User Reactions to ChatGPT and ChatGPT4 Release - A Revolution Appears to Be on the Horizon

Vitória Willani de Almeida, Mirian Wawrzyniak Hage

Vitória Willani de Almeida. Bacharela em matemática. Rua Henrique Cabral, 342 – São Luiz; 31270-760 Belo Horizonte, Minas Gerais, Brasil.
Mirian Wawrzyniak Hage
2 Universidade de São Paulo. Doutora em Economia, finanças corporativas e econometria. Av Trabalhador São Carlense, 400 – Parque Arnold Schimidt; 13566-590 - São Carlos, São Paulo, Brasil.

User Reactions to ChatGPT and ChatGPT4 Release - A Revolution Appears to Be on the Horizon

Abstract

OpenAI's ChatGPT, a chatbot released in November 2022, appears to have gone viral. It can tasks such as aid with code debugging, answering queries and composing stories. The purpose of this capstone article was to investigate the breadth and intensity of users' reactions to ChatGPT's first month of operation and to the rollout of GPT4 update. It was envisaged to verify whether popular opinion and sentiment towards ChatGPT4 had shifted over time, so that responses were compared from different time periods. This was achieved by combining RoBERTa model (which employs the transformers architecture) and VADER model (which utilizes Bag of Words architecture). These models were picked because of their track records of success in studying natural language in web texts, which is essential for sentiment analysis. This research sheds light on how the general public interprets and responds to the introduction of cutting-edge technologies such as ChatGPT. We can learn more about how people feel about these tools as a whole by analyzing natural language data through sentiment analysis models. Companies and producers can use these data to hone their product development strategies and gain a deeper understanding of their target demographic.

Keywords: Artificial Intelligence, NLP, VADER, RoBERTa

Introduction

Through the development of artificial intelligence [AI], new opportunities for human-machine interaction have emerged. AI, particularly, chatbots have gained popularity as an efficient means of automating repetitive duties and providing customer service. However, the effectiveness of a chatbot is contingent on its ability to discern user input and respond appropriately. Chat GPT and Chat GPT4 language models come into play here.

Chat GPT and Chat GPT4 are powerful language models that have been pre-trained on enormous amounts of data and are capable of generating high-quality responses to user inputs. These models have the potential to revolutionize human-machine interaction and improve the chatbot user experience. According to Liu et al. (2005), building trust with users and avoiding unexpected consequences are both aided by the openness of natural language processing [NLP] models, it is critical to prioritize explainability in NLP algorithms. The authors propose a framework that incorporates model architecture, training data, and post hoc explanations in order to evaluate the interpretability of these models.

Moreover, a recent study by Hendricks et al. (2021) emphasizes the importance of human oversight and intervention in NLP algorithms. According to authors' suggestion, human annotators are involved in training and assessing NLP models. This study demonstrates this methodology can improve accuracy and impartiality of NLP models while fostering human learning and creativity.

In January, ChatGPT surpassed 100 million monthly active users, making it the fastest-growing consumer application. It has become the most popular website ever, surpassing Instagram, Facebook, Netflix, and TikTok3 Haque et al. (2022). According to more recent evaluations Borji; Frieder et al. (2023), despite five updates as of mid-February 2023, ChatGPT still cannot accurately count the number of words in a sentence, a task that elementary school children would typically solve with ease. This is a surprising and disappointing weakness. In the education sector, ChatGPT is viewed as both a threat to academic ethics and a facilitator of academic dishonesty, as well as an opportunity to reorient instruction toward more advanced forms of writing (Ventayen, 2023; Zhai, 2022).

The purpose of this study is to dissect the sentiments expressed in two sets of tweets: those from the ChatGPT launch and those from ChatGPT4 update. By providing a comprehensive evaluation of ChatGPT's current impression, our report can help influence public discourse and guide its future growth.

Materials and Methods

Two datasets were used in this paper; the first dataset contained 219,281 observations, while the second dataset contained 12,799 observations. The first dataset contains tweets posted during the first month of ChatGPT's release, while the second dataset contains tweets posted after ChatGPT4 upgrade. Using Python programming language and its main module, we initially conducted exploratory analysis and data cleaning. Pandas, Numpy, Matplotlib, Seaborn, NLTK, and Textblob were the utilized libraries.

The analysis of emotions was conducted using two distinct approaches NLP. The first implementation was VADER (Valence Aware Dictionary for Sentiment Reasoning) model, as proposed by Hutto and Gilbert, (2014). VADER model employs the Bag of Words [BOW] technique and was specifically developed for the analysis of social media texts, with the ability to process unique characteristics of this form of text.

The second approach employed RoBERTa (Robustly Optimized BERT Pretraining Approach) model, proposed by Yinhan Liu, Myle Ott, et al. (2019). RoBERTa is based on the 2018 version of Google's BERT (Bidirectional Encoder Representations from Transformers) paradigm. RoBERTa is an enhanced variant of BERT that was designed to outperform BERT.

To achieve this, ROBERTa was trained with a much larger data corpus than its predecessor, allowing it to learn more precise representations of word relationships. In addition, during training, RoBERTa used a faster learning rate and implemented parameter normalization techniques, which contributed to its improved performance.

Results and Discussion

We loaded the first date set with 219,294 observations and eleven columns published between November 30, 2022, and December 31, 2022. The columns are 'tweet_id', 'created_at', 'like_count', 'quote_count', 'reply_count', 'retweet_count', 'tweet', 'country', 'photo_url', 'city', 'country_code' and we initially deleted the columns "created_at", "country","photo_url","city" and " country_code ".

	tweet_id	like_count	quote_count	reply_count	retweet_count	tweet
0	1598014056790622225	2	0	0	0	ChatGPT: Optimizing Language Models for Dialog...
1	1598014522098208769	12179	889	1130	3252	Try talking with ChatGPT, our new AI system wh...
2	1598014741527527435	2	0	0	1	ChatGPT: Optimizing Language Models for Dialog...
3	1598015493666766849	561	8	25	66	THRILLED to share that ChatGPT, our new model ...
4	1598015509420994561	1	0	0	0	As of 2 minutes ago, @OpenAI released their ne...

Figure 1. Data frame GPT after eliminating the unnecessary columns
Source: Twitter

Then, we individually examined a subset of the tweets.

Tweet in position 1: `'Try talking with ChatGPT, our new AI system which is optimized for dialogue. Your feedback will help us improve it. https://t.co/sHDm57g3Kr'`

Tweet in position 0: `'ChatGPT: Optimizing Language Models for Dialogue https://t.co/K9rKRygYyn @OpenAI'`

Tweet in position 372: `'im asking ChatGPT simple math questions and this legitimately is giving me useful explanations in terms i can understand lmao'`

Tweet in position 761: `'So far ChatGPT seems like a better therapist than ELIZA at least. https://t.co/77dWnlInDA'`

Tweet in position 949: "[GPT-3] This post discusses an individual's experience with ChatGPT, a chatbot that uses natural language processing to generate responses to questions. The individual felt as though the chatbot's responses were manipula [...] https://t.co/P196qLOhzM"

Tweet in position 298: 'Today was a great day for @OpenAI to release ChatGPT:\\n\\nAn A.I generated interview between @SBF_FTX and @andrewrsorkin https://t.co/1fei3pEL8q'

Tweet in position 11973: 'A bit of poetry about phages from #ChatGPT. 😆 https://t.co/6RHv5gpwhg'

Tweet in position 7681: "As part of our degree we ask the students to produce an experimental proposal on a subject. Interested what @OpenAI #ChatGPT would do with my Q - I'm amazed how coherent the proposal is. Although shallow and brief) it is certainly a good starting point #Essay #educhat https://t.co/DrO2KovPEu"

Tweet in position 12126: 'This is amazing! #ChatGPT can generate components in #VueJS with #tailwindcss and probably much more than that. 🤯 Should I be scared of losing my job, anytime soon? https://t.co/V6Pesw7ciq'

From this initial observation, it is possible to conclude that links, emoticons, hashtags, abbreviations, and all other expected characteristics of social network text are prevalent in tweets. For this reason, we selected VADER model, which can accommodate these variations of social networks.

VADER is an algorithm for NLP that integrates a sentiment lexicon approach with grammatical rules and syntactic conventions to articulate both polarity and intensity of sentiment. Vader is an open-source Natural Language Toolkit [NLTK] application. The lexicon approach indicates this algorithm created a dictionary containing an exhaustive list of sentiment characteristics. In addition to words, this lexical dictionary includes phrases (such as "the shit" and "kiss of death"), emoticons (such as ":-("), and abbreviations with emotional connotations (such as "WTF" and "OMG"). Ten independent human raters assessed polarity and intensity of each lexical feature on a scale ranging from "-4: Extremely Negative" to "+4: Extremely Positive." The average score is then used as an indicator of sentiment for each lexical feature in the dictionary. In Vader, the word "ok" has a positive rating of 0.9, "good" is

1.9, and "great" is 3.1, whereas "horrible" has a negative rating of -2.5, ":(" has a negative rating of -2.2, and "sucks" has a negative rating of -1.5.

VADER's lexicon dictionary contains approximately 7,500 sentiment features, and any word not included in the dictionary will receive a score of "0: Neutral". In addition to the lexicons of sentiment, there are neutral structures that can affect the polarity of sentiment (such as "not" and "but") or the intensity of the entire sentence (such as "very" and "extremely"). Several heuristic principles for handling punctuation, capitalization, adverbs, and contrastive conjunctions have been implemented in VADER. Below, we demonstrate how VADER responds to minor sentence modifications.

VADER is smart, handsome, and funny. compound: 0.8316, neg: 0.0, neu: 0.254, pos: 0.746, VADER is smart, handsome, and funny! compound: 0.8439, neg: 0.0, neu: 0.248, pos: 0.752,

VADER model can recognize that the interjection in this instance has a positive polarity.

VADER is very smart, handsome, and funny. compound: 0.8545, neg: 0.0, neu: 0.299, pos: 0.701, VADER is VERY SMART, handsome, and FUNNY. compound: 0.9227, neg: 0.0, neu: 0.246, pos: 0.754,

VADER model increased the positive polarity because it determined that the capitalization of the text highlighted the expressed emotion. In addition, by duplicating the same phrase with and without an exclamation point, it interprets the uppercase as being more positive than the exclaim.

VADER is VERY SMART, handsome, and FUNNY!!! compound: 0.9342, neg: 0.0, neu: 0.233, pos: 0.767, VADER is VERY SMART, really handsome, and INCREDIBLY FUNNY!!! compound: 0.9469, neg: 0.0, neu: 0.294, pos: 0.706,

With these phrases, we can observe how the model behaves prior to intensity adverbs.

VADER searches the text for recognizable emotional cues, adjusts the text's strength and polarity in accordance with predetermined rules, adds up all of the scores for these cues, and then normalizes the final score to the range [-1, 1] using the function:

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

here x represents the total score and alpha is 15, the maximum possible value of x.

In addition to the sentence's composite score, VADER returns the proportion of positive, negative, and neutral sentiment features, as demonstrated in the preceding example (Ying Ma, 2020).

Based on the heuristic rules and the normalization calculation, we can predict that VADER will average out the sentiment if the input text is relatively lengthy or has multiple tone and sentiment transitions. This is because Vader is designed to recognize microblog-like contexts, which typically contain no more than 280 words and a single sentimental theme. Algorithm validation also confirmed that VADER performs exceedingly well in the social media domain and outperforms human raters in classifying the sentiment of tweets.

VADER is computationally efficient compared to machine learning algorithms that require massive operations for word embedding and training due to rules and lexicon that are embedded. Even though sentiment features are limited to the built-in lexicon and rules, it is relatively simple to modify and extend the sentimental vocabulary while customizing VADER for particular contextual use cases.

Applying sentiment analysis with VADER model to our data, yielded a data frame containing the polarities of emotions for each tweet in addition to the compound column, which combines positive, negative, and neutral polarities.

	neg	neu	pos	compound
1598014056790622225	0.0	0.700	0.300	0.4588
1598014522098208769	0.0	0.677	0.323	0.8225
1598014741527527435	0.0	0.889	0.111	0.4588
1598015493666766849	0.0	0.597	0.403	0.9029
1598015509420994561	0.0	1.000	0.000	0.0000

Figure 2. Data frame containing polarities produced by VADER model
Source: Original search results

We created a word cloud to determine which words appeared most frequently in the text. We obtained ChatGPT, Open AI, https, AI, Google, and a variety of non-emotional words that simply convey the context of the tweets analyzed.

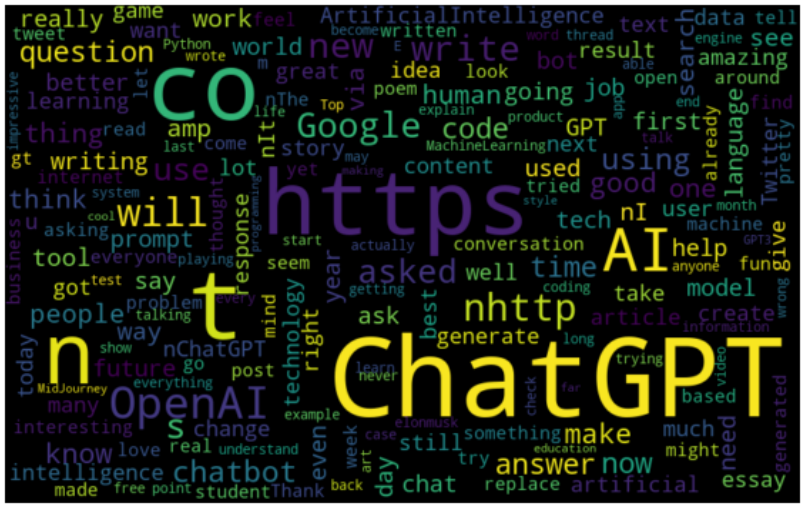

Figure 3. Word Cloud from ChatGPT tweets
Source: Original search results

We also created a chart that organized words into bigrams. The chart outcome confirmed what had been observed in the word cloud.

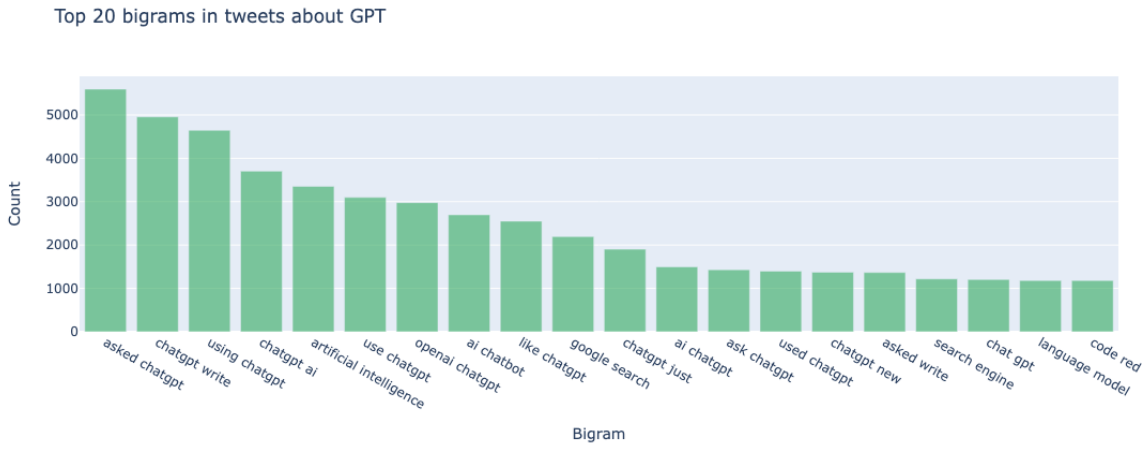

Figure 4. The top 20 bigrams in tweets about ChatGPT
Source: Original search results

By tallying the twenty most popular emoticons in tweets, the following list was compiled. Then, we inserted VADER analysis into our original data frame and reworked our sentiment analysis with RoBERTa model in order to compare the results from the two models.

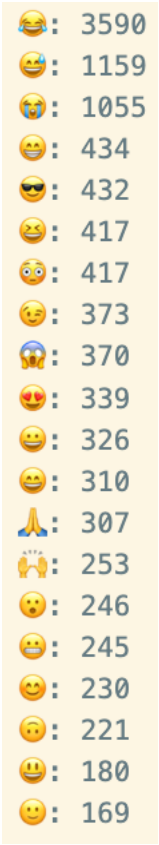

Figure 5. Most popular emoticons in tweets about ChatGPT
Source: Original search results

Hugging Face created RoBERTa, a language model that is particularly good for sentiment analysis. The model is based on transformer architecture and is an extension of the well-known BERT (Bidirectional Encoder Representations from Transformers) model, with several modifications and enhancements designed to increase its performance.

The significant enhancements to RoBERTa make it ideally adapted for sentiment analysis. RoBERTa is trained on a much larger dataset, which comprises more than 160 GB of text data compared to BERT's 34 GB. This is one of its primary advantages. This enables RoBERTa to have access to a much larger pool of knowledge and context when processing text. This is especially significant in the text, where context is essential for understanding the expressed sentiment. In contrast to BERT, dynamic masking presents the model during training with a random subset of elements from each input sentence. This enables RoBERTa to learn from a more diverse set of examples and can make it more robust to changes in input data, which is crucial when dealing with informal and inconsistent social media text.

RoBERTa also employs "No-Mask-Left-Behind" (NMLB) technique, which assures that all tokens are masked at least once during training, whereas BERT employs "Masked

Language Modeling" (MLM) technique, which masks only 15% of the tokens. This results in a more accurate representation of the input sentence and sentiment analysis.

The model can be fine-tuned for sentiment analysis, and it has demonstrated state-of-the-art performance in numerous benchmarks, making it a potent instrument for this particular NLP task.

After applying RoBERTa model's sentiment analysis and inserting findings into the data frame, the polarity of the feelings expressed by VADER and ROBERTa were then compared.

The first difficulty in comparing the results was caused by the different scales used: whereas VADER provides values of polarity of feelings between -1 and 1, with -1 being the most negative and 1 being the most positive, RoBERTa provides values from 0 to 1, where 0 is the absence of polarity and 1 is its maximum dominance. In addition, as RoBERTa is expressed in percentages, there is no compound column like VADER, which is another factor that complicates matters. To address this issue, a new column under the name "normalized_vader" was created, with values derived by adding one to "vader_compound" column and then dividing everything by two. Then, we plotted graphs to observe the distribution of polarities.

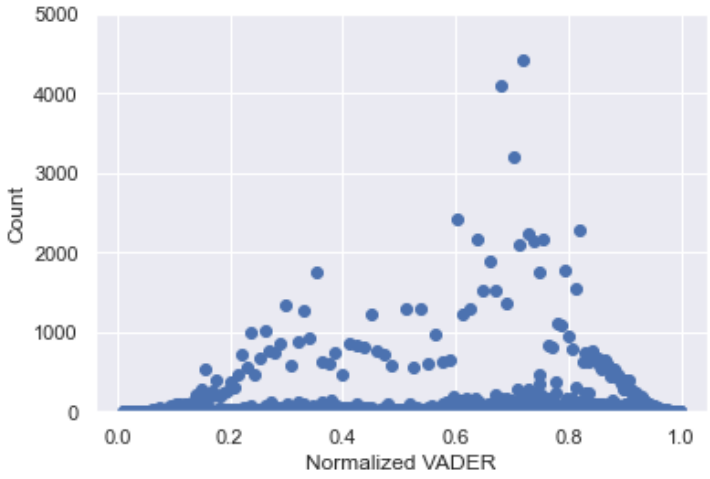

Figure 6. Distribution of polarities VADER ChatGPT
Source: Original search results

Most tweets with polarities between 0.6 and 0.8, as depicted in the preceding graph, has a positive sentiment. The graph also demonstrates that extreme polarities are uncommon. We were unable to create a dispersion graph for RoBERTa since it lacked a compound column, like we did for VADER.

Given the lack of a column that summarizes the polarity of tweets ranked by RoBERTa model, one solution was to add a label column to the results. Since our data lacked an original designation, it was impossible to calculate the analysis' accuracy. Therefore, we decided to identify the outcomes from the two models and then analyze the labels.

Whenever vader_compound is greater than or equal to 0.5, the assigned label is positive. If vader_compound fell between -0.5 and 0.5, a neutral label was assigned. If vader_compound was less than -0.5, the negative label was applied.

The most present polarity has been assigned as the designation for RoBERTa. Thus, we were able to compare the results from the analyses conducted by the two models using a data set that had been properly classified.

	VADER	RoBERTa
Positive	47.87%	33.68%
Neutral	34.01%	48.58%
Negative	18.12%	17.74%

Table 1. Labels VADER and RoBERTa
Source: Original search results

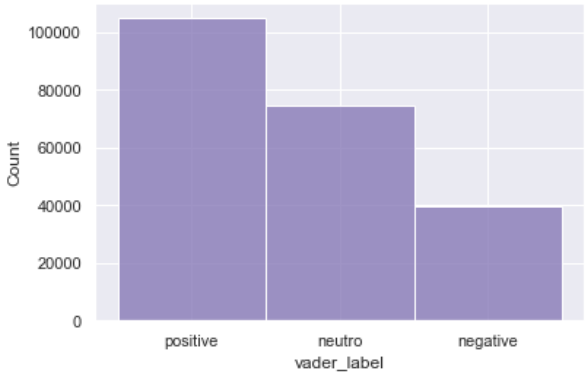

Figure 7. Labels VADER
Source: Original search results

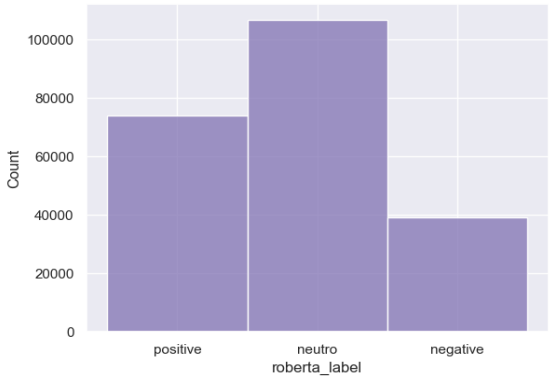

Figure 8. Labels RoBERTa
Source: Original search results

Observing the results, VADER rated the majority of tweets as positive, whereas RoBERTa rated the majority of tweets as neutral. The labels were identical in 60.45% of instances.

By manually analyzing a subset of tweets differentially ranked by the two models, it was possible to determine that VADER is extremely sensitive to certain words. Even if the remainder of the context is neutral, a single term with a positive connotation can have a

significant impact on the model. This allowed recognizing that RoBERTa's analyses tend to be more accurate, which was expected given that its technology is more recent.

'#ai Models are set to become the search engines of the future, ATM they still struggle with veracity… here is #chatgpt by @OpenAI based on #GPT3. #seo\\n\\nhttps://t.co/qqZ1G0fOTy'

This remark was rated negatively by VADER and neutrally by RoBERTa. RoBERTa seems to be more accurate.

'Damn, OpenAI is at it again. Just tried this and the implications of having an assistant like ChatGPT at your disposal is revolutionary. https://t.co/QJUQMAXdMq'

This remark was rated negatively by VADER and positively by RoBERTa. RoBERTa seems to be more accurate.

The diagrams below illustrate the degree of agreement between the models for each of the labels: positive, negative, and neutral.

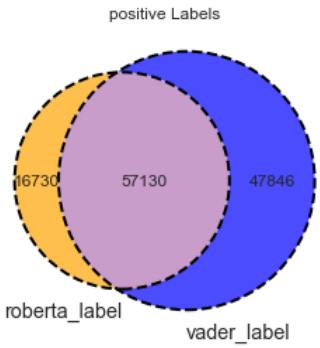

Figure 9. Positive labels

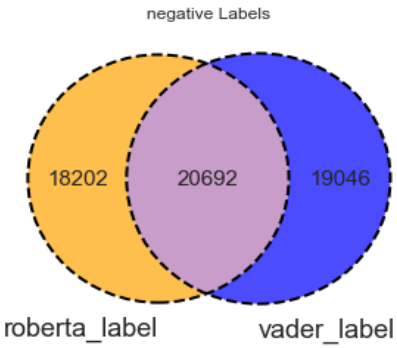

Figure 10. Negative labels

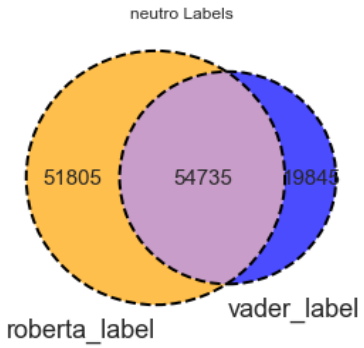

Figure 11. Neutral labels

Source: Original search results

Analyzing the diagrams reveals discrepancies between models. While VADER has a greater proportion of tweets classified as positive, RoBERTa appears to have classified these tweets as neutral.

We then proceed to the analysis of tweets published during GPT4 launch. We loaded the second date set with 12,799 observations and twelve columns published between March 14, 2023 and March 24, 2023. The columns are 'date', 'text', 'user_name', ‘user_location’, ‘user_description’, 'user_created', 'user_followers', 'user_friends', ‘user_favourites’, ‘user_verified’, ‘hashtags’, ‘source’. We then removed the columns for 'date', 'user_name', 'user_location', 'use_description', 'usu_created', 'user_followers', 'users_friends', 'User_favorites', 'USER_verified', 'hashtags' and 'source’.

	date	text	user_name	user_followers	user_friends	user_favourites
0	2023-03-24 23:25:54+00:00	Explain how important the Asian region is to t...	RawrA'tin	92	109	345
1	2023-03-24 23:21:00+00:00	Interesting. #ChatGPT4 doesn't know that it is...	darkmage	362	433	19587
2	2023-03-24 23:15:11+00:00	OMFG!!!! 😲 🤯🧠🤖 \n\nI can't wait for access to ...	Dr. Bo	204	466	3510
3	2023-03-24 23:15:04+00:00	"As part of their investment, Microsoft gained...	TechWafer Insights	14	5	4
4	2023-03-24 23:14:26+00:00	Now it's time for me to start assembling.\n\nI...	Ethan Robinson	52	114	856
...	...	...	...	...	...	...
12794	2023-03-14 13:18:03+00:00	#gpt4 also #brootswasright https://t.co/tMTUiN...	Clone ❌ God	1531	2789	16486
12795	2023-03-14 13:03:11+00:00	How much energy does it take to train #GPT4, t...	Adam Ai	219	202	25
12796	2023-03-14 12:45:23+00:00	It looks like VIDEO is dominating the GPT-4 po...	Intercept	484	646	447
12797	2023-03-14 12:40:07+00:00	The perfect storm is brewing, with the upcomin...	Dante	34	339	1387
12798	2023-03-14 12:37:17+00:00	On the horizon - #GPT4 and why it matters. | T...	MohitRajhans	7860	8651	35931

Figure 12. Data frame GPT4 after eliminating unnecessary columns
Source: Twitter

Checking the initial tweets:

Tweet in position 0: 'Explain how important the Asian region is to the global trading? @ChatGPTBot #gpt4'

Tweet in position 1: "Interesting. #ChatGPT4 doesn't know that it is ChatGPT4, nor does it know its own capabilities. #GPT4 @openai @gdb https://t.co/WqzrihbM2A"

Tweet in position 2: "OMFG!!!! 😲 🤯🧠🤖 \n\nI can't wait for access to #ChatGPT plugins! I will be able to connect my Gmail and Google calendars to the #GPT4!!! Ahhhhhhhh!!!! Not to mention the other infinite functionality!!! \n\nFind out more from @mreflow: https://t.co/VhHgh6zqiR"

Tweet in position 189: 'I don't know who created this but current feelings. #GPT4 #ChatGPT https://t.co/yHbUhOxcP4'

Tweet in position 374: '#web3 + $gpt4 = anything is possible 💪🧠 \n\n👉https://t.co/TVTQo6BEIw\n\n #airdrop #NFT $ARB $USDC $OP $SUI #ChatGPT #AI $CFX $BONE $RDNT $ARB $GPT4 #crypto #openai #GPT4 $GPT4 https://t.co/uS9nBfL0am'

We observed the same characteristics in the second data set as we did in the first: links, emoticons, hashtags, abbreviations, and an apparent tendency toward positive and/or neutral polarity.

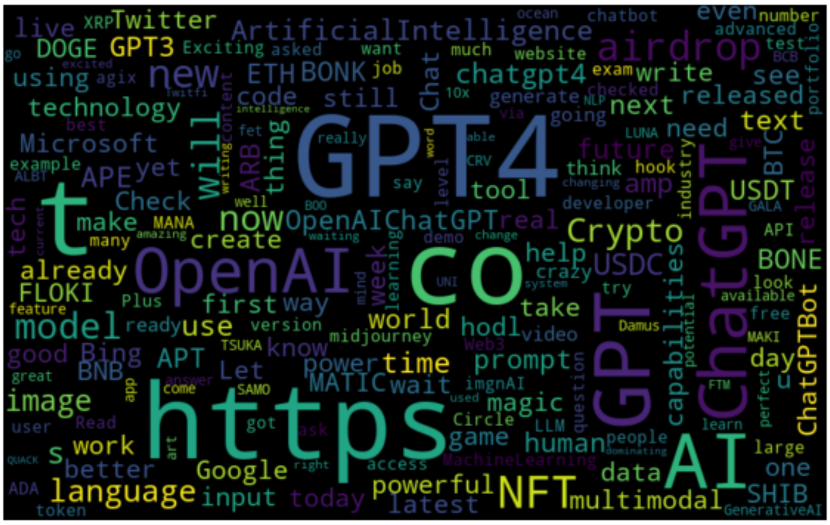

Figure 13. Word cloud ChatGPT4

Source: Original search results

We also created a chart that organized the words into bigrams. Chart outcome confirms what was observed in the word cloud.

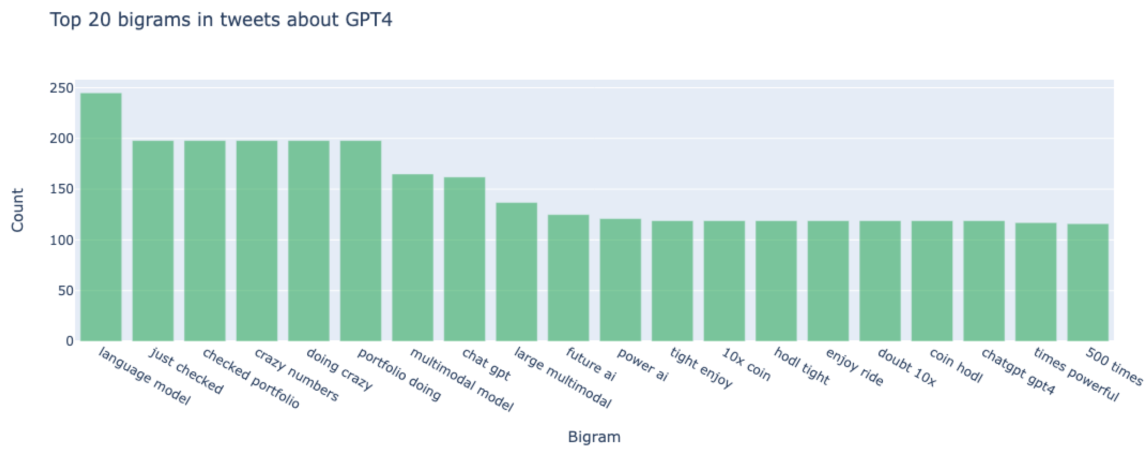

Figure 14. Top 20 bigrams in tweets about ChatGPT4

Source: Original search results

By tallying the twenty most popular emoticons in tweets, the following list was compiled.

😂: 242
😅: 75
😱: 49
😎: 37
😍: 36
😳: 31
😉: 27
🙌: 24
😄: 24
😭: 23
😑: 22
🙏: 20
😀: 20
😆: 19
😃: 18
😲: 17
😄: 17
😮: 16
😊: 15
😏: 15

Figure 15. Top 20 emoticons in tweets about ChatGPT4
Source: Original search results

Following the exact same procedures as with the initial data set, we applied VADER and RoBERTa models, modified the scale of values assigned by VADER to enable comparison with values provided by RoBERTa, and created the data label to facilitate an effective comparison between the two feeling analyses.

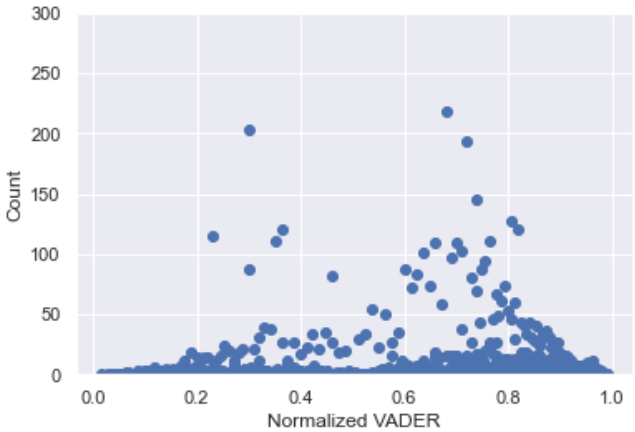

Figure 16. Distribution of polarities VADER ChatGPT4
Source: Original search results

By analyzing the distribution depicted in the preceding graph, positive polarity is observed to prevail among the data. The table below represents a numerical analysis of the percentage of positive, neutral, and negative labels provided by each model.

	VADER	RoBERTa
Positive	**50.67%**	**48.24%**
Neutral	**34.02%**	**43.05%**
Negative	**15.31%**	**8.71%**

Table 2. Labels VADER and RoBERTa
Source: Original search results

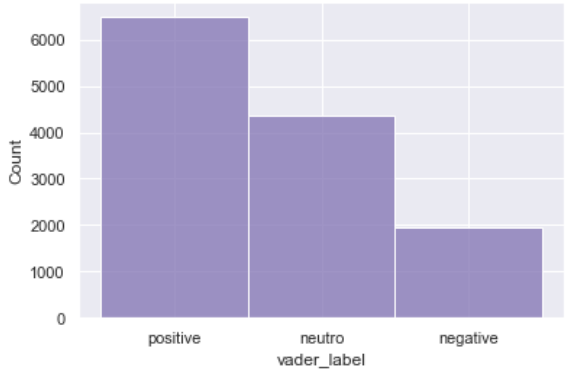

Figure 17. Labels VADER
Source: Original search results.

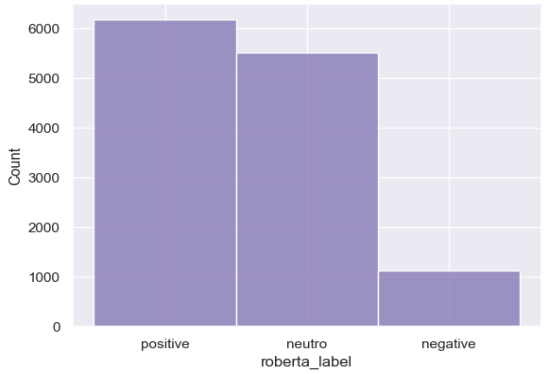

Figure 18. Labels RoBERTa
Source: Original search results

In 62.68% of the observations, the labels provided by the models were identical.

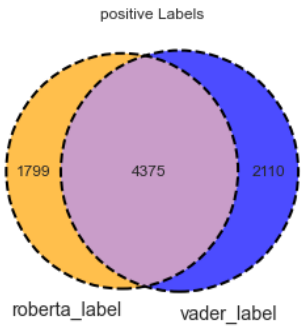

Figure 19. Positive labels

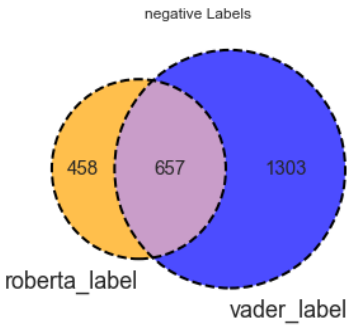

Figure 20. Negative labels

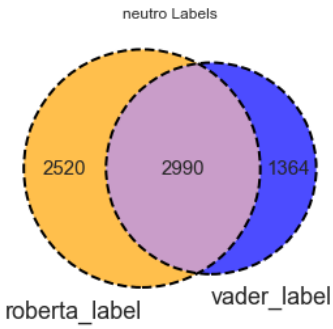

Figure 21. Neutral labels

Source: Original search results

Analyzing the diagrams reveals that the greatest discrepancy between the models was in the classification of negative tweets; VADER had a greater number of tweets classified as negative, whereas RoBERTa appeared to have classified these tweets as neutral.

By comparing the results from the launch of ChatGPT to counterparts from the release of ChatGPT4, are observed to negatives have been reduced in both models. In addition, VADER experienced minimal adjustments to its neutral rate and a slight rise in its positive rate. RoBERTa had a significant decrease in neutrals and a substantial increase in positives. The following graph illustrates those findings.

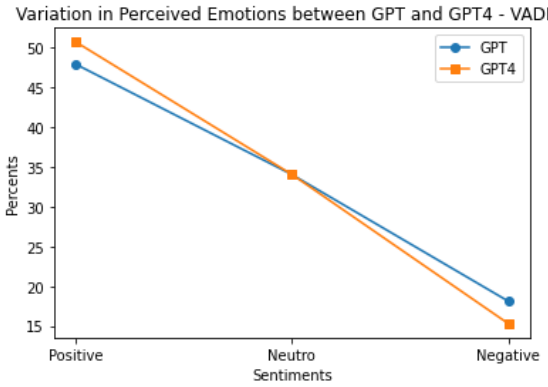

Figure 22. Variation in labels VADER
Source: Original search results

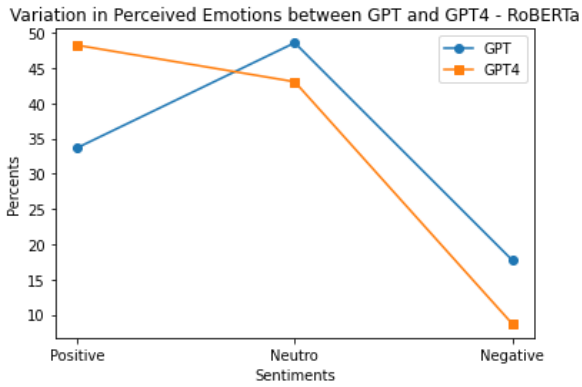

Figure 23. Variation in labels RoBERTa
Source: Original search results

Conclusion

The goal of this research was to compare and contrast the feelings stated in tweets during ChatGPT launch and CharGPT4 update.

Given the public's inexperience with the tool, we concluded that tweets posted shortly after the launch of ChatGPT were more neutral with positive trends, which we interpreted as usual. Already by the time of GPT4's release, we observed an increase in positivity, confirming public's favorable reception. Based on performance and comparison of applied models, we observe that VADER tends to be influenced by specific words, whereas RoBERTa depicts the overall polarity of the text more accurately. Within the constraints of their architectures, the models provided a comparable analysis of the polarity of tweets, corroborating the reliability of our findings. Future research should investigate changes over extended periods of time, consider the prevalence of tweets and papers (via likes and citations), investigate additional dimensions other than sentiment and emotion, and examine the expertise of social media actors and their geographic as well as demographic locations. As language models such as

ChatGPT continue to evolve and acquire more capabilities, future research will be able to evaluate their actual (rather than anticipated) impact on society, including their potential to exacerbate or mitigate existing inequalities and biases.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proc. WLSM-11s.

Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. In Proc. EMNLP-09.

Ali Borji. 2023. A categorical archive of chatgpt failures. ArXiv, abs/2302.03494.

Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. arXiv preprint arXiv:1206.6426, 2012.

Christoph Leiter and Ran Zhang and Yanran Chen and Jonas Belouadi and Daniil Larionov and Vivian Fresen and Steffen Eger. ChatGPT: A Meta-Analysis after 2.5 Months.

C.J. Hutto, Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. ICCL-10.

Dimosthenis Antypas, Asahi Ushio, Jose CamachoCollados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. Twitter topic classification. In Proceedings of the 29th International Conference on Computational Linguistics, pages 3386– 3400. Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, Marseille, France. European Language Resources Association.

Haque, M.U., Dharmadasa, I., Sworna, Z.T., Rajapakse, R.N. and Ahmad, H. (2022), “I think this is the most disruptive technology: exploring sentiments of ChatGPT early adopters using Twitter data”, arXiv. doi: 10. 48550/arXiv.2212.05856.

Hendricks, Alison Eisel; Watson-Wales, Makayla; Reed, Paul E. (2021): Student perceptions of AAE (Hendricks et al., 2021). ASHA journals. Presentation. https://doi.org/10.23641/asha.15241638.v1

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Kamps, J., Mokken, R. J., Marx, M., & de Rijke, M. (2004). Using WordNet to measure semantic orientation. In Proc. LREC-04.

Leah M. Bishop. 2023. A computer wrote this paper: What chatgpt means for education, research, and writing. SSRN Electronic Journal.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proc. WWW-05.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Mubin Ul Haque, I. Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. "i think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data. ArXiv, abs/2212.05856.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. In Journal of Artificial Intelligence Research, 37:141-188, 2010.

Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning.

Xiaomin Zhai. 2022. Chatgpt user experience: Implications for education. SSRN Electronic Journal

Ying Ma. NLP: How does NLTK. Vader Calculate Sentiment? In Medium.com Feb5, 2020.