INTRODUCTION

November 30, 2022 marks a milestone in the history of AI. Built by San Francisco based Open AI, ChatGPT, arguably the most intelligent chatbot till date, was opened for public testing. By December 4, 2022 over a million had signed up. Naturally, it provoked a spectrum of opinions from “greatest bonanza since the invention of wheel”, and “as useful as an iPhone app” to “AI-rmageddon”, and everything in between. In this project, I have attempted to capture exactly that – the diverse opinions or sentiments on the topic of ChatGPT and how those sentiments have changed since the time it was launched four months ago.

Sentiment analysis is widely applied to ‘voice of the customer’ materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. It basically categorizes a piece of text as positive, negative or neutral. It is extremely valuable for a manufacturer or a marketer to know what exactly the people are talking about their products and services. We can track the change in sentiments over time, before and after an event, on products, services, social issues, even political personalities. In this project, I have done a Topic Modelling analysis to find out relevant topics being discussed. I have also analysed the sentiments of Reddit users on the topic of ChatGPT, and have also studied the changes in sentiments on ‘ChatGPT’ over the four months, since its launch on November 30, 2022.

Since ‘ChatGPT’ and AI have ignited passionate conversations from dinner tables to Global Economic Forums, sites like Reddit.com, was abuzz with activity, with millions of comments in its subreddit (a sub thread) ‘r/ChatGPT’. According to a survey conducted by ‘foundationinc’ in 2019, Reddit was the ‘sixth most popular website in the U.S, recording 1.7 billion comments in 2019’. Given its tech-savvy users, a hot tech topic like the launch of ‘ChatGPT’ can lend itself to a lot of valuable information through the ‘comments’ section on the website.

DATA

There are two primary ways to extract ‘comments’ from Reddit: Reddit API via ‘praw’, or, PushShift API via ‘pmaw’. Though both methods have their advantages and disadvantages, I have used the PushShift API method to extract large amounts of data (roughly 550,000 comments). But for practical reasons like memory capacity and ease of use of programming languages, I have also extracted several batches of data in smaller, and more manageable sizes. Overall, all the data used were extracted from subreddit r/ChatGPT, from December 6, 2022, to April 11, 2023.

SENTIMENT ANALYSIS OVER MONTHS

DISTRIBUTION OF COMMENTS BY MONTHS

MONTH	NO. OF COMMENTS
December 2022	56,871
January 2023	84,236
February 2023	112,998
March 2023	178,038

ChatGPT was launched on November 30, 2022 and the first post /comment was made on the subreddit r/ChatGPT on the December 6, 2022. We can see the popularity for ChatGPT growing over the months. To find out what exactly was being discussed in the 'Comments' section of the subreddit, we can use a Word Cloud:

A Word Cloud of all the months shows typically what one would expect from a brand-new tech product, but it also features the range of sentiments evoked. From the functional aspects like 'code', 'content', 'internet', 'information' etc to more metaphysical concepts like 'time', 'humanity', 'hope' 'history' etc. It is evident, that one technology can evoke a gamut of

emotions from awe to disgust. There are also words that describe the early glitches and shortcomings of ChatGPT like ‘false’, ‘wrong’, ‘complicated’ etc. We can also create Word Clouds for different months to see if there is a change in the pattern of the words

December

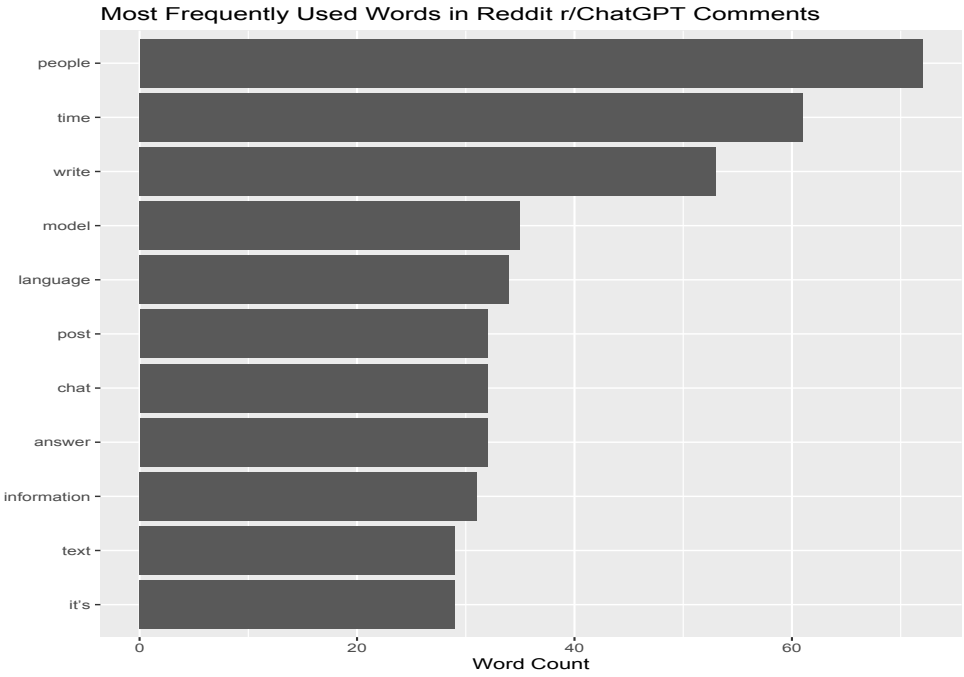
January

February

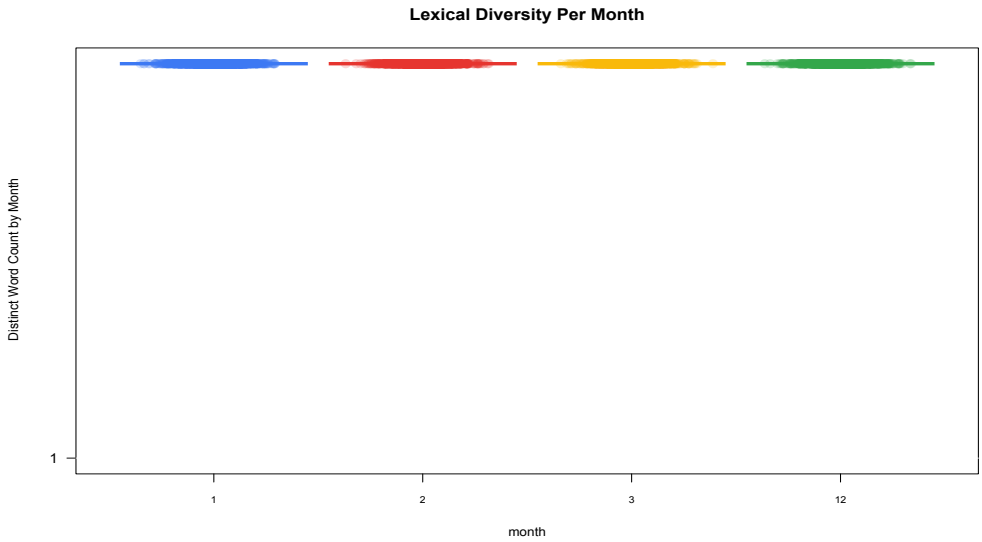
March

Though it is still early days for ChatGPT to discern any glaring patterns between the months, we can roughly see the arc of exploration with months of December, January, February showing words that explain the features of ChatGPT, when compared to March where more serious topics like ‘jobs’, ‘access’, ‘dangerous’ etc are entering the discussion.

We can also see if there is any pattern emerging, when we study the most frequently used words. The three words overall, as was evident from the Word Clouds are ‘people’, ‘time’ and ‘write’. And this pattern continues for all 4 months.

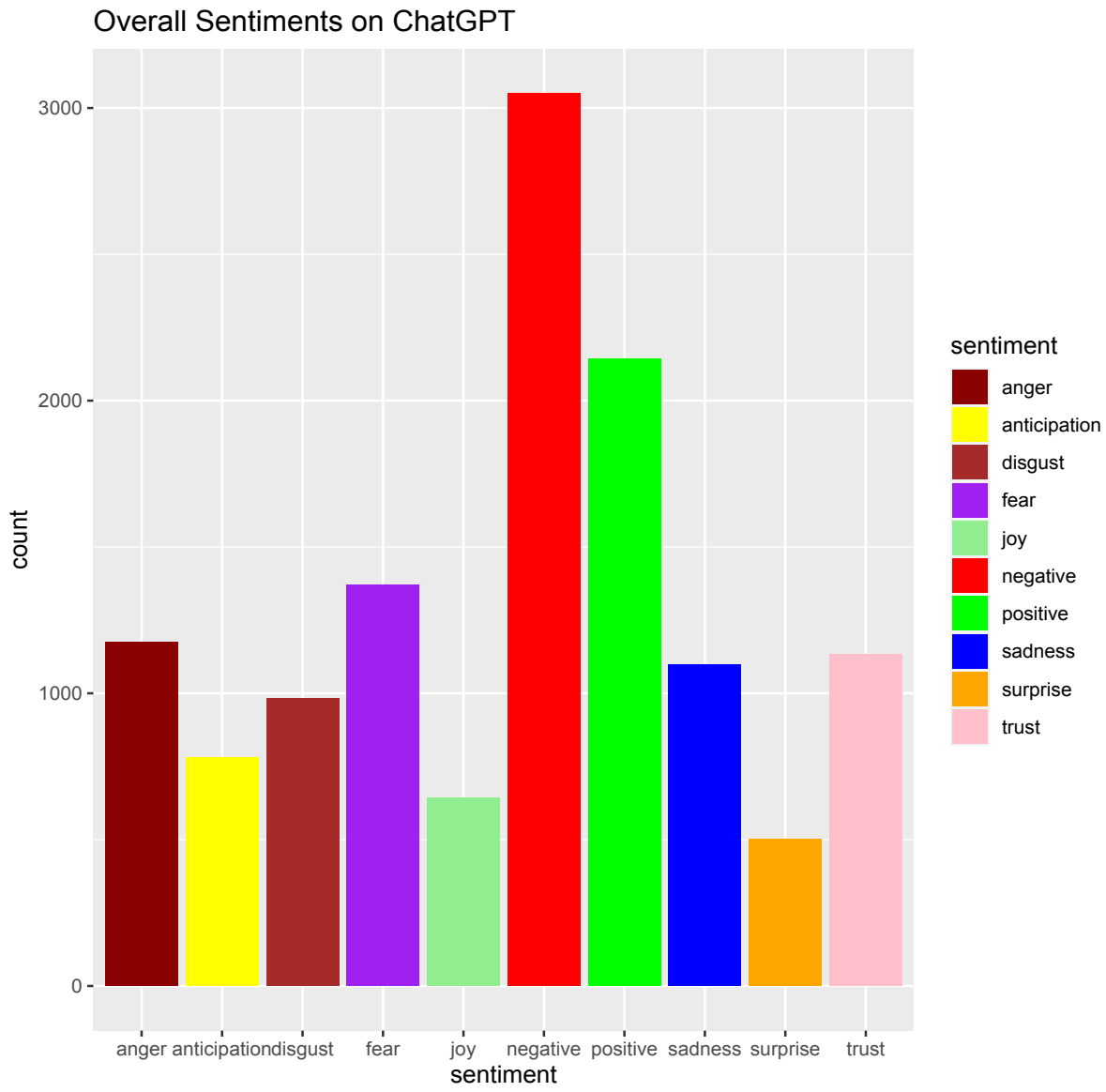


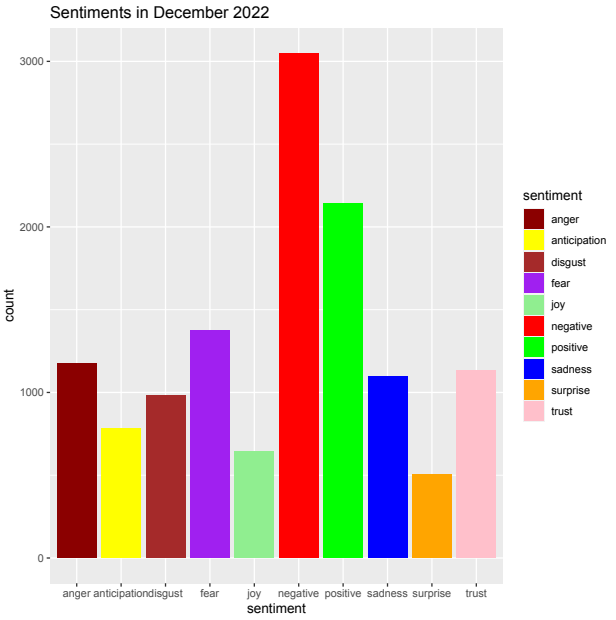
Since only four months have passed since the launch, there is not a lot of lexical diversity between the months, as it is evident from the plot below.

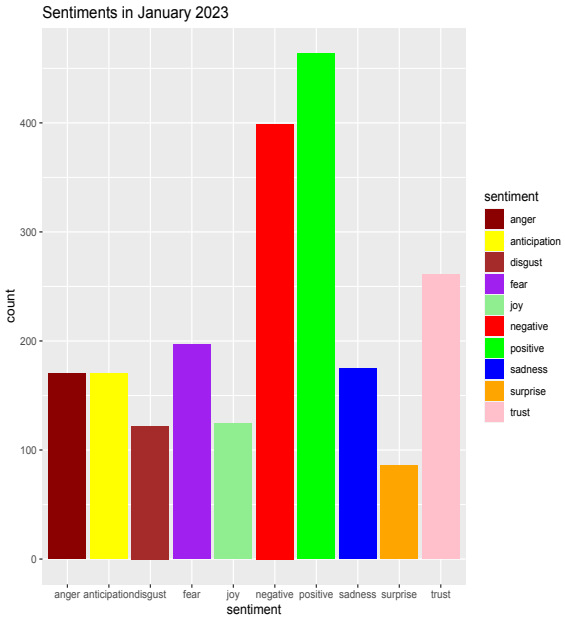


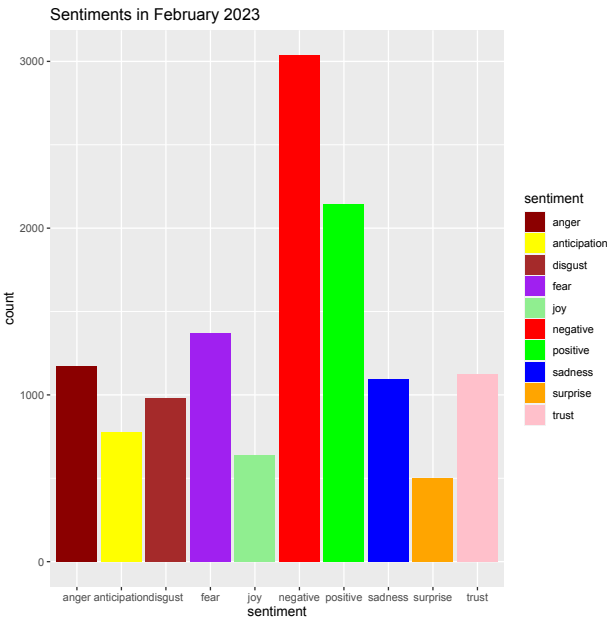
NRC Sentiment Analysis

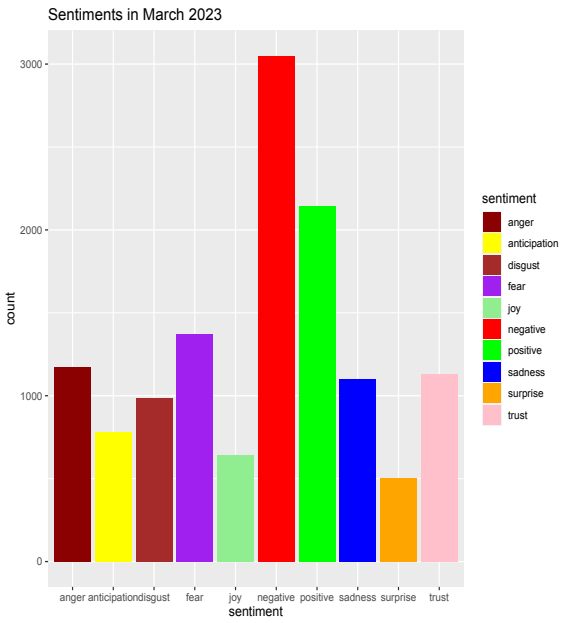
The NRC Emotion Lexicon is a list of 5,636 English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). Using the NRC Lexicon, I have analysed the overall sentiments of Reddit users towards ChatGPT and have also looked at individual months to detect any changes.

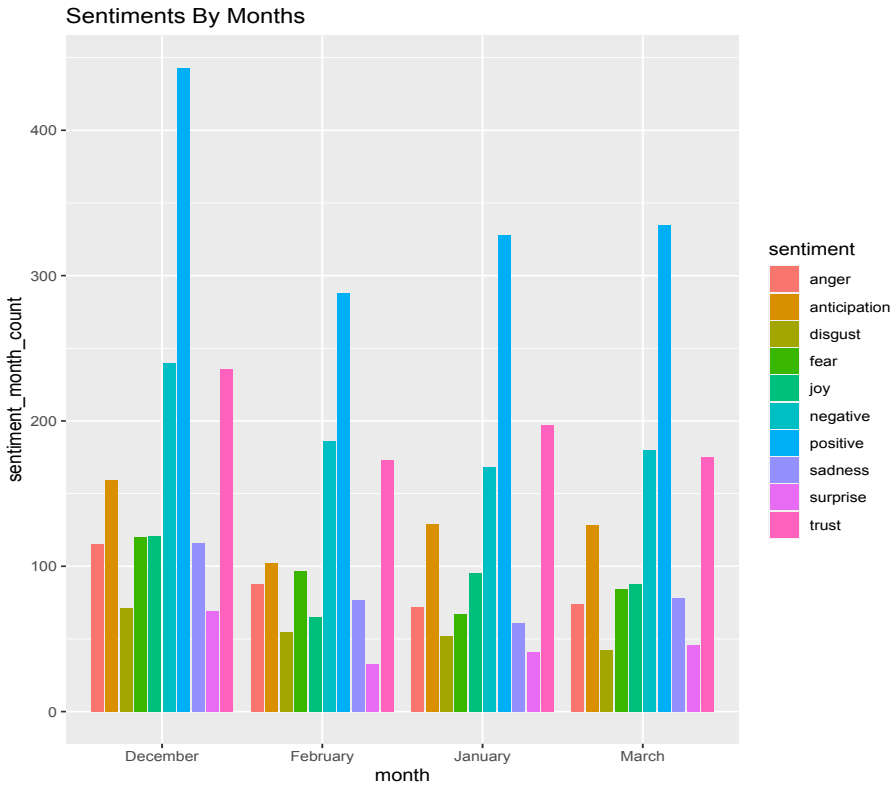



Sentiments in December 2022
count
3000
2000
1000
0
anger anticipationdisgust fear joy negative positive sadness surprise trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust

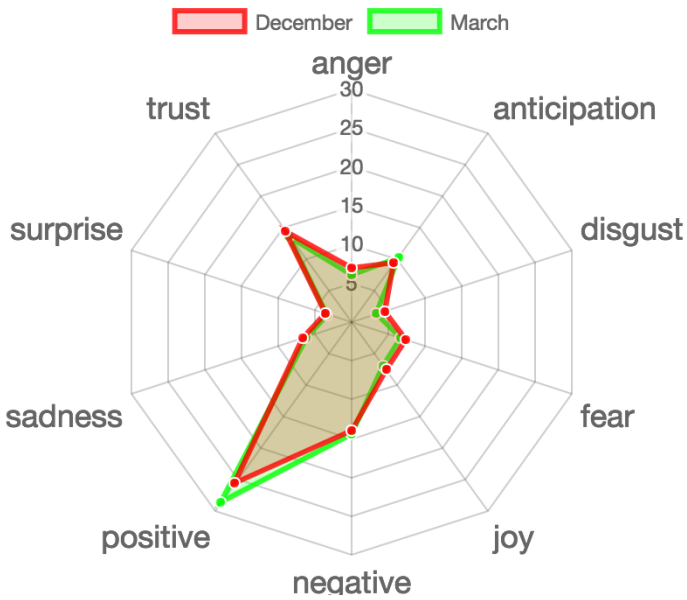

Sentiments in January 2023
count
400
300
200
100
0
anger anticipationdisgust fear joy negative positive sadness surprise trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust


Sentiments in February 2023
count
3000
2000
1000
0
anger anticipationdisgust fear joy negative positive sadness surprise trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust

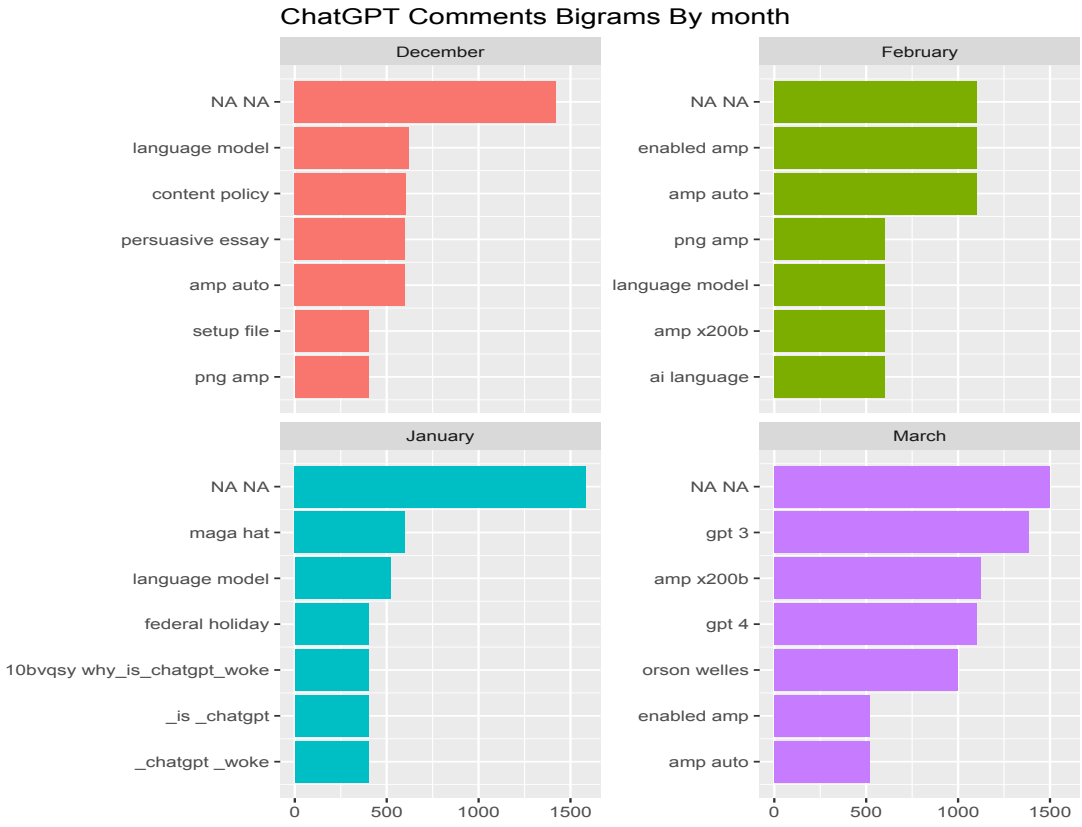

Sentiments in March 2023
count
3000
2000
1000
0
anger anticipationdisgust fear joy negative positive sadness surprise trust
sentiment
sentiment
anger
anticipation
disgust
fear
joy
negative
positive
sadness
surprise
trust



The overall sentiments of Reddit customers on ChatGPT have been negative when the words used are compared against the NRC lexicon, whilst sentiments like anticipation, fear, trust have remained fairly static over the months. Perhaps, if the data is spread over many years, instead of months, there might be a discernible pattern in the change of sentiments.



When the sentiments for the month of December and March are compared, they are almost identical, except March is slightly more positive. Bigram Analysis does not reveal a lot of differences between months.



TOPIC MODELLING

In the field of Natural Language Processing (NLP), topic modelling is typically done as an unsupervised learning task in which algorithms apply statistics to figure out what words are similar so they can be clustered into groups. The clusters represent the topics and are composed of the words that fit within the topic.

Preparing the data

With the CSV file downloaded from BERTopic installed in the Python environment, we begin by importing the dependencies and then opening the CSV file into a pandas dataframe object.

```
[3] #create a pandas dataframe
    df = pd.read_csv('chatgpt_comments.csv')
```

```
#check the first 5 dataframe rows
df.head()
```

	id	author	body	permalink	utc_datetime_str
0	j2baicq	BitzenBoy	Yes I made ginger snap cookies and they were d...	/r/ChatGPT/comments/zzfa8k/has_anyone_tried_ma...	30/12/2022 23:57
1	j2baby6	AutoModerator	In order to prevent multiple repetitive commen...	/r/ChatGPT/comments/zzfl8h/katzenfaszination_c...	30/12/2022 23:56
2	j2ba6o5	AutoModerator	In order to prevent multiple repetitive commen...	/r/ChatGPT/comments/zzfker/ive_made_a_channel_...	30/12/2022 23:55
3	j2ba5zu	AutoModerator	In order to prevent multiple repetitive commen...	/r/ChatGPT/comments/zzfkaj/i_put_chatgpt_throu...	30/12/2022 23:55
4	j2ba4cv	err_mate	CatGPT	/r/ChatGPT/comments/zxqfzv/chatgpt_pretends_to...	30/12/2022 23:55

```
[5] len(df)

12000
```

We delete unwanted rows.

```
[6] #drop the rows where the value of 'author' is equal to 'AutoModerator'
    df.drop(df[df.author == 'AutoModerator'].index, inplace=True)
```

```
[7] len(df)

10602
```

After reducing the dataframe, we're left with 10,602 rows of data to explore. And we have 355 topics.

```
topic_model.get_topic_info()
```

	Topic	Count	Name
0	-1	965	-1_mesa_uh_agree_dan
1	0	120	0_evolution_irony_which_phrases
2	1	70	1_amazing_weathering_captivating_chills
3	2	70	2_dead_love_thought_no
4	3	61	3_ha_monkey_dedguy21_vegas
...	...	...	...
350	349	11	349_emily_sassy_half_knows
351	350	10	350_parlance_hilarious_clever_indeed
352	351	10	351_subscribe_absent_suddenly_wont
353	352	10	352_remains_id_does_something
354	353	10	353_visible_structures_eye_naked

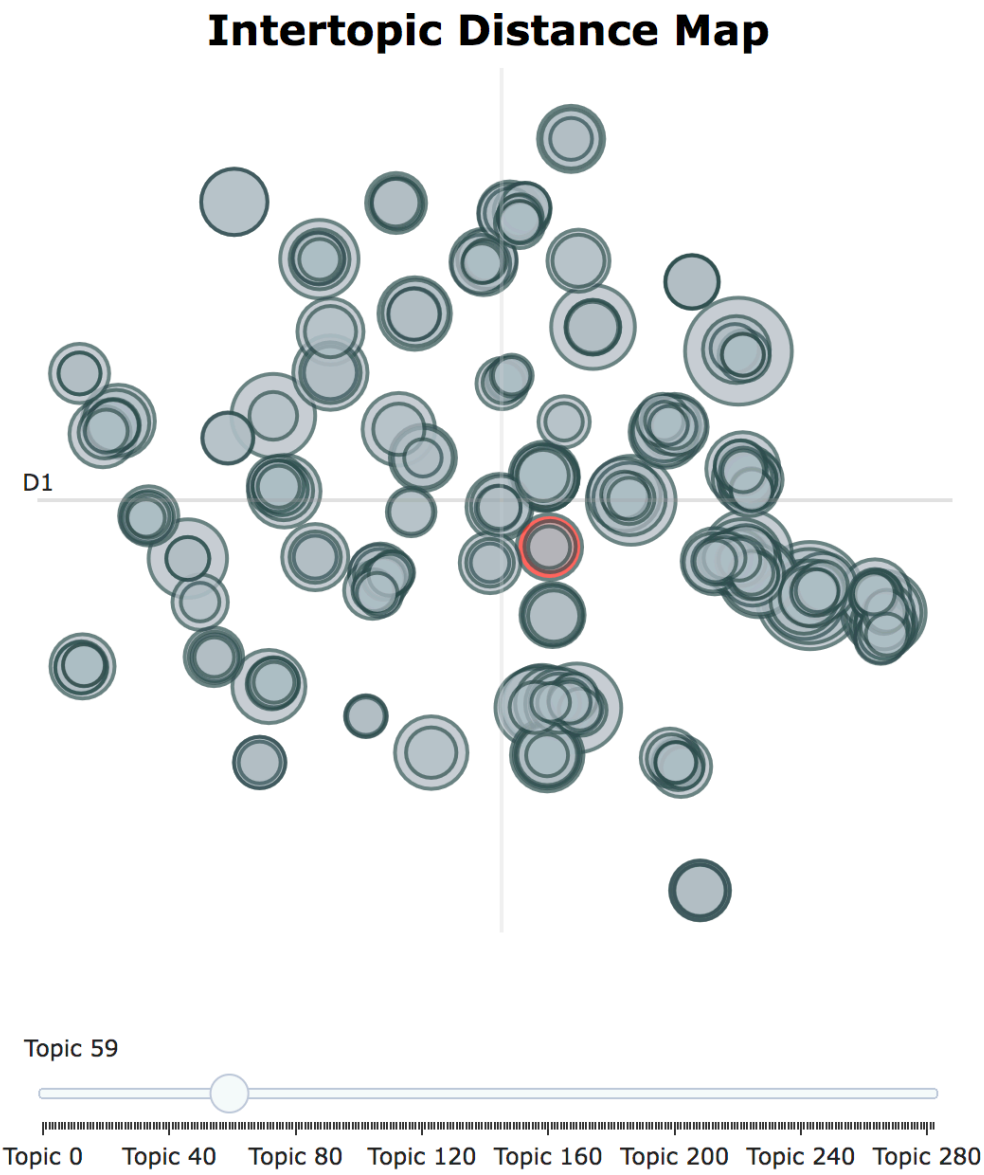
355 rows × 3 columns

This can be reduced further.

	Topic	Count	Name
0	-1	450	-1_jobs_after_spending_time_on_this_sub_im_con...
1	0	131	0_error_modelgpt_description_store
2	1	130	1_vlexld_streamable_lt_copied
3	2	101	2_input_rough_batman_proposal
4	3	90	3_lmao_daniel_whoa_uhhhh
...	...	...	...
279	278	16	278_activities_mission_existence_non
280	279	16	279_bait_click_ad_news
281	280	16	280_tempest_teapot_luck_definition
282	281	16	281_cover_today_letter_bitch
283	282	16	282_sexy_job__

284 rows × 3 columns

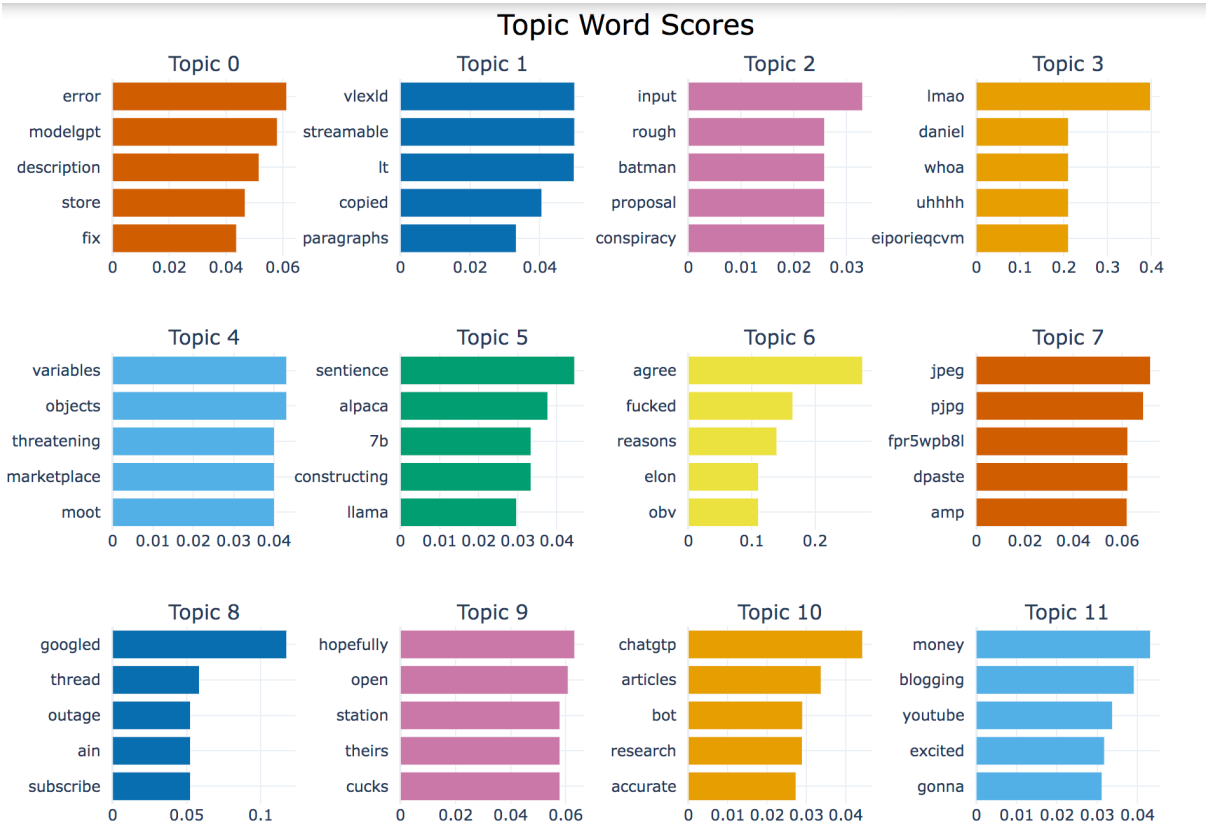
From here, we store the array of ChatGPT comments in a variable named ‘docs’ so we can pass it into BERTopic. Once the topic model is complete, we can explore the topics using the get_topic_info() method to output a dataframe containing the topic group, the count of words, and the topic name. We can then visualize the topics using visualize_topics() method as two-dimensional representations of our topics, making them easier to explore.



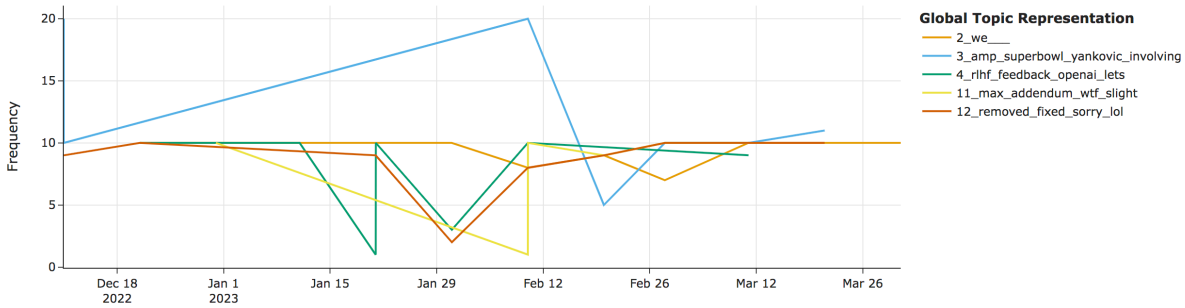
With the interactive plot we can explore the various topics from numbered 0 to 284. The size of the circles represents the popularity and distance between them gives their connectedness.

Visualizing word frequency

We can visualize the top terms topics by creating bar charts out of the c-TF-IDF scores. This makes it easier to compare what terms are within each topic so you can easily compare topic representations to each other.



Topics over Time



CONCLUSION

Whether it's a launch of a product, or occurrence of an event, people form opinions and are not shy of expressing them. More and more methods are being developed to harness this valuable information found in the comments of a forum, website, etc in the form of unstructured texts. In this project, I have extracted large amounts of data, in terms of 'comments' from the Reddit website, on the topic of 'ChatGPT'. By analysing the sentiments through the months of December 2022 to March 2023, I have studied the prevailing sentiments on the topic, its variation over time, and also topics discussed and their inter-connectivity.

Through webscraping using the API method, more than 550,000 comments on the topic of ChatGPT was analysed. Though the overall sentiment came out negative, there were revealing information as to the topics that were discussed. One of the objectives of this project, was to study if there is any variation of the sentiments over time. However, due to the short span of the study (4 months), there is not enough evidence that suggest significant differences of opinions within this time. By analysing the words used in the 'comments' section we could also see the lack of lexical diversity between the months. If this technique of extracting and processing unstructured texts is followed over a greater span of time, several meaningful insights can be gleaned.