
5 - Analysis by post hour




```
date
Wednesday    506
Thursday     475
Friday       442
Tuesday      436
Monday       391
Saturday     366
Sunday       339
Name: count, dtype: int64
```

We can see that midweek days (Wednesday to Friday) show the highest posting activity for ChatGPT-related discussions, while weekends see a drop in user engagement. Most active in discussing ChatGPT during the evening hours, particularly between 5 PM and 7 PM.

6- Correlation Matrix



The moderate correlation between score and num_comments suggests that higher-scoring posts are likely to generate more discussion. However, the weak correlations involving upvote_ratio indicate that while positive reception (high upvote ratio) contributes to scores and comments, other factors also play significant roles in driving engagement.

7- Word Cloud of most popular words

The most frequent occurring terms are the largest words in the word cloud, such as "AI," "model," "ChatGPT," "GPT," and "language,". These words are central to the discussion or content being analyzed. Clearly, since our text are discussions about GPT then they would be the most common words.

4 DATA PREPROCESSING

In my data preprocessing pipeline, I implemented a comprehensive text cleaning function aimed at preparing textual data for subsequent analysis. The preprocessing steps include the creation of an expanded set of stopwords, encompassing both common English stopwords and additional domain-specific terms such as 'x200b', 'link', 'openai', 'gpt4', 'language', 'model', 'chatgpt', 'Model', 'chatbot', 'gpt3', 'ai', and 'chat'. This augmented stopword list ensures that non-informative and very common words are effectively removed from the text.

I utilized the WordNet Lemmatizer to reduce words to their base forms, enhancing the uniformity of our text data. The `clean_text` function follows a series of steps to standardize and clean the input text: converting it to lowercase, removing newline characters, and stripping out retweets and URLs. Mentions are also removed to prevent noise from user references.

Additionally, I handle negations by capturing and marking them appropriately, ensuring that the sentiment and context of negative expressions are preserved. Punctuation and numerical values are stripped from the text, and the remaining words are tokenized. Each token is checked against the stopword list, and only those not in the list are lemmatized and retained.

Finally, the cleaned tokens are reassembled into a coherent string, with extra whitespace removed to produce a tidy and consistent output. This preprocessing framework is crucial for transforming raw textual data into a structured format suitable for analysis, enhancing the reliability and accuracy of subsequent machine learning models and analyses.

Feature extraction and token counting were also performed to gain a better understanding of the text data.

5 NETWORK ANALYSIS

1 - EXCTRACTION OF THE COMMENTS

- Objective: To understand the interaction dynamics within Reddit discussions about ChatGPT by analyzing the relationships between post authors and commenters.
- Method: Extracted comments for each Reddit post and identified the relationship between authors and commenters
- The full dataset of the relationships was a total of 136046 rows. However, For the purpose of the analysis, since doing the calculations and plotting the graph of that many rows can be very resource-intensive and may exceed the capabilities of many system, i will be using a reduced dataset of relationships of 11000 rows and doing work on that. I have tried doing it on all of it however, even with waiting for several hours, the output wouldn't come out.

	post_author	comment_author
0	Efron1234	zipperdz
1	zipperdz	Efron1234
2	Efron1234	Erophysia
3	Erophysia	Efron1234
5	Efron1234	Sudden-Anybody-6677
...	...	...
170254	Justice4Ned	NeuroXORMute
170255	Justice4Ned	SaddleSocks
170256	Endonium	hueshugh
170257	hueshugh	NeuroXORMute
170258	NeuroXORMute	hueshugh

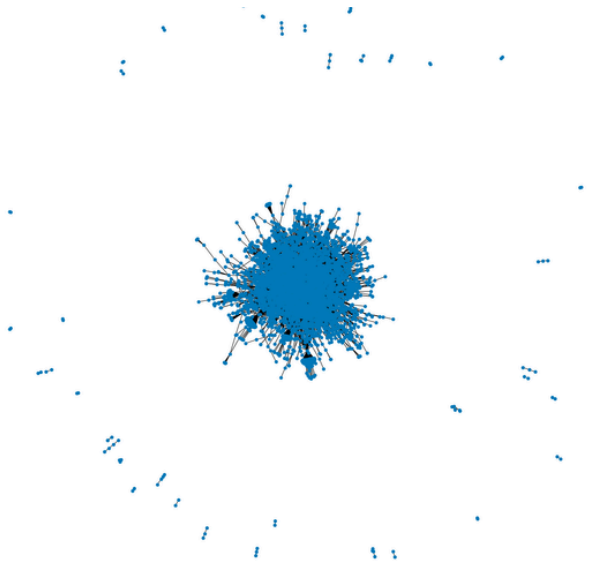
136046 rows × 2 columns

- Furthermore, Nodes with 10 or fewer connections are identified for potential removal, reducing clutter and focusing on more connected parts of the graph and removing isolated nodes to further simplify the graph and improve the clarity of the visualization.

.

2- GRAPH CREATION

- The figure on the right shows to graph generated from the reduced dataframe for relationships, We can see that there is a dense central cluster of nodes with many interconnections. This suggests that there is a core group of users who frequently interact with each other. There are several nodes on the periphery with few connections. These likely represent users who interact less frequently or are more isolated in the network.
- The nodes that are far from the central cluster are outliers, potentially indicating users who are less engaged in the overall community or who only interact sporadically.



2- MEASURES OF CENTRALITY

	node	degree_centrality	betweenness_centrality	eigenvector_centrality	closeness_centrality
6055	b-damandude	0.036366	0.077483	0.415331	0.269396
256	LinuxLover3113	0.036090	0.074736	0.307438	0.263709
4255	Biolevinho	0.035260	0.086487	0.359392	0.274206
2807	Furious_Vein	0.031250	0.077563	0.175671	0.273619
2001	SessionGloomy	0.025996	0.063980	0.067656	0.268542
3335	AnsibleAnswers	0.023092	0.064769	0.088344	0.275150
107	adt	0.022539	0.083169	0.054090	0.286166
5660	zvone187	0.020326	0.061739	0.062600	0.272834
6505	lostlifon	0.020326	0.075446	0.055627	0.282329
5102	userranger	0.019773	0.041063	0.036097	0.254999

High Influence Users:

- b-damandude has the highest eigenvector centrality (0.415331), indicating significant influence and connections with other well-connected users.
- Biolevinho and LinuxLover3113 also have high eigenvector centrality scores, suggesting they are influential within the network.

Brokers in the Network:

- b-damandude and Biolevinho have high betweenness centrality scores, indicating they play critical roles in connecting different parts of the network.

Highly Connected Users:

- b-damandude and LinuxLover3113 have the highest degree centrality scores (0.036366 and 0.036090 respectively), showing they have many direct connections within the network.

Access to Information:

- adt has a relatively high closeness centrality (0.286166), meaning this user can efficiently reach other users in the network.
- lostlifon also has a high closeness centrality (0.282329), suggesting quick access to information across the network.

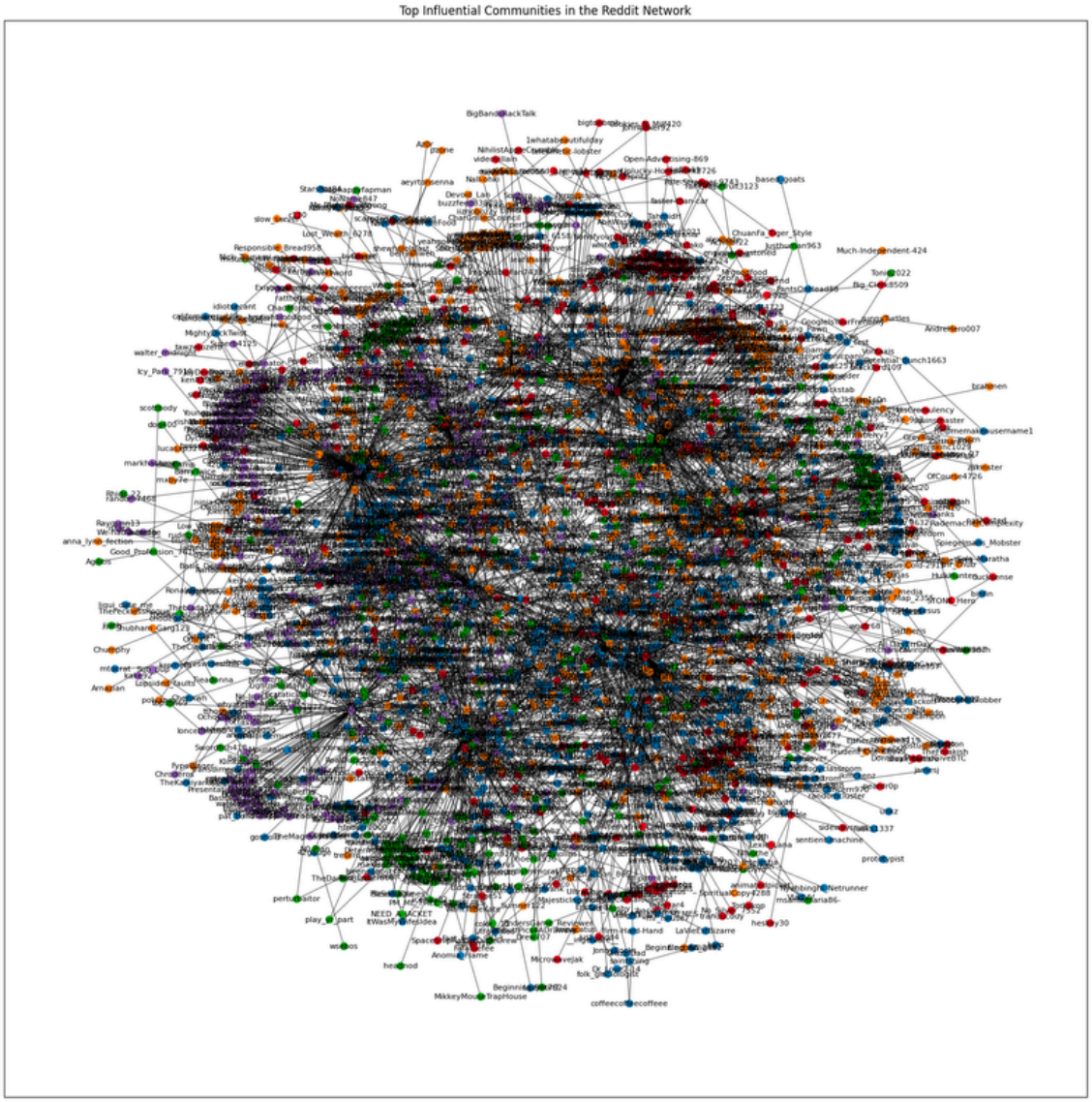
3- COMMUNITY DETECTION

The resulting table shows the top nodes based on their centrality measures (degree, betweenness, eigenvector, closeness) along with their community assignments. This analysis allows us to identify key influencers within each community and understand the network's structure more comprehensively.

community	degree_centrality	betweenness_centrality	eigenvector_centrality	closeness_centrality	overall_centrality
0	0.316233	0.514435	0.304790	134.506064	135.641522
1	0.262306	0.436499	0.928296	125.160502	126.787604
2	0.156112	0.205982	1.218158	79.466075	81.046327
4	0.129425	0.540258	1.587800	74.514758	76.772241
5	0.120299	0.165130	2.645552	71.370816	74.301797
3	0.149751	0.258655	0.353259	72.419222	73.180887
6	0.101217	0.138309	5.521062	63.711276	69.471864
8	0.092782	0.107065	6.707726	61.000653	67.908226
12	0.086283	0.098632	5.134602	57.577143	62.896660
7	0.118501	0.152898	1.013903	60.201305	61.486608

We can see that:

- Community 0 and 1 are the most influential communities in the network based on overall centrality scores.
- Community 4 stands out with high betweenness centrality, indicating it acts as a bridge within the network.
- Community 6, 7, 8, and 9 have specialized roles with notable eigenvector and closeness centrality measures but overall lower influence.



The network's central region is densely connected, indicating a high level of interaction among the nodes. This area likely represents the most influential communities, where members actively engage with each other.

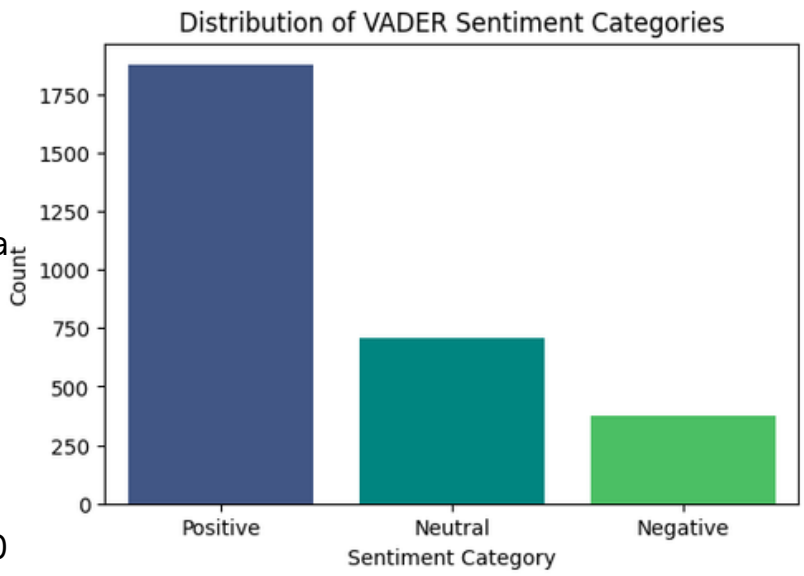
Identifying these influential users and understanding the network's structure and dynamics is crucial for analyzing the Reddit discussion about ChatGPT. This analysis can help identify key users and communities for targeted sentiment analysis. However, since I have only taken a subset of the data, which may not fully represent the entire situation, I will conduct sentiment analysis on the overall Reddit posts rather than delving deeper into individual comments and findings from this analysis (most influencial users, communities,..).

6 CONTENT ANALYSIS - SENTIMENT ANALYSIS

1 - VADER

To understand the sentiment of the discussions on Reddit about ChatGPT, I performed sentiment analysis using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool. The VADER sentiment analyzer is designed to analyze the sentiment of social media texts and is effective at capturing the emotional tone conveyed in short text formats.

The results were stored in the DataFrame with columns for each sentiment score (pos_score_vader, neg_score_vader, neu_score_vader, compound_score_vader) and the classified sentiment category (vader_sentiment).



- Positive Sentiment: The majority of the posts exhibit a positive sentiment, with approximately 1,750 posts classified as positive. This indicates a general positive reception or discussion around ChatGPT on Reddit.
- Neutral Sentiment: The second most common sentiment category is neutral, with around 750 posts. These posts do not convey a strong positive or negative sentiment, suggesting factual or balanced discussions.
- Negative Sentiment: The least common sentiment category is negative, with approximately 500 posts. This shows that while there is some criticism or negative discussion, it is relatively less frequent compared to positive and neutral sentiments.

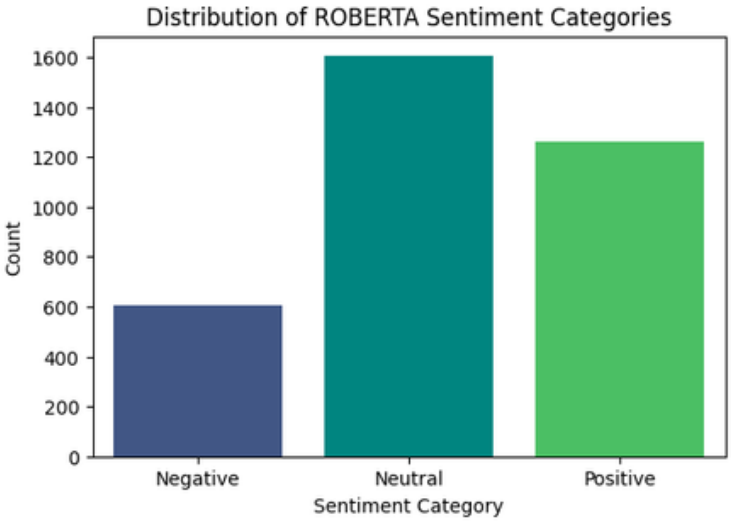
2 - roBERTa

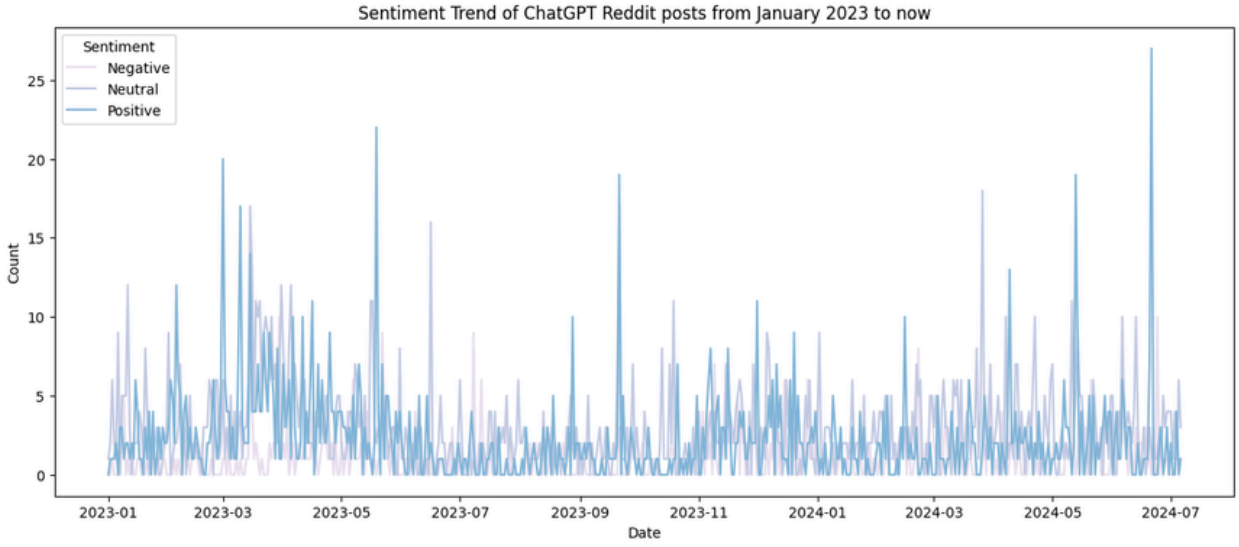
In addition to the VADER sentiment analysis, I performed sentiment analysis using the RoBERTa model to gain a deeper understanding of the sentiment expressed in Reddit discussions about ChatGPT. RoBERTa (Robustly optimized BERT approach) is a transformer-based model pre-trained on a large corpus of data, making it highly effective for natural language understanding tasks.

RoBERTa model works better with minimalistic amount of manipulation to the data, so instead of the original data preprocessing that was performed before, I only anonymized user mentions and simplified URLs in each text.
The sentiment scores were added to the DataFrame, and a composite polarity score was computed using weighted sums of the sentiment scores. The polarity score was then scaled and categorized into "Positive", "Neutral", and "Negative" sentiments.

The chart shows the count of posts in each sentiment category:

- Negative Sentiment: Approximately 600 posts were classified as negative, indicating some level of criticism or negative discussion.
- Neutral Sentiment: The majority of the posts, around 1,600, were classified as neutral. These posts likely contain factual or balanced discussions without strong emotional content.
- Positive Sentiment: About 1,200 posts were classified as positive, reflecting a generally favorable or supportive sentiment towards ChatGPT.



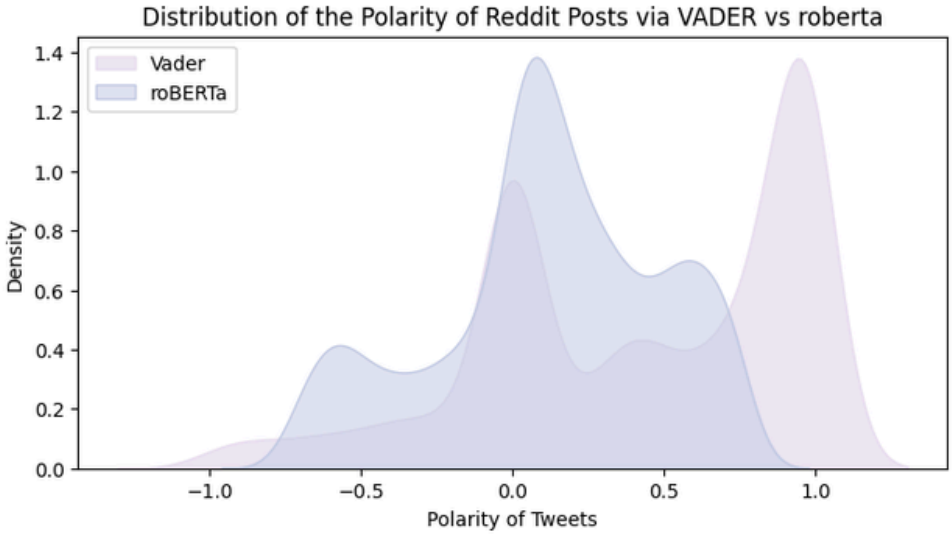


- Positive Reception: The predominance of positive sentiment suggests that Reddit users generally view ChatGPT favorably. This can be attributed to successful updates, good user experiences, or effective communication from the developers.
- Community Engagement: The consistent volume of neutral sentiment posts indicates that users are continuously engaging with ChatGPT, discussing its functionalities, sharing experiences, or seeking help, without necessarily expressing strong opinions.
- Critical Feedback: The presence of negative sentiment, although lower in volume, highlights areas where users have concerns or criticisms. These posts are crucial for identifying potential issues or areas for improvement.

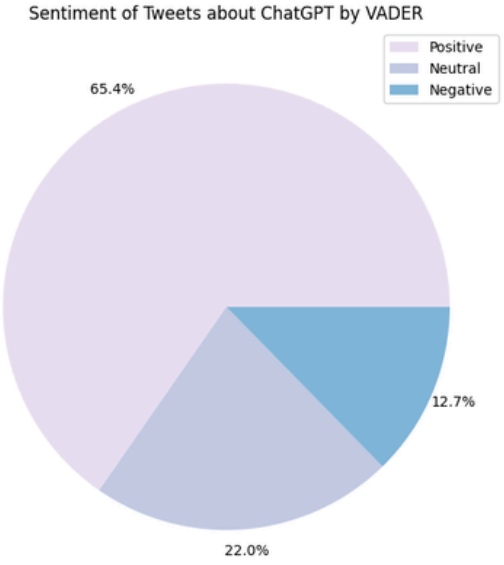
7 COMPARISON OF THE 2 MODELS

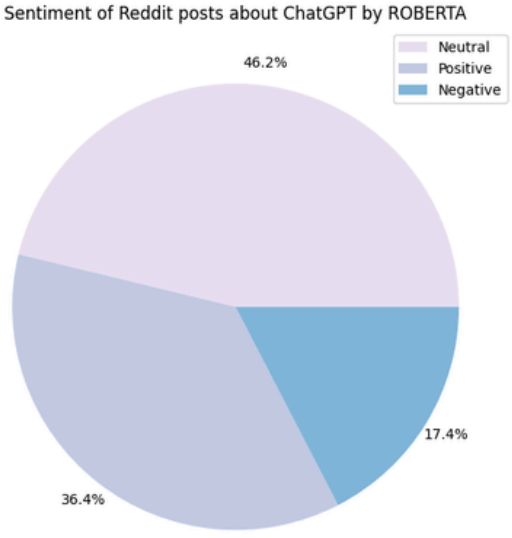
By comparing these two models, we can gain insights into their performance and effectiveness in analyzing sentiment in Reddit discussions about ChatGPT. This comparison will help us understand the strengths and limitations of each model and guide us in selecting the most appropriate tool for our sentiment analysis needs.

To compare the performance of VADER and RoBERTa sentiment analysis models, I conducted an analysis on Reddit posts discussing ChatGPT. Below are the results and visualizations of this comparison.



The density plot shows that VADER's sentiment polarity scores are more concentrated around the positive end of the spectrum, with a significant peak around 1.0.There are fewer posts classified as strongly negative, and the neutral sentiments are less pronounced compared to RoBERTa.
RoBERTa's sentiment polarity scores exhibit a more balanced distribution, with peaks at both the positive and neutral ends. This model captures a broader range of sentiment, indicating its ability to discern more nuanced emotional tones within the text. The density for negative polarity is higher in RoBERTa compared to VADER, suggesting that RoBERTa is more sensitive to negative sentiments.





The pie charts further illustrate the difference in sentiment classification between the two models:

- Positive sentiments: VADER classifies a substantial majority of the posts as positive (65.4%). This suggests that VADER may be overly optimistic or more sensitive to positive expressions. While with RoBERTa 36.4% of the posts are classified as positive, a more balanced perspective.
- Neutral Sentiment: The VADER model identifies 22.0% of the posts as neutral whilst RoBERTa classifies the highest proportion of posts as neutral (46.2%), indicating a more conservative approach to sentiment classification.
- Negative Sentiment: In Vader, only 12.7% of the posts are classified as negative, while roBERTa identifies 17.4% of the posts as negative, showing its better capability to detect negative sentiments.