Text Analysis on ChatGPT and GPT-4

Shuxiang Sui

Columbia University – Graduate School of Arts and Sciences

Natural Language Processing Project Report

Instructor: Professor. Houlihan

1. Abstract

This study aims to compare the sentiment scores of two large language models, ChatGPT and GPT-4, within the same time window. We analyzed a dataset of online text content and found that both models showed a decreasing trend in sentiment scores over time, with GPT-4 exhibiting a more stable decline while ChatGPT displayed more fluctuation. Additionally, we observed that GPT-4 had higher overall sentiment scores than ChatGPT during the analyzed period. These findings provide insights into the performance and behavior of language models in analyzing sentiment in online text content. With the increasing usage of AI-based technologies, it is crucial to understand how the public perceives and responds to these technologies. This study attempts to provide insights into public sentiment and discusses topics surrounding these two popular language models and compares them with each other.

Keywords: Sentiment Analysis, Topic Selection, LDA, ChatGPT, GPT-4, Twitter

2. Introduction

As the world's most advanced AI Chatbot, ChatGPT has garnered widespread attention and admiration from a diverse range of stakeholders worldwide. Social media discussions have been particularly lively, with people expressing their opinions on ChatGPT’s capabilities and its

potential to take on professional roles. Developed by OpenAI and released in November 2022, ChatGPT is built on top of OpenAI's GPT-3.5 and GPT-4 families of large language models (LLMs), and has been fine-tuned using both supervised and reinforcement learning techniques. A version based on GPT-4, the newest OpenAI model, was released on March 14, 2023, and is currently available for paid subscribers on a limited basis.

ChatGPT is known as a prototype dialogue-based AI chatbot capable of understanding natural human language and generating impressively detailed human-like written text. Early users have described this technology as an alternative to Google due to its ability to provide descriptions, answers, and solutions to complex questions, including offering guidance on coding, resolving layout issues, and optimizing queries. The practical applications of this technology extend to generating content for websites, responding to customer inquiries, providing recommendations, and creating automated chatbots. (Lock, 2022)

While ChatGPT has demonstrated impressive capabilities, it also has its limitations. First, the model sometimes generates responses that sound plausible but need to be corrected or nonsensical. Second, ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. Additionally, the model is often excessively verbose and overuses certain phrases, such as restating that it’s a language model trained by OpenAI. Ideally, the model would ask clarifying questions when the user's query is ambiguous. Instead, the current model relies on guessing what the user intended, which may result in inaccurate responses as well. Although OpenAI is making efforts to address these issues, it remains a challenging task. (*Introducing ChatGPT*, 2022)

As demonstrated above, ChatGPT presents numerous benefits to a wide range of users, yet its limitations remain a topic of controversy. Being a new technology, identifying public

sentiments especially those who are early users is of high importance for their opinions and sentiments can help to shape the broader perception of new technology. (Haque et al., 2022) Sentiment refers to an attitude, thought, or judgment prompted by feeling. Sentiment analysis, which is also known as opinion mining, studies people's sentiments towards specific entities. (Fang & Zhan, 2015) In this paper, we aim to investigate public perceptions of ChatGPT and GPT-4 through sentiment analysis of messages extracted from Twitter. Twitter is a popular social media platform where users share their thoughts and opinions as well as connect, communicate, and contribute to certain topics using short, 140-characters microblogs, known as tweets. (Qi & Shabrina, 2023) Due to its popularity and time-efficiency, Twitter serves as an ideal source for our analysis.

Our sentiment analysis will identify the emotional tone of each message, generate a sentiment score, track sentiment trends, and extract keywords used to express sentiment. We will also compare the outcomes of our analysis on ChatGPT and GPT-4 to highlight any differences. The analysis aims to generate a general understanding of public sentiment towards ChatGPT and the newly launched GPT-4, and provide insights into product optimization and development.

3. Data

3.1 ChatGPT Dataset

We found two datasets on Kaggle where people were talking about ChatGPT on Twitter. Because the two datasets overlap in time, we extracted the data, which includes over 350000 unique values from 11/30/2022 to 4/6/2023 as the analysis object for our subsequent discussion on ChatGPT. This dataset is a collection of tweets with the hashtags #chatgpt: discussions about the chatbot language model, sharing experiences with using ChatGPT or asking for help with ChatGPT-related issues. This dataset contains 12 columns, such as "Date" - Date of the tweet,

“Tweet” - Contents of the tweet, “Location” - Location of User, etc. The dataset also includes links to articles or websites related to ChatGPT, as well as images, videos, or other media. Overall, a collection of tweets with the hashtag #chatgpt would provide a glimpse into the online conversation surrounding ChatGPT. We conduct sentiment analysis on the content of these tweets, and combine some other columns, such as posting time, to study the changes in people's discussions on ChatGPT over time. In addition, we also have the dataset discussed for GPT4. We compare the two and find the similarities and differences. These will help us better understand people's attitudes toward ChatGPT and GPT4.

3.2 GPT-4 Dataset

This study uses GPT4 - the tweets data to conduct a sentiment analysis investigating the sentiment trend surrounding GPT-4 on Twitter. The dataset collects tweets with the hashtag #GPT4 including discussions about the GPT-4 language model, experiences using GPT-4, and requests for assistance with GPT-4-related issues. These tweets may also feature links to articles or websites pertaining to GPT-4, as well as images, videos, or other media. Overall, a collection of tweets with the hashtag #GPT4 would provide insights into the online conversation surrounding GPT4. The dataset covers the period from March 14, 2023, the day GPT-4 was launched, to April 12, 2023.

4. Methodology

This section presents our research methodology, where we discuss our research questions, preprocessing steps, identification of discussed ChatGPT and GPT-4 sentiment analysis, and topics.

4.1 Research questions

The following RQs motivated our empirical study.

RQ1. ChatGPT and GPT-4 Sentiments - What are the sentiments' changes over the time being expressed about ChatGPT and GPT-4 on Twitter?

RQ2. ChatGPT and GPT-4 Topics - What are the main topics being discussed about ChatGPT and GPT-4 on Twitter using LDA and high-frequency word cloud?

4.2 Preprocessing Steps

We pre-processed our ChatGPT and GPT-4 Tweet dataset using the following steps:

1) Lowercasing: We lowercased the tweets that represent words in different cases to the same lowercase form.

2) Punctuation Removal: We removed punctuation marks to retain only the alphanumerical data for cleaning our ChatGPTTweet dataset. In this removal step, we also removed URLs and Emojis.

3) Stop words Removal: Particularly, in the Sentiment Analysis Part (RQ1), we removed stop-words that appear frequently (e.g., this, are, and a) but do not help to distinguish one tweet from another. And at the same time kept the tweets that included the words that are negative auxiliaries like don't and doesn't. By doing that, the sentiment wouldn't be affected, like from negative to positive. In the Topic Selection Part (RQ2), we removed all the stop-words using the NLTK English stop-word list.

4) Lemmatization: We performed WordNet-based lemmatization using NLTK. We used lemmatization to represent a word's inflected forms to its dictionary-based root form.

4.3 Sentiment Analysis

To answer RQ1, we performed sentiment analysis for each dataset using Vader (Valence Aware Dictionary and sEntiment Reasoner), since it is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. We grouped the tweets by day and week to get the average sentiment score and visualize the daily/weekly sentiment change to better analyze the trend.

4.4 Topic Selections and Word Cloud

To answer RQ2, we identified a set of ChatGPT and GPT-4 key topics using the Latent Dirichlet Allocation (LDA) modeling technique. LDA is used to group tweets of our two datasets separately into a set of topics using word co-occurrence and frequency. A set of probabilities are assigned to each tweet by LDA. Here, the probabilities refer to the chances of a tweet being related to a specific topic. After getting a generalized theme through LDA, we then analyzed high-frequency words through word clouds and combined the LDA themes and word clouds to come up with our final analysis results.

5. Results

5.1 ChatGPT's Sentiment Analysis

The popularity and public opinion of ChatGPT, as analyzed through their sentiment scores, are undergoing significant changes.

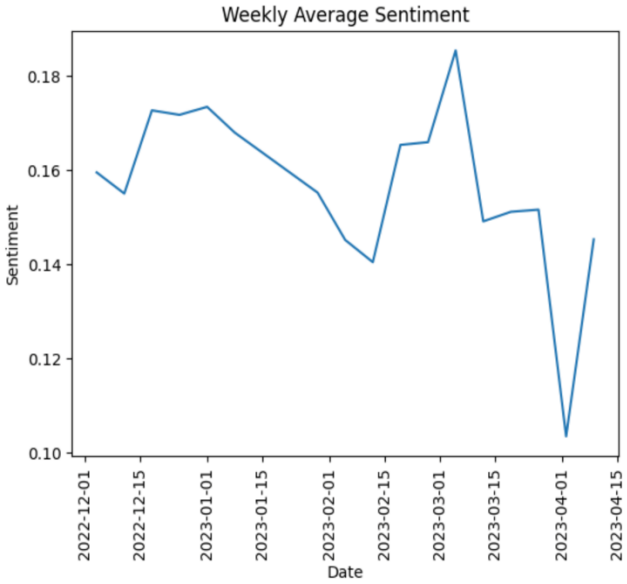


Fig. 1 Weekly average sentiment score for tweets about ChatGPT posted between 2022/12/01 to 2023/04/15

Our team conducted an in-depth analysis of the ChatGPT datasets, which are relatively large and encompass a diverse range of conversations and discussions. To ensure the accuracy of our analysis, we performed sentiment analysis on both weekly and monthly units. The weekly graph revealed significant fluctuations in the sentiment scores, making it difficult to identify a clear trend [Figure 1]. Therefore, we performed the analysis on a monthly basis, which revealed a noticeable downward trend [Figure 2]. This trend indicates that people are increasingly evaluating ChatGPT negatively over time.

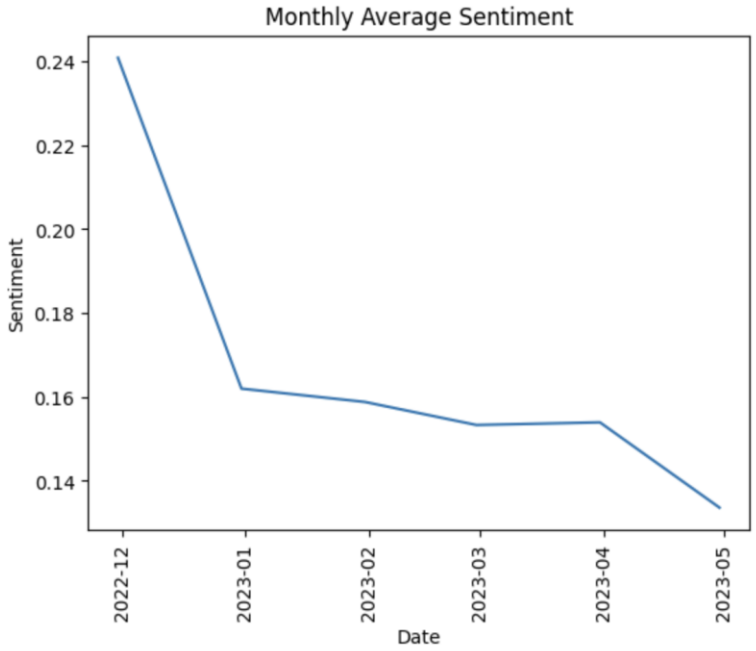


Fig. 2 Monthly average sentiment score for tweets about ChatGPT posted between 2022/12/01 to 2023/04/15

This finding has important implications for the future development and deployment of ChatGPT. If the negative trend continues, it could have significant consequences for its user base, market share, and overall success. Therefore, it is important to understand the underlying factors contributing to this trend and take appropriate measures to address them.

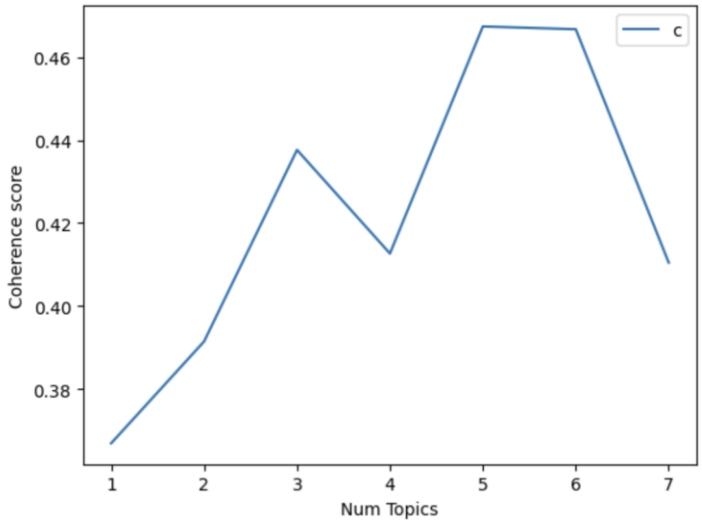
5.2 ChatGPT's Topic Selection

To better understand the topics discussed when mentioning ChatGPT, we analyzed the data using Latent Dirichlet Allocation (LDA). In the curve [Figure 3], we can observe that the coherence score gradually improves as the number of topics increases. At the point where the number of topics equals 3, we can see a clear "elbow" where the coherence score drops sharply. This number of topics at the "elbow" is typically considered as the optimal number of topics,

indicating that the program believes that when three topics are selected, they can best summarize the text accurately while maintaining a high coherence score.

Fig. 3 Relation between the number of topics and coherence score

However, when we experimented with different topic numbers, such as five, seven, and nine, we found something interesting and meaningful. The most interesting results were generated when the topic number was seven. Our analysis revealed several topics that people



often discuss when mentioning ChatGPT.

N_Topic = 7:

(0, '0.095"chatgpt" + 0.065*"ai" + 0.031*"openai" + 0.021*"new" + 0.019*"googl" + 0.016*"chatbot" + 0.013*"technolog"')*

(1, '0.056"chatgpt" + 0.025*"stori" + 0.025*"fun" + 0.023*"tweet" + 0.015*"ai" + 0.013*"style" + 0.012*"write"')*

(2, '0.058"chatgpt" + 0.057*"openai" + 0.055*"twitter" + 0.050*"power" + 0.039*"midjourney" + 0.035*"dall" + 0.030*"ai"')*

(3, '0.084"chatgpt" + 0.030*"use" + 0.020*"write" + 0.014*"code" + 0.012*"generat" + 0.011*"ai" + 0.009*"prompt"')*

(4, '0.115"chatgpt" + 0.027*"ask" + 0.014*"openai" + 0.011*"answer" + 0.010*"know" + 0.010*"one" + 0.010*"question"')*

(5, '0.165"chatgpt" + 0.064*"ai" + 0.062*"openai" + 0.024*"elonmusk" + 0.018*"job" + 0.014*"replac" + 0.013*"bot"')*

(6, '0.112"gpt" + 0.078*"chatgpt" + 0.033*"web" + 0.031*"ai" + 0.029*"openai" + 0.022*"crypto" + 0.020*"bitcoin"')*

As the results shown above, one of the most commonly mentioned topics is other artificial intelligence programs or services besides ChatGPT, such as Midjourney, an artificial

intelligence program and service created and hosted by a San Francisco-based independent research lab, Midjourney, Inc. Midjourney generates images from natural language descriptions, called "prompts," similar to OpenAI's DALL-E and Stable Diffusion. There is also a topic class called ‘dall’. The fact that Midjourney is often discussed in the same context as ChatGPT suggests that there may be significant overlap in their respective user bases and areas of application. Another frequently mentioned topic from category five was Elon Musk, the tech

Fig. 4 Word Cloud of the Top 150 words that people mention the most

entrepreneur and CEO of Tesla and SpaceX. Musk has been an outspoken critic of artificial intelligence and has made several controversial statements about the potential dangers of AI. Therefore, it is not surprising that people often discuss Musk when talking about ChatGPT, which is an AI-powered language model. We also found that people often discuss digital coins, such as Bitcoin. ChatGPT has the potential to play a significant role in the development and adoption of these currencies by providing users with a more natural and intuitive way to interact with them.

In addition to these topics, we also create word clouds, which provide a more intuitive view of the most commonly mentioned words [Figure 4]. We created a word cloud to show the most frequently mentioned top 150 words [*'education', 'code', 'resume', 'cv', 'crypto', 'nft', 'openai', 'elonmusk', 'chatbot', 'language model', 'google', 'tweet', 'search engine', 'amp', 'person', 'people', 'learn', 'love', 'human'...*]. All these words can be separated into several categories. One category was education and work, where people use ChatGPT to help them with assignments, to improve resumes and other tasks related to education and work. Another category is mentions of other search engines, indicating that ChatGPT is often compared to other popular search engines such as Google, tweet, and amp. Finally, we found some keywords that can relate with the debates about whether AI will replace humans. People often mentioned the words "human," "love," and "learning" when discussing this topic, indicating that there is significant concern about the potential impact of AI on society and the future of work.

In conclusion, our analysis of ChatGPT datasets revealed several important findings about the sentiment trends, topics, and categories of words discussed when mentioning ChatGPT. These findings provide valuable insights into the current state and potential future of ChatGPT, as well as the broader landscape of artificial intelligence and natural language processing. By understanding these factors, we can better anticipate and address the challenges and opportunities that lie ahead.

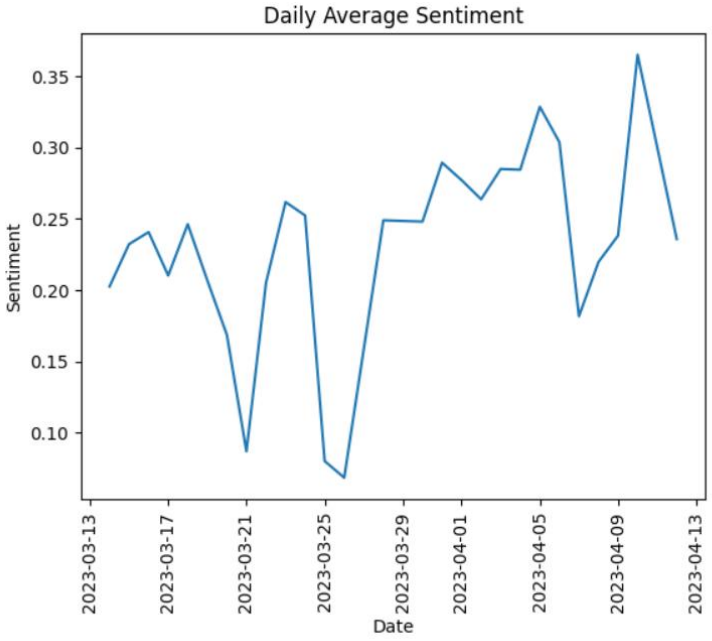
5.3 GPT-4's Sentiment Analysis

The sentiment analysis of GPT-4 reveals daily fluctuations in public opinion. Upon examining the original data source, we concluded that these sentiment waves are normal and may not be entirely accurate due to various factors. Firstly, it takes time for people to learn about the new features and make comparisons with other products. Secondly, technical issues and

version updates can result in multiple complaints appearing simultaneously. Finally, the physical location of Twitter users can lead to time differences, which may affect the accuracy of daily sentiment scores.

In total, we collected 30 days' worth of data using an API key from Twitter, starting from the launch date of GPT-4 on March 14, 2023. The daily average sentiment results (Figure 5)

Fig. 5 Daily average sentiment score for tweets about GPT-4
posted between 2023/03/14 to 2023/04/12



exhibited strong fluctuations each week. To further explore this trend, we calculated and visualized the average weekly sentiment score (Figure 6).

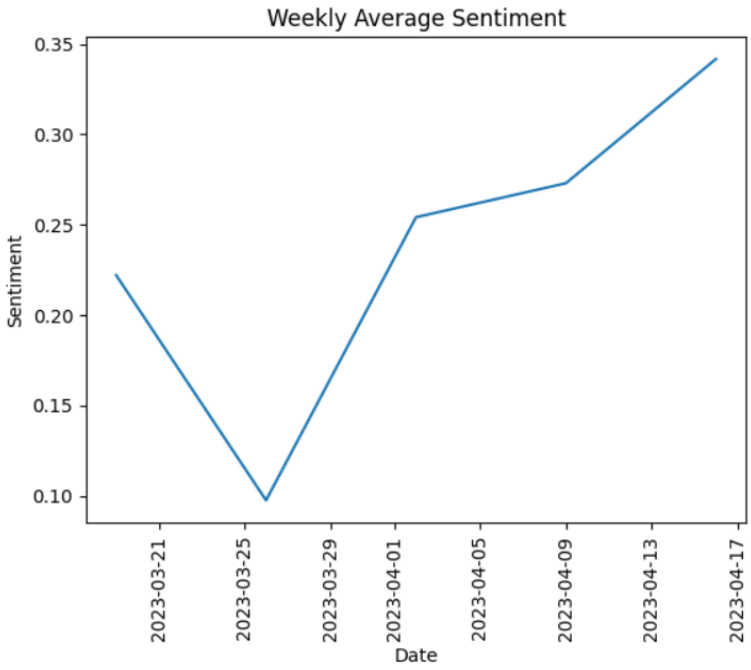


Fig. 6 Weekly average sentiment score for tweets about GPT-4 posted between 2023/03/14 to 2023/04/12

The analysis of the weekly average sentiment score of GPT-4 revealed a sudden drop in public attitude towards the model after its launch, followed by a continuous increase in sentiment score for the next 20 days. The sentiment score reached its peak at 0.34, which is an inspiring result for the sentiment analysis of GPT-4. This trend indicates that the public sentiments continued to enhance over time after the negative issues caused by the launch mistakes and misunderstandings were resolved.

5.4 GPT-4's Topic Selection

In our study on GPT-4 topic selection, we employed the Latent Dirichlet Allocation (LDA) method to identify the best topics. However, we observed signs of overfitting or heteroskedasticity, possibly due to the limited data available. Instead of delving into the causes

behind the optimal topic count of 2, as selected by the algorithm (Figure 7) with a Coherence Score of 0.5255, we decided to explore further.

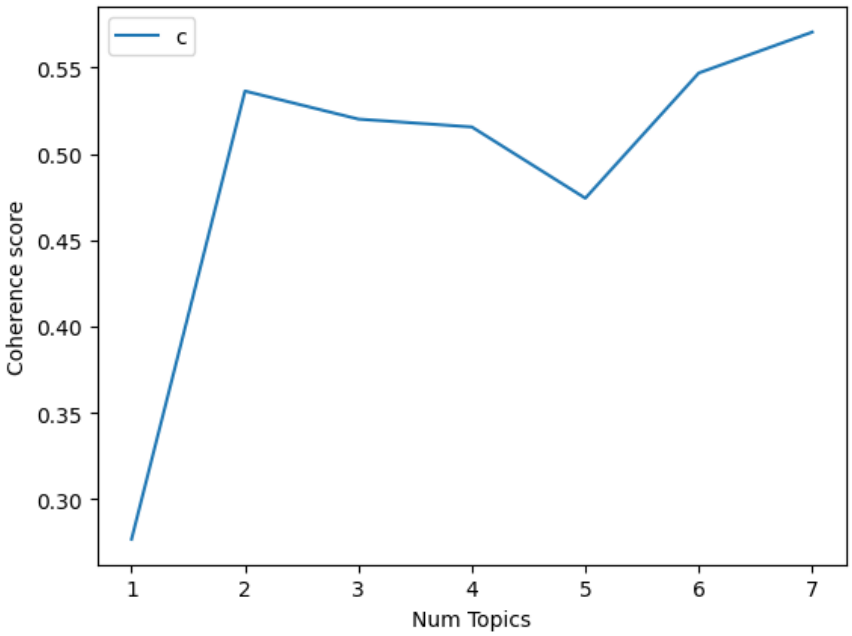


Fig. 7 Relation between the number of topics and coherence score

Upon increasing the number of topics in the LDA method, we found that when the topic count reached 7, our interpretations improved significantly, especially when cross-referenced with external information available online. This result allowed us to move beyond the first two topics, which primarily focused on GPT, OpenAI, and ChatGPT, and uncover findings more closely aligned with our previous LDA analysis of ChatGPT.

N_Topic = 7:

(0, '0.084*"ape" + 0.049*"high" + 0.042*"buy,token" + 0.038*"crazy,airdrop"')

(1, '0.140*"ai" + 0.122*"chatgpt" + 0.080*"crypto" + 0.064*"world"')

(2, '0.024*"but" + 0.023*"not" + 0.010*"like" + 0.009*"good"')

(3, '0.063*"openai" + 0.061*"c" + 0.056*"anything,possible" + 0.055*"haven,may"')

(4, '0.111*"openai" + 0.080*"ai" + 0.060*"doge" + 0.053*"chatgpt"')

(5, '0.035*"ai" + 0.031*"chatgpt" + 0.026*"model" + 0.022*"openai"')

(6, '0.071*"ai" + 0.064*"chatgpt" + 0.052*"openai" + 0.017*"new"')

Interestingly, we noticed that the texts highlighted in orange were predominantly published and shared by third-party blockchain websites or related advertising accounts.

To gain a deeper understanding of the public's perception of GPT-4, we analyzed the top 200 most frequent words used in Twitter posts on the subject. After filtering out words with ambiguous meanings, we found that the top 150 words were primarily positive and conveyed a sense of excitement about GPT-4.

Exhibition of TOP 150 words:

*['ai', 'chatgpt', 'openai', 'life', 'doge', 'airdrop', 'week', 'world', 'bonk', 'web', 'ready', 'possible', 'save', 'anything', 'high', 'token', 'dog', 'buy', '**productivity**', 'may', 'miss', 'triple', 'haven', 'openaichatgpt', '**opportunity**', 'contract', 'millionaire', 'new', 'bull', 'like', 'future', 'only', 'model', 'crazy', 'language', 'see', '**portfolio**', 'checked', 'technology', '**microsoft**', '**code**', 'text', 'next', 'time', '**google**', 'chat', 'beauty', 'tech', 'today', 'one', 'chatgptbot', 'prompt', 'human', 'powerful', 'data', 'know', '**bing**', 'create', 'write', 'check', 'twitter', 'real', 'help', 'image', 'intelligence', 'take', 'better', 'game', 'latest', 'think', 'news', 'work', '**magic**', 'would', 'content', 'people', 'free', 'first', 'power', 'could', 'need', 'good', 'access', 'day', 'read', 'wait', 'art', 'generate', 'via', 'artificial', 'large', 'innovation', 'video', '**multimodal**', 'live', 'powered', 'potential', 'advanced', 'available', 'learn', 'much', 'poem', 'level', 'business', 'many', 'glad', 'got', '**blockchain**', 'ask', 'want', 'open', 'best', 'used', 'writing', '**tool**', '**exciting**', 'try', 'really', 'learning', 'well', 'right', 'bard', 'industry', 'mind', 'great', 'version', 'development', 'poet', 'female', 'punk', 'release', '**amazing**', 'aesthetic', 'based', 'go', 'research', 'able', 'link', 'plus', 'generative', 'domain', 'give', 'k', '**python**', 'e', 'say', 'soon', 'times', 'look', '**marketing**']*

After the highly frequent word selection process, we then proceed with many generations of WordCloud graphics from the word list that only includes meaningful vocabulary (Figure 8):

Fig. 8 Word Cloud of the Top 150 words that people mention the most

By synthesizing the most frequent and correlated words into an information summarization derived from 150 keywords, this paragraph encapsulates the essential reflections on people's expectations and concerns regarding GPT and ChatGPT. The captured key terms emphasize OpenAI's ChatGPT as a transformative language model revolutionizing various sectors, not only because its multimodal approach has improved productivity and efficiency in areas such as content generation, writing, and marketing, but also due to its exceptional capabilities, including generating text, images, and poetry. The evaluation reveals trends of collaboration with leading technology corporations like Google and Microsoft, which have facilitated sophisticated features and enhanced accessibility for users. Moreover, the summary delves into the burgeoning interest surrounding blockchain technology, with a mention of Doge tokens and crediting Elon Musk for the strong connection between Twitter and Dogecoin. Ultimately, this concise overview effectively reflects public sentiment about GPT and ChatGPT, while showcasing the considerable potential of artificial intelligence.

5.5 Sentiment Score Comparison

In order to compare the differences and similarities between ChatGPT and GPT-4, we conducted a sentiment score comparison within the same time window. The 30-day sentiment score for ChatGPT displays a decreasing trend, whereas the GPT-4 sentiment score exhibits a larger fluctuation but still presents a statistically significant difference compared to the ChatGPT score (Figure 9).

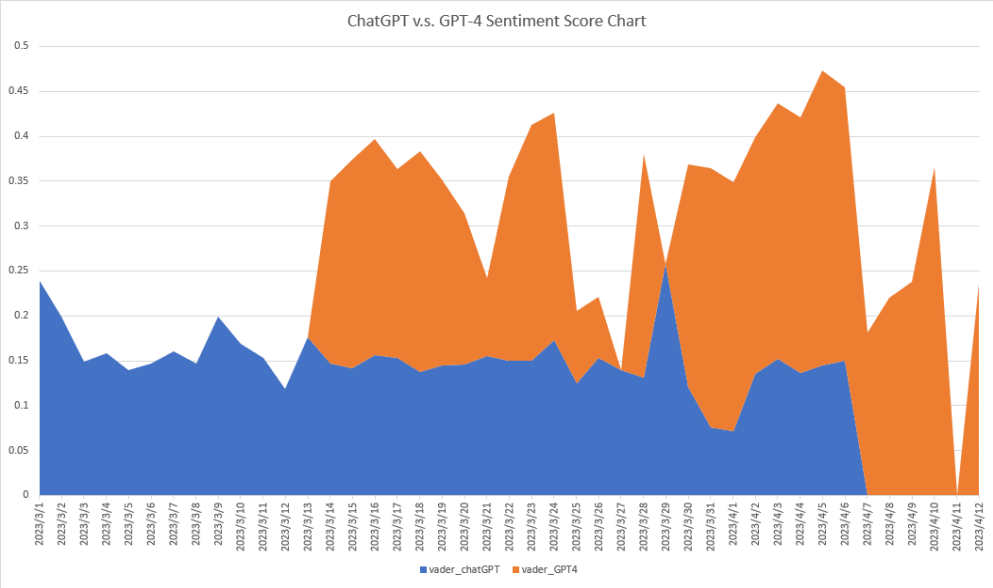


Fig. 9 Sentiment Score Comparison between ChatGPT and GPT-4 from Tweets

The average GPT-4 sentiment score is 0.231, which is 50% higher than the ChatGPT sentiment score of 0.151 for the recorded month, excluding the missing data. The variances for GPT-4 and ChatGPT data are 0.0049 and 0.0012, respectively, both of which are acceptable. Using a two-sample t-test with a degree of freedom of 62, the critical t-value for a two-tailed test is 1.998, and the p-value is 1.00369E-07. This result implies that we can reject the null hypothesis of no difference between ChatGPT and GPT-4 sentiment score trends at nearly all significance levels (Figure 10).

t-Test: Two-Sample Assuming Equal Variances

	vader_chatGPT	vader_GPT4
Mean	0.151452794	0.230900789
Variance	0.001159049	0.004864008
Observations	37	27
Pooled Variance	0.002712742	
Hypothesized Mean	0	
df	62	
t Stat	-6.026587843	
P(T<=t) one-tail	5.01847E-08	
t Critical one-tail	1.669804163	
P(T<=t) two-tail	1.00369E-07	
t Critical two-tail	1.998971517	

Fig. 10 T-Test for sentiment scores for ChatGPT and GPT-4 from Tweets

6. Limitations and Solutions

The data collected from Twitter may not represent a fully representative sample of the population, as not all individuals use Twitter, and those who do may not be representative of the general population. To address the issue of data bias, multiple resources could be considered, such as other social media platforms, surveys, or some interviews with individuals who do not use social media.

While sentiment analysis tools have improved over time, they are not perfect and may misclassify certain tweets. This could impact the accuracy of the results obtained. Maybe cross-validate the sentiment analysis could be conducted using multiple sentiment analysis tools or manual reviews could also be considered.

7. Conclusion

In conclusion, our study began with text data extraction from Twitter utilizing a public API, followed by the reorganization of the text data chronologically into a database. We applied

a comprehensive range of text-cleaning processes to the ChatGPT and GPT-4 datasets, including character filtering, URL deletion, English word selection, negative forms identification, stop word removal, token stemming, sentiment intensity analysis, and Latent Dirichlet Allocation methods. As a result, we created a time-series-based dataset for various generations of GPTs, conducted correlated sentiment score analyses, and generated two WordCloud graphics by extracting key topics and corroborating them manually using LDA methods.

Our comparison revealed that the public reaction to GPT-4 is significantly higher than that of ChatGPT on average over time. While tweets indicate a decreasing sentiment trend for ChatGPT, expectations within the Twitter community for GPT-4 show a contrasting increase. The two-tailed t-test further substantiates the statistical significance at all levels. Ultimately, our study offers valuable insights into the sentiment analysis of Twitter users regarding different generations of GPT products and highlights the potential and recommendations for future applications of GPTs' iterative and secondary products.

References

Fang, X., & Zhan, J. (2015, June 16). Sentiment analysis using product review data. *Journal of Big Data*.

Haque, M. U., Dharmadasa, I., & Sworna, Z. T. (2022, December 12). Exploring Sentiments of ChatGPT Early Adopters using Twitter Data.

Introducing ChatGPT. (2022, November 30). OpenAI. Retrieved April 17, 2023, from https://openai.com/blog/chatgpt

Lock, S. (2022, December 5). What is AI chatbot phenomenon ChatGPT and could it replace humans? *The Guardian*. https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans

Qi, Y., & Shabrina, Z. (2023, January 20). Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. https://link.springer.com/article/10.1007/s13278-023-01030-x