Project description

The aim of the project is to analyze the sentiment of Reddit posts related to ChatGPT and AI to gain insights into public perception of it. By examining these posts, we can understand how users feel about ChatGPT, what aspects they appreciate, and what concerns they might have. The analysis will involve:

- Data Collection: The gathering of the posts was done previously using the reddit API.
- EDA and Data Preprocessing : Cleaning and preparing the data as well as understanding the datam, engagement trends over time, including the number of posts, scores, and upvote ratios.
- Social anlysis: Building a graph based on relationships between authors and commentors, calculating centrality measures and community detection.
- Content analysis: Sentiment Analysis: Classifying the posts into positive, negative, or neutral sentiments.

Data Scraping from Reddit

The following cells demonstrate how to scrape data from Reddit using the Reddit API. The goal is to collect posts from specific subreddits and filter them based on certain keywords. I will utilize the PRAW (Python Reddit API Wrapper) library to interact with the Reddit API and extract valuable information from the posts.
The subreddits and keywords we are focusing on are as follows:
- Subreddits: OpenAI, ChatGPT, artificial, machinelearning
- Keywords: ChatGPT, OpenAI GPT-3, GPT-3, GPT-4, OpenAI, AI chatbot, language model, natural language processing, AI assistant, conversational AI.
The praw.Reddit object reddit will be used to interact with the Reddit API.
Function get_reddit_posts:

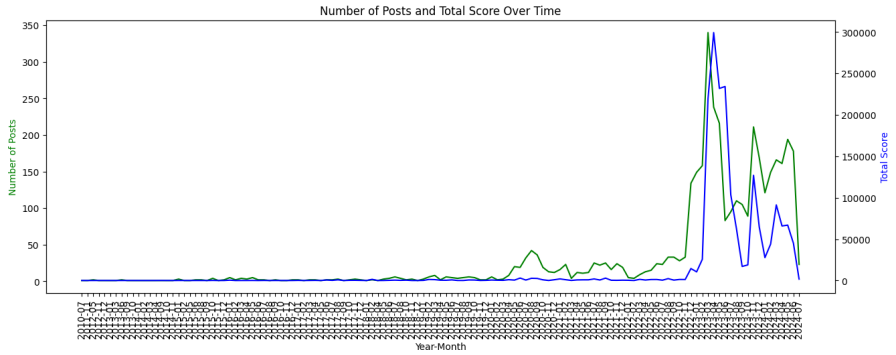
- **Loops through the list of subreddits.**
- **For each subreddit, it fetches the 1000 newest posts.**
- **Filters posts based on the presence of any of the specified keywords in the post title or body.**
- **Collects relevant information from the posts and appends it to a list.**
- **Returns a pandas DataFrame containing the collected posts.**

EDA

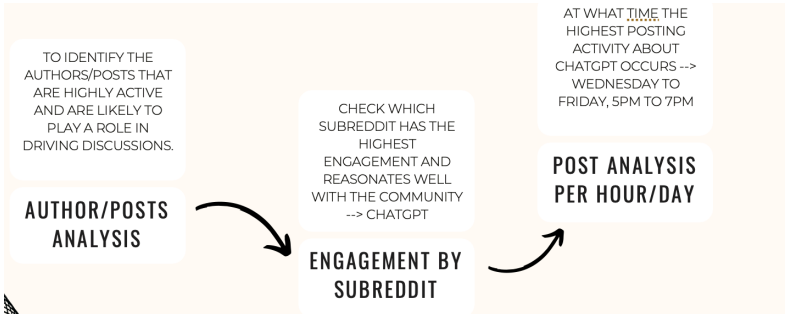
The aim of the EDA is to Gain a deeper understanding of my data by summarizing its main characteristics, and visualizing them. I checked my dataframe columns and immediately identified my text columns. I noticed that since some reddit posts don't have a selftext (simply just posted as a title), I will be combining the title and the text together into one column and considering that column as my target. I checked the timeframe of my dataset and decided to focus on after the year 2023 when chatgpt started becoming popular.
Then, I checked the statistics of

	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

The dataset summary indicates significant variability in Reddit post engagement, with an average score of 594.9 and a high standard deviation of 2614.2, suggesting the presence of both low and highly popular posts. The number of comments also varies widely, averaging around 100 comments per post but ranging from 0 to 8979, with a median of 15, indicating a skewed distribution towards fewer comments. Most posts have a positive reception, as shown by the upvote ratio's median of 0.86, with values ranging from 0.09 to 1.0, reflecting a generally favorable audience response.



The green line represents the number of posts, which experienced a significant spike, peaking at around 350 posts in a given month. Similarly, the blue line, indicating the total score, also shows a corresponding increase, reaching up to 300,000 at its peak. This trend suggests a growing interest and interaction with the topics discussed, reflected by both the higher number of posts and the increased cumulative scores. 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.



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). These authors likely play a key role in driving discussions and shaping the community's perception of ChatGPT.
These authors may provide valuable insights into what type of content resonates most with the community. They create posts that generate high levels of engagement (scores and comments). These authors can be considered as key content creators whose posts have a substantial impact on the community's engagement and sentiment.
The ChatGPT subreddit shows the highest engagement with average scores and comments significantly above others, indicating it as the central hub for discussions about ChatGPT. The OpenAI subreddit also has high engagement, suggesting strong community interest in OpenAI's developments, including ChatGPT.
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.

The correlation matrix visualizes the relationships between the score, number of comments, and upvote ratio of Reddit posts.

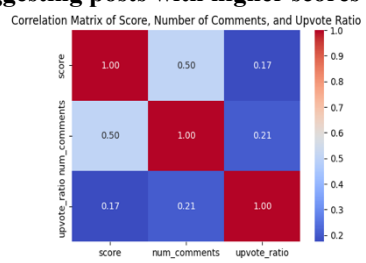
- The diagonal elements represent the correlation of each variable with itself, which is always 1.0. Hence, the values on the diagonal are all 1.0, indicating perfect correlation.

- The correlation coefficient between the score and the number of comments is 0.50. This positive correlation indicates a moderate relationship, suggesting that posts with higher scores tend to have more comments, but it's not a perfect relationship. There are other factors affecting the number of comments as well.

- The correlation coefficient between the score and the upvote ratio is 0.17. This weak positive correlation implies that posts with higher scores are slightly more likely to have a higher upvote ratio, but the relationship is not strong. The score is influenced by other factors beyond just the upvote ratio.

- The correlation coefficient between the number of comments and the upvote ratio is 0.21. This weak positive correlation indicates that posts with more comments tend to have a slightly higher upvote ratio, but again, the relationship is not strong.

- Between score and number of comments (0.50), suggesting posts with higher scores generally receive more comments.



- While there is some relationship between these metrics, no single metric can fully explain the others. This indicates the complexity and multifaceted nature of user engagement on Reddit.

The correlation matrix helps in understanding how these metrics are interrelated and highlights the need to consider multiple factors when analyzing post performance and engagement on Reddit.

Text cleaning:

The following function will perform the following:

- Convert to lowercase and remove newlines to ensure uniform text.
- Apply mapping for text replacement: Substitutes specified substrings based on the provided mapping.
- Handle Negations by using a regex to find negation words and adds a hyphen between them to avoid loss of information.
- Removing numbers from texts.
- Tokenization: splitting the text into individual words (tokens).
- Remove Stopwords and Punctuation: filters out stopwords and punctuation
- Lemmatization: lemmatizes the remaining tokens.
- And finally, join the cleaned tokens back into a single string.

After that, the extraction of linguistic features using TextBlob to extract noun phrases, part-of-speech tags, and sentiment.

 Each entry in the linguistic_features column is a list of dictionaries, where each dictionary contains detailed linguistic information about a word (or token) from the text.

 The key fields in each dictionary include:

- **token**: The original word as it appears in the text.
- **lemma**: The base or root form of the word (useful for linguistic analysis).
- **pos**: The part of speech tag for the word (e.g., VERB, NOUN, ADJ for verb, noun, adjective, respectively).

- **Linguistic Features**:
  - Each token in the text is analyzed to provide its lemma and part of speech, which helps in understanding the grammatical structure and semantics of the content.
- **Common Words**:
  - The frequency of words helps identify the main topics and recurring themes in the dataset, providing a snapshot of what users are talking about most frequently.

Together, these analyses help in understanding the content at both a granular (word-level) and holistic (topic-level) scale, facilitating deeper insights into the dataset's characteristics.

SOCIAL NETWORK ANALYSIS

1) Extracting Author Relationships from Reddit Comments

In this section, we will process the cleaned Reddit posts DataFrame to build a network of interactions between users based on their comments. The goal is to identify relationships where one user comments on another user's post or comment. This will be achieved by iterating through each post, extracting the author of the post, and then extracting the authors of the top-level, second-level, and third-level comments.

I CREATED A DATAFRAME OF THE RELATIONSHIPS OF COMMENTORS AND POST OWNERS.

For the purpose of the graph, since doing the calculations and plotting the graph of 122898 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.

- the **degree** of a node is the number of edges connected to it. For undirected graphs, it's simply the count of connections a node has.
- **Degree Centrality** measures the importance of a node based on how many connections it has. Nodes with higher degrees are often more central or influential in the network.

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.

I reduced the size of the graph from 122898 to 9406

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.

Centrality Measures

Centrality measures help identify the most important nodes within a graph based on various criteria. Here are the four centrality measures calculated:

1. **Degree Centrality**:
   - Measures the number of edges connected to a node.
   - Indicates how connected or active a node is in the network.
2. **Betweenness Centrality**:
   - Measures the extent to which a node lies on the shortest paths between other nodes.
   - Indicates the node's role in facilitating communication between other nodes (acting as a bridge).
3. **Eigenvector Centrality**:
   - Measures the influence of a node based on the influence of its neighbors.
   - A high eigenvector centrality means the node is connected to other nodes that are themselves highly connected.
4. **Closeness Centrality**:
   - Measures the average length of the shortest path from the node to all other nodes in the network.
   - Indicates how close a node is to all other nodes in the network, reflecting its ability to spread information quickly.

Insights from the Table

1. **b-damandude**:
   - Has the highest degree centrality (0.036366), suggesting this node has the most connections.
   - Also shows high eigenvector centrality (0.415331), indicating it is connected to other influential nodes.
   - High betweenness centrality (0.077483) suggests it plays a significant role in connecting different parts of the network.
   - Closeness centrality (0.269396) indicates it is relatively close to all other nodes.
2. **LinuxLover3113**:
   - Similar characteristics to b-damandude, but slightly lower in all centrality measures.
3. **Biolevinho**:
   - High in both betweenness (0.086487) and eigenvector centrality (0.359392), indicating strong influence and connectivity.
4. **Furious_Vein** and **SessionGloomy**:
   - Moderate degree and betweenness centrality, but lower eigenvector centrality, suggesting less influence through their neighbors.
5. **adt**:
   - Notably high betweenness centrality (0.083169), suggesting a critical role in the network's connectivity, despite lower degree and eigenvector centrality.
6. **userranger**:
   - Lower in all centrality measures, indicating less centrality in terms of connections, bridging, influence, and closeness.

Community detection

 greedy_modularity_communities(G_reduced): This function uses a greedy algorithm to detect communities in the graph G_reduced. Communities are groups of nodes that are more densely connected internally than with the rest of the network.

Then, I calculated the overall centrality of each community by summing the centrality measures of all nodes within each community, and then ranks the communities based on their overall centrality.

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 greedy modularity algorithm for community detection works by iteratively optimizing the modularity score of the graph. Modularity is a measure of the strength of the division of a network into communities. A high modularity score indicates dense connections within communities and sparse connections between communities.

Afterwards, I identified the top 5 communities based on a centrality measure, Extracted the nodes belonging to these top communities, created a subgraph of these nodes, Assigned unique colors to each community and Visualized the subgraph with nodes colored according to their community.

Based on the visualization, here are some observations and insights about the top influential communities in the Reddit network:

1. **Dense Connectivity**:
   - The graph appears to be densely connected, indicating a high level of interaction between nodes within the top communities.
   - Many nodes have multiple connections, suggesting that these communities are well-integrated and active.
2. **Community Clustering**:
   - Different colors represent different communities, and we can observe distinct clusters of nodes that are densely connected within the same color. This signifies that the algorithm successfully identified groups of nodes that interact more frequently with each other.
   - Some clusters are more tightly knit, showing strong intra-community connections, while others are more dispersed, indicating looser community structures.
3. **Influential Nodes**:
   - Certain nodes are more central within their clusters, indicating higher influence or activity within their communities. These nodes might act as hubs or connectors within their respective groups.
   - Nodes that are highly connected across different clusters could be influencers that bridge multiple communities.
4. **Community Sizes**:
   - The sizes of the clusters vary, indicating that some communities are larger and more active than others. Larger clusters suggest communities with more members or more interactions.
   - Smaller clusters might represent niche communities with specialized interests.
5. **Inter-Community Connections**:
   - There are edges connecting nodes across different colors, showing inter-community interactions. These edges are crucial for information flow and interaction between different groups.
   - The presence of these inter-community connections suggests that while communities are distinct, there is still significant communication and interaction between them.
6. **Layout and Visualization**:
   - The spring layout algorithm used here positions nodes in a way that minimizes edge crossings and reveals the underlying structure of the network. This makes it easier to visually identify clusters and influential nodes.
   - The clear separation of colors helps in distinguishing different communities, making the analysis of community structure more intuitive.

Insights and Applications

- **Community Influence**: The visualization can help identify which communities are most influential within the Reddit network. This information can be useful for targeted marketing, content dissemination, and understanding user behavior.
- **Network Structure**: Understanding the structure of these communities can provide insights into how information spreads within the network. It can help in designing strategies to enhance or control information flow.
- **Key Players**: Identifying key nodes (influential users) within and between communities can help in engaging with these users for various purposes, such as community management or promotional activities.
- **Community Health**: The structure and connectivity of communities can be indicators of their health and activity levels. Highly connected and dense communities might be more robust and active, while sparse or fragmented communities might need engagement efforts to revive or sustain them.

Overall, this visualization offers a comprehensive overview of the top influential communities in the Reddit network, highlighting the structure, connectivity, and key players within the network.

Content analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed to analyze social media text. VADER is particularly adept at understanding the context and intensity of emotions expressed in short pieces of text, such as tweets, reviews, and comments. Here's a detailed explanation of how VADER works:

Key Components of VADER

1. **Lexicon**:
   - VADER uses a dictionary that maps words to their sentiment intensity scores. Each word in the lexicon is associated with a sentiment score ranging from -4 (extremely negative) to +4 (extremely positive).
   - The lexicon includes common words, slang, abbreviations, and emoticons typically used in social media.
2. **Rules**:
   - VADER uses a set of grammatical and syntactical rules to adjust the sentiment scores based on the context in which the words appear. These rules help account for the nuances of natural language.

The chart shows a predominantly positive sentiment in the text data, with a substantial neutral sentiment and a smaller portion of negative sentiment. This suggests that the overall tone of the data is positive, with many texts expressing positive emotions, some being neutral, and fewer expressing negative sentiments. This distribution can help in understanding the general mood and opinions reflected in the data and guide further, more detailed analyses.

ROBERTA

Text data is preprocessed to replace user mentions and URLs with placeholders.The preprocessed text is tokenized using RoBERTa's tokenizer, converting it into a format suitable for input to the model. The tokenized text is passed through the RoBERTa model to obtain sentiment scores for negative, neutral, and positive sentiments.The raw sentiment scores are processed using softmax to obtain probabilities, which are then used to determine the overall sentiment of the text.

 RoBERTa is trained on a large corpus of text data using the masked language modeling (MLM) objective. In MLM, a percentage of the input tokens are randomly masked, and the model learns to predict the original tokens based on the context provided by the unmasked tokens. After pre-training, RoBERTa is fine-tuned on specific tasks, such as sentiment analysis, fine-tuning, the model weights are adjusted based on the labeled data for the target task.

The analysis shows that the sentiment around ChatGPT on Reddit is predominantly neutral and positive, with significant positive sentiment peaks indicating favorable events or discussions. Negative sentiment is less frequent but still present, suggesting areas for monitoring and potential improvement. This comprehensive sentiment analysis provides valuable insights into the community's perception and engagement with ChatGPT over time.

COMPARISON

The comparison between VADER and RoBERTa highlights differences in sentiment analysis approaches:

- VADER tends to classify more text as positive and less as neutral or negative, which might be due to its lexicon-based nature.
- RoBERTa, being a more sophisticated model leveraging deep learning, captures a broader range of sentiments with a more balanced distribution.

These differences suggest that while VADER provides a quick and straightforward analysis, RoBERTa offers a deeper and more nuanced understanding of sentiment, making it more suitable for complex and diverse textual data like Reddit posts.