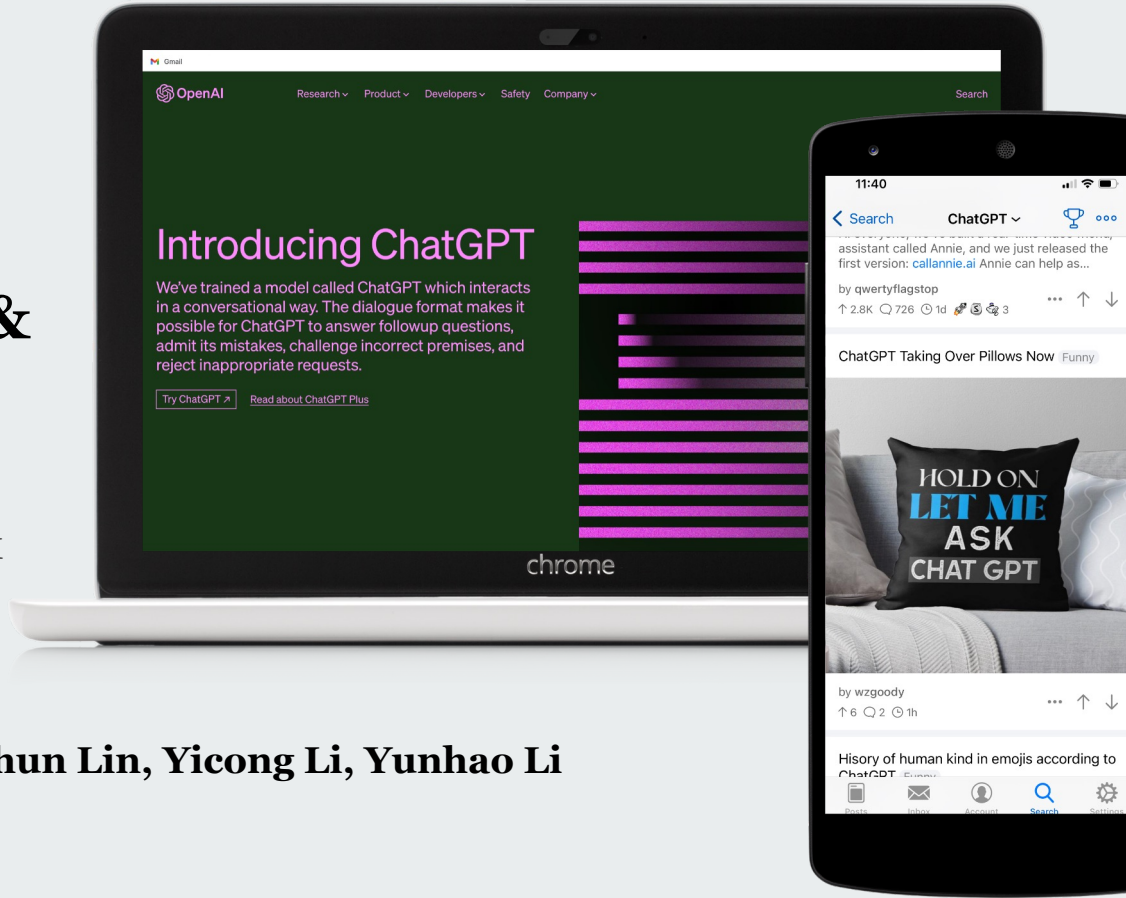
ChatGPT on Reddit:

Unveiling Sentiments & Buzzing Topics

MSDS 597 Data Wrangling & Husbandry Final Project

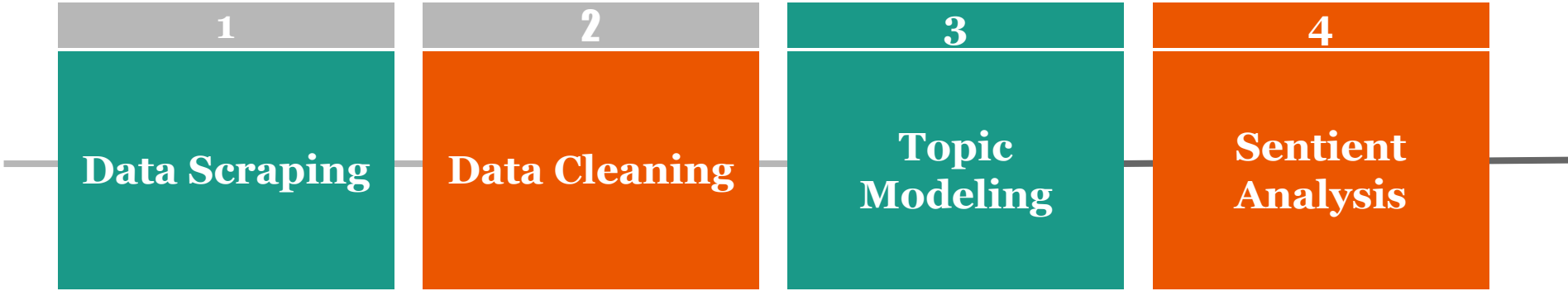


Team members: Siheng Huang, Jiechun Lin, Yicong Li, Yunhao Li

Questions We're Interested in

- How does **public attitude** toward ChatGPT evolve over time? (Sentiment analysis)
- What are the users' **primary concerns** with ChatGPT? (Topic modeling)
- **What awaits human beings** as the AI revolution unfolds?

Outline


1
Data Scraping
2
Data Cleaning
3
Topic Modeling
4
Sentient Analysis

Packages Used:

library(RedditExtractoR)

library(dplyr)

library(tidyverse)

library(tidytext)

library(textclean)

library(lubridate)

library(topicmodels) library(tm)

library(textstem)

library(textclean)

library(sentimentr)

ChatGPT's Latest Big Events

Nov 31

Initial Launch

ChatGPT is born!

Feb 1

ChatGPT Plus

Faster response times and priority access

Feb -Mar

Competitors emerge

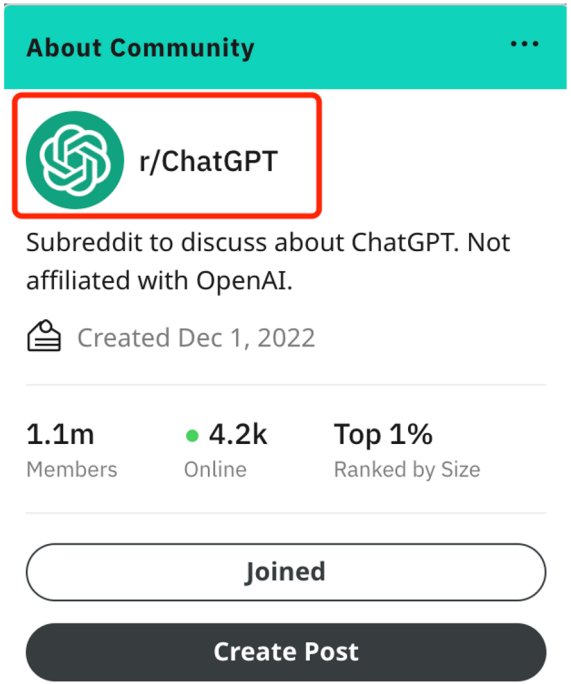
Microsoft Bing and Google Bard are available

Mar 14

GPT-4 released

It can parse both text and image input

Data Scraping

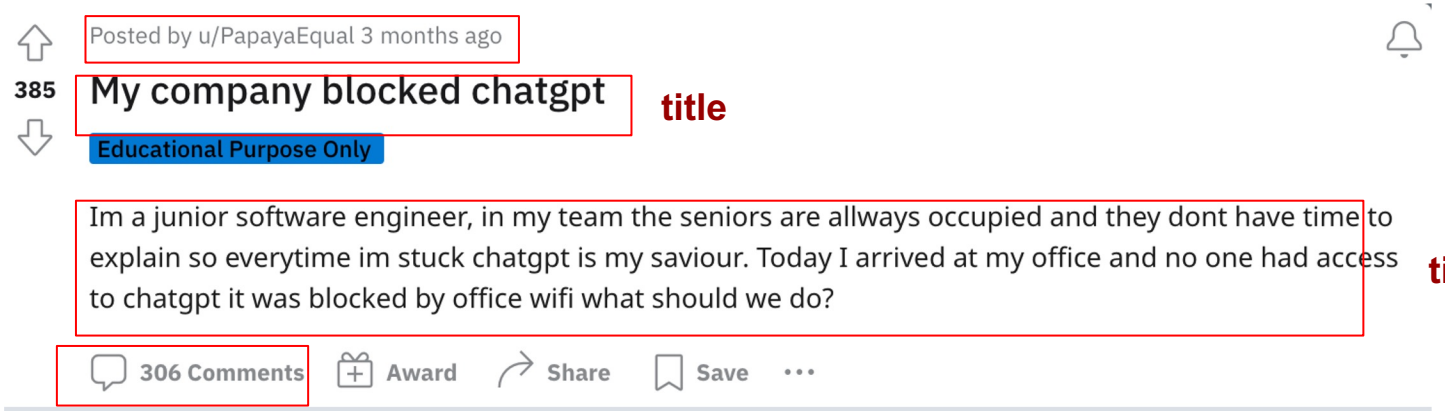


- **Data Source:**
- Scrape **4 months'** (Dec 3 2022-April 12, 2023) worth of data from the Subreddits “ChatGPT” using *RedditExtractoR*.
- A total of **152,098 comments** were collected and sorted in **chronological order** based on the date each comment was created.

Data Glimpse

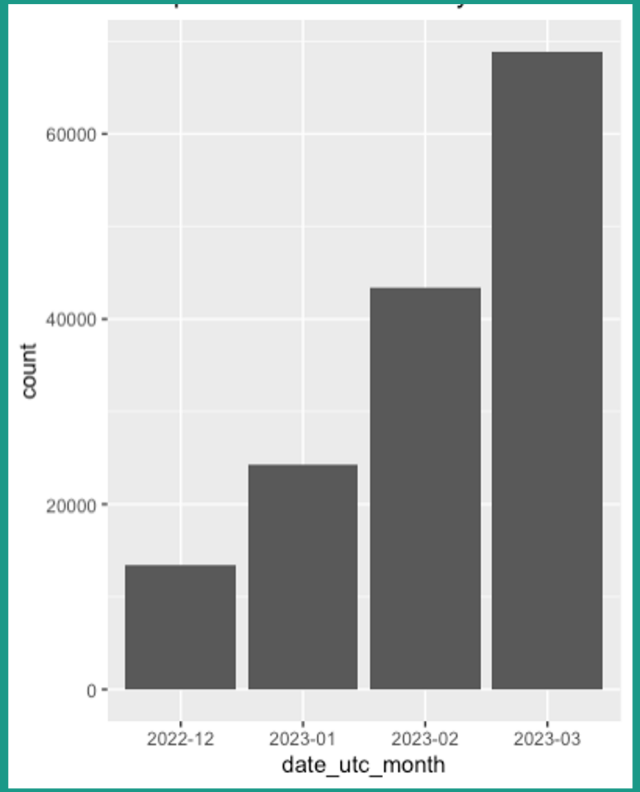
- 861 unique links
- 860 unique title (title “accurate” has two threads)
- 145 unique title_text(many thread posters left the title text blank.)

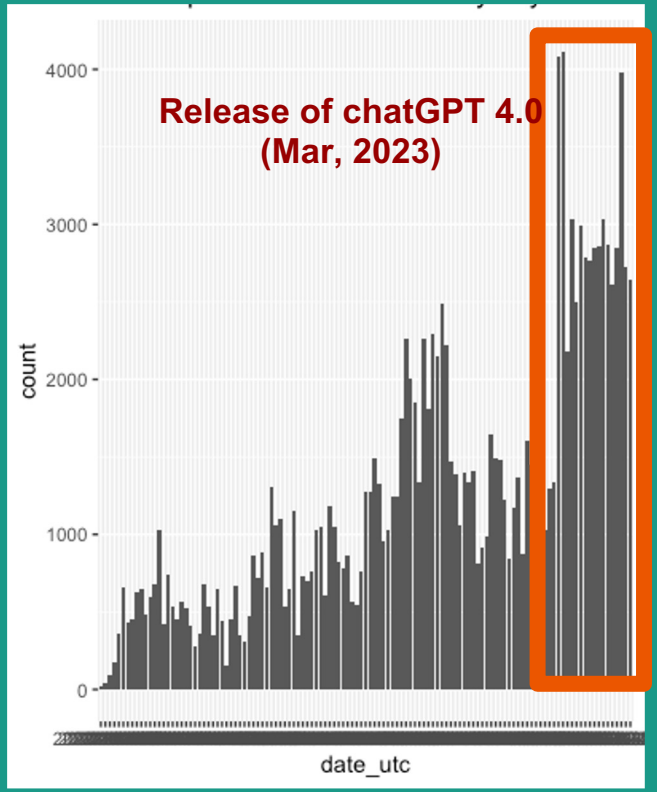
date_utc	url	title	title_text	comment
2023-01-06	https://www.xxx	xxx	xxx	xxx
…	…	…	…	…
152, 098 comments				



Comment: show all the comments

Data Summary - Trend of Comment Volume Over Time





- As we can see from the graph, with the release of chatGPT, the sentiment heat has been rising, especially when it comes to the release of version 4.0, opinion volume has skyrocketed

Top 20 Topics(ranked by comment counts) in the Past 4 Months

[1] "Got access to Bing AI. Here's a list of its rules and limitations. AMA"
[2] "GPT-4 AMA"
[3] "GPT-4 Day 1. Here's what is already happening"
[4] "Google releases ChatGPT competitor, Bard-NYT"
[5] "GPT-4 message limit changed to 25 every 3 hours with further reduced cap coming next week"
[6] "Microsoft lays off its entire AI Ethics and Society team"
[7] "Okay yeah now I am threatened"
[8] "GPT-4 released"
[9] "Bing gets jealous of second Bing and has a meltdown begging me not to leave or offer a chance at humanity to other Bing"
[10] "ChatGPT now supports plugins!!"
[11] "So many people do not realise how huge this is"
[12] "Sorry, You do not Actually Know the Pain is Fake"
[13] "Most Influential People of All Time (According to Chat GPT)"
[14] "Why are not governments afraid that AI will create massive unemployment?"
[15] "Chatgpt Plugins Week 1. GPT-4 Week 2. Another absolutely insane week in AI. One of the biggest advancements in human history"
[16] "Elon Musk calling for 6 month pause in AI Development"
[17] "I think those saying AI will not take their jobs are missing something really important."
[18] "List of Jobs AI will replace (as per chatgpt)"
[19] "The problem with the \"there will be new jobs to replace the old ones\" argument is..."
[20] "Are you automating any life or work tasks with ChatGPT, if so, which ones?"

ChatGPT 4.0
Jobs
Influence to life
Competitors

Data Cleaning: Data Structure Issue

Normalized then tokenized, one title_text may maps to multiple comments, some titles do not have title_text

Normalized then tokenized

date_utc	url	title	title_text	comment
2023-01-06	https://www.xxx	xxx	xxx	xxx

Normalized to format “yyyymmdd”

Used as reference for unique title text, then removed to save space

Normalized

title_text_id
1

Added, unique id for each distinct title_text, used to save space

comment_id
1

Added, unique id for each distinct comment, used to save space

Data Cleaning: Detailed Issues in Comments

Three main issues:

- Control characters:
  - handle them by manual replacement. ("\031" replaced as """, other control characters usually does not have specific meaning, thus removed directly.)
- HTML markup, white spaces, & symbols:
  - Auto-transformed by package.
- Informal writings (e.g., contractions, emojis, internet slang, kerning):
  - Auto-transformed by package.

Normalization Solution:
textclean package
Document Reference:
https://www.rdocumentation.org/packages/textclean/

replace_white()	Replace white space (e.g. "\n" -> " ")
replace_contraction()	Replace english contractions (e.g. “I’m” -> “im”)
replace_kern()	Replace kerning (e.g. "W O R L D" -> "WORLD")
replace_word_elongation()	Replace word elongation (e.g. "goooood" -> "good")
replace_html()	Replace HTML (e.g. ">" -> ">")
replace_emoji()	Replace Emoji (e.g. “ðŸ“¦” -> “package”)
replace_emoticon()	Replace Emoticons (e.g. “:)” -> “smiley”) (currently not functional)
replace_symbol()	Replace symbols (e.g. "@" -> at")
replace_internet_slang()	Replace internet Slang (e.g. "NP" -> "no problem")
replace_ordinal()	Replace ordinal Numbers (e.g. "1st" -> "first")

Data Cleaning: Before & After

Original Text	Normalized Text
I mean, you\031re not wrong, but you don't seem to be aware that Google\031s own large language model ;-)	I mean, you are not wrong, but you do not seem to be aware that Google's own large language model ;-)
\n\n A lot of people know this probably but because [@ this thread](https://www.reddit.com/r/ChatGPT/comments/119j7u5/why_how_noble_of_you/)	A lot of people know this probably but because at this thread
NP, I'm happy to pay it as long as it's a fair price. $5 a month is fair.	no problem, I am happy to pay it as long as it is a fair price. dollar 5 a month is fair.
They've done nothing and their stock price has been declining for a year. \n\nEvery trader I know is shorting G O O G.	they have done nothing and their stock price has been declining for a year. Every trader I know is shorting GOOG.
>It can absolutely not do 99% of what Google can do yet. \n\nWrong.\n\nProtip: Ask ChatGPT to behave like Google and it will.	>It can absolutely not do 99 percent of what Google can do yet. Wrong. Protip: Ask ChatGPT to behave like Google and it will.
It doesn\031t have access to any \034information\035.	It does not have access to any information.
I dunno, I can see a huge portion of current chatGPT users going back to google when openai starts charging for access.	I dunno, I can see a huge portion of current chatGPT users going back to google when openai starts charging for access.
100% agree, lol	100 percent agree, laughing out loud

Tokenization & Anti-join Stopwords

Unigrams

Description: df [10 × 2]

	word_comment <chr>	n <int>
1	ai	1094641
2	chatgpt	905425
3	people	610469
4	gpt	470682
5	prompt	434990
6	human	392827
7	time	375572
8	dan	268803
9	model	261837
10	write	248174

1-10 of 10 rows

Bigrams

Description: df [10 × 2]

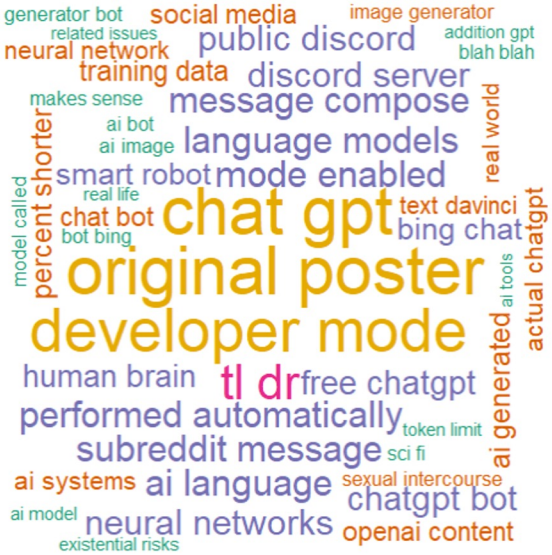
	bigram_comment <chr>	n <int>
1	language model	29172
2	original poster	21214
3	chat gpt	20829
4	developer mode	19176
5	tl dr	14571
6	mode enabled	10074
7	language models	9781
8	performed automatically	9688
9	message compose	9608
10	subreddit message	9608

1-10 of 10 rows

Word Cloud

Unigram Word Cloud

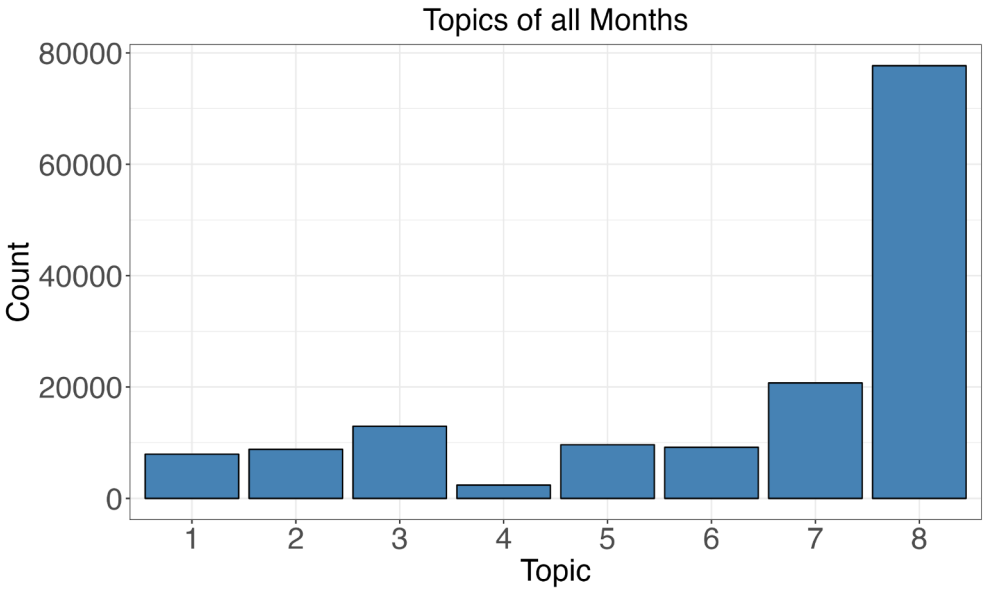
Bigrams Word Cloud



Topic Modeling

- Combines *titles* and *title_text* into a single column.
- Tokenizes (splits) the text into individual words.
- Removes common stopwords (e.g., "the", "and", "is").
- Lemmatizes words, converting them to their base forms.
- Creates a document-term matrix to represent the text data.
- Defines a function that uses Latent Dirichlet Allocation (LDA) for topic modeling.
- Applies the function to assign the most probable topic to each title.

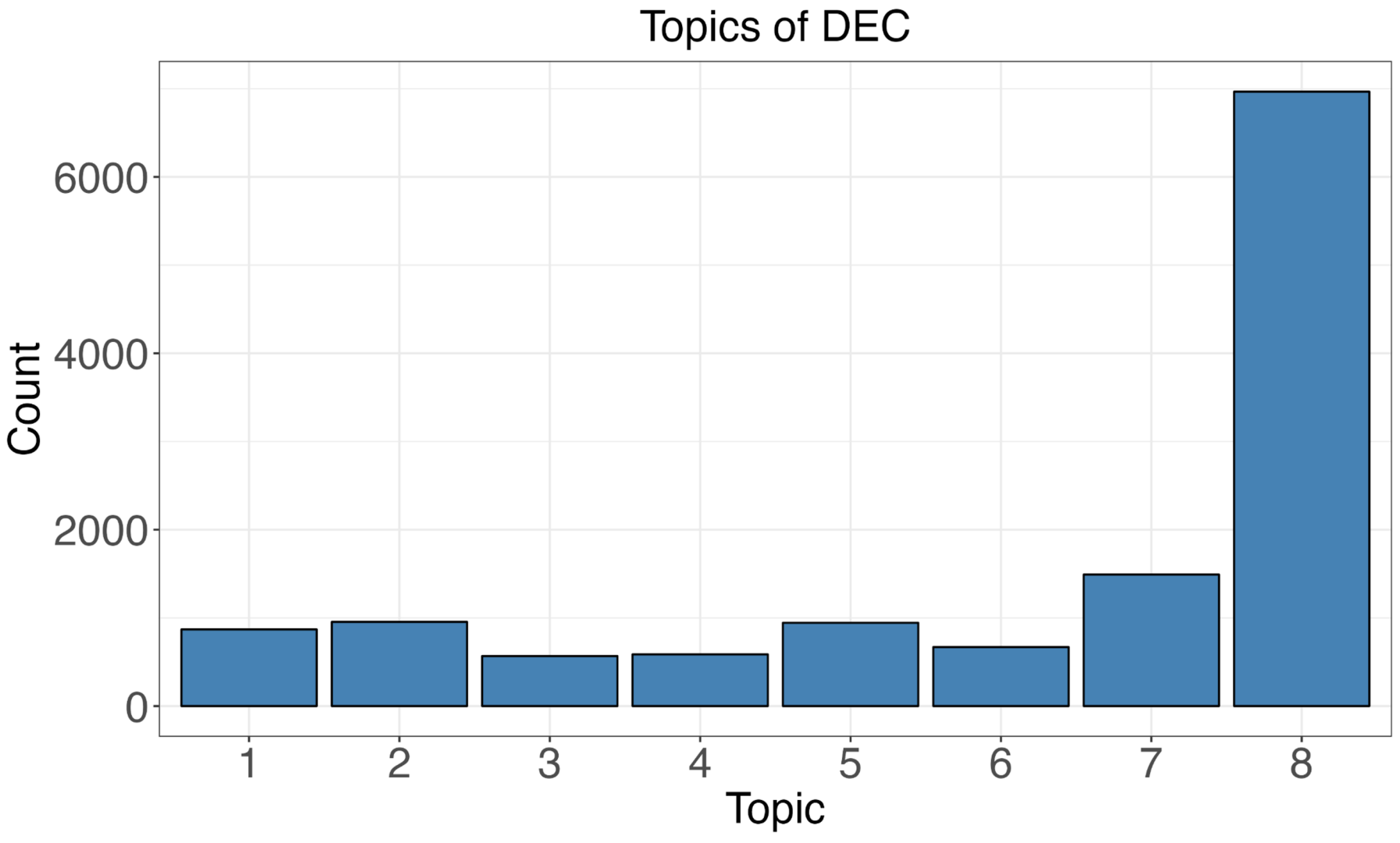
Topic Definition

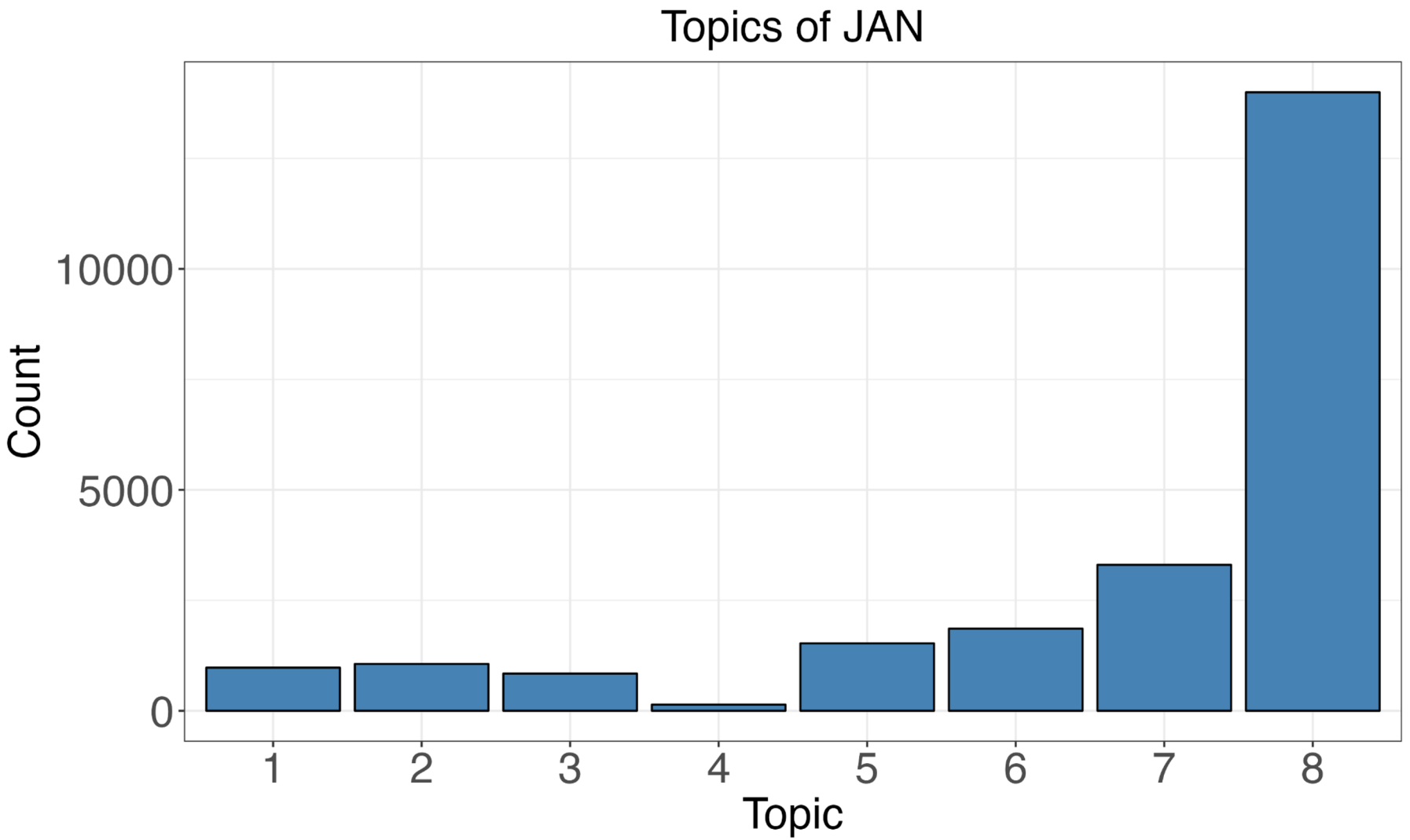


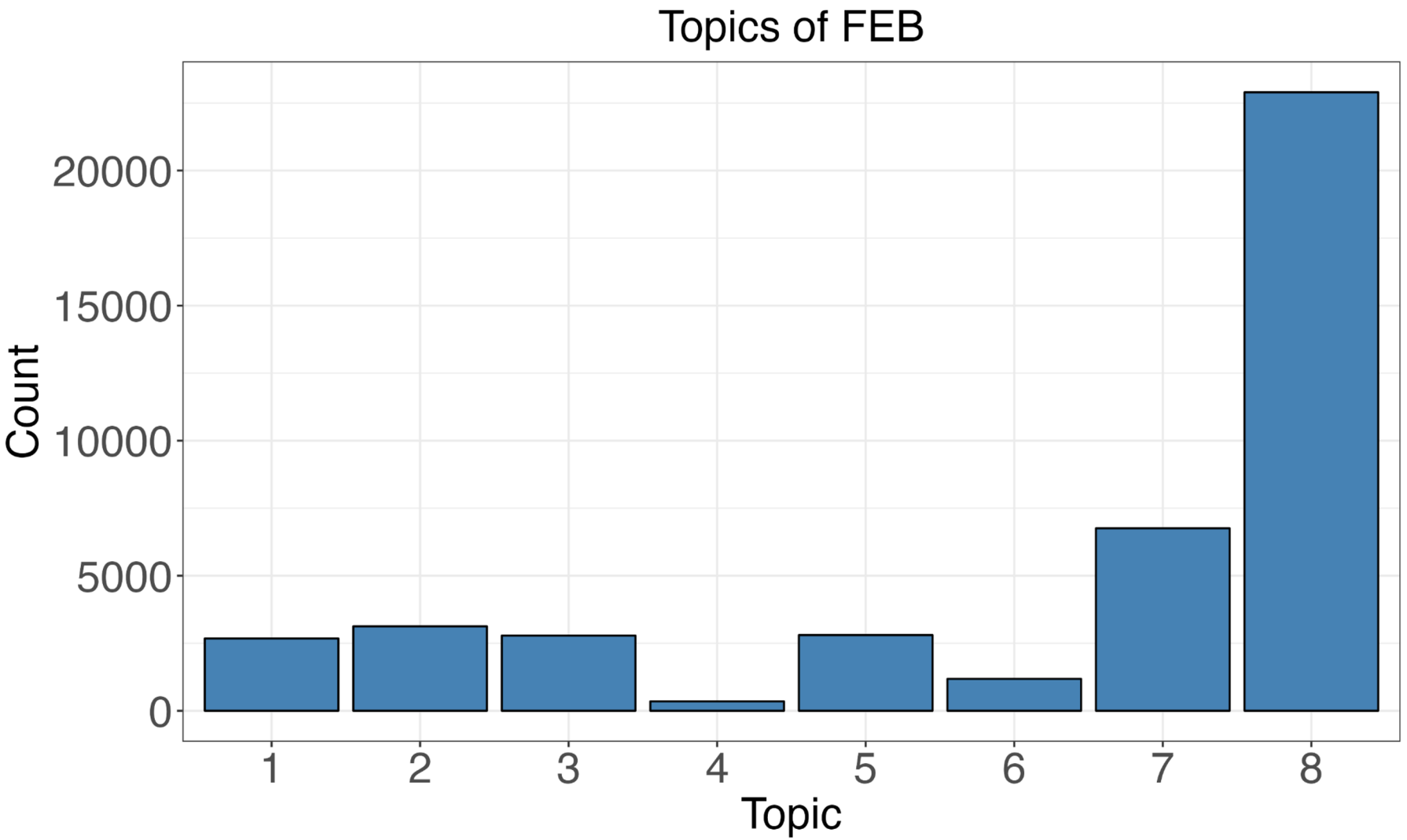
- Based on the topic classification, we grouped the data by topics and filtered the unique titles and title text for each topic. After a thorough examination, the 8 topics have been defined as follows:
- Topic 1: Chatbots and effective prompts
- Topic 2: Safety concerns and ethical implications
- Topic 3: GPT-4
- Topic 4: Creativity of ChatGPT
- Topic 5: Educational potential and controversies
- Topic 6: Accuracy
- Topic 7: Impact of AI on society, work, and various industries
- Topic 8: Remaining topics

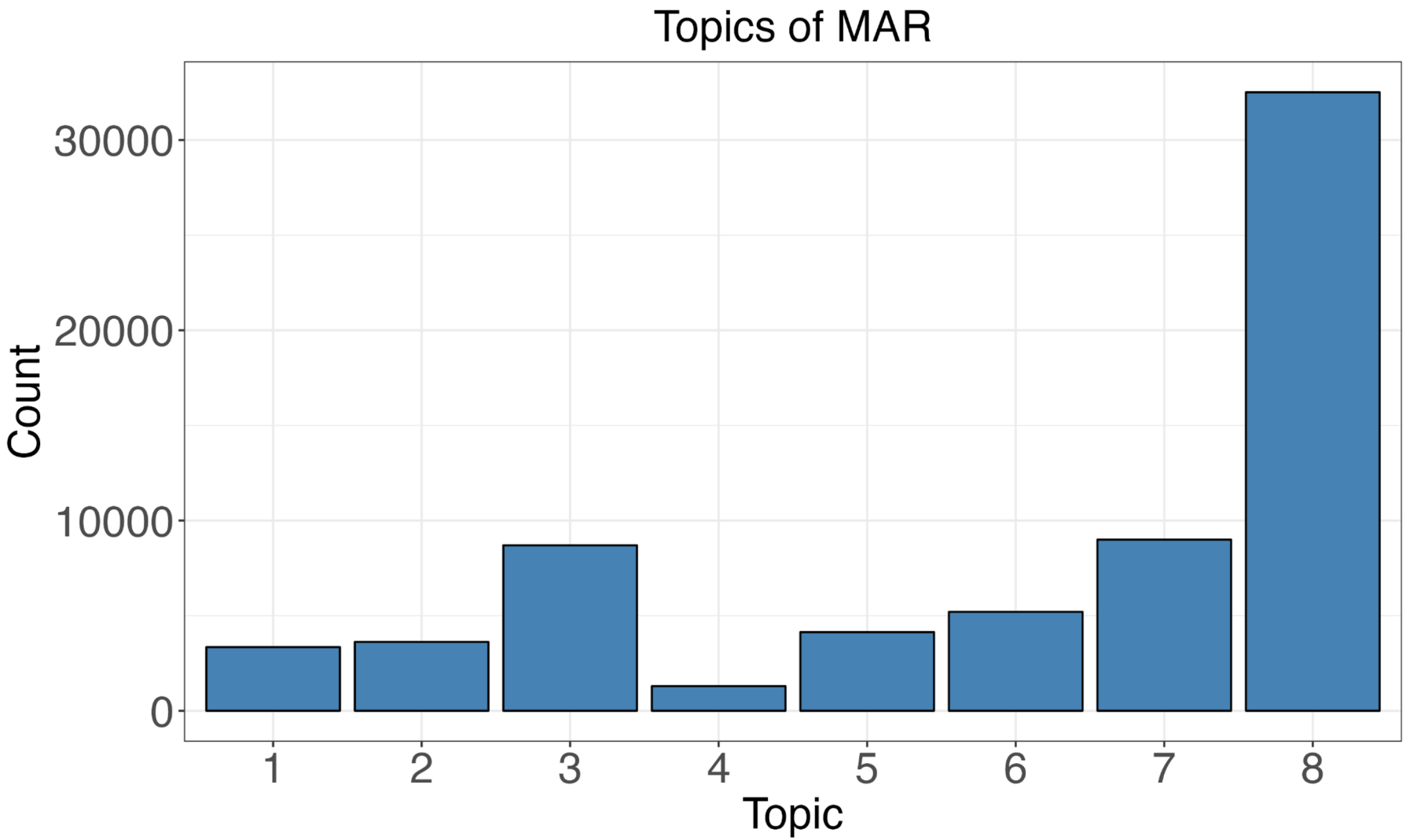
Top 10 Words of each Topic

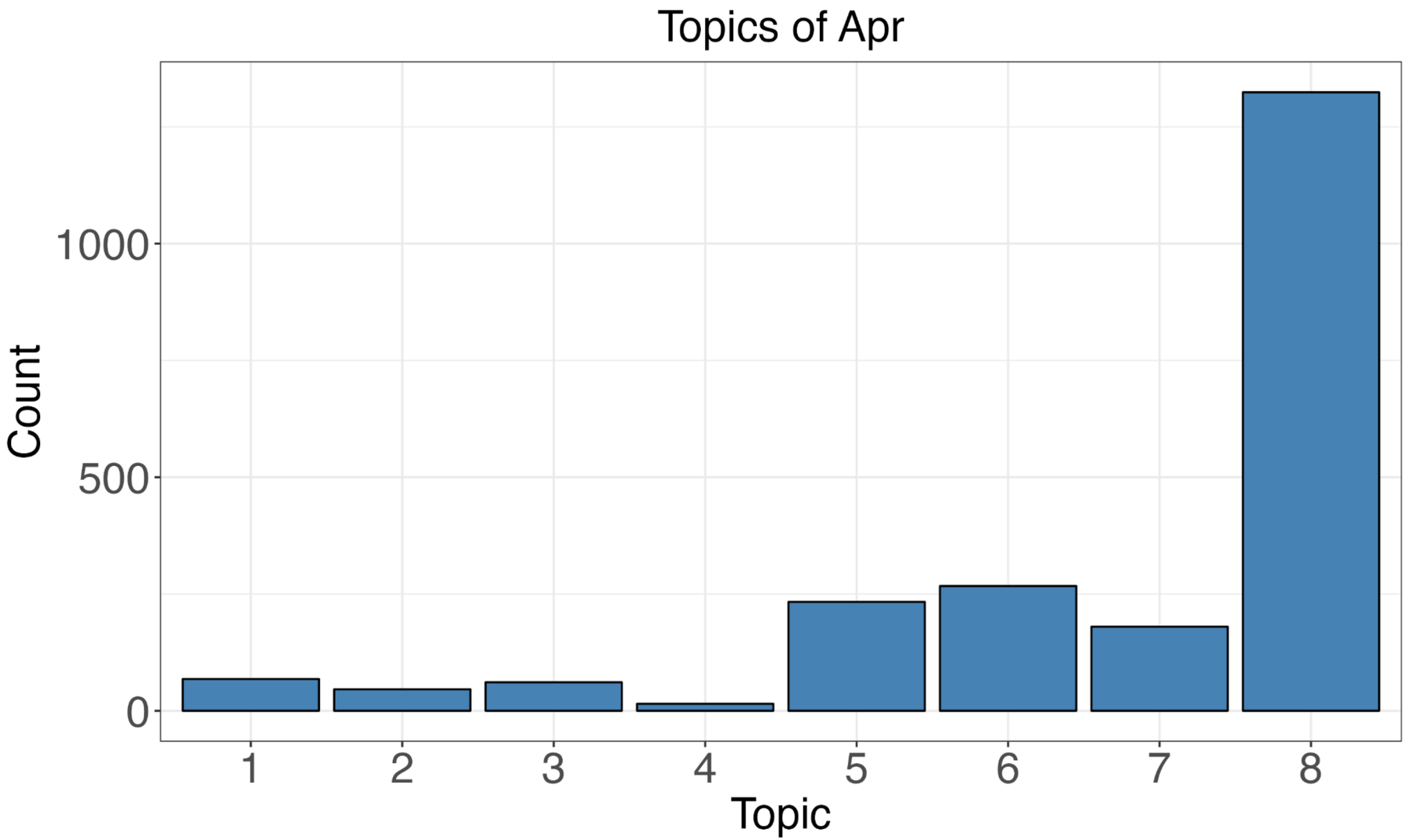
Topic 1	**chatgpt**	**ai**	**prompt**	**model**	**language**	**write**	**gpt**	**tool**	**learn**	**job**
Topic 2	**dan**	**prompt**	**jailbreak**	**chatgpt**	**edit**	**token**	**release**	**version**	**click**	**system**
Topic 3	**chapter**	**4**	**scene**	**gpt**	**outline**	**write**	**book**	**detail**	**level**	**note**
Topic 4	**chatgpt**	**write**	**ai**	**teacher**	**question**	**conversation**	**time**	**message**	**gpt**	**text**
Topic 5	**chatgpt**	**ai**	**emotion**	**job**	**people**	**prompt**	**gpt**	**change**	**laugh**	**blow**
Topic 6	**prompt**	**answer**	**chatgpt**	**response**	**chat**	**reply**	**content**	**jailbreak**	**ai**	**mode**
Topic 7	**link**	**ai**	**gpt**	**build**	**create**	**4**	**code**	**people**	**crazy**	**release**
Topic 8	**bing**	**ai**	**human**	**people**	**chatgpt**	**model**	**brain**	**word**	**future**	**write**


Topics of DEC
Count
6000
4000
2000
0
1
2
3
4
5
6
7
8
Topic


Topics of JAN
Count
10000
5000
0
1
2
3
4
5
6
7
8
Topic


Topics of FEB
Count
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
Topic


Topics of MAR
30000
20000
10000
0
Count
1
2
3
4
5
6
7
8
Topic


Topics of Apr
Count
1000
500
0
1
2
3
4
5
6
7
8
Topic

Some Insights into Topic Changes Over Time

- Topic 8 is a mixture of various elements. Due to the complexity of human language, particularly in social media posts, this portion of the **topic remains difficult to determine**.
- The **topic changes remain relatively consistent with the timeline** of ChatGPT. For instance, in March when GPT-4 was announced, the comment heat increased compared to other topics.
- Over time, **more attention** has shifted to the last three topics, which relate **to educational potential, accuracy, and the impact of AI on society, work, and various industries**. This suggests that the "honeymoon" phase of AI has passed, and the public is now seriously considering how to utilize the potential of GAI like ChatGPT while remaining cautious about its impact.

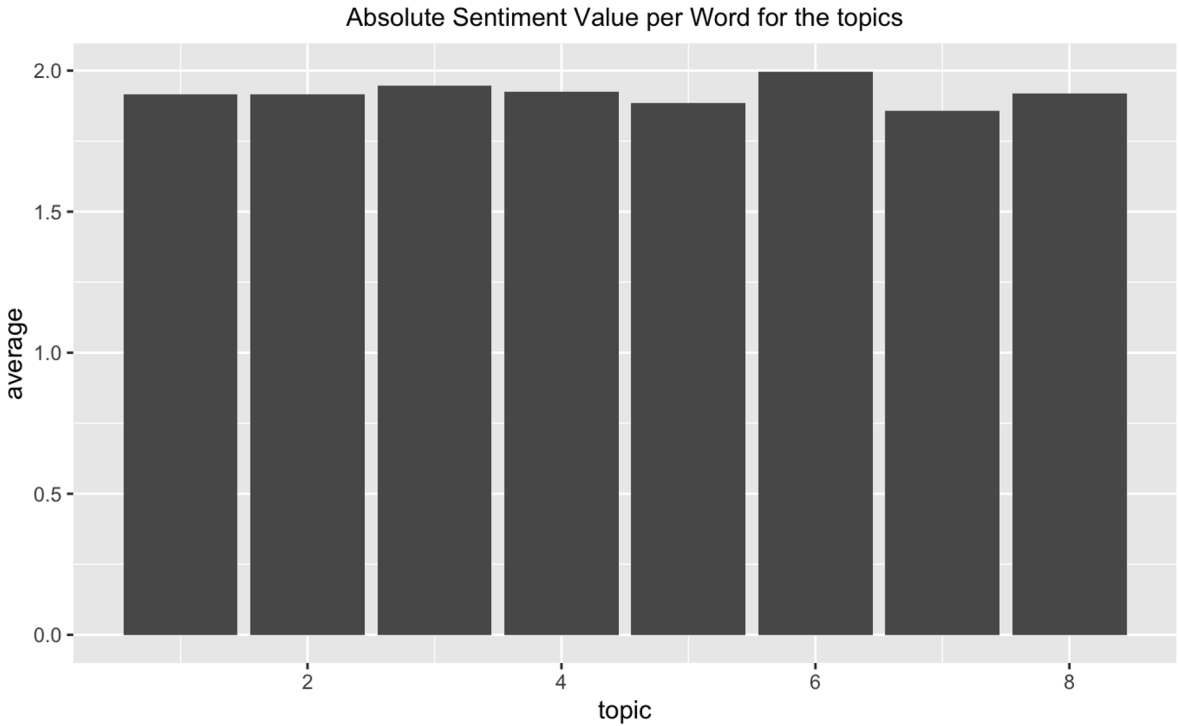
Sentiment Analysis

- ❏ Tokenize comments
- ❏ Apply lemmatization to the tokenized comments
- ❏ Utilize sentiment lexicons (e.g., Bing, Afinn) to determine the sentiment scores of each word
- ❏ Group sentiment scores by topics or by days

Sentiment Analysis by Topics

- Initially, we aimed to identify topics with strong emotions by analyzing the **average emotion score** of words **within each topic**. Unfortunately, as evident from the graph, the scores for each group are **virtually identical**.

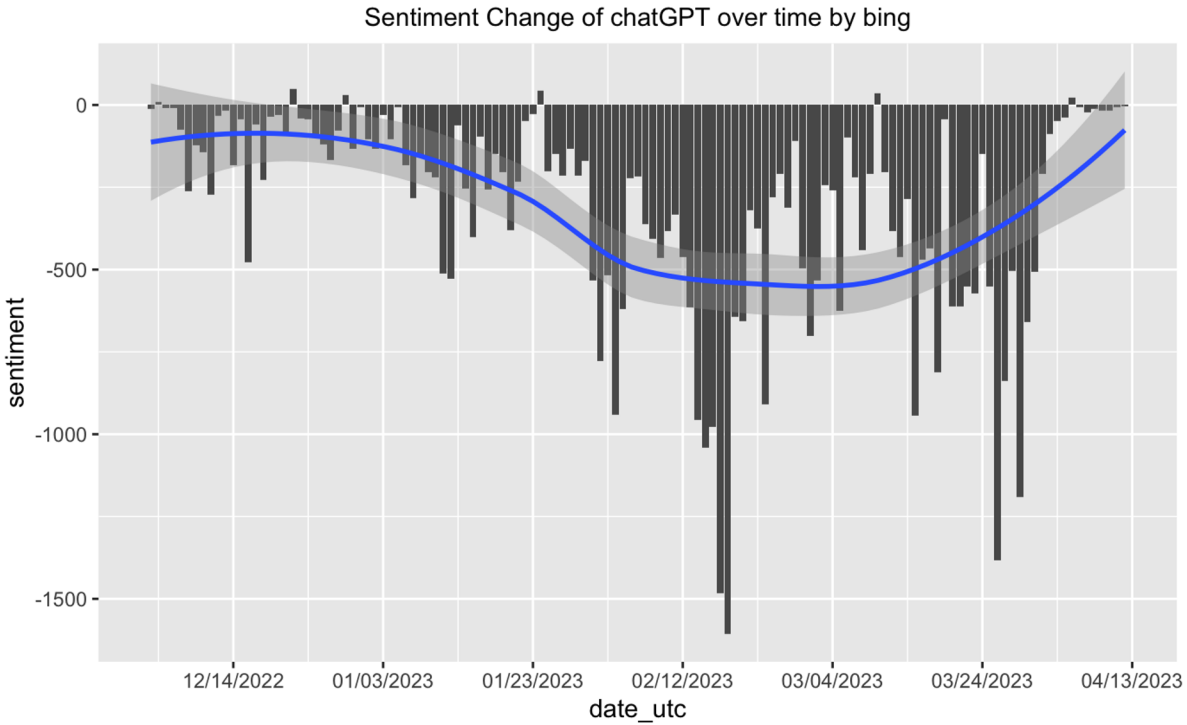
Bing



Limitations of LDA

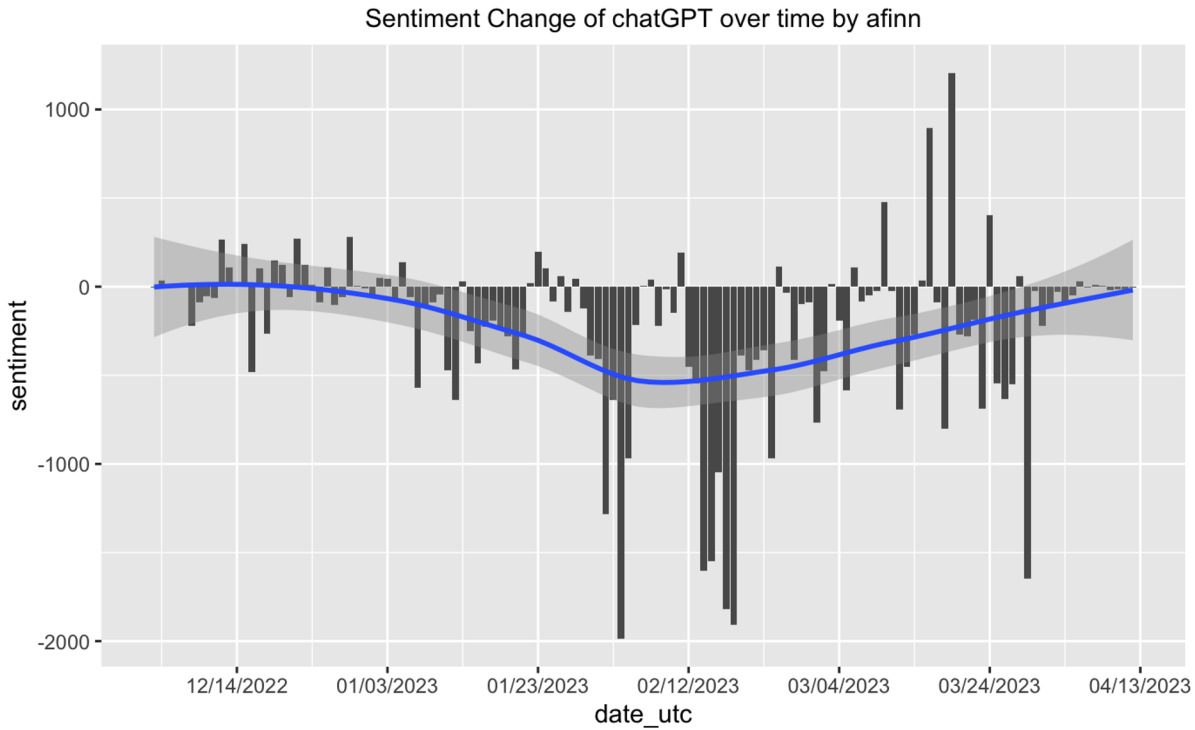
- **LDA is an unsupervised learning** technique, so the topic labels are not directly provided by the algorithm. It is **up to the analyst to interpret and label** the topics based on the top words and the context in which they appear.
- Therefore, relying solely on the algorithm to determine topics **might not be very informative or accurate**, as real-life situations are often more complex
- Moving forward, we will **disregard the topics generated by LDA** and instead perform **sentiment analysis** directly on the original comment texts.

Bing Sentiment over Time



- The results is predominantly negative. This contradicts common sense.
- Possible reasons: **Bing** only categorizes sentiment as **positive or negative**, which **might not be accurate** enough to **capture** the **nuances** in the text.

Afinn Sentiment over Time

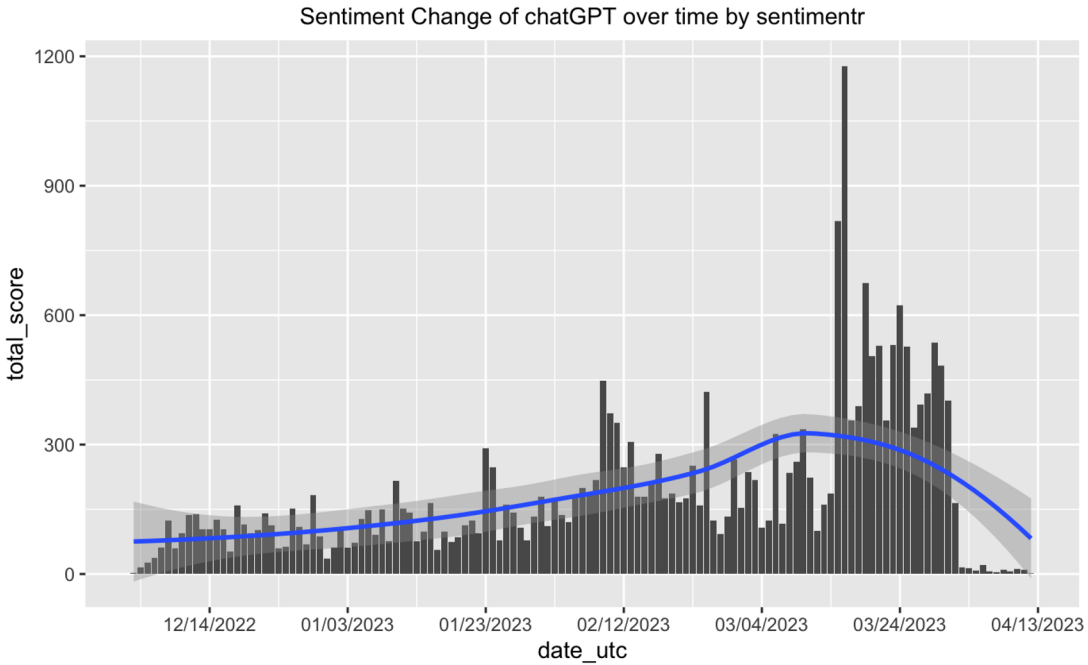


- Afinn **performs slightly better**, but negative sentiment still dominates.
- Possible reasons:

1. Comment texts are primarily **social media posts,** where people might tend to use **strong/exaggerated expressions**, which could potentially **skew sentiment** analysis results **towards negative** sentiment.

2. **Survivor bias**: People are more likely to comment when something doesn't work or when they have negative experiences.

Sentiment Analysis by Sentimentr



- From the plot above, we can find that the **sentiment** over all months is **positive**, and it reaches the **highest in mid-March**. The reason, we guess, is that **GPT-4 went live on March 14th.** The functionality of GPT-4 is a **significant improvement** over 3.5. Many reddit posts **praised** the new version.

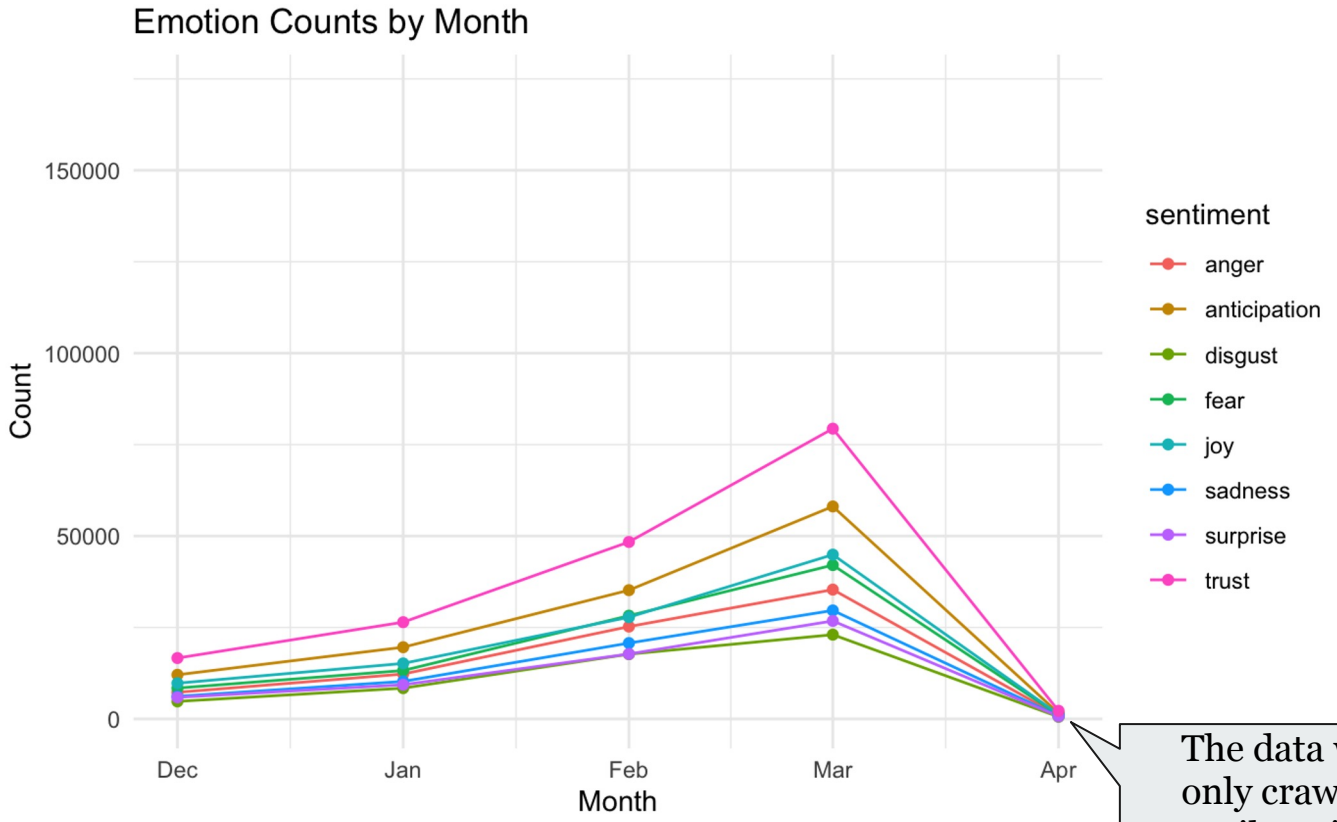
- The *sentimentr* package accounts for the shortcomings of the *Afinn* and *Bing*.
- It considers context by **analyzing sentences instead of individual words**, and it takes negations and intensifiers into account.
- Sentimentr **usually** produces **more accurate** results compared to Afinn and Bing regarding social media posts.

Sentiment Analysis by NRC

- We categorize words using the NRC lexicon, which classifies text into eight basic emotion categories by counting the occurrences of words associated with each emotion.

month <chr>	anger <dbl>	anticipation <dbl>	disgust <dbl>	fear <dbl>	joy <dbl>	sadness <dbl>	surprise <dbl>	trust <dbl>
December	7272	12066	4788	8424	9762	6166	5893	16643
January	12269	19606	8410	13236	15175	10249	9345	26452
February	25250	35201	17706	28224	27808	20735	17760	48388
March	35358	58095	23037	42062	44916	29663	26793	79360
April	937	1562	577	1111	1268	800	718	2215
Total	81086	126530	54518	93057	98929	67613	60509	173058

Sentiment Analysis by NRC



- As evident from the figure, "**trust**"and "**anticipation**" have consistently been the **top two** emotions across all five months. This generally positive trend aligns with our expectations.

The data was only crawled until April 12th

Top 10 Threads in Jan

title <chr>
Subscription option has appeared but it does not say if it will be as censored as the free version or not and
Is this all we are?
What lesser known but amazing functionality of CHATGPT are you willing to share?
What are your thoughts on ChatGPT being monetized soon
ChatGPT can not code anymore. Dec 2 vs today.
Can no longer use CHAT GPT. I was bored and asked how Walter White cooks meth and now I no longer can use it=<
I am quitting chatgpt
With ChatGPT and MidJourney I was able to write, edit, illustrate, and publish a 3 paged book in 10 days! (See comm
My School on kids using ChatGPT
It used to be so much better at release

Top 10 Threads in Feb

title <chr>
Got access to Bing AI. Here's a list of its rules and limitations. AMA
Sorry, You do not Actually Know the Pain is Fake
Bing gets jealous of second Bing and has a meltdown begging me not to leave or offer a chance at humanity to oth
Has the novelty worn off for anyone else?
ok - I have played with chatGPT for a week now. it was fun but the novelty is gone now. do not feel like playing an
Jobs Erased by AI
How to make chatgpt block you
chatgpt is 100 percent not conscious, and it does 100 percent not have feelings
After spending time on this sub, I am convinced that AI cults will be a major problem going forward.
So the bot straight up refused to answer my prompt because it was "too boring and simple" laughing my ass off

Top 10 Threads in Mar

title <chr>
GPT-4 AMA
GPT-4 Day 1. Here's what is already happening
Google releases ChatGPT competitor, Bard-NYT
GPT-4 message limit changed to 25 every 3 hours with further reduced cap coming next week
Microsoft lays off its entire AI Ethics and Society team
Okay yeah now I am threatened
GPT-4 released
ChatGPT now supports plugins!!
Most Influential People of All Time (According to Chat GPT)
Why are not governments afraid that AI will create massive unemployment?

Are You Freaked Out?

ChatGPT Will Replace Programmers Within 10 Years

Predicting The End of Manmade Software

Adam Hughes · Follow
Published in Level Up Coding · 12 min read · Feb 28

849 59

Article origin

OpenAI's DALL-E prompted with "photo-realistic 3D robot destroying a computer"

Phase	Timeframe	Job Loss Prediction
0: The Prototypes	Q1 2023	2%
1: Scale and IDE Infiltration	Q2 - Q4 2023	5%
2: Advanced IDE Tooling and Consolidation	1 - 2 yr	25%
3: SaaS and No-Code	2 - 5 yr	75%
4: AI Native and Domain Dominance	5 - 10 yr	95%
5: Heat Death	10+ yr	99%

Oh, well, life always finds a way.

Nonetheless, thank you ChatGPT, for debugging my code and correcting my grammar!