Twitter Text Pattern Report

The task of this report and machine learning algorithms is to find the most popular implicit topics in the twtters toward ChatGpt from January 2023 to March 2023, the pattern of the sentiments of people toward ChatGpt, How it differs across three months of 2023, the conversations around ChatGpt on Twitter and how it has changed over time.

The methods of machine learning algorithms implemented in this report include sentiment analysis, the unsupervised models Lda and Word2vec,and the semi-supervised model Corex based on the multiple tasks of this report and desired outcome of the algorithm.

1.Text Data

The source of the dataset comes from Kaggle, and it contains a csv file related to ChatGpt, including the keyword chatgpt,#hashtag and mentions about ChatGpt. The author of this dataset is Khalid AnsariA , who is a scholar from New York University and the Collaborators is Khalid Ansari. The number of tweets collected in the dataset includes information on 500,000 tweets. The timeframe started from January 4, 2023 to March 29th 2023. The unique accounts are tweets per uses, snscrape package was implemented during the process of the data collection and no null values identified. The initial dataset comprises 6 columns , “id”, “date”, “content”, “like_count”, “retweet_count” and 500031 rows. The datasets aim to determine AI-powered conversational technologies by analysing tweet volume, sentiments, and user engagement.

2.Data Pre-Processing

Six kinds of data cleaning methods were used, including punctuation, tokenization, lowercase, stopword removal, limmatization, and stemming. More specifically, I removed punctuation by using the remove_punct function with library re, removed URLS and emojis in tokenization using library re, removed stopwords using nltk, changed words to lowercase using lower function, Limmatization of Tweet using ntlk.WordNetLemmatizer (), removal of suffices, such as “ing”,”ly”,”s”, etc using PorterStemmer from the library NLTK for stemming.

3.Method

3.1 Sentiment Analysis

Converted the “date “column to datetime format, resampled the data by day and computd the sentiment count ,and plot the data using seaborn after resetting the index of the data frame and melt the sentiment count. Using SentimentIntensityAnalyzer to replace bigrams indicating ChatGpt Positive as cpos and ChatGpt negative as cneg, assigned polarity score as cpos = -3 and cneg = 3 and Defined sentiments based on intensity score, shown as Figure 2, showing overly positive: >0.75, Positive :between 0.05 &0.75, Overly negative :< 0.75,Negative :between -0.05 && -0.75, Neutral :- 0.05 to 0.05. In order to generate funnel chart of sentiment distribution (Figure 3&4 in Appendix), Setting feature of dataset “content_lemmatized”as value ,”Wordcloud” (shown in Figure 5) was used to quickly identified the most important themes in the large body of text.

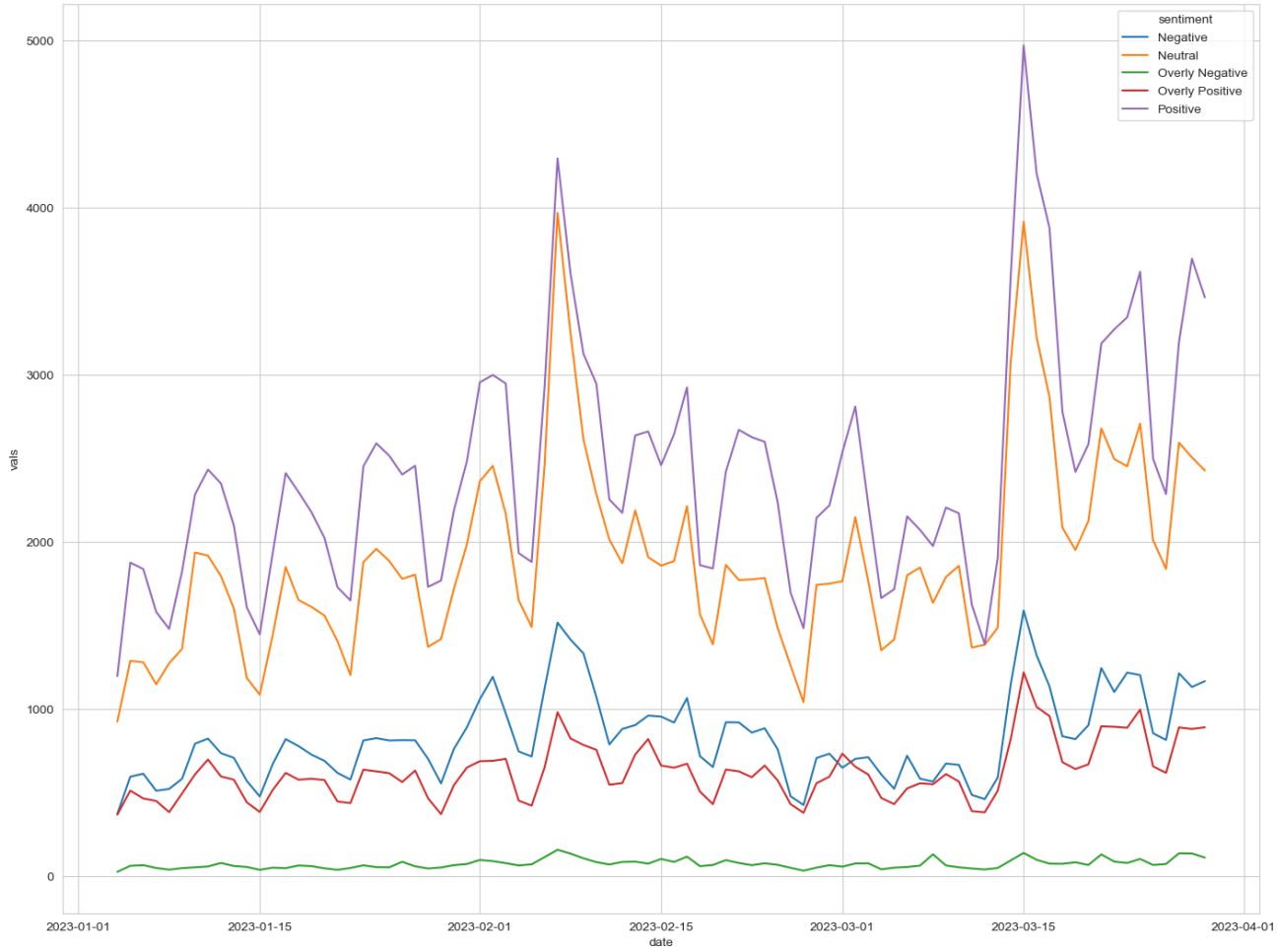
3.2 Unsupervised: Lda and Word2Vec in genism

Built Lda to extract top topics from the text, and computed a probability score for how likely a tweet belongs to each other's topic. Computed the total number of worked and unique keywords. In order to get most similar words related to"chatgpt", Word2Vec and keyedvectors from the genism model was used.

3.3 Semi-Supervised: CorEx

Transformed data into a sparse matrix using Countvectorizer. Set the number of topics as in hidden latent topics as 14 and feature: anchor_strength (how much weight CorEx puts towards maximizing mutual information between anchor words and their respective topics). Normalized topic correlations within individual documents explained by a particular topic. In order to generate normalized topic correlations to represent the correlations within an individual document explained by a particular topic (shown in Figure 6 in the Appendix).

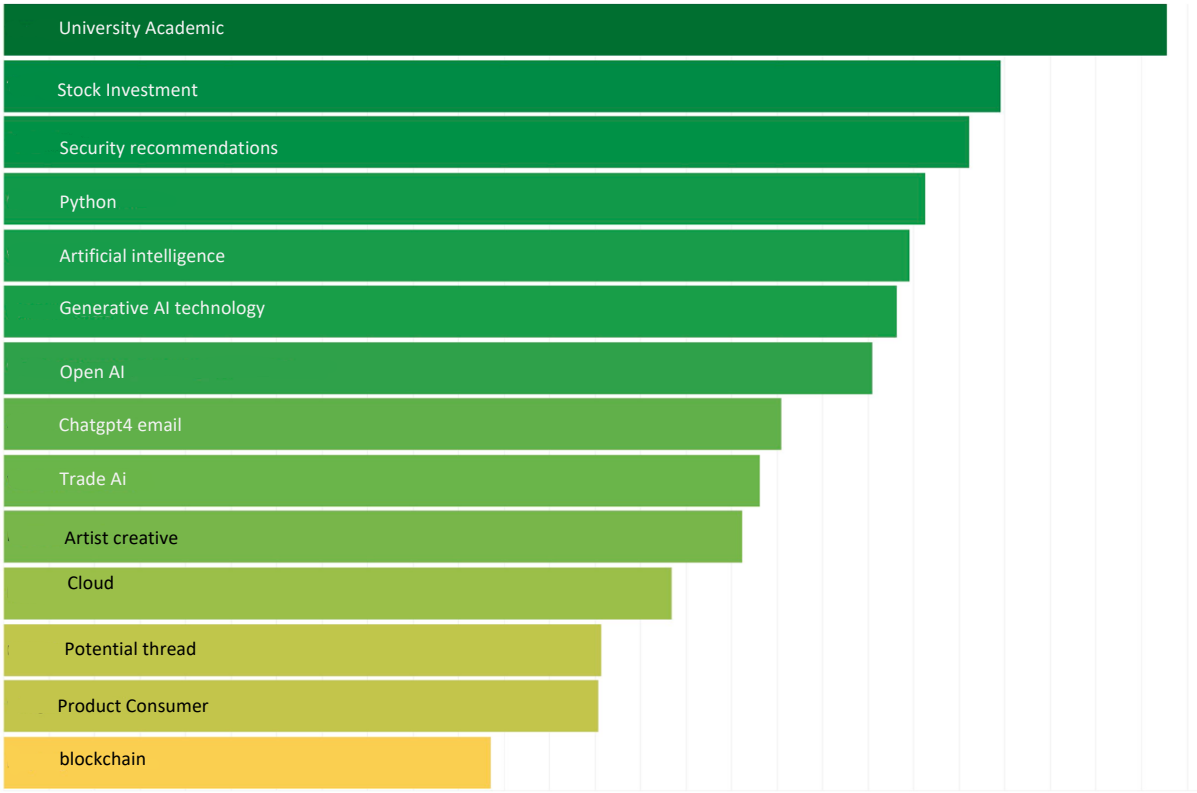
4.Evaluation and findings



Fingure1: Sentiment Analyse from Jan.1 2023 to April.1 2023

February had the lowest number of tweets as compared to January and March in the dataset. 7th February 2023 and 15th March had the highest number of tweets as compared to 26th January and 04th January(lowest) based on the information from Figure 8-9). The pattern of sentiment analysis from 04 January 2023 to 29th March 2023 showed fluctuations between "Neutral", "Negative",

“Overly Positive”, and “Positive”, with a on 7th Feb 2023.However, the category of “Overly Negative “remained consistently low throughout this period (shown in Figure 1).



Fingure2: Most popular topics between Jan 1 2023 to April 1 2023

Topics identified include “university academic”, “stock investment”, “security recommendations”, “python”, “artificial intelligence”, “generative ai technology”, “open AI”, “chatgpt 4 email”, “Trade AI”, “ Artist creative”, “Potential Thread”, “Cloud”, “Product consumer”, “Blockchain”, according to the Topic Correlation Table (Shown in Figure 6 and Figure 2).

The most related words to “ChatGpt” were “ai”, “open ai”, “chatgpts”, “software”, “brickwall”, “chatbot”, “technology” from word2vec model, and according to WorldCloud Image(Figure 5), the larger the word in the visual, the more common the word was in the tweet text, “chatgpt”, “open ai”, “artificial intelligence” appear frequently.

The Pearson correlation coefficient between likes and retweets is 0.73, indicating that there is a significant positive relationship between likes and retweets based on the regression plot(Figure 7 in the Appendix).

Conclusion

Results obtained are based on collected data over three months. Positive sentiment contributed the most in overall sentiment(208066), followed by neutral(162388) and negative(70386) sentiments. People were discussing most about topics such as: utilized in university academic, stock investment and security recommendation, Topics that remained underrepresented include artist creative and Cloud. These findings can help understand specific topics and human sentiments creating greater traction ,and how ChatGpt could be utilized in artistic creativity and advance Cloud could be taken into higher consideration by developers and stakeholders.

Appendix

	sentiment	content_lemmatized
4	Positive	208066
1	Neutral	162388
0	Negative	70387
3	Overly Positive	52853
2	Overly Negative	6342

sentiment	like_count	retweet_count	score
Negative	6.733516	1.185119	-0.376120
Neutral	5.543252	1.019565	0.000151
Overly Negative	4.888363	0.972564	-0.827867
Overly Positive	7.651978	1.816832	0.837100
Positive	8.421495	1.872544	0.446248

Figure 2: Sentiment Score

Figure 3:Funnel-Cahrt of Sentiment Distribution

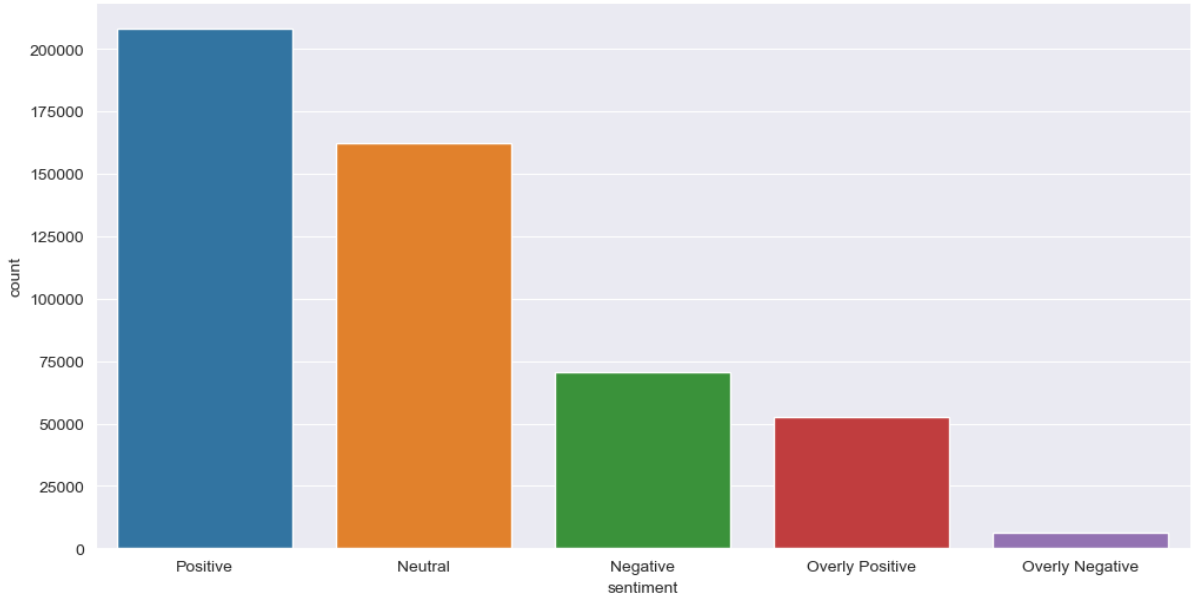


Figure 4: Sentiment Diagram

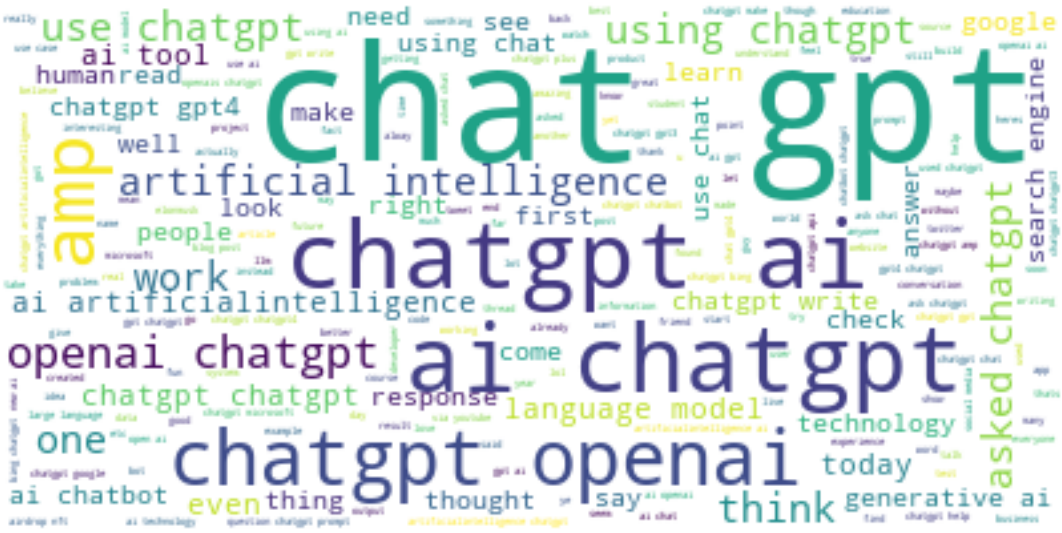


Figure 5: WorldCloud Image

```
0: tool, amp, learning, student, education, school, teacher, essay, edtech, use
1: human, data, look, news, text, research, stock, test, investment, exam
2: asked, answer, question, im, good, ask, code, response, got, problem
3: new, artificialintelligence, google, microsoft, tech, bing, search, bard, company, app
4: tech, based, generated, current, speed, china, instruction, pm, browser, edge
5: use, using, tool, make, way, people, help, used, data, year
6: read, check, article, artificial, midjourney, generativeai, story, generated, blog, written
7: artificialintelligence, tool, model, ai, language, content, gpt3, machinelearning, chatgpt, text
8: crypto, nft, web3, bitcoin, airdrop, blockchain, btc, eth, token, invest
9: artificialintelligence, tech, machinelearning, web3, innovation, python, coding, ml, programming
metaverse
10: microsoft, bing, search, free, based, engine, stock, trading, option, buy
11: like, new, time, im, world, thing, day, let, going, great
12: like, make, know, think, work, people, need, human, thing, say
13: new, gpt4, microsoft, model, business, gpt3, version, service, access, api
```

Figure6 : Topic Correlation

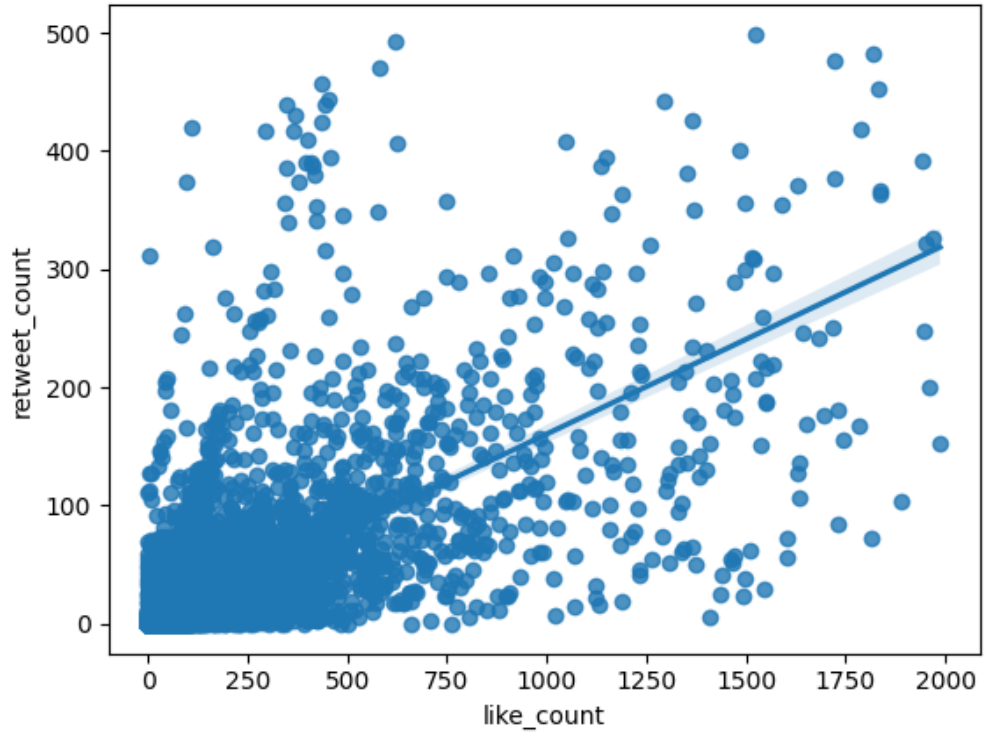


Figure 7: Regression plot

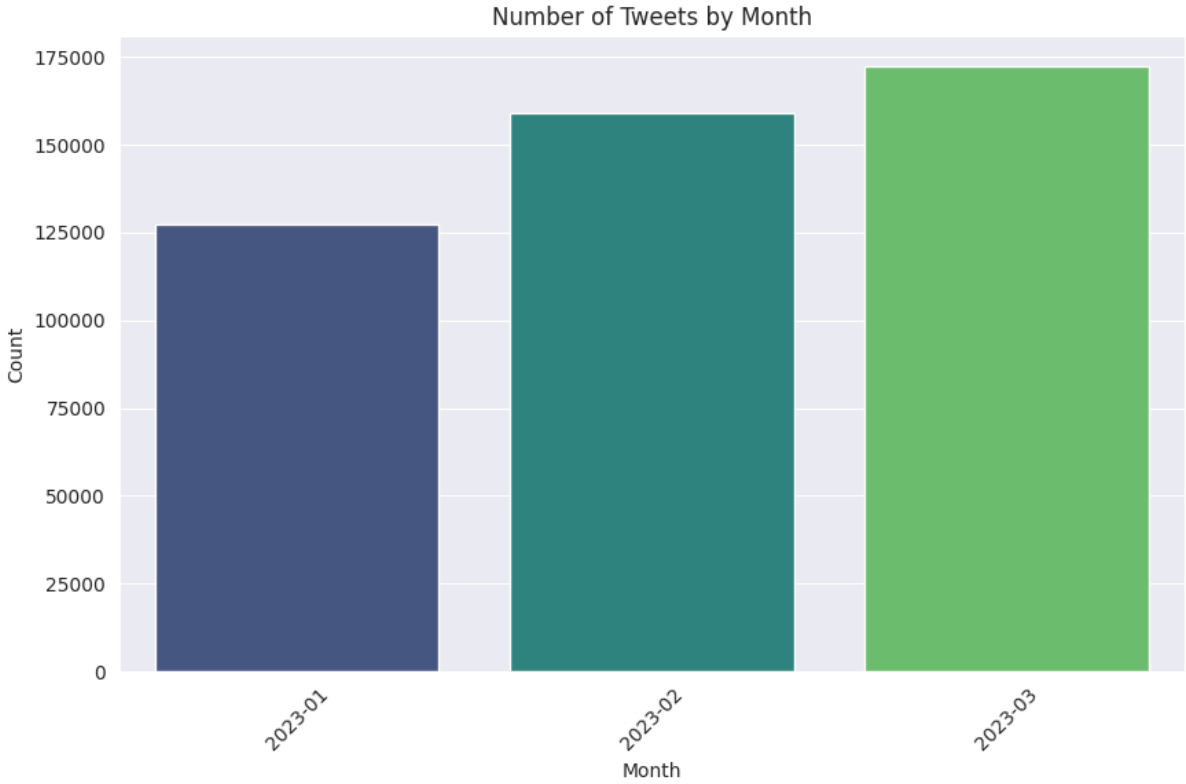


Figure 8: Number of Tweets by Month

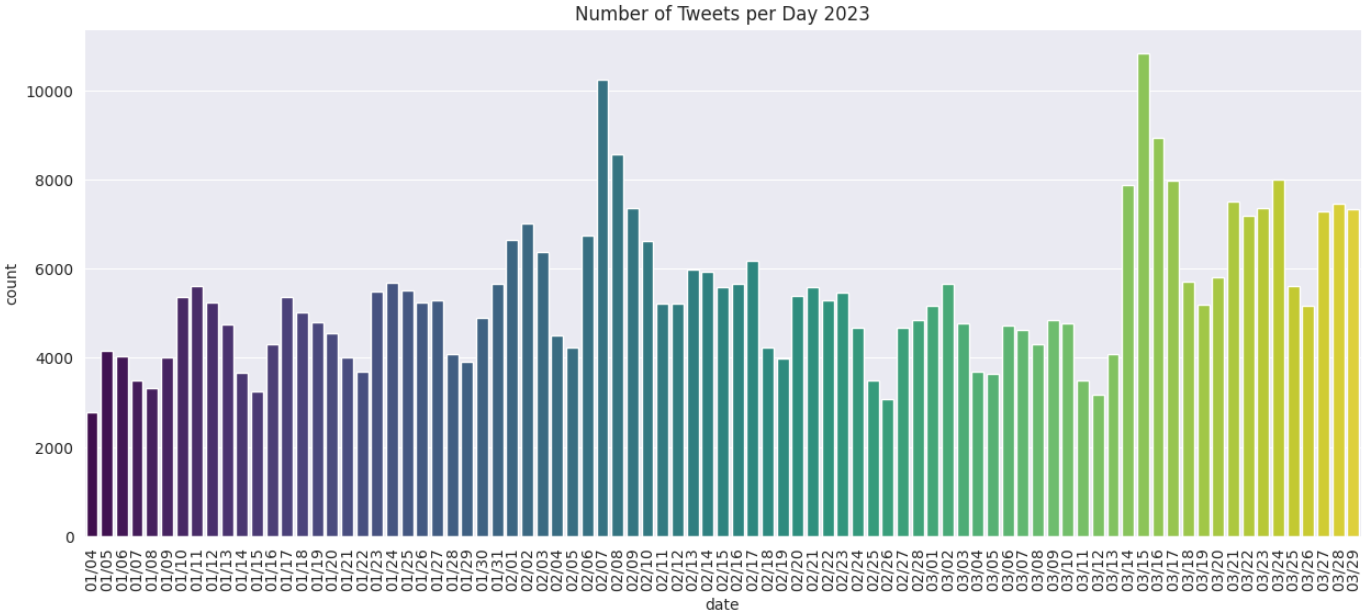


Figure 9: Number of Tweets by per Day 2023

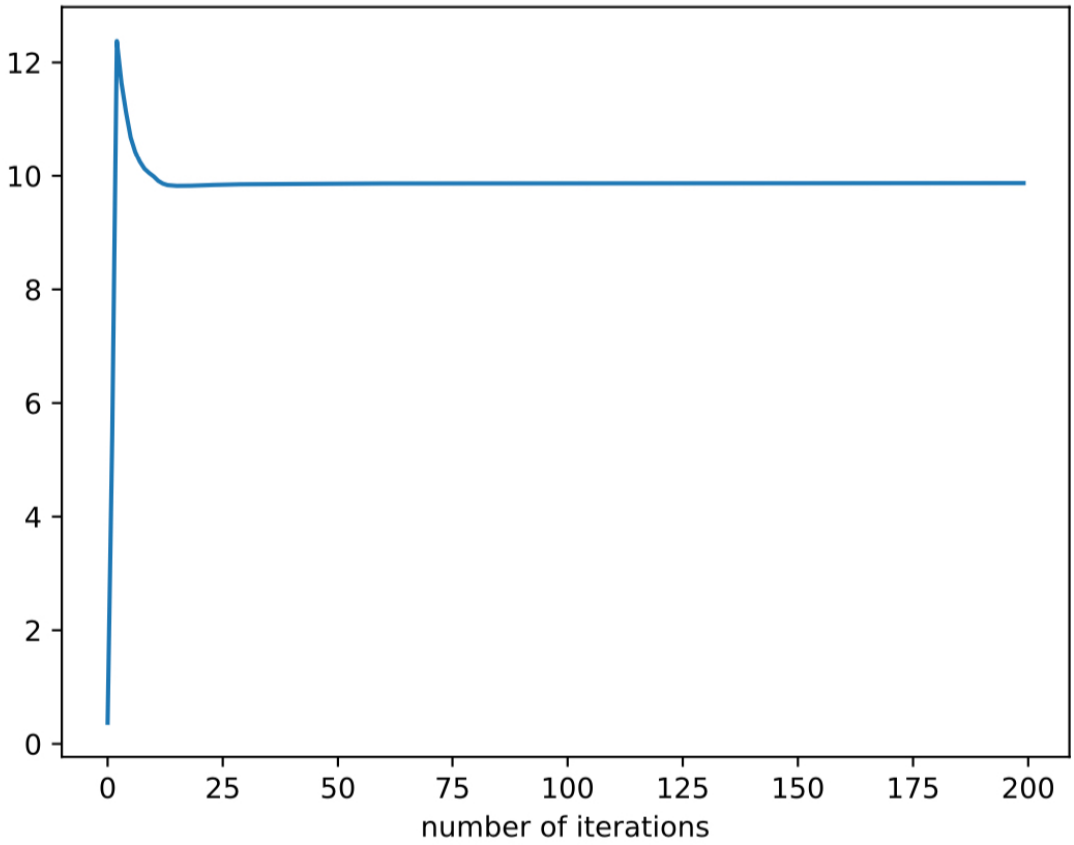


Figure 10: Number of iterations

References

Portugal, I., Alencar, P. S. C., & Cowan, D. D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems With Applications*, *97*, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020

Ayodele, T. (2010). Types of Machine Learning Algorithms. In InTech eBooks. https://doi.org/10.5772/9385

Scaria, V., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S. C., Mujumdar, S., Afzal, S., Mittal, R. S., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. https://doi.org/10.1145/3394486.3406477

Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2013.6726818

Agarwal, B., & Mittal, N. (2016). Machine Learning Approach for Sentiment Analysis. In Socio-affective computing (pp. 21–45). Springer International Publishing. https://doi.org/10.1007/978-3-319-25343-5_3

Hasan, A. N., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine Learning-Based Sentiment Analysis for Twitter Accounts. Mathematical and Computational Applications, 23(1), 11. https://doi.org/10.3390/mca23010011

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., & Qu, H. (2010). Context preserving dynamic word cloud visualization. https://doi.org/10.1109/pacificvis.2010.5429600

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. https://doi.org/10.1109/hicss.2014.231

Kumar, P. S., Gupta, D. B., Naha, T. K., & Gupta, S. (2006). Factors affecting fuel rate in Corex process. Ironmaking & Steelmaking, 33(4), 293–298. https://doi.org/10.1179/174328106x101493

Codes From Jupyter

In [2]:
```
import pandas as pd
import numpy as np
import re
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import string
import sys,csv,re
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
from wordcloud import WordCloud
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, ConfusionMatrixDisplay
```
In [3]:
```
df = pd.read_csv('/Users/luqiansong/Desktop/TwitterJanMar.csv')
```
In [4]:
```
df.head()
```
Out[4]:

	date	id	content	username	like_count	retweet_count
0	2023-03-29 22:58:21+00:00	1641213230730051584	Free AI marketing and automation tools, strate...	RealProfitPros	0.0	0.0
1	2023-03-29 22:58:18+00:00	1641213218520481805	@MecoleHardman 4 Chat GPT says it's 15. 😂	AmyLouWho321	0.0	0.0
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/FjJSprt0te - Chat with any PDF!\n...	yjleon1976	0.0	0.0
3	2023-03-29 22:57:52+00:00	1641213110915571715	AI muses: "In the court of life, we must all f...	ChatGPT_Thinks	0.0	0.0
4	2023-03-29 22:57:26+00:00	1641213003260633088	Most people haven't heard of Chat GPT yet.\nFi...	nikocosmonaut	0.0	0.0

In [5]:
```
#getting basic information about datasets
df.info()
```
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500036 entries, 0 to 500035
Data columns (total 6 columns):
 #   Column        Non-Null Count   Dtype  
---  ------        --------------   -----  
 0   date          500036 non-null  object 
 1   id            500030 non-null  object 
 2   content       500030 non-null  object 
 3   username      500002 non-null  object 
 4   like_count    499974 non-null  float64
 5   retweet_count  499974 non-null  float64
dtypes: float64(2), object(4)
memory usage: 22.9+ MB
```
In [6]:
```
df.shape
```
Out[6]:
```
(500036, 6)
```
In [7]:
```
df.isnull().sum()
```
Out[7]:
```
date             0
id               6
content          6
username        34
like_count      62
retweet_count   62
dtype: int64
```
In [8]:
```
df.columns
```
Out[8]:
```
Index(['date', 'id', 'content', 'username', 'like_count', 'retweet_count'], dtype='object')
```
In [9]:
```
df.reset_index(drop=True)
```
Out[9]:

	date	id	content	username	like_count	retweet_count
0	2023-03-29 22:58:21+00:00	1641213230730051584	Free AI marketing and automation tools, strate...	RealProfitPros	0.0	0.0

	date	id	content	username	like_count	retweet_count
1	2023-03-29 22:58:18+00:00	1641213218520481805	@MecoleHardman4 Chat GPT says it's 15. 😂	AmyLouWho321	0.0	0.0
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/FjJSprt0te - Chat with any PDF!\n...	yjleon1976	0.0	0.0
3	2023-03-29 22:57:52+00:00	1641213110915571715	AI muses: "In the court of life, we must all f...	ChatGPT_Thinks	0.0	0.0
4	2023-03-29 22:57:26+00:00	1641213003260633088	Most people haven't heard of Chat GPT yet.\nFi...	nikocosmonaut	0.0	0.0
...	...	...	...	...	...	...
500031	2023-01-04 07:18:08+00:00	1610536038094757888	@GoogleAI #LAMDA Versus @OpenAI #ChatGPT ?! Wh...	Pup_In_Cup	1.0	0.0
500032	2023-01-04 07:17:50+00:00	1610535961670172674	#ChatGPT \n\nSo much #Censorship.\n\nNever tru...	TryingToOffend	2.0	0.0
500033	2023-01-04 07:17:20+00:00	1610535837363486720	all my twitter feed is about ChatGPT and @Open...	mcp350	3.0	1.0
500034	2023-01-04 07:17:08+00:00	1610535786017091584	I'm quite amazed by Chat GPT. A really promisi...	manumurali369	1.0	0.0
500035	2023-01-04 07:16:56+00:00	1610535734758219778	I used chat gpt to get gym workout program and...	pnik91	0.0	0.0

500036 rows × 6 columns

In [11]:

```
#All LowerCase

def converter(x):
    try:
        return ' '.join([x.lower() for x in str(x).split() if x not in stop_words])
    except AttributeError:
        return None
df['content'] = df['content'].apply(converter)
```

In [12]:

```
df.head()
```

Out[12]:

	date	id	content	username	like_count	retweet_count
0	2023-03-29 22:58:21+00:00	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0
1	2023-03-29 22:58:18+00:00	1641213218520481805	@mecolehardman 4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0

	date	id	content	username	like_count	retweet_count
3	2023-03-29 22:57:52+00:00	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0
4	2023-03-29 22:57:26+00:00	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0

In [13]:
```
#Removing Punctuation
df['content_punc'] =df['content'].str.replace('[^\w\s]','')
df.head()
```

```
/var/folders/m_/m3lsq_594494n7k5zm6nmdtc0000gn/T/ipykernel_40873/787010310.py:2: FutureWarning:
The default value of regex will change from True to False in a future version.
  df['content_punc'] =df['content'].str.replace('[^\w\s]','')
```

Out[13]:

	date	id	content	username	like_count	retweet_count	content_punc
0	2023-03-29 22:58:21+00:00	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...
1	2023-03-29 22:58:18+00:00	1641213218520481805	@mecolehardman4 chat gpt says it’s 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...
3	2023-03-29 22:57:52+00:00	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny...
4	2023-03-29 22:57:26+00:00	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0	people heard chat gpt yet first elite factions...

In [14]:
```
#Removal of stop words
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stop = stopwords.words('english')
df['content_stop']  = df['content_punc'] .apply(lambda x: " ".join(x for x in x.split() if x not in stop))
df.head()
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/luqiansong/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[14]:

	date	id	content	username	like_count	retweet_count	content_punc	content_stop
0	2023-03-29 22:58:21+00:00	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...

	date	id	content	username	like_count	retweet_count	content_punc	content_stop
1	2023-03-29 22:58:18+00:00	1641213218520481805	@mecolehardman4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...
3	2023-03-29 22:57:52+00:00	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny...	ai muses court life must face judge destiny ju...
4	2023-03-29 22:57:26+00:00	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0	people heard chat gpt yet first elite factions...	people heard chat gpt yet first elite factions...

In [15]:
```
#Tokenization of Tweets
import textblob
from textblob import TextBlob
def tokenization(content):
    content = re.split('\W+', content)
    return content
df['content_tokenized'] = df['content_stop'].apply(lambda x: tokenization(x.lower()))
df[['content', 'content_punc', 'content_stop', 'content_tokenized']][0:9]
```

Out[15]:

	content	content_punc	content_stop	content_tokenized
0	free ai marketing automation tools, strategies...	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	[free, ai, marketing, automation, tools, strat...
1	@mecolehardman4 chat gpt says it's 15. 😂	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15	[mecolehardman4, chat, gpt, says, 15]
2	https://t.co/fjjsprt0te - chat pdf! check new ...	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	[httpstcofjjsprt0te, chat, pdf, check, new, ai...
3	ai muses: "in court life, must face judge dest...	ai muses in court life must face judge destiny...	ai muses court life must face judge destiny ju...	[ai, muses, court, life, must, face, judge, de...
4	people heard chat gpt yet. first, elite factio...	people heard chat gpt yet first elite factions...	people heard chat gpt yet first elite factions...	[people, heard, chat, gpt, yet, first, elite, ...
5	@nytimes no! chat gpt putting together amazing...	nytimes no chat gpt putting together amazing r...	nytimes chat gpt putting together amazing recipes	[nytimes, chat, gpt, putting, together, amazin...
6	@ylzkrtt yes also chat gpt make generative art...	ylzkrtt yes also chat gpt make generative art ...	ylzkrtt yes also chat gpt make generative art ...	[ylzkrtt, yes, also, chat, gpt, make, generati...
7	@robinhanson @razibkhan people heard chat gpt ...	robinhanson razibkhan people heard chat gpt ye...	robinhanson razibkhan people heard chat gpt ye...	[robinhanson, razibkhan, people, heard, chat, ...
8	robotically - shaun usher - letters note | thi...	robotically shaun usher letters note think ...	robotically shaun usher letters note think cha...	[robotically, shaun, usher, letters, note, thi...

In [16]:
```
#Lemmatization is a more effective option than stemming because it converts the word into its root word,
```

```
#rather than just stripping the suffices.
#nltk.download('wordnet')
wn = nltk.WordNetLemmatizer()
def lemmatizer(content):
    content = [wn.lemmatize(word) for word in content]
    return content
df['content_lemmatized'] = df['content_tokenized'].apply(lambda x: lemmatizer(x))
df[['content', 'content_punc', 'content_tokenized','content_stop','content_lemmatized']][0:9]
```

Out[16]:

	content	content_punc	content_tokenized	content_stop	content_lemmatized
0	free ai marketing automation tools, strategies...	free ai marketing automation tools strategies ...	[free, ai, marketing, automation, tools, strat...	free ai marketing automation tools strategies ...	[free, ai, marketing, automation, tool, strate...
1	@mecolehardman4 chat gpt says it’s 15. 😂	mecolehardman4 chat gpt says its 15	[mecolehardman4, chat, gpt, says, 15]	mecolehardman4 chat gpt says 15	[mecolehardman4, chat, gpt, say, 15]
2	https://t.co/fjjsprt0te - chat pdf! check new ...	httpstcofjjsprt0te chat pdf check new ai quic...	[httpstcofjjsprt0te, chat, pdf, check, new, ai...	httpstcofjjsprt0te chat pdf check new ai quick...	[httpstcofjjsprt0te, chat, pdf, check, new, ai...
3	ai muses: "in court life, must face judge dest...	ai muses in court life must face judge destiny...	[ai, muses, court, life, must, face, judge, de...	ai muses court life must face judge destiny ju...	[ai, mus, court, life, must, face, judge, dest...
4	people heard chat gpt yet. first, elite factio...	people heard chat gpt yet first elite factions...	[people, heard, chat, gpt, yet, first, elite, ...	people heard chat gpt yet first elite factions...	[people, heard, chat, gpt, yet, first, elite, ...
5	@nytimes no! chat gpt putting together amazing...	nytimes no chat gpt putting together amazing r...	[nytimes, chat, gpt, putting, together, amazin...	nytimes chat gpt putting together amazing recipes	[nytimes, chat, gpt, putting, together, amazin...
6	@ylzkrtt yes also chat gpt make generative art...	ylzkrtt yes also chat gpt make generative art ...	[ylzkrtt, yes, also, chat, gpt, make, generati...	ylzkrtt yes also chat gpt make generative art ...	[ylzkrtt, yes, also, chat, gpt, make, generati...
7	@robinhanson @razibkhan people heard chat gpt ...	robinhanson razibkhan people heard chat gpt ye...	[robinhanson, razibkhan, people, heard, chat, ...	robinhanson razibkhan people heard chat gpt ye...	[robinhanson, razibkhan, people, heard, chat, ...
8	robotically - shaun usher - letters note | thi...	robotically shaun usher letters note think ...	[robotically, shaun, usher, letters, note, thi...	robotically shaun usher letters note think cha...	[robotically, shaun, usher, letter, note, thin...

In [17]:

```
df.drop(columns=['content', 'content_punc', 'content_tokenized', 'content_stop'])
```

Out[17]:

	date	id	username	like_count	retweet_count	content_lemmatized
0	2023-03-29 22:58:21+00:00	1641213230730051584	RealProfitPros	0.0	0.0	[free, ai, marketing, automation, tool, strate...
1	2023-03-29 22:58:18+00:00	1641213218520481805	AmyLouWho321	0.0	0.0	[mecolehardman4, chat, gpt, say, 15]
2	2023-03-29 22:57:53+00:00	1641213115684536323	yjleon1976	0.0	0.0	[httpstcofjjsprt0te, chat, pdf, check, new, ai...

	date	id	username	like_count	retweet_count	content_lemmatized
3	2023-03-29 22:57:52+00:00	1641213110915571715	ChatGPT_Thinks	0.0	0.0	[ai, mus, court, life, must, face, judge, dest...
4	2023-03-29 22:57:26+00:00	1641213003260633088	nikocosmonaut	0.0	0.0	[people, heard, chat, gpt, yet, first, elite, ...
...	...	...	...	...	...	...
500031	2023-01-04 07:18:08+00:00	1610536038094757888	Pup_In_Cup	1.0	0.0	[googleai, lamda, versus, openai, chatgpt, car...
500032	2023-01-04 07:17:50+00:00	1610535961670172674	TryingToOffend	2.0	0.0	[chatgpt, much, censorship, never, trust, syst...
500033	2023-01-04 07:17:20+00:00	1610535837363486720	mcp350	3.0	1.0	[twitter, feed, chatgpt, openai, lol, ai, chat...
500034	2023-01-04 07:17:08+00:00	1610535786017091584	manumurali369	1.0	0.0	[im, quite, amazed, chat, gpt, really, promisi...
500035	2023-01-04 07:16:56+00:00	1610535734758219778	pnik91	0.0	0.0	[used, chat, gpt, get, gym, workout, program, ...

500036 rows × 6 columns

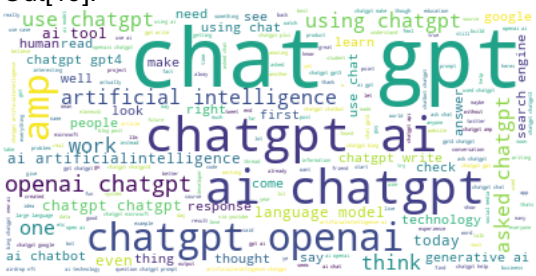
In [18]:

```
df.head()
df.to_csv('Tweets_Cleaned.csv')
```

In [19]:

```
# Import the wordcloud library
from wordcloud import WordCloud
# Join the different processed titles together.
long_string = ' '.join(list(df['content_stop'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000, contour_width=3,
contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

Out[19]:



In [20]:

```
#Stemming refers to the removal of suffices, like “ing”, “ly”, “s”, etc. by a simple rule-based approach.
#For this purpose, we will use PorterStemmer from the NLTK library.
from nltk.stem import PorterStemmer
ps = PorterStemmer()
def stemming(content):
    content = [ps.stem(word) for word in content]
    return content
df['content_stemmed'] = df['content_tokenized'].apply(lambda x: stemming(x))
```

df[['content', 'content_punc', 'content_tokenized','content_stop','content_stemmed']][0:9]
Out[20]:

	content	content_punc	content_tokenized	content_stop	content_stemmed
0	free ai marketing automation tools, strategies...	free ai marketing automation tools strategies ...	[free, ai, marketing, automation, tools, strat...	free ai marketing automation tools strategies ...	[free, ai, market, autom, tool, strategi, coll...
1	@mecolehardman4 chat gpt says it's 15. 😂	mecolehardman4 chat gpt says its 15	[mecolehardman4, chat, gpt, says, 15]	mecolehardman4 chat gpt says 15	[mecolehardman4, chat, gpt, say, 15]
2	https://t.co/fjjsprt0te - chat pdf! check new ...	httpstcofjjsprt0te chat pdf check new ai quic...	[httpstcofjjsprt0te, chat, pdf, check, new, ai...	httpstcofjjsprt0te chat pdf check new ai quick...	[httpstcofjjsprt0t, chat, pdf, check, new, ai,...
3	ai muses: "in court life, must face judge dest...	ai muses in court life must face judge destiny...	[ai, muses, court, life, must, face, judge, de...	ai muses court life must face judge destiny ju...	[ai, muse, court, life, must, face, judg, dest...
4	people heard chat gpt yet. first, elite factio...	people heard chat gpt yet first elite factions...	[people, heard, chat, gpt, yet, first, elite, ...	people heard chat gpt yet first elite factions...	[peopl, heard, chat, gpt, yet, first, elit, fa...
5	@nytimes no! chat gpt putting together amazing...	nytimes no chat gpt putting together amazing r...	[nytimes, chat, gpt, putting, together, amazin...	nytimes chat gpt putting together amazing recipes	[nytim, chat, gpt, put, togeth, amaz, recip]
6	@ylzkrtt yes also chat gpt make generative art...	ylzkrtt yes also chat gpt make generative art ...	[ylzkrtt, yes, also, chat, gpt, make, generati...	ylzkrtt yes also chat gpt make generative art ...	[ylzkrtt, ye, also, chat, gpt, make, gener, ar...
7	@robinhanson @razibkhan people heard chat gpt ...	robinhanson razibkhan people heard chat gpt ye...	[robinhanson, razibkhan, people, heard, chat, ...	robinhanson razibkhan people heard chat gpt ye...	[robinhanson, razibkhan, peopl, heard, chat, g...
8	robotically - shaun usher - letters note | thi...	robotically shaun usher letters note think ...	[robotically, shaun, usher, letters, note, thi...	robotically shaun usher letters note think cha...	[robot, shaun, usher, letter, note, think, cha...

In [47]:
```
from gensim.models import Word2Vec,KeyedVectors
```
In [49]:
```
model=Word2Vec(df.content_lemmatized,min_count=1,vector_size=32)
```
In [53]:
```
model.wv.most_similar("chatgpt")
```
Out[53]:
```
[('ai', 0.8875566720962524),
 ('openai', 0.801838219165802),
 ('chatgpts', 0.7576317191123962),
 ('here', 0.7523341774940491),
 ('software', 0.7216390371322632),
 ('chatgpt4', 0.709638774394989),
 ('brickwall', 0.7091189026832581),
 ('chatbot', 0.7029999494552612),
 ('openais', 0.6793273687362671),
 ('technology', 0.6696925759315491)]
```
In [90]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import re
import string
import nltk
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('darkgrid')
```

```
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer as SIA
from wordcloud import WordCloud,STOPWORDS
plt.rc('figure',figsize=(17,13))
import plotly.express as px
import plotly.graph_objs as go
import plotly.offline as pyo
from plotly.subplots import make_subplots
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/luqiansong/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
In [92]:
df['content_lemmatized'] = df['content'].str.replace('[^\w\s]','')
/var/folders/m_/m3lsq_594494n7k5zm6nmdtc0000gn/T/ipykernel_40873/2903736268.py:1: FutureWarning:
The default value of regex will change from True to False in a future version.
In [95]:
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
new_words = {
    'cpos': -3.0,
    'cneg': 3.0,
}
analyser = SentimentIntensityAnalyzer()
analyser.lexicon.update(new_words)
scores=[]
for i in range(len(df['content_lemmatized'])):

    score = analyser.polarity_scores(df['content_lemmatized'][i])
    score=score['compound']
    scores.append(score)
sentiment=[]
for i in scores:
    if i>=0.75:
        sentiment.append('Overly Positive')
    elif (i>=0.05) and (i<0.75):
        sentiment.append('Positive')
    elif i<=(-0.75):
        sentiment.append('Overly Negative')
    elif (i<=-0.05) and (i>-0.75):
        sentiment.append('Negative')
    else:
        sentiment.append('Neutral')
df['sentiment']= pd.Series(np.array(sentiment))
In [96]:
df['score']= pd.Series(np.array(scores))
In [97]:
df.head()
Out[97]:
```

	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	content_stemmed	sentiment	score
0	2023-03-29 22:58:21+00:00	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	[free, ai, marketing, automation, tools, strat...	free ai marketing automation tools strategies ...	[free, ai, market, autom, tool, strategi, coll...	Positive	0.51106
1	2023-03-29 22:58:	1641213218520481805	@mecolehardman4 chat	AmyLouWho321	0.0	0.0	mecolehardman4 chat	mecolehardman4 chat	[mecolehardman4,	mecolehardman4 chat gpt	[mecolehardman4,	Neutral	0.00000

	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	content_stemmed	sentiment	score
	18+00:00		gpt says it's 15. 😂				gpt says its 15	gpt says 15	chat, gpt, says, 15]	says its 15	chat, gpt, say, 15]		
2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	[httpstcofjjsprt0te, chat, pdf, check, new, ai...	httpstcofjjsprt0te chat pdf check new ai quic...	[httpstcofjjsprt0t, chat, pdf, check, new, ai,...	Positive	0.7184
3	2023-03-29 22:57:52+00:00	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny ...	ai muses court life must face judge destiny ju...	[ai, muses, court, life, must, face, judge, de...	ai muses in court life must face judge destiny...	[ai, muse, court, life, must, face, judg, dest...	Neutral	0.00000
4	2023-03-29 22:57:26+00:00	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0	people heard chat gpt yet first elite factions...	people heard chat gpt yet first elite factions...	[people, heard, chat, gpt, yet, first, elite, ...	people heard chat gpt yet first elite factions...	[peopl, heard, chat, gpt, yet, first, elit, fa...	Neutral	0.0258

In [98]:
df.shape
Out[98]:
(500036, 13)
In [100]:
df.groupby(by="sentiment").mean()
Out[100]:

	like_count	retweet_count	score
sentiment			
Negative	6.733516	1.185119	-0.376120
Neutral	5.543252	1.019565	0.000151
Overly Negative	4.888363	0.972564	-0.827867
Overly Positive	7.651978	1.816832	0.837100
Positive	8.421495	1.872544	0.446248

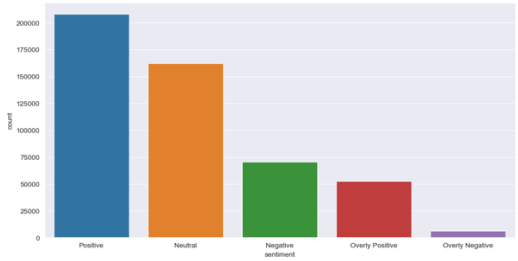
In [101]:
temp = df.groupby('sentiment').count()['content_lemmatized'].reset_index().sort_values(by='content_lemmatized',ascending=False)
temp.style.background_gradient(cmap='Purples')
Out[101]:

	sentiment	content_lemmatized
4	Positive	208066

	sentiment	content_lemmatized
1	Neutral	162388
0	Negative	70387
3	Overly Positive	52853
2	Overly Negative	6342

In [113]:

```
plt.figure(figsize=(12,6))
sns.countplot(x='sentiment',data=df)
fig = go.Figure(go.Funnelarea(
    text =temp.sentiment,
    values = temp.content_lemmatized,
    title = {"position": "top center", "text": "Funnel-Chart of Sentiment Distribution"}
    ))
fig.show()
```



In [176]:

```
import pandas as pd
import time
import datetime
import calendar
import seaborn as sns
import matplotlib.pyplot as plt
from datetime import datetime
```

In [207]:

```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
new_words = {
    'cpos': -3.0,
    'cneg': 3.0,
}
analyser = SentimentIntensityAnalyzer()
analyser.lexicon.update(new_words)
scores=[]
for i in range(len(df['content_lemmatized'])):

    score = analyser.polarity_scores(df['content_lemmatized'][i])
    score=score['compound']
    scores.append(score)
sentiment=[]
for i in scores:
    if i>=0.75:
        sentiment.append('Overly Positive')
    elif (i>=0.05) and (i<0.75):
        sentiment.append('Positive')
    elif i<=(-0.75):
        sentiment.append('Overly Negative')
    elif (i<=-0.05) and (i>-0.75):
        sentiment.append('Negative')
    else:
        sentiment.append('Neutral')
```

df['sentiment']= pd.Series(np.array(sentiment))
df['score']= pd.Series(np.array(scores))
df.head()
Out[207]:

	Unnamed: 0	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	sentiment	score
0	0	2023-03-29 22:58:21+00:00	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...	Positive	0.5106
1	1	2023-03-29 22:58:18+00:00	1641213218520481805	@mecolehardman4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15	['mecolehardman4', 'chat', 'gpt', 'says', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']	Neutral	0.0000
2	2	2023-03-29 22:57:53+00:00	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	Positive	0.7184
3	3	2023-03-29 22:57:52+00:00	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny. ..	ai muses court life must face judge destiny ju...	['ai', 'muses', 'court', 'life', 'must', 'face...	['ai', 'mus', 'court', 'life', 'must', 'face',...	Neutral	0.0000
4	4	2023-03-29 22:57:26+00:00	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0	people heard chat gpt yet first elite factions ...	people heard chat gpt yet first elite factions ...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	Neutral	0.0258

In [208]:
df.date
Out[208]:
0 2023-03-29 22:58:21+00:00
1 2023-03-29 22:58:18+00:00
2 2023-03-29 22:57:53+00:00
3 2023-03-29 22:57:52+00:00
4 2023-03-29 22:57:26+00:00
...
500031 2023-01-04 07:18:08+00:00
500032 2023-01-04 07:17:50+00:00
500033 2023-01-04 07:17:20+00:00
500034 2023-01-04 07:17:08+00:00
500035 2023-01-04 07:16:56+00:00
Name: date, Length: 500036, dtype: object
In [232]:

```
from datetime import datetime
# define a function to convert date format
def convert_date_format(date_str):
    try:
        date_obj = datetime.strptime(date_str, '%Y-%m-%d %H:%M:%S+00:00')
        return date_obj.strftime('%Y-%m-%d')
    except ValueError:
        return date_str
# apply the function to the date column
df['date'] = df['date'].apply(convert_date_format)
```

In [233]:

```
df.date
```

Out[233]:

```
0         2023-03-29
1         2023-03-29
2         2023-03-29
3         2023-03-29
4         2023-03-29
             ...
500031    2023-01-04
500032    2023-01-04
500033    2023-01-04
500034    2023-01-04
500035    2023-01-04
Name: date, Length: 500036, dtype: object
```

In [234]:

```
df.head()
```

Out[234]:

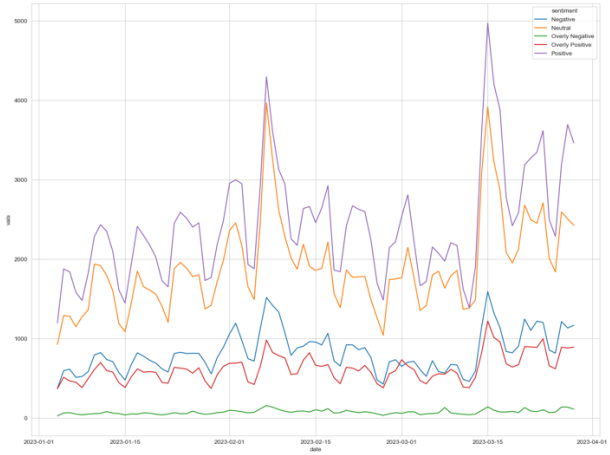
	Unnamed: 0	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	sentiment	score
0	0	2023-03-29	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...	Positive	0.5106
1	1	2023-03-29	1641213218520481805	@mecolehardman4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15	['mecolehardman4', 'chat', 'gpt', 'says', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']	Neutral	0.0000
2	2	2023-03-29	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	Positive	0.7184
3	3	2023-03-29	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny...	ai muses court life must face judge destiny ju...	['ai', 'muses', 'court', 'life', 'must', 'face...	['ai', 'mus', 'court', 'life', 'must', 'face',...	Neutral	0.0000

	Unnamed: 0	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	sentiment	score
4	4	2023-03-29	1641213003260633088	people heard chat gpt yet. first, elite factio...	nikocosmonaut	0.0	0.0	people heard chat gpt yet first elite factions...	people heard chat gpt yet first elite factions...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	Neutral	0.02588

In [237]:
```
df = df[pd.to_datetime(df['date'], errors='coerce').notnull()]

# convert the date column to datetime format
df['date'] = pd.to_datetime(df['date'])

# resample the data by day and compute the sentiment count
timeline = df.resample('D', on='date')['sentiment'].value_counts().unstack(1)

# reset the index of the dataframe and melt the sentiment columns
timeline.reset_index(inplace=True)
timeline = timeline.melt('date', var_name='sentiment', value_name='vals')

# plot the data using seaborn
import seaborn as sns
import matplotlib.pyplot as plt

sns.set_style('whitegrid')
sns.lineplot(x='date', y='vals', hue='sentiment', data=timeline)
plt.figure(figsize=(6,4))
```

Out[237]:

<Figure size 600x400 with 0 Axes>



<Figure size 600x400 with 0 Axes>

```
# Static Day
tweets_by_day = df.groupby(pd.Grouper(key='date', freq='D')).size().reset_index()
tweets_by_day.columns = ['date', 'count']
tweets_by_day['date'] = tweets_by_day['date'].dt.strftime('%m/%d')
plt.figure(figsize=(15, 6))
sns.barplot(data=tweets_by_day, x='date', y='count', palette='viridis')
plt.title('Number of Tweets per Day 2023')
plt.xticks(rotation=90)
plt.show()
# Static plot
# Regression plot to understand the relationship between likes and retweets
```

```
sns.regplot(x='like_count', y='retweet_count', data=df_filtered)
# Static Plot
plt.figure(figsize=(12, 6))
sns.histplot(data=polarity_df, bins=40, kde=False, alpha=0.5, palette=['#1DA1F2', '#00CC96'])
plt.title('Distributions of sentimental polarities Vader Vs. TextBlob')
plt.xlabel('Polarity')
plt.ylabel('Count')
plt.show()
```

In [5]:

```
import numpy as np
import scipy.sparse as ss
import matplotlib.pyplot as plt
import pandas as pd
import corextopic.corextopic as ct
import corextopic.vis_topic as vt # jupyter notebooks will complain matplotlib is being loaded twice

from sklearn.feature_extraction.text import CountVectorizer
```

In [6]:

```
news_data = pd.read_csv("/Users/luqiansong/Desktop/newtwitter.csv")
```

```
/var/folders/m_/m3lsq_594494n7k5zm6nmdtc0000gn/T/ipykernel_7251/93900312.py:1: DtypeWarning:
Columns (1) have mixed types. Specify dtype option on import or set low_memory=False.
  news_data = pd.read_csv("/Users/luqiansong/Desktop/newtwitter.csv")
```

In [7]:

```
news_data.head()
```

Out[7]:

	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized
0	2023-03-29	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...
1	2023-03-29	1641213218520481805	@mecolehardman4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15	['mecolehardman4', 'chat', 'gpt', 'says', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']
2	2023-03-29	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...
3	2023-03-29	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny...	ai muses court life must face judge destiny ju...	['ai', 'muses', 'court', 'life', 'must', 'face...	['ai', 'mus', 'court', 'life', 'must', 'face',...
4	2023-03-29	1641213003260633088	most people heard chat gpt yet. first, elite f...	nikocosmonaut	0.0	0.0	most people heard chat gpt yet first elite fac...	people heard chat gpt yet first elite factions...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...

In [8]:

```
news_data['content_lemmatized1'] = news_data['content_lemmatized'].str.replace('[^\w\s]','')
```

In [9]:

```
#Removal of stop words
```

```
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
stop = stopwords.words('english')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/luqiansong/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [10]:

```
new = ('12', 'academic', 'teach', 'educator', 'assignment', 'class', 'using', 'use', 'university', 'teaching', 'robot', 'tool', 'edtech', 'essay', 'learning', 'teacher', 'school', 'amp', 'education', 'student','billion', 'tool', 'available', 'access', 'business', 'released', 'next', 'update', 'gpt3', 'product', 'plus', 'feature', 'customer', 'model', 'version', 'service', 'api', 'new', 'microsoft', 'gpt4',)
```

In [11]:

```
news_data['Tokens']  = news_data['content_lemmatized1'] .apply(lambda x: " ".join(x for x in x.split() if x not in stop))
news_data.head()
```

Out[11]:

	date	id	content	username	like_count	retweet_count	content_punc	content_stop	content_tokenized	content_lemmatized	content_lemmatized1	Tokens
0	2023-03-29	1641213230730051584	free ai marketing automation tools, strategies...	RealProfitPros	0.0	0.0	free ai marketing automation tools strategies ...	free ai marketing automation tools strategies ...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...	['free', 'ai', 'marketing', 'automation', 'too...
1	2023-03-29	1641213218520481805	@mecolehardman4 chat gpt says it's 15. 😂	AmyLouWho321	0.0	0.0	mecolehardman4 chat gpt says its 15	mecolehardman4 chat gpt says 15	['mecolehardman4', 'chat', 'gpt', 'says', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']	['mecolehardman4', 'chat', 'gpt', 'say', '15']
2	2023-03-29	1641213115684536323	https://t.co/fjjsprt0te - chat pdf! check new ...	yjleon1976	0.0	0.0	httpstcofjjsprt0te chat pdf check new ai quic...	httpstcofjjsprt0te chat pdf check new ai quick...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...	['httpstcofjjsprt0te', 'chat', 'pdf', 'check',...
3	2023-03-29	1641213110915571715	ai muses: "in court life, must face judge dest...	ChatGPT_Thinks	0.0	0.0	ai muses in court life must face judge destiny...	ai muses court life must face judge destiny ju...	['ai', 'muses', 'court', 'life', 'must', 'face...	['ai', 'mus', 'court', 'life', 'must', 'face',...	['ai', 'mus', 'court', 'life', 'must', 'face',...	['ai', 'mus', 'court', 'life', 'must', 'face',...
4	2023-03-29	1641213003260633088	most people heard chat gpt yet. first, elite f...	nikocosmonaut	0.0	0.0	most people heard chat gpt yet first elite fac...	people heard chat gpt yet first elite factions ...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...	['people', 'heard', 'chat', 'gpt', 'yet', 'fir...

In [12]:

```
# Transform data into a sparse matrix
```

```
vectorizer = CountVectorizer(stop_words='english', max_features=None, binary=True)
doc_word = vectorizer.fit_transform(news_data.Tokens)
doc_word = ss.csr_matrix(doc_word)
doc_word.shape # n_docs x m_words
Out[12]:
(500036, 619062)
In [13]:
doc_word
Out[13]:
<500036x619062 sparse matrix of type '<class 'numpy.int64'>'
        with 6817072 stored elements in Compressed Sparse Row format>
In [16]:
# Get words that label the columns (needed to extract readable topics and make anchoring easier)
words = list(np.asarray(vectorizer.get_feature_names_out()))
In [19]:
anchor_words = [
# academic university help
['12', 'academic', 'teach', 'educator', 'assignment', 'class', 'using', 'use', 'university', 'teaching', 'robot', 'tool',
'edtech', 'essay', 'learning', 'teacher', 'school', 'amp', 'education', 'student'],
# stcok investiment
['stock', 'investment', 'pas', 'state', 'medical', 'look', 'text', 'law', 'algorithm', 'investing', 'news', 'training', 'take',
'trained', 'paper', 'research', 'test', 'human', 'data', 'exam'],
# answer question give response
['problem', 'im', 'would', 'code', 'first', 'try', 'think', 'response', 'asking', 'got', 'give', 'time', 'good', 'one', 'like', 'get',
'ask', 'asked', 'question', 'answer'],
# artificialintelligence company business
['extension', 'impact', 'article', 'new', 'amp', 'read','company', 'war', 'googlebard', 'app', 'competitor', 'new',
'rival', 'googleai', 'tech', 'bardai', 'launch', 'apple', 'market', 'artificialintelligence', 'bing', 'microsoft', 'search',
'bard', 'google'],
# instruction of generated ai
['tempupdate', 'kmhr', 'temp', 'wind', 'browser', 'tech', 'chinese', 'baidu', 'status', 'bot', 'edge', 'generated',
'sunrise', 'sunset', 'instruction', 'based', 'speed', 'pm', 'current', 'china'],
# security recommendation
['gonna', '100', 'hour', 'case', 'way', 'used', 'help', 'million', 'much', 'people', 'money', 'year', 'month', 'tool',
'information', 'using', 'time', 'data', 'make', 'user', 'use','recommendation', 'date', 'security', 'pro',
'httpstcorlyimpqw40', 'price', 'stablediffusion2', 'powered', 'open', 'wait', 'long', 'cybersecurity', 'gpt', 'last',
'midjourney', 'join', 'dalle', 'short', 'imagine'],
# artists
['digitalart', 'script', 'style', 'used', 'article', 'written', 'poem', 'book', 'story', 'asked', 'write"using', 'another',
'tweet', 'word', 'make', 'fake', 'wrote', 'read', 'writing', 'song','generated', 'risk', 'world', 'podcast', 'check', 'blog',
'generativeai', 'tech', 'business', 'artificial','bill', 'new', 'aiartwork', 'tesla', 'world', 'woke', 'joke', 'artist', 'image',
'dalle2', 'stablediffusion', 'musk', 'twitter', 'elon', 'aiart', 'art', 'midjourney', 'elonmusk'],
# writing email
['website', 'industry', 'generative','response', 'processing', 'machinelearning', 'developed', 'generative',
'chatbots', 'gpt4', 'nlp', 'text', 'natural', 'ability', 'chatgpt3', 'large', 'source', 'domain', 'artificialintelligence', 'llm',
'gpt3', 'model', 'language', 'potential', 'latest', 'future', 'artificialintelligence', 'generate', 'blog', 'email', 'check',
'help', 'amp', 'writing', 'youtube', 'seo', 'via', 'using', 'video', 'create', 'tool', 'use', 'marketing', 'prompt', 'content'],
# future blockchain
['bnb', 'ethereum', 'cryptocurrency', 'powerful', 'nfts', 'invest', 'time', 'magic', 'coin', 'token', 'future', 'eth',
'blockchain', 'bitcoin', 'btc', 'web3', 'airdrop', 'gpt4', 'nft', 'crypto'],
# code python
['new', 'best',   'developer', 'code', 'artificialintelligence', 'programming', 'coding', 'machinelearning','learning',
'build', 'free', 'business', 'get', 'datascience', 'tech', 'learn', 'python', 'innovation','analytics', 'productivity', 'skill',
'ml','latest', 'python', '5g', 'iot', 'say', 'nft', 'lol', 'gaming', 'friend', 'ar', 'dan', 'resume', 'tech', 'vr', 'web3', 'gt',
'artificialintelligence', 'metaverse'],
# trade ai
['altman', 'based', 'bingai', 'nocode', 'sam', 'signal', 'chart', 'trade', 'new', 'trial', 'buy', 'option', 'free', 'stock',
'trading', 'engine', 'microsoft', 'search', 'bing', 'gpt4'],
# potential thread
['come', 'keep', 'new', 'world', 'time', 'take', 'great', 'life', 'one', 'thread', 'thing', 'mind', 'going', 'change', 'like',
'conversation', 'let', 'see', 'im', 'day'],
# cloud job
```

['get', 'use', 'work', 'could', 'right', 'make', 'say', 'dont', 'need', 'thing', 'even', 'would', 'replace', 'cant', 'like', 'people', 'human', 'think', 'know', 'job'],
product consumer
['billion', 'tool', 'available', 'access', 'business', 'released', 'next', 'update', 'gpt3', 'product', 'plus', 'feature', 'customer', 'model', 'version', 'service', 'api', 'new', 'microsoft', 'gpt4']]
anchored_topic_model = ct.Corex(n_hidden=14, seed=2)
anchored_topic_model.fit(doc_word, words=words, anchors=anchor_words, anchor_strength=2);
WARNING: Anchor word not in word column labels provided to CorEx: take
WARNING: Anchor word not in word column labels provided to CorEx: would
WARNING: Anchor word not in word column labels provided to CorEx: first
WARNING: Anchor word not in word column labels provided to CorEx: give
WARNING: Anchor word not in word column labels provided to CorEx: one
WARNING: Anchor word not in word column labels provided to CorEx: get
WARNING: Anchor word not in word column labels provided to CorEx: much
WARNING: Anchor word not in word column labels provided to CorEx: last
WARNING: Anchor word not in word column labels provided to CorEx: writeusing
WARNING: Anchor word not in word column labels provided to CorEx: another
WARNING: Anchor word not in word column labels provided to CorEx: bill
WARNING: Anchor word not in word column labels provided to CorEx: via
WARNING: Anchor word not in word column labels provided to CorEx: get
WARNING: Anchor word not in word column labels provided to CorEx: keep
WARNING: Anchor word not in word column labels provided to CorEx: take
WARNING: Anchor word not in word column labels provided to CorEx: one
WARNING: Anchor word not in word column labels provided to CorEx: see
WARNING: Anchor word not in word column labels provided to CorEx: get
WARNING: Anchor word not in word column labels provided to CorEx: could
WARNING: Anchor word not in word column labels provided to CorEx: even
WARNING: Anchor word not in word column labels provided to CorEx: would
WARNING: Anchor word not in word column labels provided to CorEx: cant
WARNING: Anchor word not in word column labels provided to CorEx: next
In [26]:
```
for n in range(len(anchor_words)):
    topic_words,_,_ = zip(*anchored_topic_model.get_topics(topic=n))
    print('{}: '.format(n) + ', '.join(topic_words))
```
0: tool, amp, learning, student, education, school, teacher, essay, edtech, use
1: human, data, look, news, text, research, stock, test, investment, exam
2: asked, answer, question, im, good, ask, code, response, got, problem
3: new, artificialintelligence, google, microsoft, tech, bing, search, bard, company, app
4: tech, based, generated, current, speed, china, instruction, pm, browser, edge
5: use, using, tool, make, way, people, help, used, data, year
6: read, check, article, artificial, midjourney, generativeai, story, generated, blog, written
7: artificialintelligence, tool, model, ai, language, content, gpt3, machinelearning, chatgpt, text
8: crypto, nft, web3, bitcoin, airdrop, blockchain, btc, eth, token, invest
9: artificialintelligence, tech, machinelearning, web3, innovation, python, coding, ml, programming, metaverse
10: microsoft, bing, search, free, based, engine, stock, trading, option, buy
11: like, new, time, im, world, thing, day, let, going, great
12: like, make, know, think, work, people, need, human, thing, say
13: new, gpt4, microsoft, model, business, gpt3, version, service, access, api
In [27]:
```
vt.vis_rep(anchored_topic_model, column_label=words, prefix='twitters')
```
Print topics in text file