Sentiment Analysis on Generative AI Using The Reddit API

Introduction

Generative artificial intelligence (GenAI) has made large strides within the past few years. The convergence of a number of reasons is responsible for the rise in popularity of GenAI. Some of the most influential reasons are improved hardware, enhanced algorithms and architectures, and increased data availability. Although GenAI has numerous applications, it is important to be aware of the potential risks it poses, some of which have already become very apparent. The implementation of GenAI in the entertainment industry, specifically in the animation and visual effects industry, are expected to cause heavy disruptions to jobs in the coming years. In addition, extremely realistic, synthetically generated images and videos, known as “deepfakes,” have begun to be used as tools of deception. This project aims to use the Reddit API to conduct sentiment analysis of two different subreddits, r/GenerativeAI and r/artificial, to understand how public sentiment towards GenAI has changed over time.

Data Collection

The data collection was done using the Python Reddit API Wrapper (PRAW). The API was used to fetch the 1000 most recent posts from the r/GenerativeAI and r/artificial subreddits. These subreddits were identified as containing the posts that relate most closely to GenAI. The original goal was to collect all posts from the range of 06/01/2022-06/01/2024. However the PRAW does not allow the filtering of posts by date. After extracting the 1000 most recent posts from the two subreddits, the date of the earliest post was 08/25/2023. In order to keep the analysis limited to the months for which complete data exists, the data was filtered to the range 09/01/2023-05/31/2024. The type of data included the title of the post, the text of the post, and the top comment of the post. The data for both subreddits was stored in CSV files.

Data Preprocessing

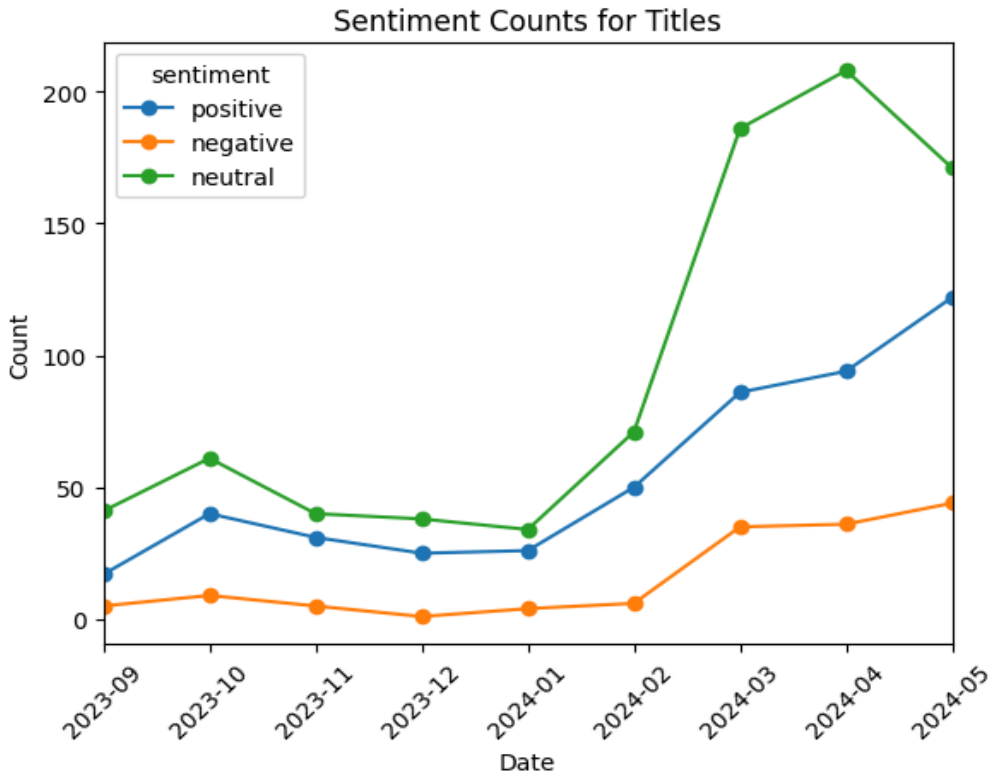
To clean the raw data, duplicates were first removed and the data type of the title, text, and comment columns were changed to strings. A function was created to prepare the data for sentiment analysis. This function converts the text to lowercase, performs tokenization to split the text into individual tokens, removes punctuation and other special characters, removes stopwords which are a set of commonly used words that do not contribute to sentiment, and performs lemmatization to reduce word variants to their base form.

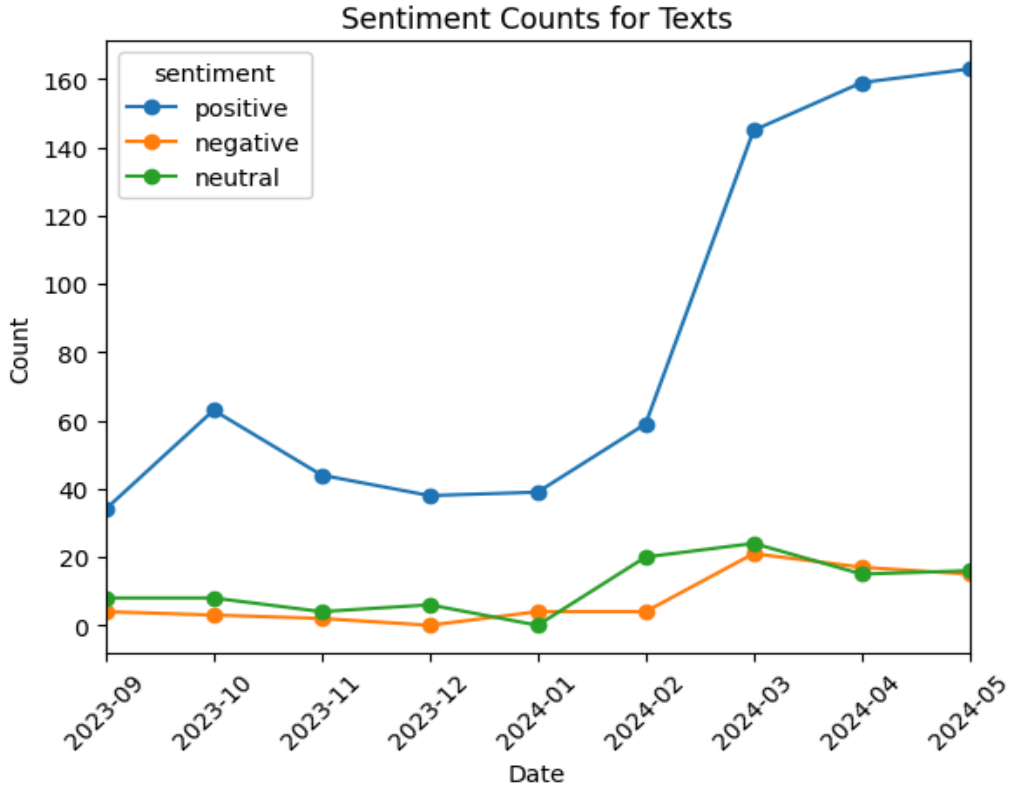
Sentiment Analysis

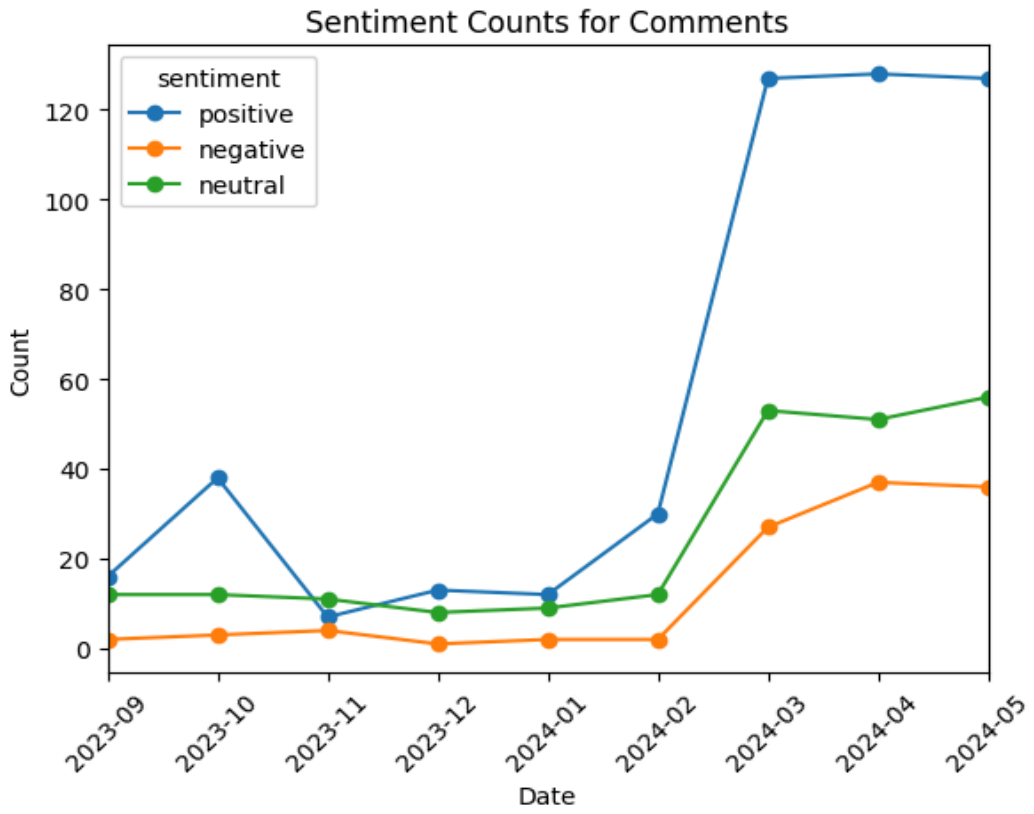
The sentiment analysis portion of this project was done with Valence Aware Dictionary for Sentiment Reasoning (VADER). VADER was chosen as the model for evaluation because it is a tool specifically designed to handle sentiments expressed in social media. As the posts written on Reddit will contain conversational language such as slang and acronyms similar to that which is found in social media posts, VADER will be more effective for the purpose of this project. A function was created to get the sentiment score of each preprocessed post and then classify the post as positive (>= 0.05), negative (<= -0.05), or neutral (anything else).

Data Visualization

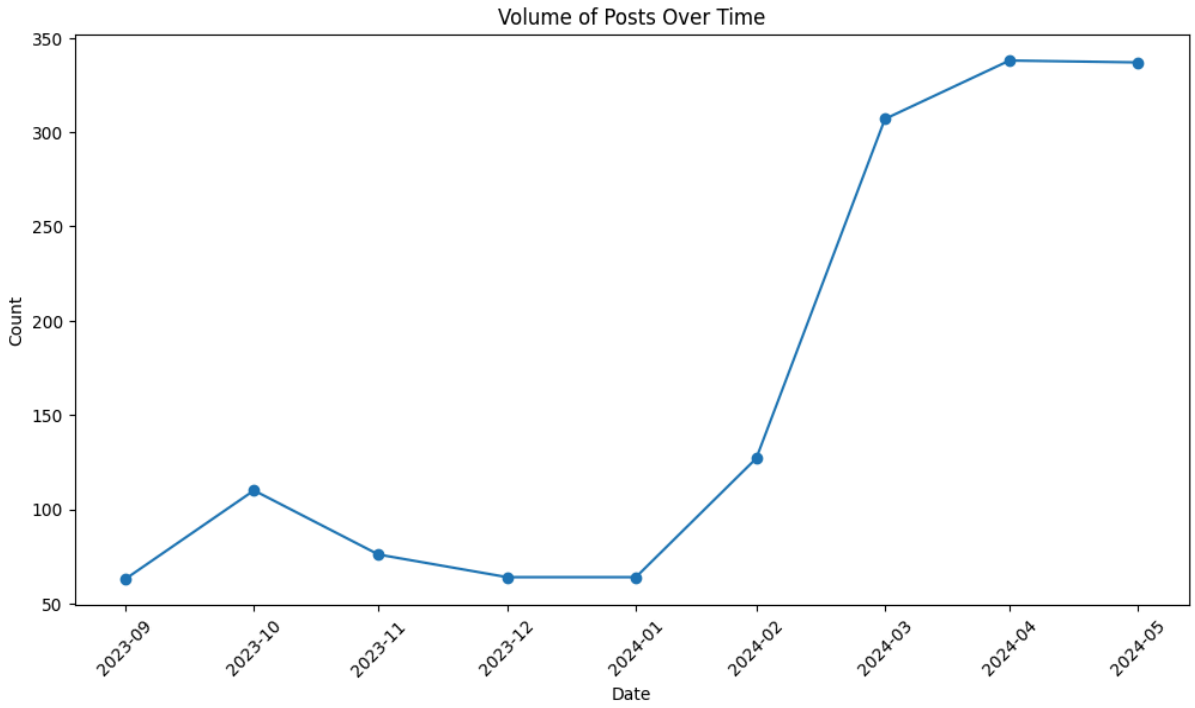
Distribution of positive, negative, and neutral sentiments over time for titles, texts, and comments:


Sentiment Counts for Titles
sentiment
positive
negative
neutral
Count
Date
0
50
100
150
200
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05

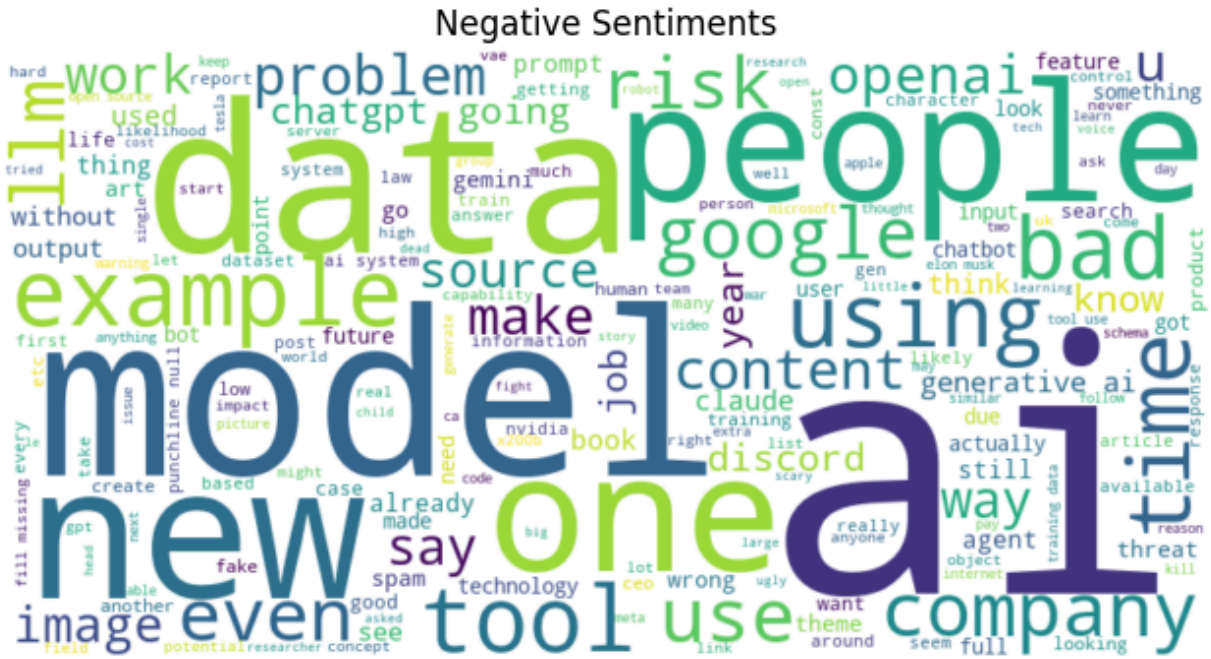

Sentiment Counts for Texts
sentiment
positive
negative
neutral
Count
Date
0
20
40
60
80
100
120
140
160
2023-09
2023-10
2023-11
2023-12
2024-01
2024-02
2024-03
2024-04
2024-05



Volume of posts over time:



Word clouds of positive and negative sentiments:



Analysis and Interpretation

From the plots of the distribution of positive, negative, and neutral sentiments over time for all post types, the general trend seems to be that the number of posts of all three sentiment types is growing over time. This is very likely due to the fact that the volume of posts over time is also increasing. The trends for the sentiment distribution plot and volume plot are closely linked, showing a small spike in October 2023, then staying about constant until starting to increase dramatically in February 2024. Although I'm unsure of what event caused the spike in posts in October 2023, I believe the increase that started February 2024 is due to the start of more widespread usage of GenAI applications for generating text, videos, and other media. From the plot showing the volume of posts, it appears that after reaching a high in April 2024, the volume plateaued and remained about the same for the following month. When new developments occur in the field of GenAI, it is expected that there will be an increase in overall post volume.

To answer the initial research question of how public sentiment towards GenAI has changed over time, within the last 9 months all sentiments have generally been rising, with positive sentiments having much more dramatic increases than neutral or negative sentiments. An interesting observation that can be seen from the plots showing the sentiment distributions is that for the title posts, there are more neutral sentiments than positive or negative sentiments, while for text and comment posts, there are more positive sentiments, and this gap is especially apparent from March 2024 to May 2024. There could be more neutral sentiments in title posts due to the fact that people are more careful with their words when they write titles, while being more free with their speech in the body text of their post or within the comments.

Conclusion

Sentiment analysis of the title, text, and comment posts from the r/GenerativeAI and r/artificial subreddits from the period September 2024 to May 2024 has shown that GenAI has become a very prominent topic. Not only has the volume of posts on the topic been increasing, but also the positive sentiments towards the topic. Starting March 2024, the gap between positive

compared to negative and neutral posts has been quite large. Originally, the goal was to conduct sentiment analysis over a period of two years as GenAI has been a greatly trending topic not only within the past nine months, but within the past few years. However, due to the limitations of the PRAW, the posts over the desired time window were not able to be acquired. In order to make the analysis of this project more detailed, more data can be extracted from other related subreddits such as r/MachineLearning, r/ArtificialIntelligence, and r/deeplearning to name a few. For this project I combined the data for two subreddits, but for further research it would be interesting to analyze sentiment trends across different subreddits to see if trends remain the same or vary.