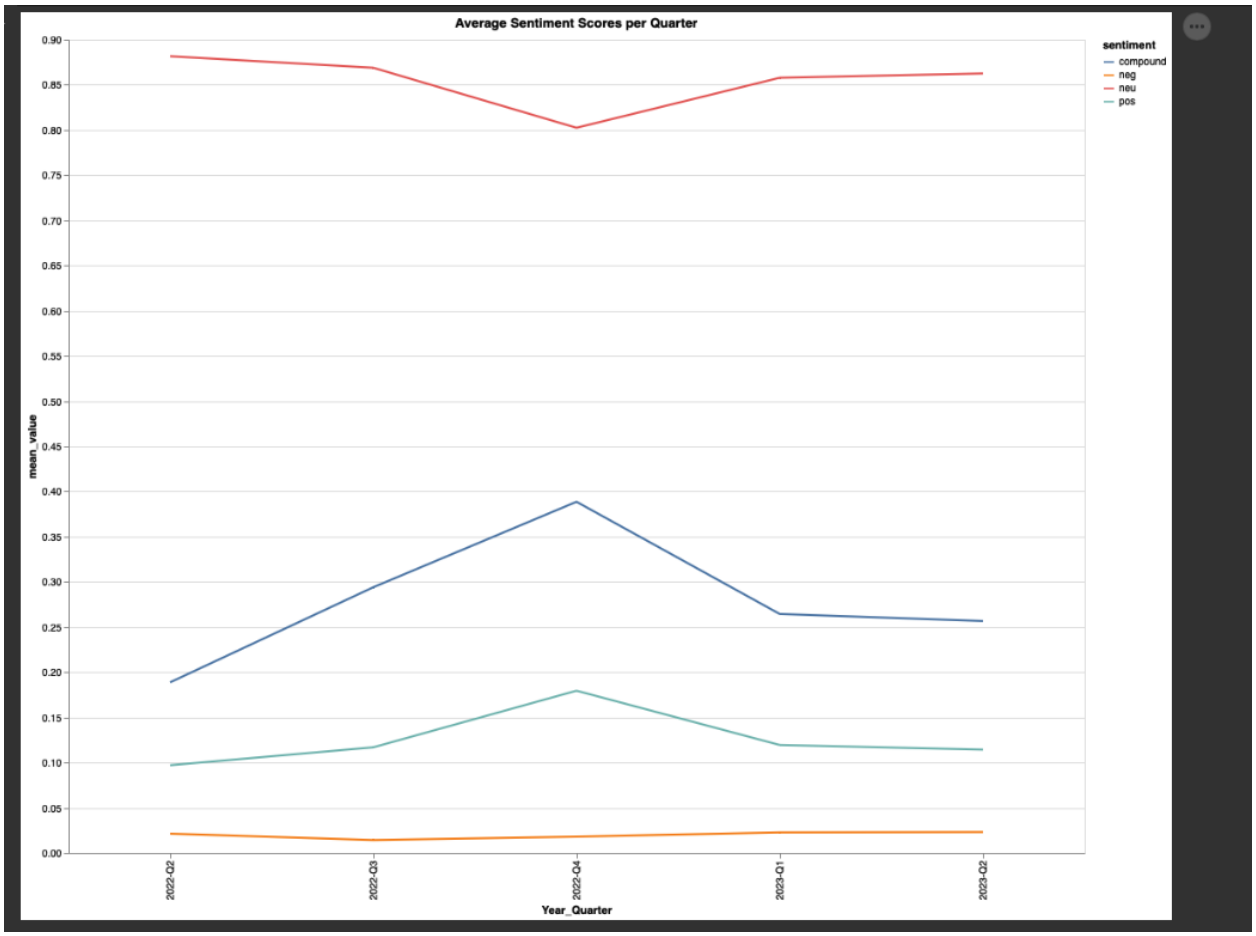
Exploratory Data Analysis

Abdulaziz Macabangon
Amanda D'Alessio
Amit Gattadahalli

W209 - Section 3

Hypothesis 1:

The velocity of tweets about GenAI and the sentiment of tweets over business quarters.



What's informative about this view:

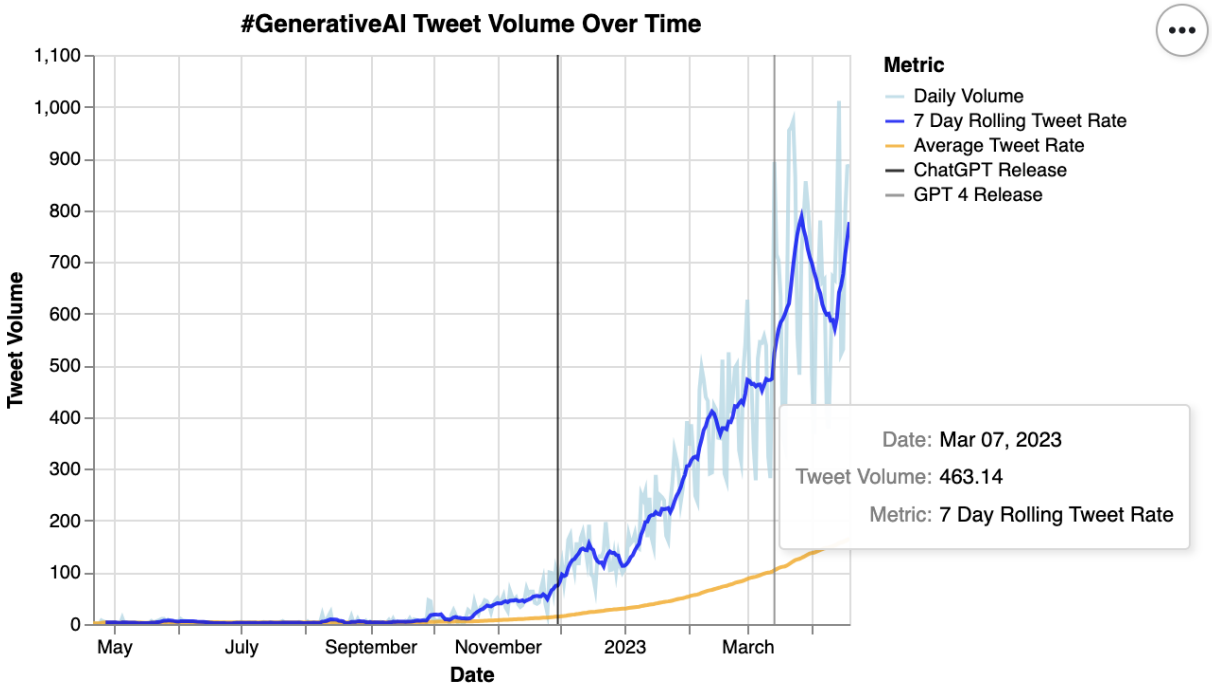
Here we can see from the compound score that the mean value of tweets were under .5 which when it comes to using the vader NLKT model closest to +1 is a positive and -1 is a negative. From what I can conclude on this the sentiment was getting better and it's all time high in 2022 Q4 which was around the time chatGPT came out which was November 2022. This could just be indicative of more volume of tweets about AI in general and with its mainstream breakthrough and newfound users realizing the capabilities of AI, may have caused this rise in sentiment.

What could be improved about this view:

This could definitely do better by showing a shadow of the volume of tweets and showing a weight to each line to get an idea of the total amount of tweets used to generate this visualization.

Hypothesis 2:

Tweet volume with the #GenerativeAI hashtag has generally increased over time between April 2022 and April 2023

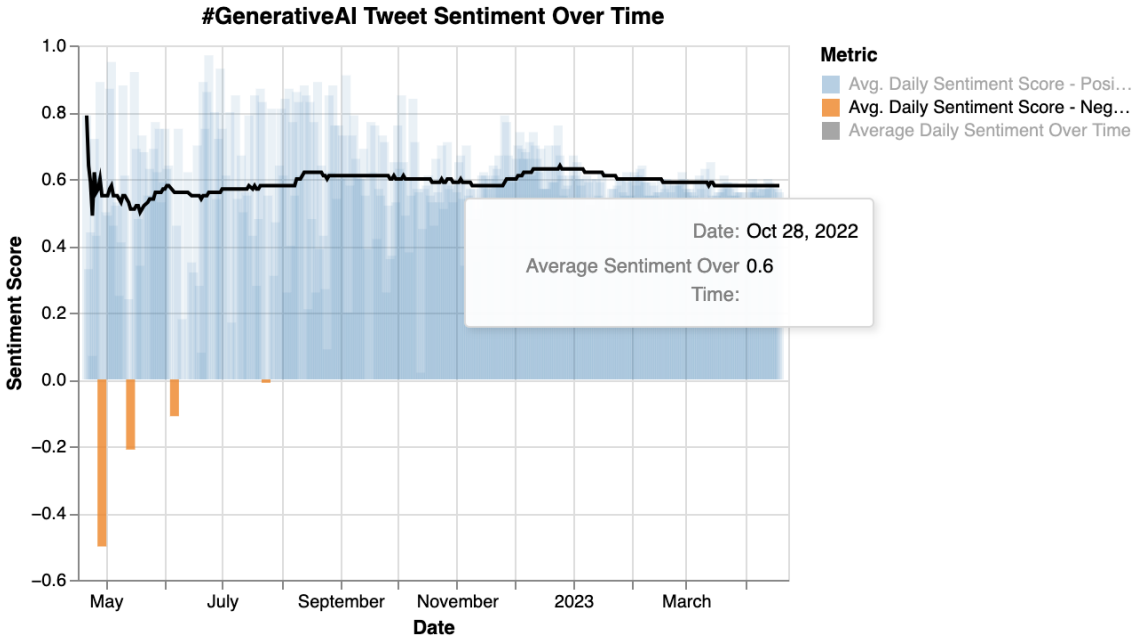


Visualization Notes:

- Daily Volume, 7 Day Rolling Average Tweet Rate, and Cumulative Average Tweet Rate Time
- General increase over time can be seen by the monotonically increasing trend of average tweet rate over time, which is echoed by both daily volume and 7 Day Rolling Average Tweet Rate having a consistent tendency to increase over time
- Major inflection points in tweet volume appear to occur near ChatGPT's initial release (11/30/2022) and GPT4 release (3/14/2023)
- IMPROVEMENTS:
    - Add in additional superimpositions of mainstream GenAI related news to the graph (OpenAI <> MS Partnership announcement, alternate tools like BARD or Llama, OpenAI plugins announcement, AI code freeze, etc.)

Hypothesis 3:

Sentiment towards Generative AI has collectively improved over time, especially around temporal regions where major Generative AI announcements were made. (We can replace this viz with the visual provided above by Abdulaziz Macabangon if that one is preferred, as well as the quick blurb on the VADER sentiment analysis technique).

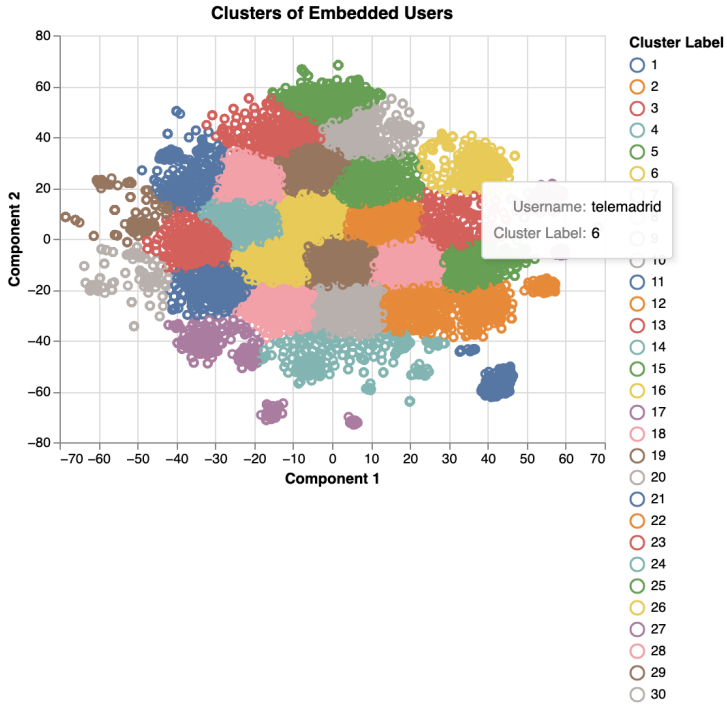


Visualization Notes:

- Sentiment for tweets is assigned by a deep learning model that we trained on the sentiment140 dataset of ~1.6 million positive and negative sentiment tweets. Huggingface's sentence transformer provides an embedding representation of the tweet as model input, and a score between -1 (highly negative) and 1 (highly positive) is returned where scores near 0 indicate a neutral sentiment
- Visualization shows Average Daily Sentiment, color coded by direction (positive / negative) and Cumulative average sentiment over time.
- While there are slight inflections in sentiment over time, such as the region near the end of November / early December 2023 that sees an increase in cumulative average sentiment over time (0.58 → 0.63), generally sentiment has appeared to be fairly static over time hovering near an average of near 0.6 representing a moderately positive sentiment.
- There are very individual few days with overall negative sentiments on average, and most days where this occur are very early in the data collection timeline (perhaps indicating skewed results that have a larger impact due to lower tweet volumes)
- Average Daily Sentiment has much less volatility as a metric post ChatGPT release (variation in values reduces with minor oscillations around 0.6)
- IMPROVEMENTS:
    - Add in superimpositions of major events,similar to the tweet volume chart
    - x axis showing datetime has an odd xtick (showing 2023) after November 2022

Hypothesis 4:

There are natural clusters of both users and tweets that showcase users / tweets that share similar content in terms of topics, semantic meaning, and sentiment.



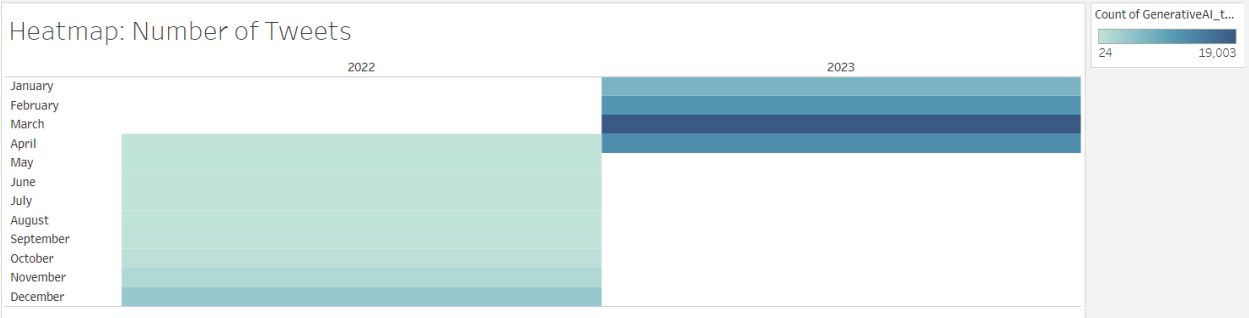
Visualization Notes:

- TSNE to reduce dimensionality of user embeddings (centroid of tweet embeddings) to 2 dimensions → KMeans clustering (30 clusters) on lower dimensional user embeddings
- There appear to be some natural clusters, with some especially nice separation around the borders
- IMPROVEMENTS:
    - Clusters are uninterpretable at the moment, would be good to inspect clusters to see what kind of attributes are shared amongst similar groups of data points (topics, sentiment, hashtags)
    - Dimensionality reduction (385 → 2) and then clustering may not be the optimal approach for identifying similar user / tweet groups. Could be worth trying alternative approaches such as directly clustering on more interpretable feature representations of tweets / users (sentiment + specific hashtags or keywords + lower dimensional tweet embeddings)
    - Different clustering algorithms (KMeans IDs spherical clusters, DBSCAN or Hierarchical could be better?)
    - Similarities between representations of users / tweets could be visualized as a network, which could be better suited for this as long as we can identify logical labels for the groups that we see

Hypothesis 5:

More tweets on the topic would correlate to a more polarizing view (either positive or negative).

Started by gathering information about the dataset, including seeing around what time over the year people were tweeting about this the most and the average number of tweets per user that could be expected. The assumption was that there would be more tweets after March 2023 (when ChatGPT GPT4 was released), and that people would tweet about this once or twice a month.

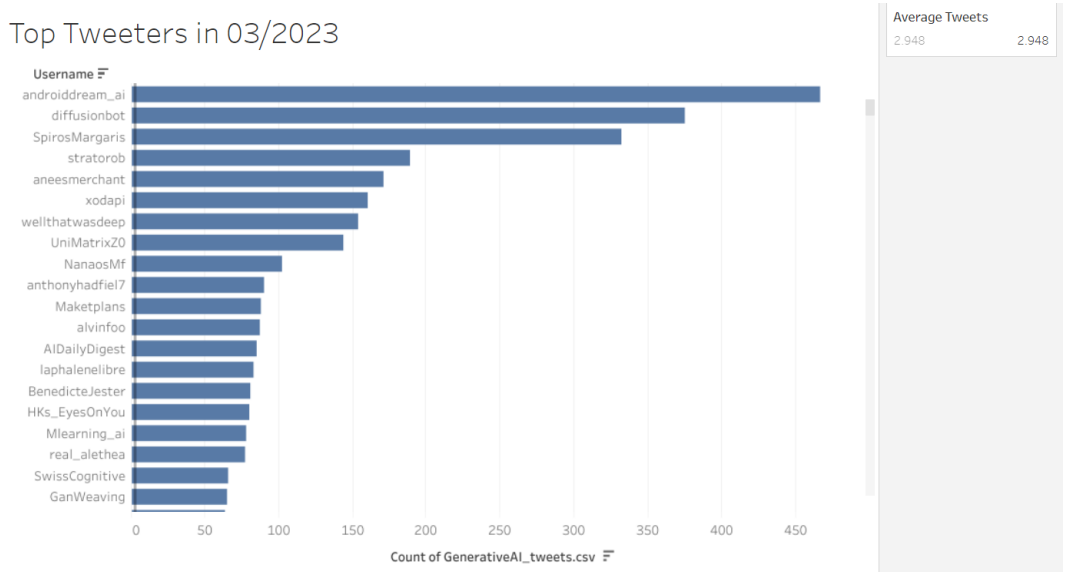


What's informative about this view:

This confirms the idea that March 2023 was when tweets about this topic became more frequent.

What could be improved about this view:

It would be interesting to know how many tweets the average user in our dataset put out in March 2023. Strange values here (too high) might lead to more exploratory analysis to see if some bots weren't removed properly in the data cleaning.

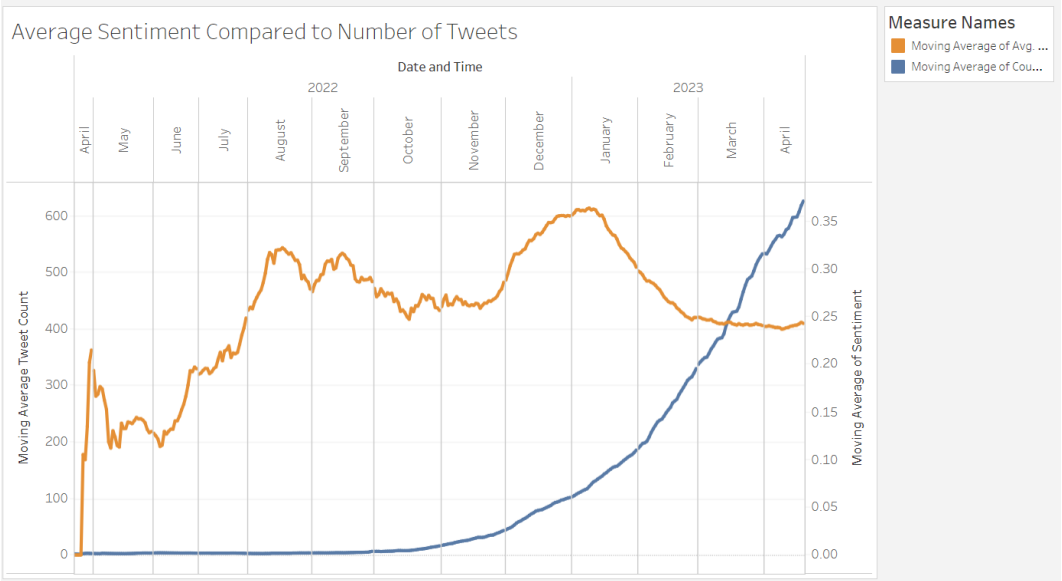


What's informative about this view:

This shows me that the average user included in the list was tweeting about this an average of 2-3 times during the busiest month of this dataset. That aligns with expectations and so now more confident exploration of this dataset can begin.

What could be improved about this view:

This provides the second piece to the puzzle, but now I would like to combine the average number of tweets with the general sentiment of GenAI.



What's informative about this view:

The orange line is the moving average of the sentiment analysis of the tweets and corresponds to the axis on the right, the blue line is the moving average of the number of tweets and corresponds to the axis on the left. Right around March 2023 you can see something interesting happen; the number of tweets starts to increase quickly while the general sentiment plateaus.

This correlates ChatGPT's GPT4 release (3/2023).

What could be improved about this view:

I think this could lead us to the conclusion that views were starting to become more polarized with GPT4's release, even though the data plateaus. My thought is that with more positive and negative sentiments the average remains the same. It would be interesting to add some lines in the background that show the general range of sentiments of that month as well, to see if the ranges become larger over time.

Conclusion:

Though the original hypothesis is unconfirmed, this does lead to some interesting topics to be explored in the final project. More analysis on the ranges of sentiment over time would be useful here.