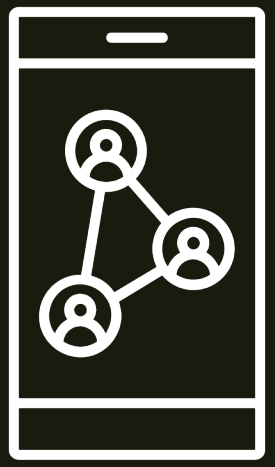


TWITTER SENTIMENT ANALYSIS

Project Background

Originally completed in Spring 2023 in Data Mining class

Worked with two other group members

Objectives

- Complete text wrangling of twitter data (provided by professor)
- Construct Naïve Bayesian model that would predict whether the tweets have negative or positive sentiment
- Predict on a pre-made test set submitted to a Kaggle competition

Initial Results

#	△	Team	Score
1	—	Team 5	0.74058
2	—	Team 2	0.74051
3	▲ 1	Team 4	0.73444
4	▼ 1	**Team 3**	0.73284

- Applied text wrangling
- Training accuracy: 0.746
- Testing accuracy: 0.733
- We placed fourth in the Kaggle competition ☹

Project Enhancements

Re-approach text wrangling of the data

Alter text wrangling sequence
Apply alternate Regex functions

Adjust Naïve Bayes algorithm to increase accuracy of the trained model

Apply findings to *ChatGPT* Tweet datasets

Data Cleaning: Text Wrangling Sequence

"@Anjeebaby it's on the stove for u, coming up http://twitpic.com/4h20u"

Modeling: Naïve Bayes Algorithm Re-Evaluation

Lower Frequency Bound	Accuracy Without Laplace	Accuracy With Laplace
25	0.742	0.745
20	0.744	0.747
15	0.744	0.748
10	0.746	0.752
5	0.746	0.756
4	0.746	0.758

- Conclusion: the lower the frequency bound, the higher the accuracy
  - Why? Naïve Bayes algorithm design
- Drawback: increased processing time and computational power…

```
> finalTrain <- apply(trainDTMFreq, MARGIN = 2, convertCounts)
Error: vector memory exhausted (limit reached?)
```

Why I Decided to Enhance My Enhancements

- Lots of unknowns:
  - *Where did this data come from?*
  - *How was the sentiment determined?*
  - *When are these tweets from?*
- ***I don't know the "answers" to the test dataset from the Kaggle competition, so there is nothing to confirm if the model improves or not in application!***
- More extensive text analysis piqued my interest

Expanded Plans

Use ChatGPT twitter data from two Kaggle sources

One dataset extracted using Twitter API from November 30th to December 31, 2022[6]

One dataset extracted using Python package "snsscrape" from January to March 2023[7]

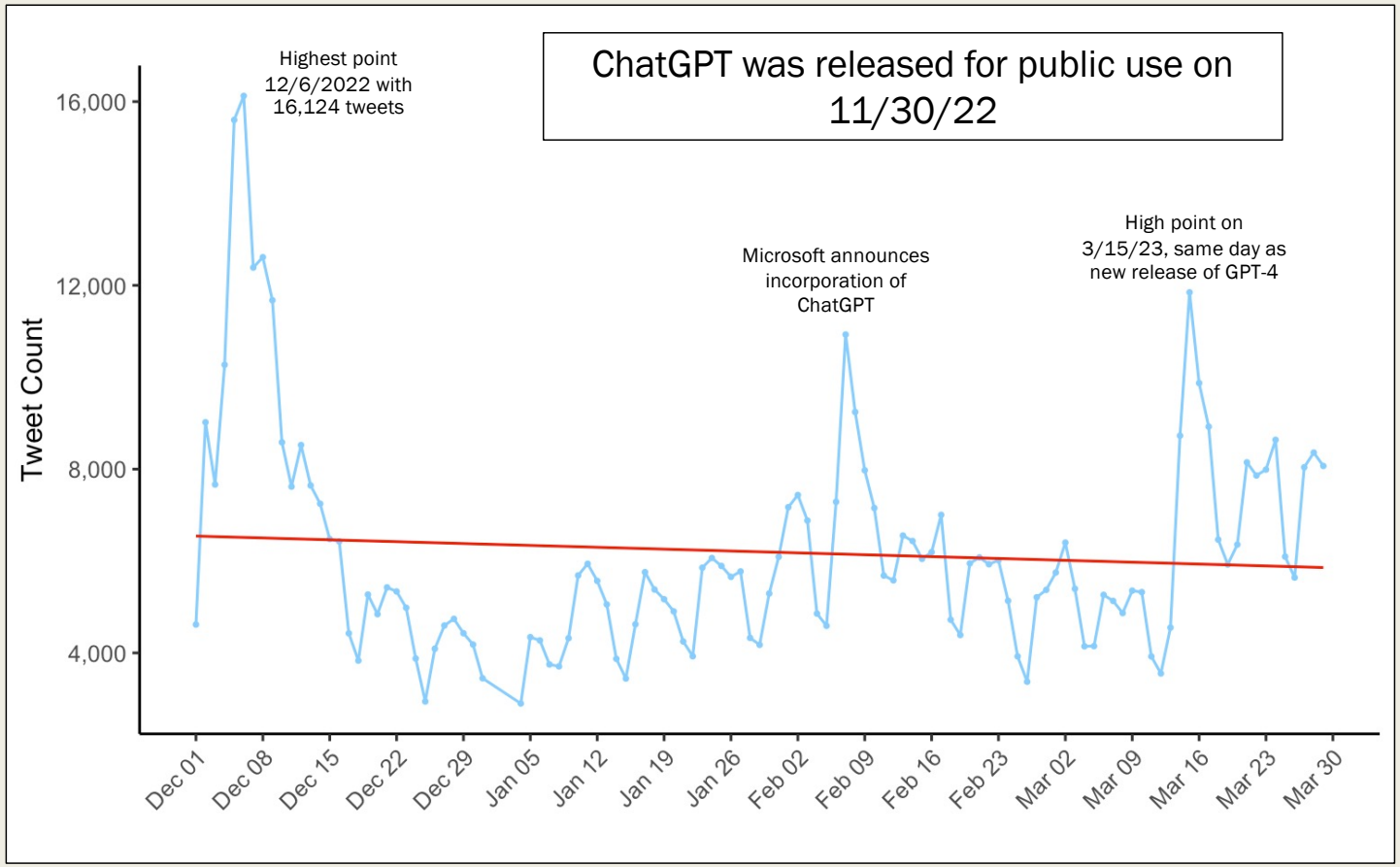
Apply refined text wrangling sequence from previous project

Visualize tweet rate over time and word frequency

Compare sentiment/emotion of tweets about ChatGPT from December 2022 to March 2023

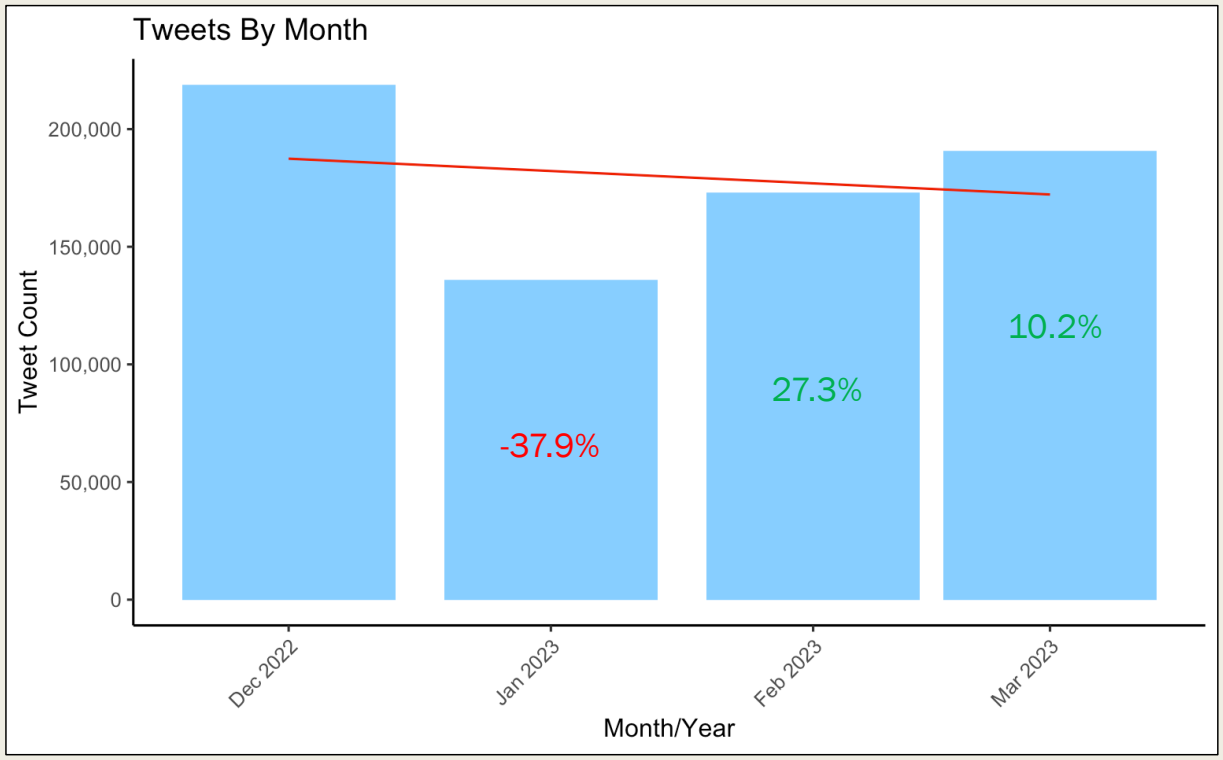
[6] https://www.kaggle.com/datasets/pcminh0505/chatgpt-twitter
[7] https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023, https://medium.com/@ka2612/effortlessly-scraping-massive-twitter-data-with-snscrape-a-guide-to-scraping-1000-000-tweets-in-d01c38e82d18

Data Exploration: Tweet Rate by Day

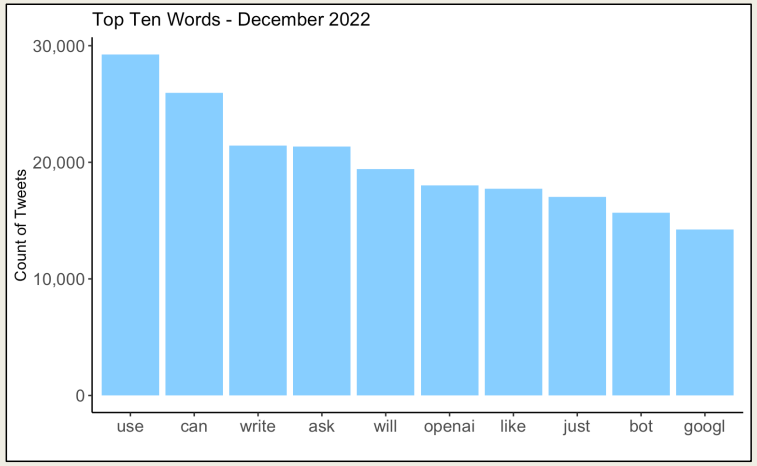


[8] https://techcrunch.com/2023/07/13/chatgpt-everything-you-need-to-know-about-the-open-ai-powered-chatbot/#:~:text=November%2030%2C%202022%20is%20when%20ChatGPT%20was%20released%20for%20public%20use.
[9] https://www.cnbc.com/2023/02/07/microsoft-open-ai-chatgpt-event-2023-live-updates.html
[10] https://www.axios.com/2023/03/15/gpt4-openai-chatgpt-new-version

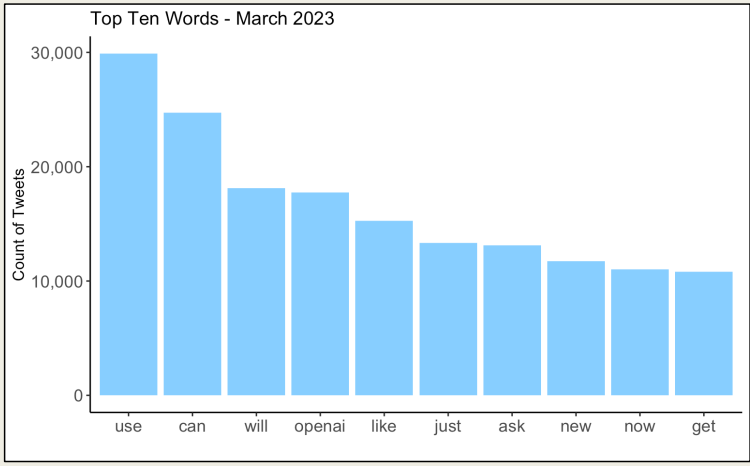
Data Exploration: Tweet Rate by Month

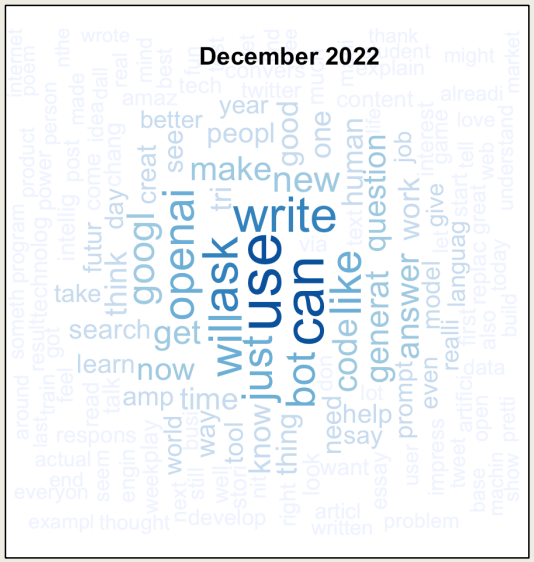


High volume on release, followed by sharp decline and steady increase thereafter

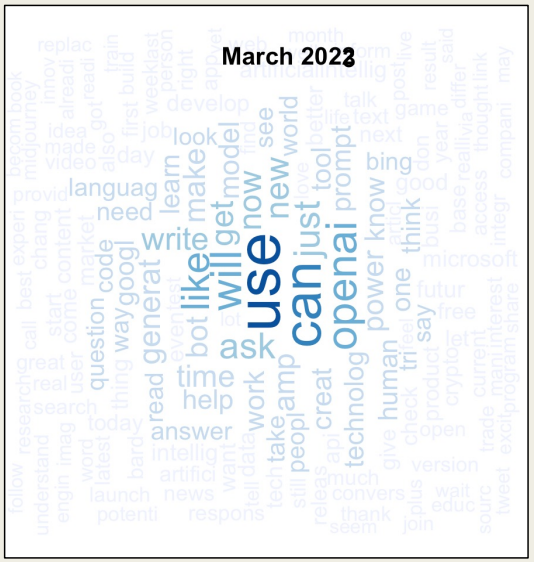


Top 10 Words

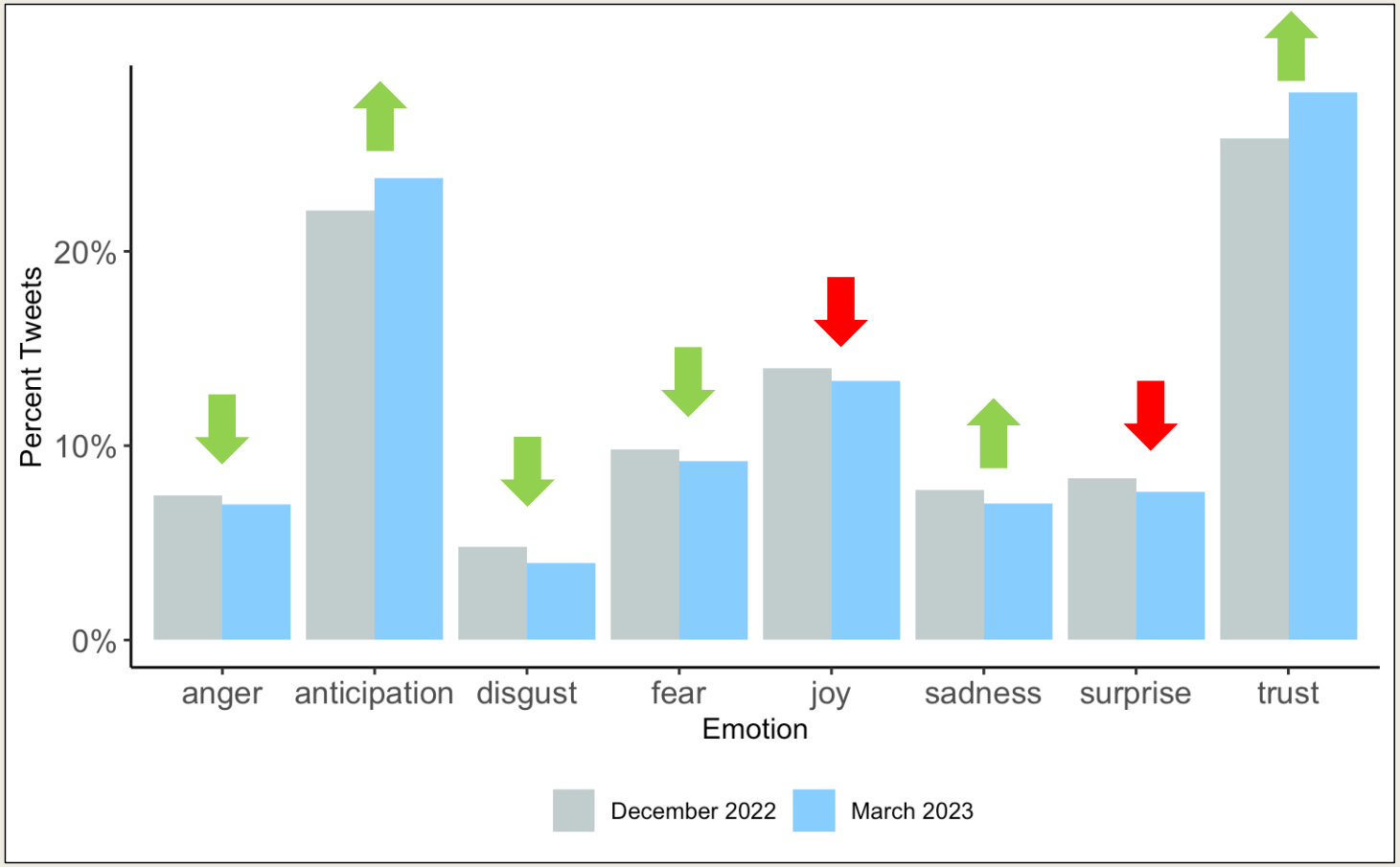




Word Cloud

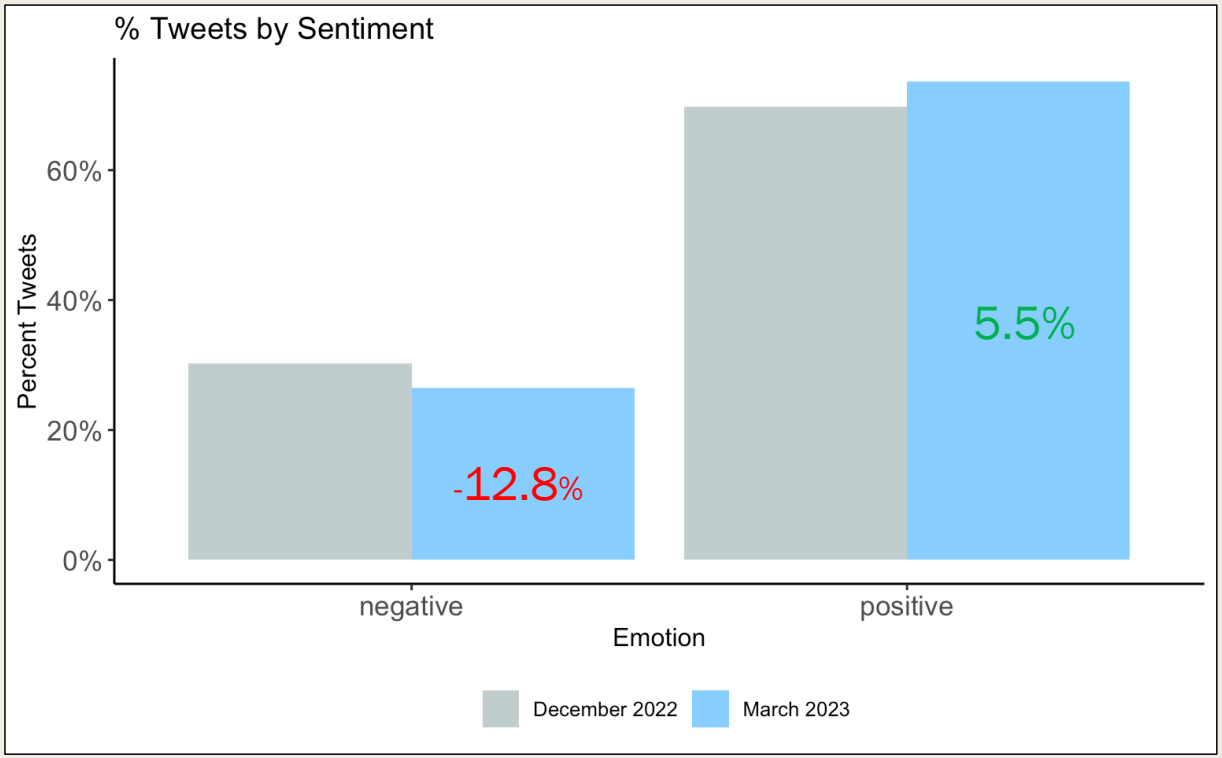


Comparing Emotion: R Package *syuzhet*[11]



[11] https://cran.r-project.org/web/packages/syuzhet/index.html

Comparing Sentiment: R Package *syuzhet*



Good news for *OpenAI*!

Challenges and Further Work

Twitter API is no longer available for free use[12]

Dataset is large and requires long processing time for sentiment

Text wrangling is not perfect

So much room for improvement and extra analysis!

[12] https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected/?sh=2fb5f6762664