**Big Data Infrastructures Project Report**

***Twitter Analysis of ChatGPT***

BADM 558  
University of Illinois Urbana-Champaign  
Spring 2023

May 7, 2023

***Team 6***

Bill Bi, Catherine Wu, Crystal Wang, Jiahe Yu, Pranavi Kakani, and Xiying Zhao

***Faculty***

Professor Zilong Liu

# Table of Contents

[**Abstract 3**](#_av97ffxz2zoq)

[**Introduction 4**](#_lsjxfsw8jqqf)

[**Methods 5**](#_una67vqos8yg)

[**Results and Discussion 8**](#_rr4vd5q70pa6)

[**Conclusion 27**](#_6685wkmpnk9f)

[**References 28**](#_9tt9hch3qvkf)

# 

# Abstract

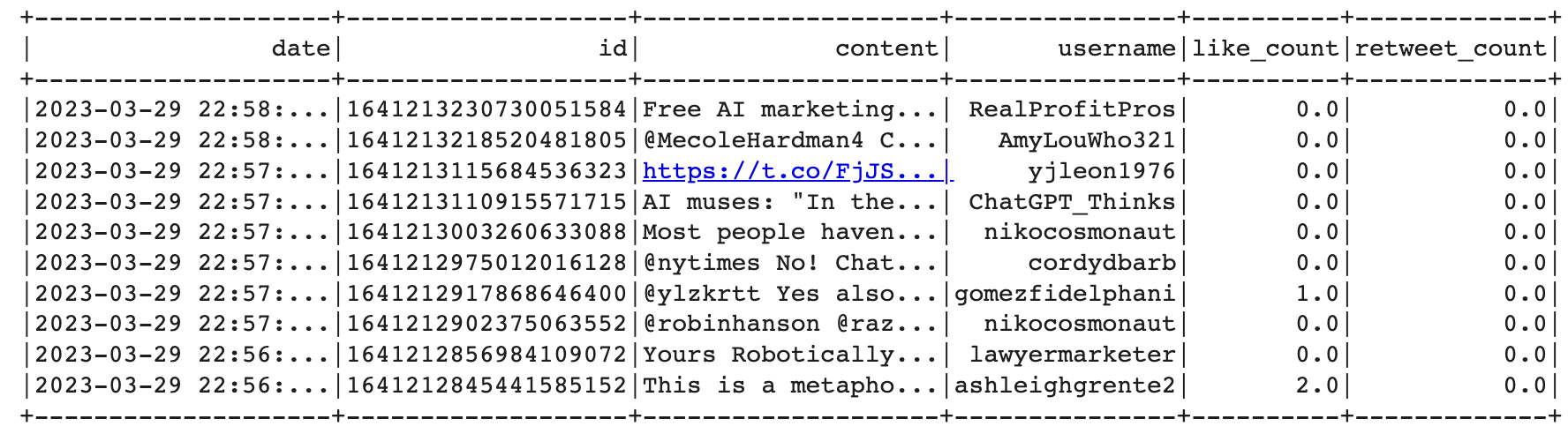
As the application with the fastest user-growing speed in history (reaching 1 million users in 5 days), ChatGPT has been a hot topic in 2023. As millions of tweets discuss such a disruptive technology, this project conducted a sentiment analysis to find trend topics and people’s attitudes towards this innovative technology. This project used a dataset from Twitter containing 500,000 tweets about ChatGPT from January to March 2023. The data was preprocessed by changing the date format and removing unnecessary symbols and redundancies. Several analysis tools were used on the data set, including Latent Dirichlet Analysis (LDA) and Exploratory Data Analysis (EDA) for uncovering hidden topics and understanding popular trends, unigram, bigram, and trigrams for most frequently used words, and Vader for sentiment analysis. Several keywords related to artificial intelligence and machine learning were counted, top influencers were discovered, and public opinions toward ChatGPT were analyzed in this study. Possible assumptions and implications from these trends and attitudes were also discussed.

# 

# Introduction

Developed by OpenAI and launched in November 2022, ChatGPT was born as an artificial intelligence chatbot based on the GPT-3 architecture. ChatGPT was explicitly trained for natural language processing tasks with a massive dataset of text from the internet. The training data was used to teach the model how to understand and generate natural language. Given the advantage of ChatGPT over traditional chatbots, ChatGPT has attracted more than 100 million users in 2 months after its launch. It has been shown that ChatGPT can produce high-quality responses to various challenges on professional occasions.

To discover the public’s attitudes and general trends toward using ChatGPT, this project selected a dataset related to ChatGPT, including keywords, #hashtags, and @mentions about GhatGPT. The dataset contains 500,000 tweets about ChatGPT from January to March 2023. By analyzing tweet volume, sentiment, user engagement, and the influence of important AI events, the dataset aims to help understand public opinion, trends, and potential applications of ChatGPT.



The dataset contains five variables, as explained below:

* Data: The date and time when the tweet was posted
* ID: A unique identifier for each tweet
* Content: The text of the tweet
* Like count: The Twitter username of the person who posted the tweet
* Retweet count: The number of times the tweet was retweeted

Using the dataset, this project will offer valuable insights into ChatCPT-related tweets and allow people to make informed decisions and shape the future of AI-powered conversational technologies.

# 

# 

# 

# 

# Methods

In this project, we used several analytics methods to extract insights and information from our Twitter dataset. Broadly, we divide them into two sections - Data manipulation and Data analysis. Data manipulation techniques include data upload, inspection, and preprocessing, all essential to preparing data for analysis. Data analysis methods used in this project include exploratory data analysis, topic modeling with LDA, text analysis, and sentiment analysis using tools like Vader and TextBlob. These methods helped us understand and derive insights from large and complex datasets, like the Twitter dataset we used for our project. All the methods are explained in detail in this section.

* **Data Manipulation**

Refers to transforming and manipulating the raw data to make it suitable for analysis. It involves tasks like data cleaning, preprocessing, transformation, and visualization.

* + ***Data Upload:*** The process of loading data into a software program for analysis. This can be done by uploading a file containing the data or connecting to a database to extract the data.
  + ***Data Inspection:*** It checks and examines the data for inconsistencies, errors, missing values, and outliers. This step is essential to ensure the quality and reliability of the data.
  + ***Data Preprocessing*:** refers to cleaning, transforming, and preparing the data for analysis. This step involves removing missing values, handling outliers, scaling the data, and encoding categorical variables.
    - *Date formatting:* The software can easily interpret the process of converting date and time data, by default in a string, into a standard date format. This is important for tasks like time series analysis and data visualization.
    - *Removing unnecessary characters:*It removes punctuation marks, special characters, and emojis from the text data. This helps to standardize the text and make it easier to analyze.
    - *Removing duplicates/redundancy:* It identifies and removes duplicate or redundant records from the dataset. This helps to reduce the dataset's size and improve the analysis's efficiency.
* **Data Analysis**

It uses statistical and computational methods to extract insights and information from data. This involves tasks like descriptive statistics, inferential statistics, and machine learning.

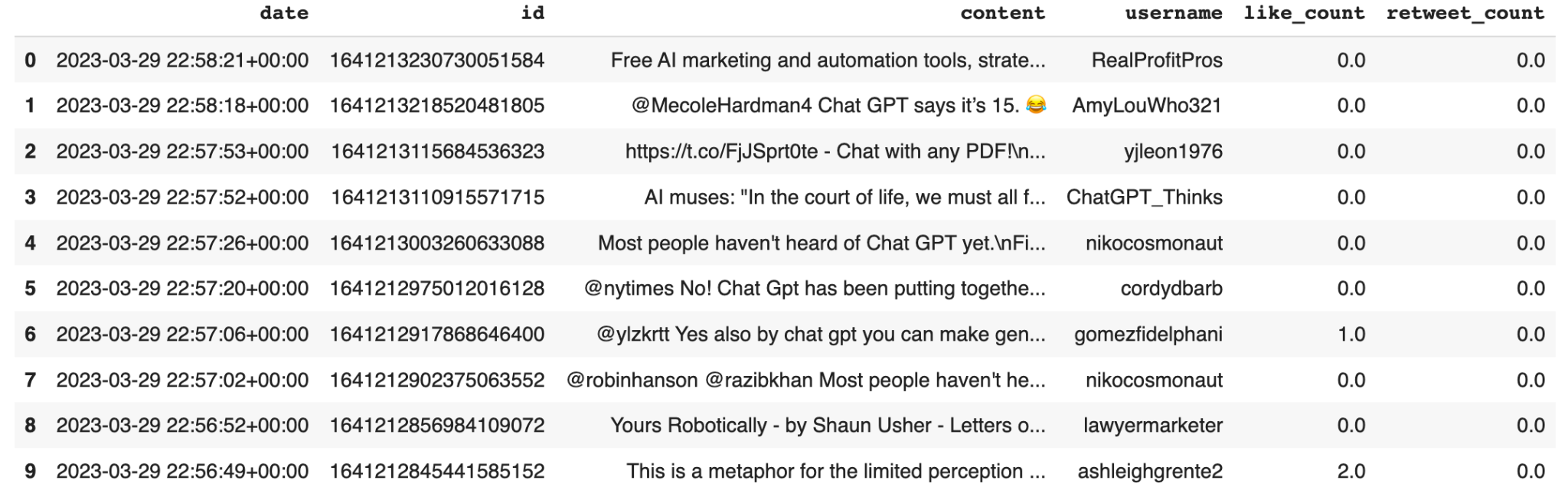
* + ***Topic Modeling with LDA:*** A technique for identifying topics and themes in a large corpus of text data. LDA stands for Latent Dirichlet Allocation, which is a probabilistic model that assumes each document is a mixture of topics and each topic is a mixture of words.
  + ***Exploratory Data Analysis:*** It analyzes and summarizes data sets to gain insights into the data. This includes summarizing the data, creating visualizations, identifying patterns and relationships, and generating hypotheses.
  + ***Text Analysis - Unigrams, Biograms, Trigrams:*** These are techniques for analyzing text data by breaking it down into individual words, pairs of words, or sets of three words. This helps to identify common patterns and themes in the text.
  + ***Sentiment Analysis:*** A technique for identifying the sentiment or emotion in text data. It involves tasks like identifying positive, negative, or neutral sentiments and quantifying the strength of the sentiment.
    - *Vader:* It is a sentiment analysis tool that uses a lexicon of words to score the sentiment of a piece of text. It can identify positive, negative, or neutral sentiments and quantify the strength of the sentiment.
    - *TextBlob:* It is a Python library for processing textual data. It includes tools for natural language processing tasks like sentiment analysis, part-of-speech tagging, and noun phrase extraction. It also includes a sentiment analysis module that can be used to score the sentiment of a piece of text.

# Results and Discussion

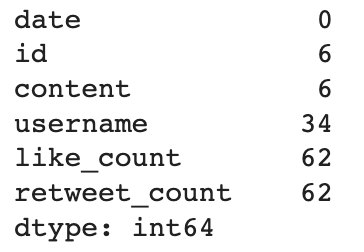
This section presents the step-by-step findings from the analysis of the Twitter dataset using various data manipulation, data analysis, and text/sentiment analysis techniques. It includes code snippets and output screenshots for each method, providing a detailed and transparent analysis account. The section summarizes the results and patterns identified through the analysis and provides a comprehensive dataset overview. It will also help demonstrate the effectiveness of the analysis techniques and provide a basis for conclusions.

1. **Data Load and Inspection**
   1. *Data upload*

**First ten rows** of the dataset -



Check for **missing** values -

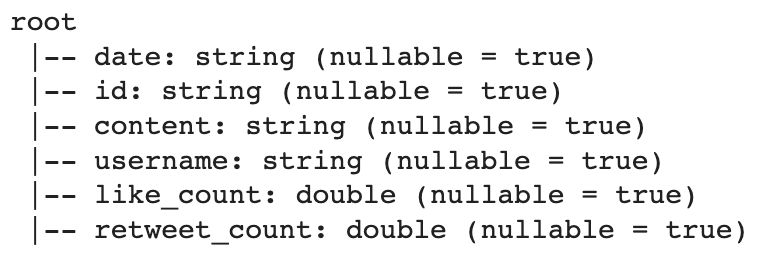


**Length** of the dataset after removing missing values -

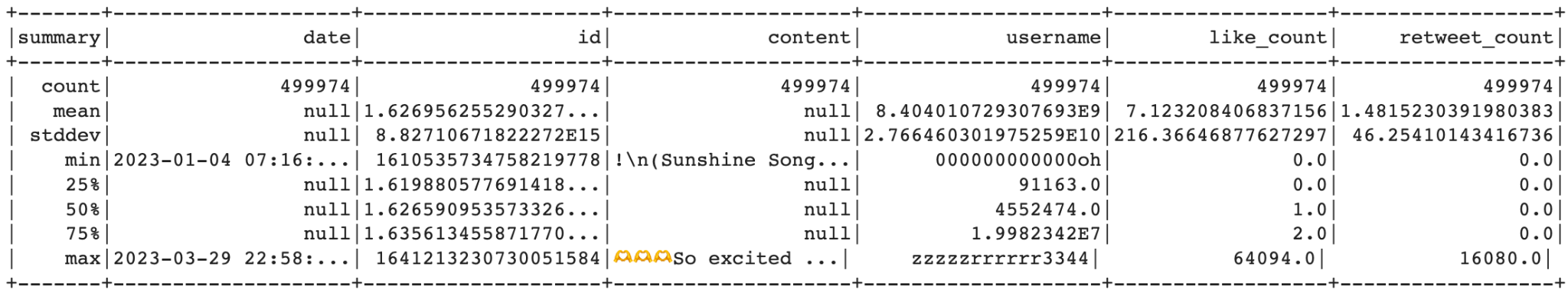


* 1. *Data Inspection*

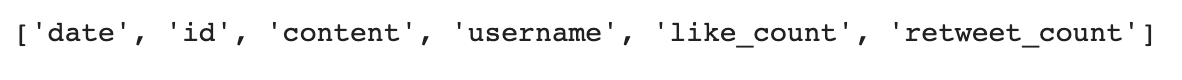
**Schema** of the data -



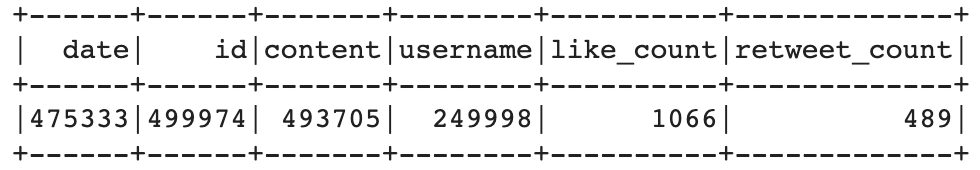
**Summary statistics** of data -



Examining column names -



Checking the number of **unique values** in each column

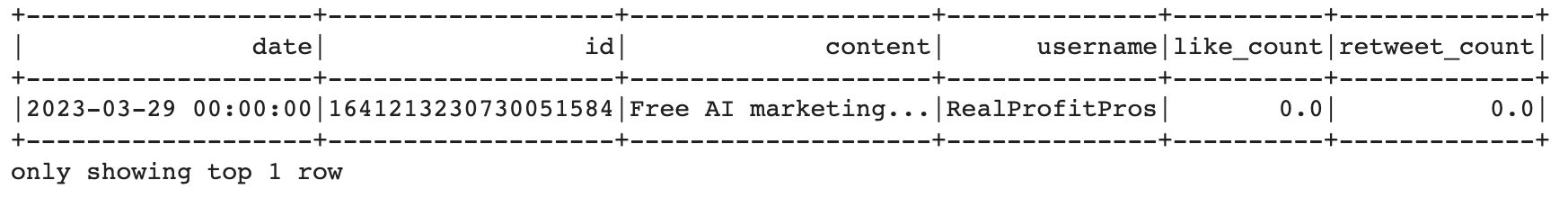


Average **tweet length** -

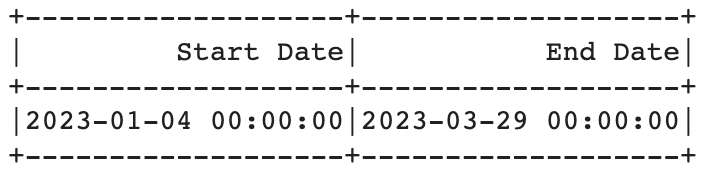


1. **Data Preprocessing**
   1. *Date column values to date time object*

Converting the 'date' column to a **DateTime object** -

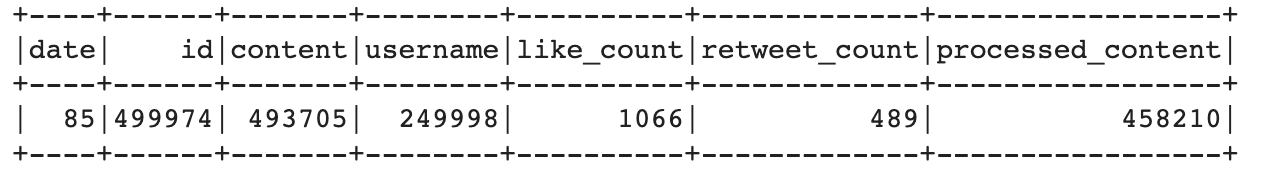


Checking **a range of dates** -



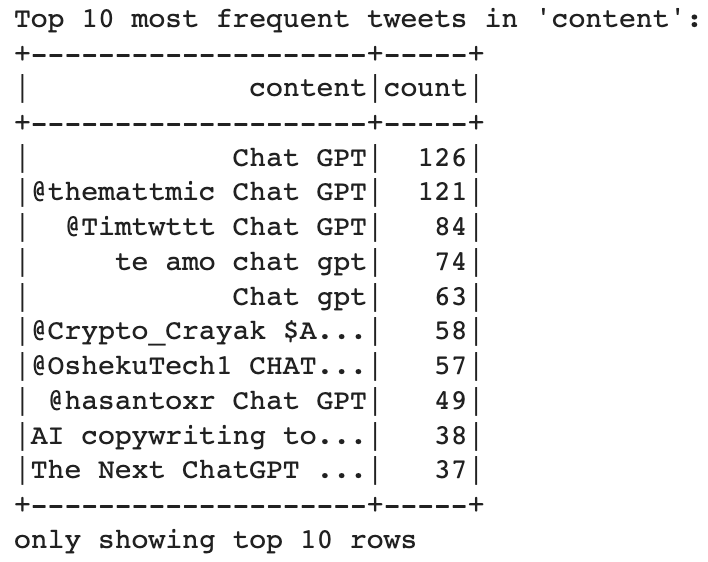
* 1. *Remove links, #, @, and unwanted characters*

Checking the number of **unique values** in each column -



* 1. *Removing Duplicates, Spam, and Redundancy*

**Value counts** for the 'content' column -



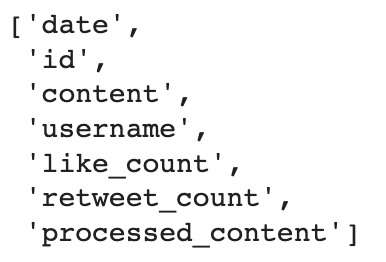
**Value counts** for the 'processed\_content' column -



Examining the number of rows **after removing duplicates** -

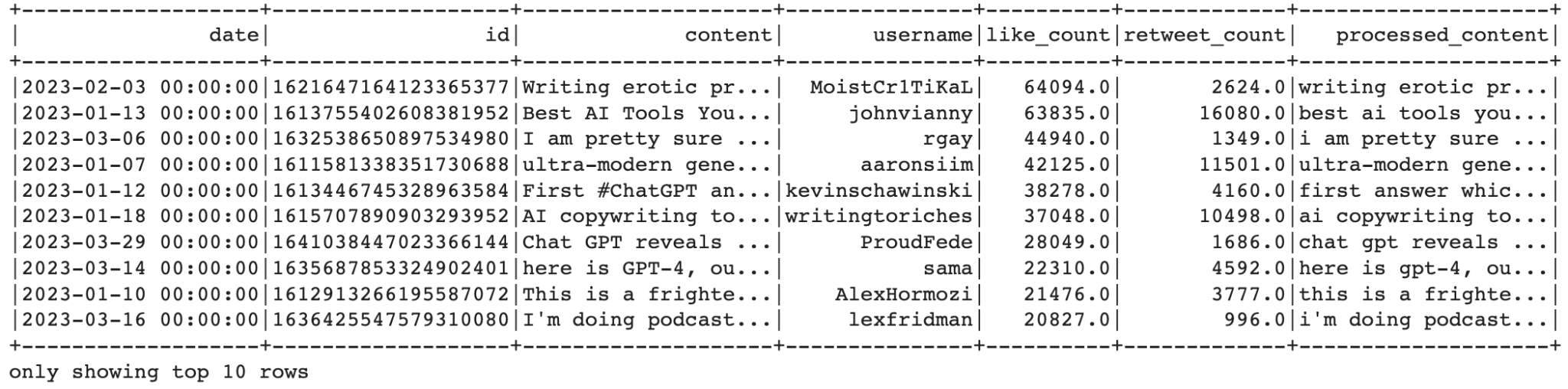


Addition of ‘**processed\_content’** column -

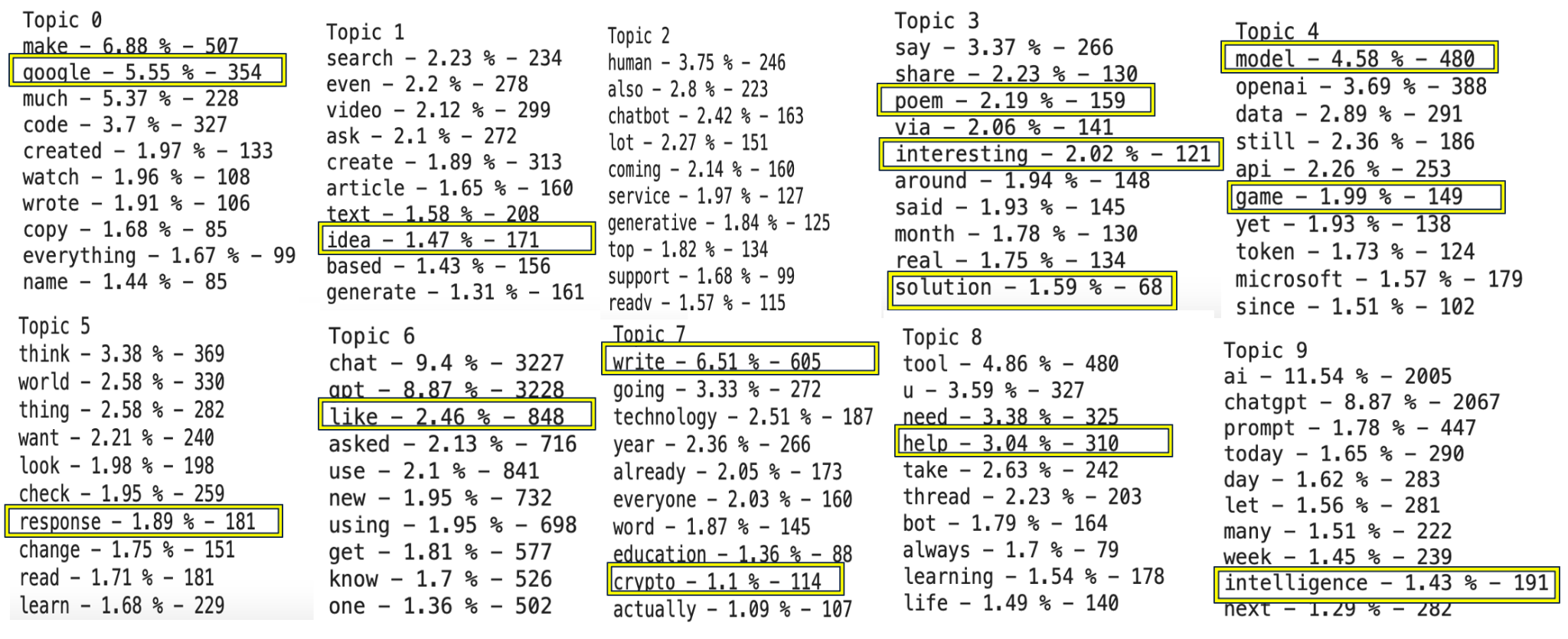


1. **Topic Modeling with LDA using Spark**

Sorting the data frame and getting the top 10000 **most liked tweets** -



**Creating topics** and printing the **top 10 words** for each subject with their frequency -

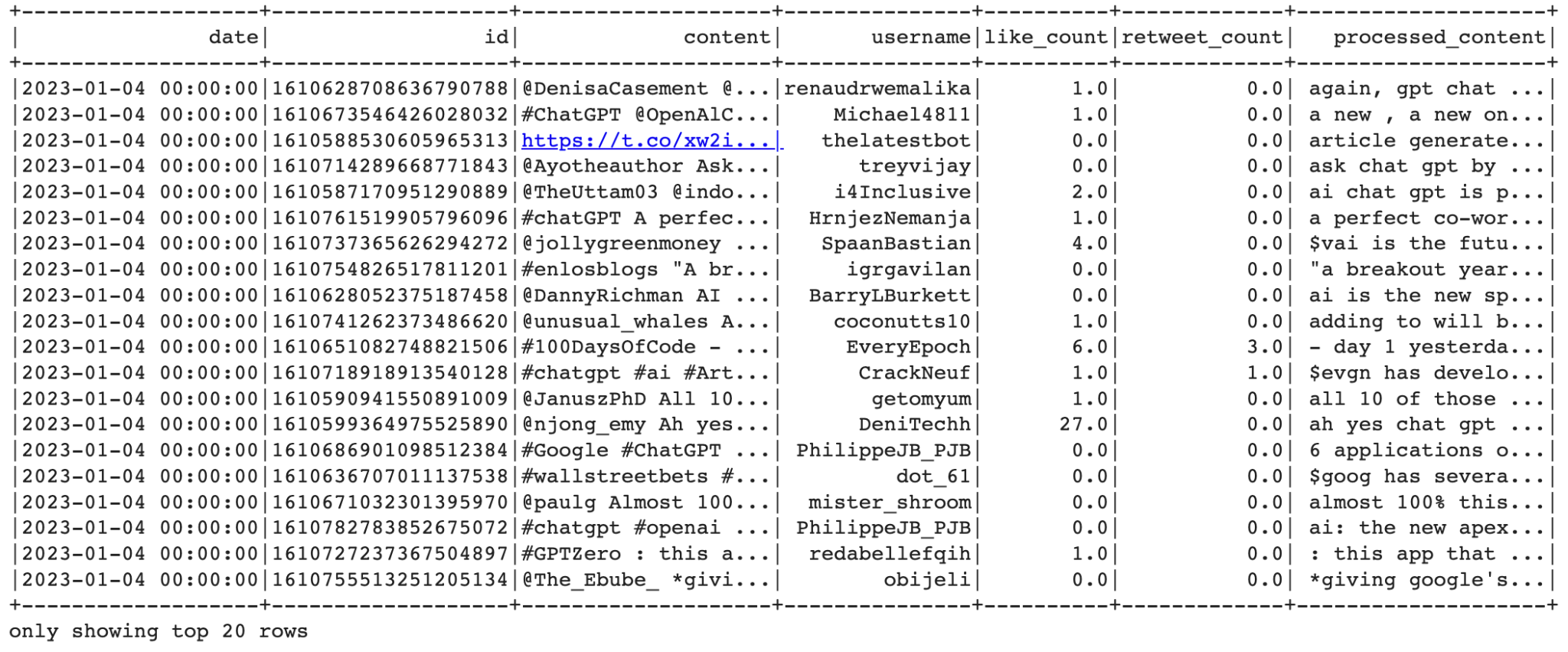


Discussion

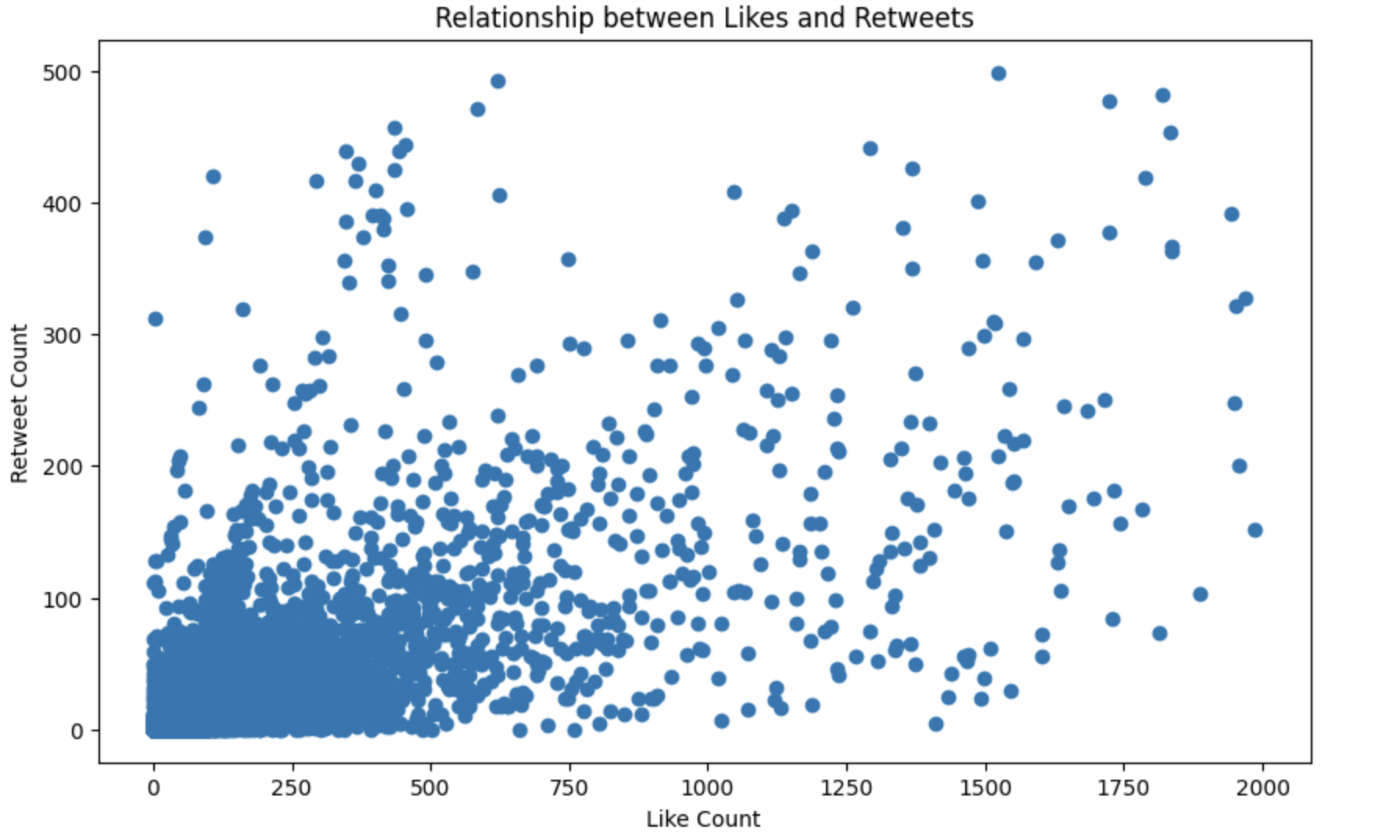
Top Modeling with LDA uncovers hidden topics within a collection of documents. Some of the keywords were unexpected, with the top ten topics filtered. For example, Google, Crypto, poem, and game appeared in the tweets frequently. It is interesting to discover the reason why people tweet these topics. Some people might be concerned about whether ChatGPT would replace Google or other search engines, while others might see ChatGPT as a tool in crypto or financial investment. ChatGPT might also be able to compose poems or play video games. These topics reflect people’s concern and curiosity about ChatGPT.

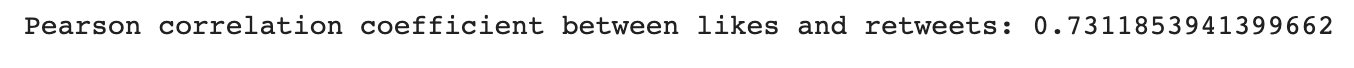
1. **Exploratory Data Analysis (EDA)**
   1. *Trends - Popularity, Timeline*
      1. Popularity

Filtering the df to include only rows where like\_count and retweet\_count exist -



Using Matplotlib to create a plot of the **relationship between Likes and Retweets -**

**Pearson correlation coefficient** between like\_count and retweet\_count -



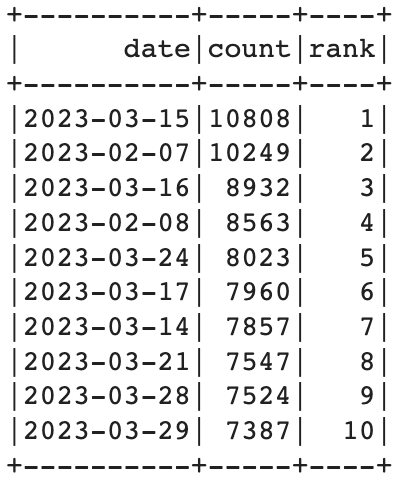
Discussion

According to the scatter plot above, most tweets have a like count of 750 and a retweet amount of 200 (the densest part). While tweets above these numbers are dispersed in the rest of the scatter plot. The Pearson coefficient measures the strength of a linear association between likes and retweets. In this case, the coefficient is approximately 0.73, indicating a moderate correlation between likes and retweets.

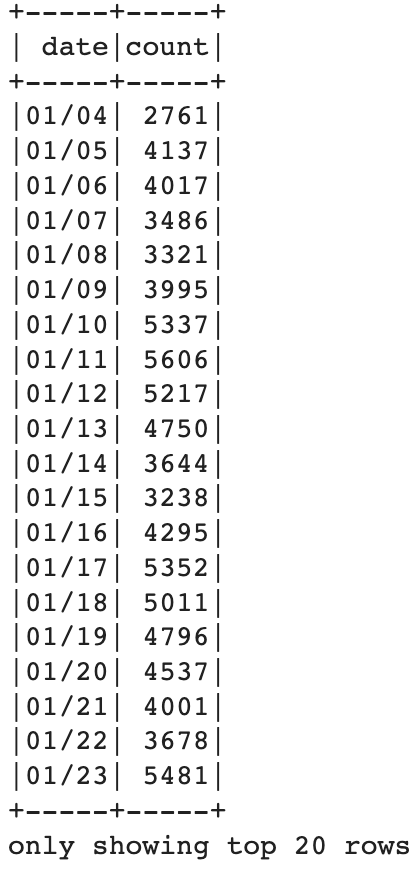
* + 1. Timeline

1. Tweets/day - Three-month (Jan, Feb, Mar) timeline

Top 10 dates with the **highest tweet counts** -



Grouping by date and counting the number of tweets -

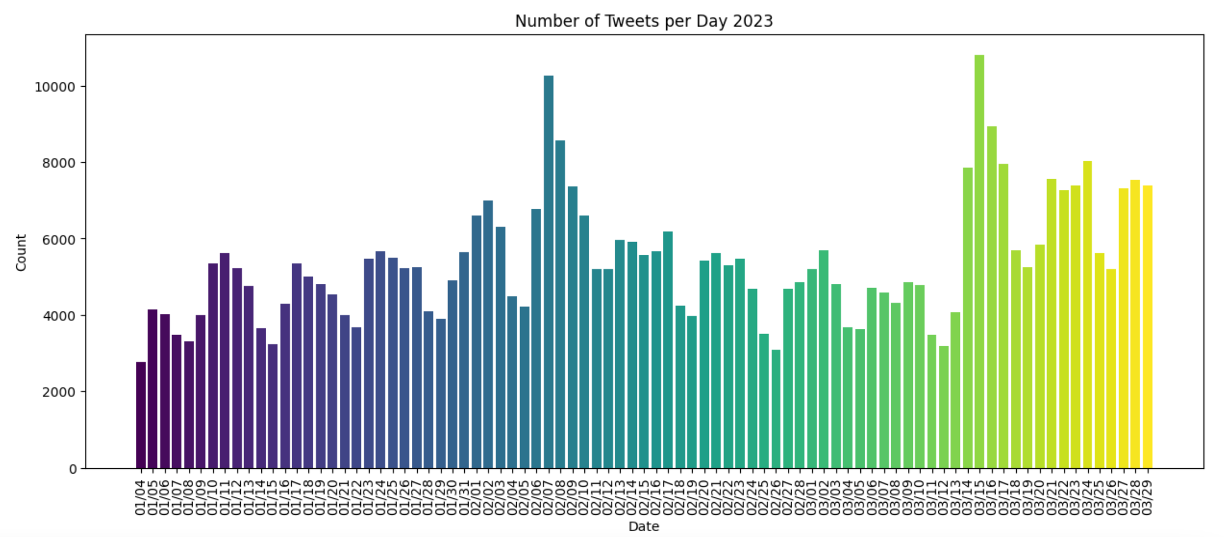


Discussion

According to the table of the top 10 dates with the highest number of retweet amount, all of the dates can be grouped into two significant events: February 7th and March 15th. On February 7th, Microsoft launched Bing Chat with inbuilt ChatGPT, allowing users to create PowerPoint slides, Excel tables, and Word documents with a single instruction. On March 14th, OpenAI launched GPT-4 for ChatGPT Plus users, and one week later, on March 23rd, the ChatGPT plugin was launched. It is reasonable to assume the reason why these dates have the most numbers of retweets - people are excited about the new launch of ChatGPT.

1. Tweets/week & Tweets/month

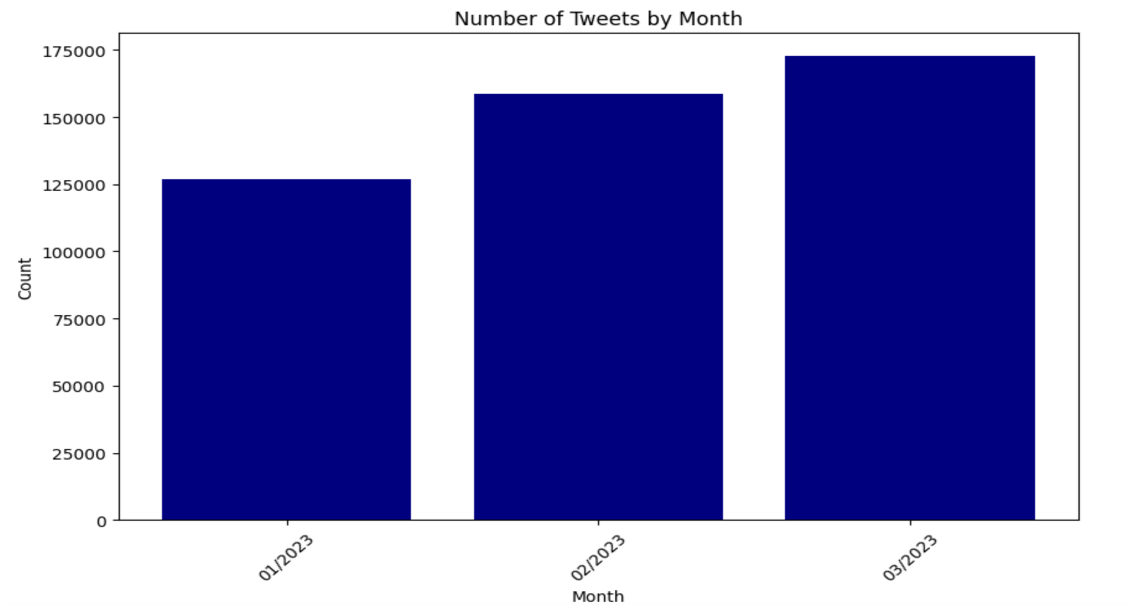
Plotting the number of tweets per **day** as a bar graph using Matplotlib -



Plotting the number of tweets per **week** as a bar graph using Matplotlib -



Plotting the number of tweets per **month** as a bar graph using Matplotlib -

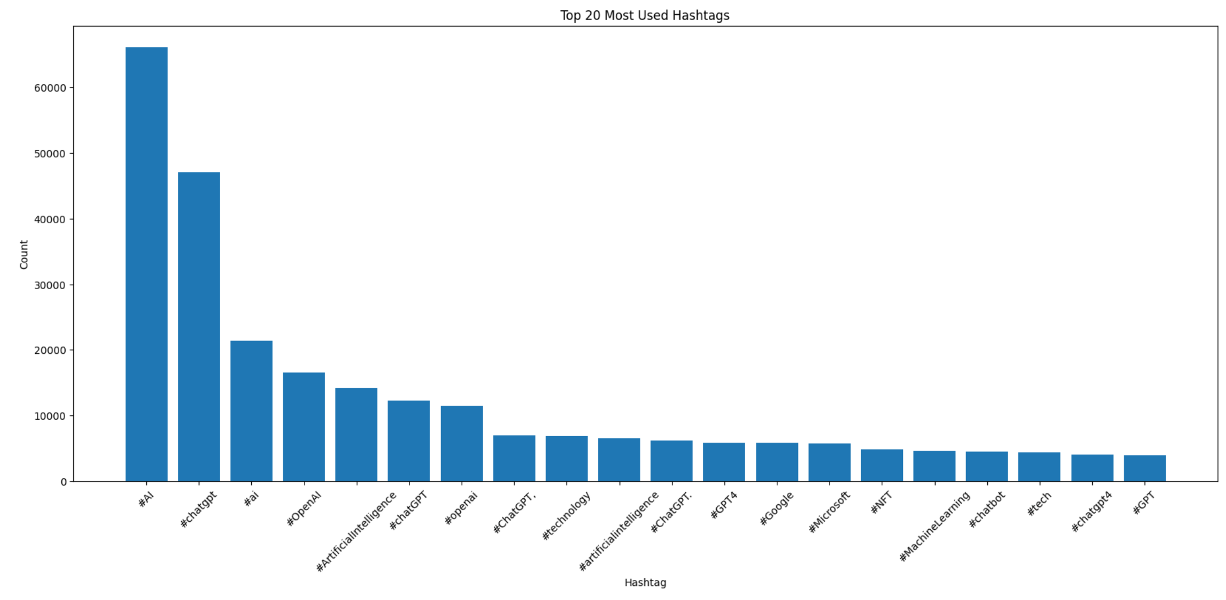


Discussion

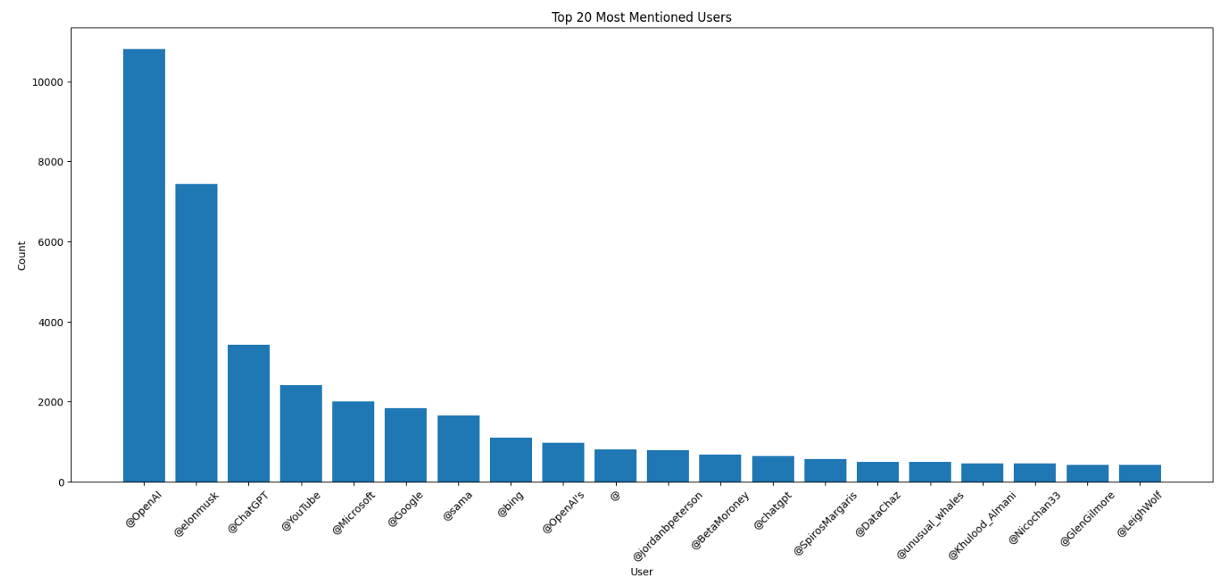
As discussed earlier, because of the new launch of Microsoft Bing Chat and ChatGPT 4.0, the number of tweets spiked around the date these new technologies were launched. Overall, with the growing popularity of ChatGPT, the total number of tweets has increased from January to March.

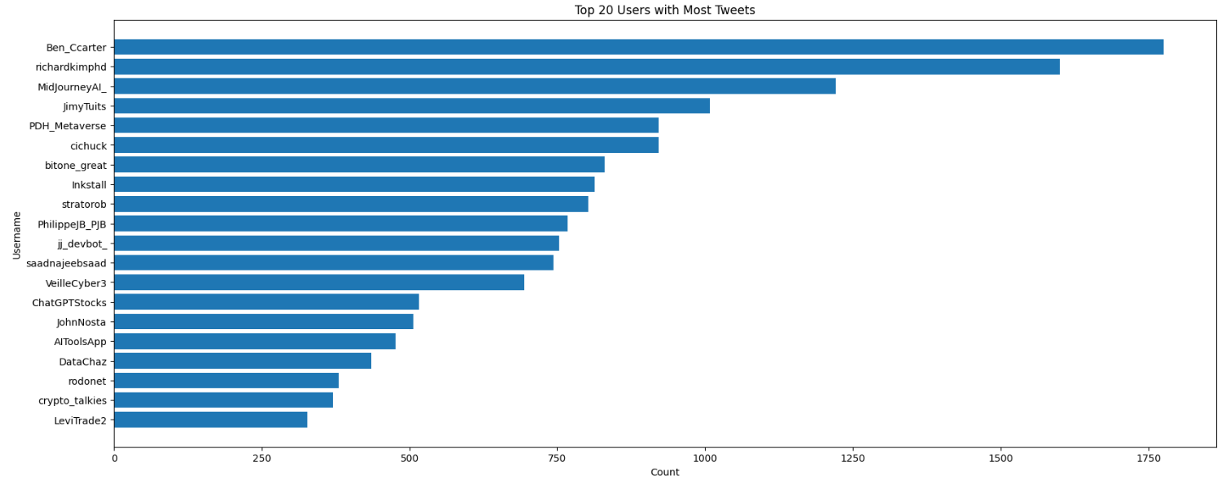
* 1. *Top #Hashtag @mention and usernames*

Extracting **hashtags** from the processed\_content column and plotting the top 20 most used hashtags-



Extracting **mentions** from the content column and plotting the top 20 most-used mentions -

Extracting **Usernames** from the content column and plotting the top 20 most used usernames -

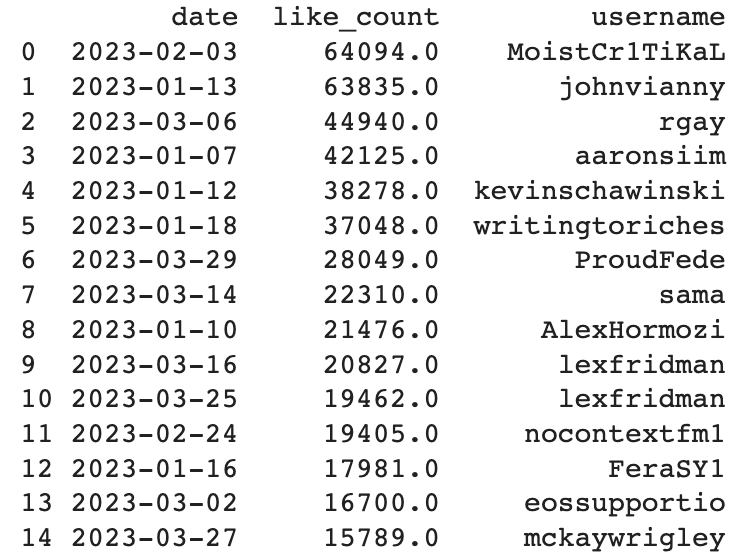


Discussion

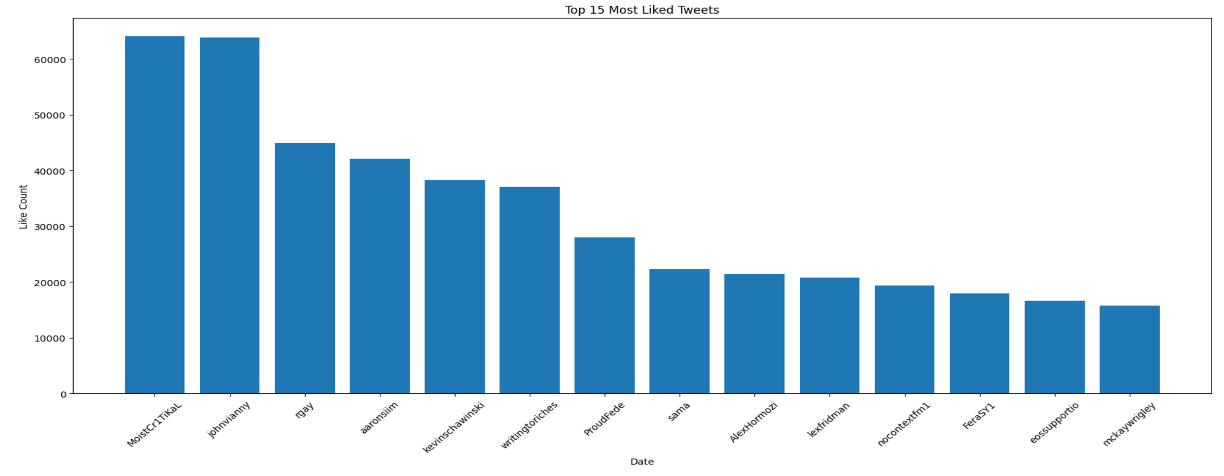
According to these three charts, AI, ChatGPT, and any topic related to artificial intelligence are hot topics that are searched the most in both hashtags and mentions. Top users can tweet more than 1,700 tweets daily with topics about ChatGPT.

* 1. *Most liked tweets & Most Influential Users.*
     1. Top 15 most-liked tweets

**Sorting** DataFrame by **like\_count** and printing top 15 -



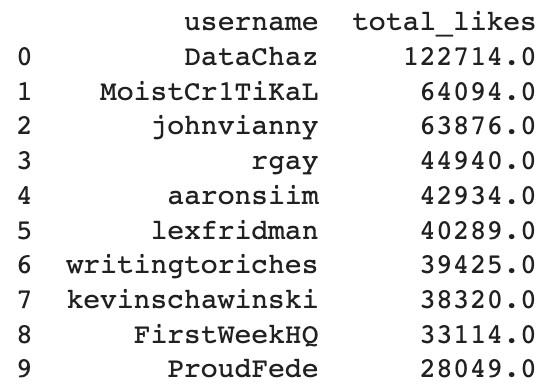
**Plotting** the top most liked tweets by username using a **bar chart** through Matplotlib



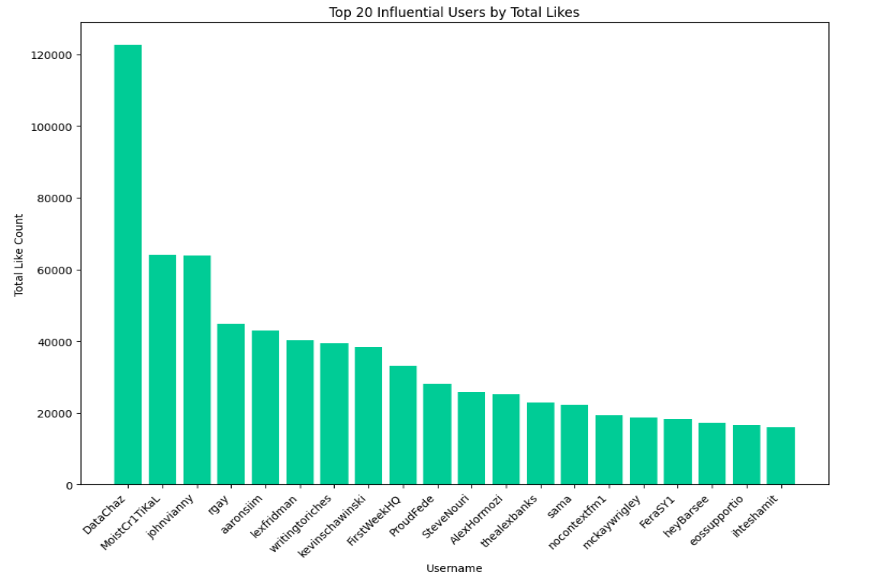
* + 1. The power users!

Most influential users by the total number of tweet likes and retweets

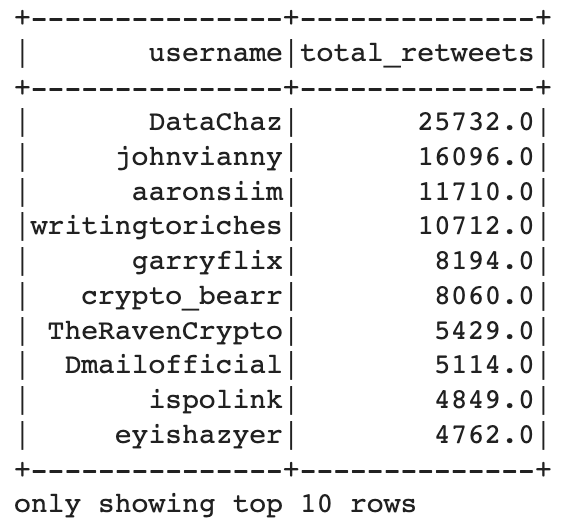
**Sorting** the DataFrame by **total likes** and printing top 10 -



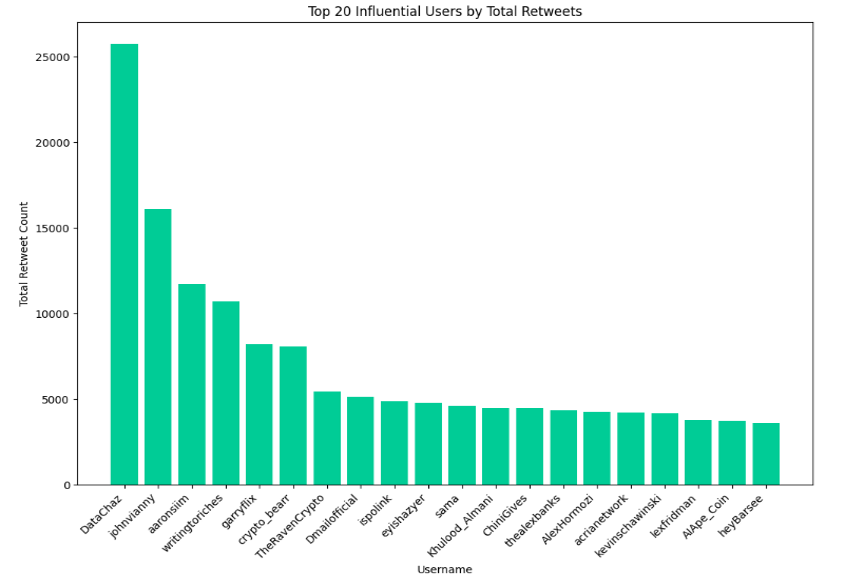
**Plotting** top influential users by **total\_likes** using a **bar chart** through Matplotlib



**Sorting** the DataFrame by **total retweets** and printing top 10 -



**Plotting** top influential users by **total\_retweets** using a **bar chart** through Matplotlib -

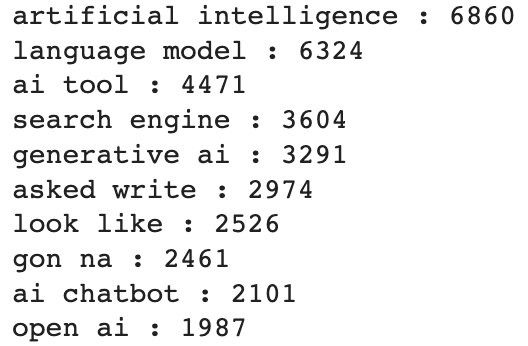


Discussion

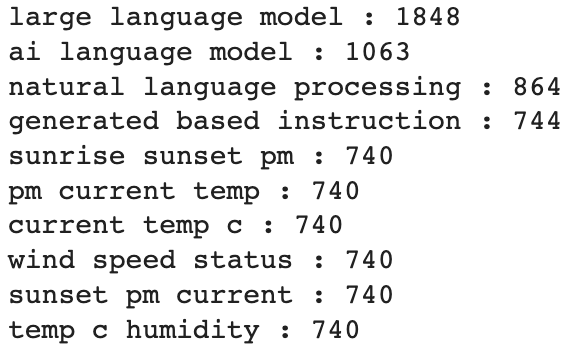
According to the analysis, the users with the most liked tweets were Moistcr1tikal, a live streamer known for his commentary videos, and lexfridman, a computer scientist and podcaster, and AI researcher. The users with the most likes were DataChazGPT, the developer of Streamlit, and moistcr1tikal. In the meantime, DataChazGPT has the most numbers of retweets. A commonality of these people with the most likes, followers, and retweets is that they specialize in technology or artificial intelligence-related fields and have already established a large follower base. The public generally accepts their comments and tweets because of their popularity.

1. **Text Analysis**
   1. *Top Bigrams and Trigrams*

Top 10 **bigrams** -



Top 10 **trigrams** -



Discussion

Bigrams and trigrams detect the frequency of two or three-word phrases in texts. While unsurprisingly, many phrases are related to artificial intelligence or natural language models; some key phrases are unexpected, especially from the trigram models. For example, phrases related to weather are mentioned frequently in this context. One assumption is that ChatGPT has integrated with a weather forecasting plugin, which has become a useful tool in people’s daily life.

* 1. *WordClouds*

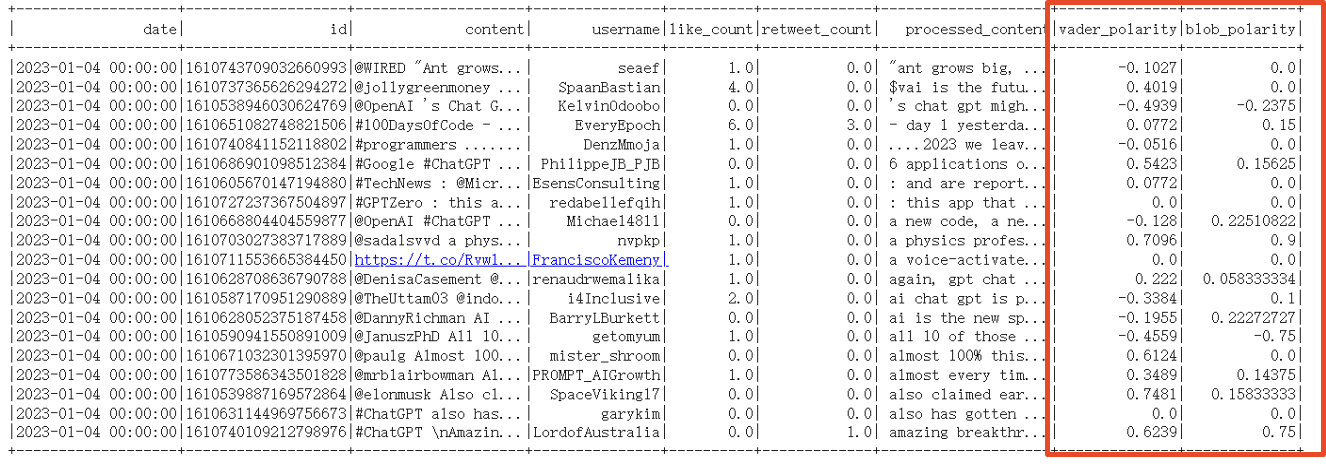
Plotting a word cloud of **top words used** (unigram) -



Discussion

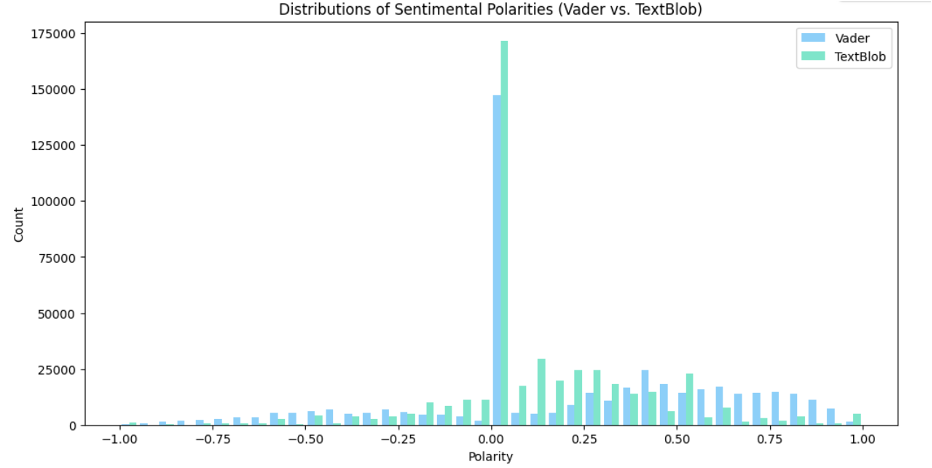
The word cloud shows keywords that most frequently occur in tweets. Many words related to machine learning, daily work routine, academics, innovation, and technology frontiers.

1. **Sentiment Analysis**
   1. *Sentiment polarity*

Using **Vader** and **text blob** polarity methods to generate polarity of tweet sentiment - 

* 1. *Sentiment Distribution*

Plotting a **histogram** of the **overall sentiment** polarities using both methods -



Discussion

Since Vader is a rule-based sentiment analysis tool designed specifically for social media text and capable of handling informal language, slang, and emojis, it is used for sentiment analysis in Twitter’s case. According to the distribution, most users have a neutral or positive sentiment toward ChatGPT.

# Conclusion

To conclude, our project used various data manipulation, data analysis, and text analysis techniques to analyze a Twitter dataset about ChatGPT. The analysis revealed several trends and insights, including popular topics related to artificial intelligence and machine learning, top influencers, and public attitudes towards ChatGPT. Through the use of sentiment analysis, the project was able to identify the predominant sentiment expressed in the tweets. These findings have several implications for the development and adoption of ChatGPT and its impact on society. Overall, this project demonstrates the effectiveness of using data analysis techniques to gain insights into large and complex datasets, and highlights the importance of understanding public attitudes and perceptions towards emerging technologies like ChatGPT.

# References

1. Ansari, K. (2023, April 11). *500k chatgpt-related tweets Jan-Mar 2023*. Kaggle. Retrieved May 7, 2023, from https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023
2. Ansari, K. (2023, April 19). *Cracking the CHATGPT code: A deep dive into 500,000 tweets using advanced NLP techniques*. Medium. Retrieved May 7, 2023, from https://medium.com/@ka2612/the-chatgpt-phenomenon-unraveling-insights-from-500-000-tweets-using-nlp-8ec0ad8ffd37
3. Blog. (n.d.). Retrieved May 7, 2023, from https://openai.com/blog
4. Hetler, A. (2023, March 17). *What is chatgpt? everything you need to know*. WhatIs.com. Retrieved May 7, 2023, from https://www.techtarget.com/whatis/definition/ChatGPT#:~:text=ChatGPT%20is%20a%20form%20of,request%20clarification%20to%20ChatGPT's%20replies
5. Hxycorn. (n.d.). *Hxycorn/twitter-sentiment-analysis-about-CHATGPT: A quantitative study on over 1.25 million tweets about chatgpt, employed data scrapping, data cleaning, EDA, topic modeling, and sentiment analysis.* GitHub. Retrieved May 7, 2023, from https://github.com/hxycorn/Twitter-Sentiment-Analysis-about-ChatGPT#background
6. Ruby, D., & About The Author Daniel Ruby Content writer with 10+ years of experience. I write across a range of subjects. (2023, May 2). *57+ CHATGPT statistics for 2023 (New Data + GPT-4 facts)*. Demand Sage. Retrieved May 7, 2023, from https://www.demandsage.com/chatgpt-statistics/