Analyze the sentiment of TikTok Posts and Users' Comments with Hashtags Related to Covid-19

member Yao,Yi-TZU/ Chen,Ying-Yu/ Li,Chen-Hsuan/ Chen,Ting-Yun

***Abstract*—With the outbreak of covid19 and the slowdown so far, many people have expressed any thoughts and emotions about the epidemic through tiktok. The focus of this report is that we collected data on eight hashtags related to COVID-19 during the outbreak. We analyzed this data using NLP and Vader sentiment analysis, aiming to observe the mood swings of TikTok users during this time. Initially, we sorted out the data and found that 2020 is the year when people pay the most attention to the epidemic. Therefore, we chose to conduct an in-depth study on the emotional changes of TikTok users during the epidemic based on the data from 2020. Results showed that despite the challenges posed by the pandemic, positive sentiment was obvious, with 1 in 4 reviews being positive. Surprisingly, the frequency of negative reviews is very low. In addition, this study also provides interesting findings for analyzing changes in public sentiment towards the outbreak.**

I. INTRODUCTION

Since the outbreak of the COVID-19, approximately three years have passed. During this time, various regions worldwide have experienced different levels of outbreak and response measures. In the initial stages, the situation was severe, causing global attention and concern. As time went on and countries implemented prevention and control measures, the epidemic gradually came under control, and many areas witnessed a downward trend in the number of infections and death rates.

Social media platforms, such as Facebook, Twitter, Tiktok, have played a crucial role throughout this process. People have utilized these social media platforms to express their opinions, share personal experiences, disseminate information, and access news.

TikTok is a popular social media platform that allows users to create and share short videos. With its user-friendly interface and a wide range of creative tools, TikTok has gained immense popularity worldwide.

During the COVID-19 pandemic, TikTok served as a significant source of entertainment, connection, and information for many individuals. It provided a platform for people to stay connected, express themselves, and alleviate boredom during the periods of lockdowns and social distancing.

As time has passed, people's emotions and perspectives towards the pandemic have evolved. The initial panic and anxiety gradually transformed into contemplation of the pandemic's impact and discussions about response strategies. Posts and comments on social media platforms reflected the public's opinions on government response measures, vaccine distribution, economic repercussions, and societal recovery.

In this paper, we have chosen several hashtags related to the COVID-19 pandemic and collected relevant data from 2019 to 2023 that utilized these hashtags. Our goal is to conduct sentiment analysis on these data to deeply study the changes of TikTok users' sentiment ,which classes are Negative, Positive and Neutral, during the epidemic. Such analysis can help us understand and explain the evolution of people's emotions and perspectives towards the pandemic over time.

II. RELATED WORKS

Social media is a prevalent way for the public to express opinion and emotions. To identify people's feelings with social media text, such as posts and comments, several studies developed sentiment analysis methods to classify public emotion. Researchers [3] took advantage of long short-term memory (LSTM), a deep learning-based approach to natural language processing (NLP), to gain insight into the development of Twitter user sentiment during the pandemic. In addition, they conduct sentiment analysis on specific topics such as social justice, mental health, disease prevention, and misinformation. Researchers [1] obtain data related to COVID19 via twitter with the keyword "covid19 OR corona" and carry out a sentiment analysis using the Naive Bayes method. Researchers [4] tested the effectiveness of a Bidirectional Long Term Short Memory (BLTSM) model of a Recurrent Neural Network (RNN) to predict sentiment class which are Negative, Positive and Neutral in tweets. The BLTSM model achieved a high accuracy rate of 86.15% in predicting sentiment classes, demonstrating its adaptability and accuracy in analyzing text sentiment. Ngoc Thien An NGUYEN et al.[6] devised a survey employing a combination of quantitative and qualitative questions in order to examine user habits and well-established content categories on Vietnamese TikTok.

During COVID-19, some research adopted sentiment analysis to design systems analyzing the fluctuation of public sentiment in social media. X. Yu et al present Senti-COVID19, an interactive visual analytic system for reflecting and analyzing public sentiment and detecting sentiment fluctuation triggers on social media [2]. Xinran Yu et al. direct their focus towards the daily sentiment distribution of news and public opinion on Weibo concerning the keyword COVID-19. They undertake a comparative analysis of the sentiment trends, thereby presenting a novel avenue for the analysis of social public opinion [5]. Researchers [7] reviewed various research efforts using social media data to understand people's concerns and awareness about the COVID-19 pandemic, including collecting data, identifying topics of discussion, and conducting sentiment and emotion analysis. It also analyzed the application of AI methodologies, specifically machine learning and deep learning, discussing their results, limitations, and strengths. Researchers [8] examines the discourse between South African government officials and the public on social media during the COVID-19 pandemic, analyze the keywords and unique topics. The findings inform government communication strategies, highlighting key themes such as lockdown measures, COVID-19 information, government officials' roles, fake news, PPE, healthcare, school closures, and job losses.

III. FRAMEWORK

The proposed framework for sentiment analysis of COVID-19 related posts on Tiktok is shown in Figure1. The framework consists of mainly two main parts, data collection and sentiment analysis.

In the data collection part, we use apify to crawl the data we need, save it into four tables with SQL, and then use pandas for data preprocessing. Then we checked the frequency of words and made the Comments Wordcloud. In addition, we made Sentiment Scatterplot and Sentiment Pieplot by using Vader for sentimental analysis.

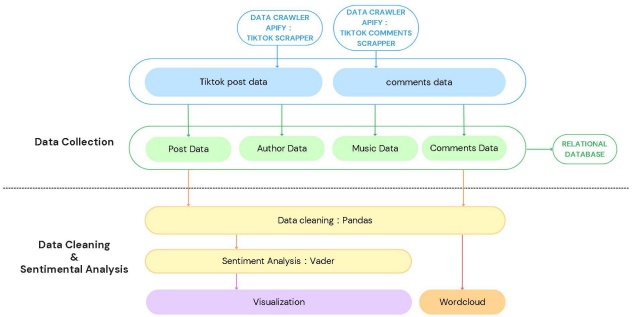

Fig. 1.The framework for sentiment analysis of COVID-19 related posts on Tiktok

VI. EXPERIMENT DESIGN

A. Data collection

Some hashtags appear due to the breakout of COVID-19, like #socialdistancing, #stayhome, #staysthome, #lockdown and also #covid19. Some hashtags were used more frequently since the COVID-19, like #staysafe #staystrong #stayhomestaysafe. We chose all of them and typed in into Apify "Tiktok Scrapper" to crawl the COVID-19 related posting data. In the data that we crawled, contains the informations that make up a post, including three parts, which is postMeta, authorMeta, and musicMeta.Then, we use "WebVideoURL" in " postMeta" to crawl the comments data of each post by posting "WebVideoURL"on "Tiktok comments scrapper" of Apify.

"Tiktok Scraper" uses the unregistered Tiktok website to search for the entered hashtag, and reaches the page of the hashtag (https://www.tiktok.com/tag/ + "hashtag name" is a website created by Tiktok for each hashtag.) to crawl the posts in order from left to right and from top to bottom. Since Tiktok limits the number of posts crawled by unregistered web users to a maximum of 1000, "Tiktok Scraper" crawled 800 to 900 posts.

Also, we use Draw SQL in Fig 2 to divide our data into four tables, including Posts table, authorMeta table, musicMeta table and comments table. The following four tables shown in Table I are the information corresponding to each of our fields.

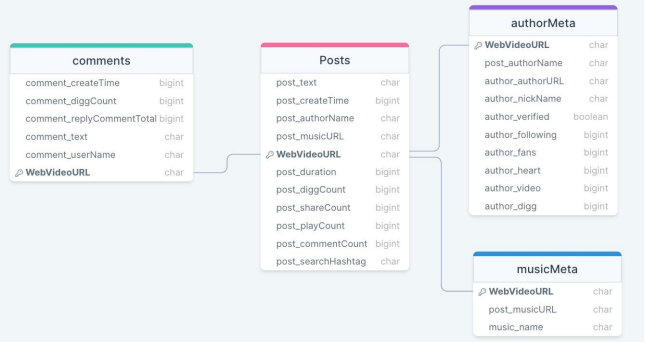

Fig. 2. Dividing data into tables by Draw SQL

TABLE I
DEFINITION OF DIFFERENT FIELDS IN FOUR DATA TABLES

post

post_text	The text of the post.
post_createTime	The time when the post was created.
post_authorName	The author's account of the post.
post_musicURL	The link of the music used in posted videos.
post_WebVideoURL	The link to the post.
post_duration	The video duration of the post, presented in seconds.
post_diggCount	The number of users who liked the post.
post_shareCount	The number of times that the post was shared.
post_playCount	The number of times that the post was played.
post_commentCount	The number of comments in the post.
post_searchHashtag	The hashtag used for searching the post.

musicMeta

music_musicURL	The link of the music used in posted videos.
music_name	The name of the music.

authorMeta

author_authorName	The author's account of the post.
author_authorURL	The link to the author's profile of the post.
author_nickName	The nickname displays on the author's profile.
author_verified	The booleans show whether the profile is verified or not.
author_following	The number of users that the author follows.
author_fans	The number of users following the author.
author_heart	The total likes on the profile's posts.
author_video	The number of videos shown on the author's profile.
author_digg	The number of videos and comments the author has liked.

Comments

comment_createTime	The time when the comment was posted.
comment_diggCount	The number of users who liked the comment.
comment_replyCommentTotal	The total replies to the comment.
comment_text	The text of the comment.
comment_userName	The user's account of the comment.
comment_WebVideoURL	The link to the post.

B. Data preprocessing

Before implementing sentiment analysis and NLP, we perform data cleaning to enhance and ensure greater reliability in the final outcome.

1. Data cleaning

The process of data cleaning includes removing duplicates, rows with missing values in specific columns, and rows that are missing a primary key. In the stage of removing missing values, since the missing values in authorMeta and musicMeta tables may not affect the

analysis of text, we only delete the rows containing missing values in the posts and comments table.

The detailed data cleaning steps are as shown in Fig 3. First, we clean the raw data of the comments table. We then clean the raw data of the remaining tables after merging posts, authorMeta, and musicMeta tables in a relational database based on the primary key, WebVedioURL. Finally, we verify that there are no missing or inconsistent primary keys, WebVedioURL, in both cleaned datasets. Table II indicates the total rows removed in each step. Table III illustrates the number of rows in each hashtag's CSV file for each year from 2019 to 2023.

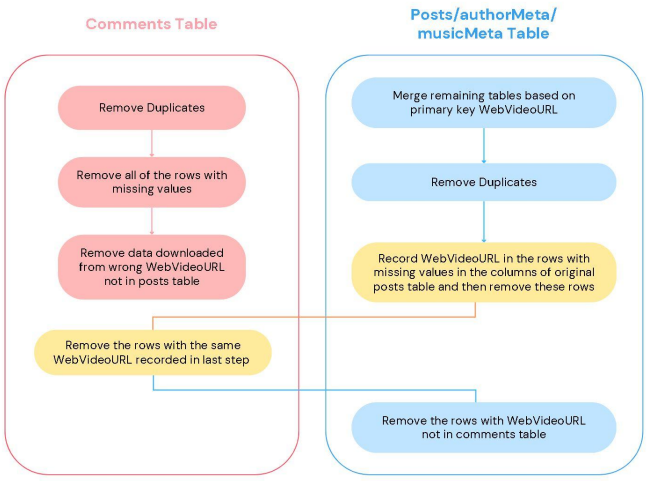

Fig. 3. The data cleaning process.

TABLE II
NUMBER OF REMOVE ROWS

tables\hashtags		#covid19	#stayhome staysafe	#stayhome	#stayathome	#staysafe	#socialdistancing	#lockdown	#staystrong
comments	the total rows of raw data	327563	154195	347626	388532	351349	334741	296450	156983
	the duplicate rows dropped	23735	38	231	1896	3102	3372	7183	1718
	total missing values	527	25	59	26128	432	2332	68	2885
	the missing data rows dropped	246	25	59	24525	432	2125	63	2859
	the rows missing primary keys dropped	138	536	986	2007	13485	29570	1315	437
	the total rows of cleaned data	288768	153596	346350	360104	334330	299674	287889	150002
others	the total rows of raw data	916	924	911	914	902	950	905	898
	the duplicate rows dropped	0	0	0	0	0	0	0	0
	total missing values	66	22	30	31	25	34	30	39
	the missing data rows dropped	2	1	28	0	21	2	2	15
	the rows missing primary keys dropped	78	163	37	18	34	40	228	338
	the total rows of cleaned data	836	760	874	896	868	908	675	545

TABLE III
NUMBER OF ROWS IN EACH HASHTAG'S CSV FILE FOR EACH YEAR FROM 2019 TO 2023

		#covid19	#stayhome staysafe	#stayhome	#stayathome	#staysafe	#socialdistancing	#lockdown	#staystrong
comments	2019	0	460	0	324421	1012	0	460	5943
	2020	162508	190161	318386	23239	192233	269918	190161	34144
	2021	87665	71892	21422	1010	74256	22590	71892	41106
	2022	26681	15721	5449	10010	48246	4815	15721	48716
	2023	11914	9655	1093	2434	18583	2351	9655	20048
others	2019	0	0	0	0	3	0	1	20
	2020	480	732	813	814	518	828	464	143
	2021	260	26	47	53	183	60	156	140
	2022	79	1	12	24	119	12	36	173
	2023	17	1	2	5	45	8	18	69

2. Vader

VADER sentiment analysis is sensitive to textual content, particularly punctuation, emoji, and emoticons[2]. The analysis result is also influenced by the case of words in the text[2]. We compare the sentiment analysis result of original comments and comments conducting text processing with NLTK. The text preprocessing includes converting all text to lowercase, removing stop words and deleting punctuations, which we will describe carefully in the next section, word cloud. After conducting text processing, the total neutral comments in hashtag #socialdistancing increased 15454 rows from 190334 to 205788. It indicates that the ability of Vader to classify positive and negative comments declines, as shown in Table IV. Also we want to gain insight into the words of positive and negative comments by visualizing the sentiment analysis result by word cloud. Thus, we do not implement any text preprocessing before performing the VADER lexicon for sentiment analysis. So that we not only get more information about emotional comments but also won't impact the performance of the sentiment analysis result.

TABLE IV
COMPARE SENTIMENT ANALYSIS CATEGORY OF COMMENTS PERFORM AND NOT PERFORM TEXT PROCESSING

original text	sentiment analysis category	text after text processing	sentiment analysis category
Please don't make these vids :(	negative	pleas don't make vid	neutral
"Don't close no doors in my house!" 😡 "thank you" 😃	negative	"don't close door house" 😡 "thank you" 😃	neutral
Your dog is a menace	negative	dog menac	neutral
this is totally funny but Jesus is watching😂😂😂😂😂	positive	total funni jesus watching😂😂😂😂😂	neutral
Amazing 🥺❤️	positive	amaz 🥺❤️	neutral
he said thats not how you treat a king👑	negative	said that treat king👑	positive
He not playing w you 💀💀💀	negative	play w 💀💀💀	positive
Now that's turning a bad situation into something memorable and amazing!	positive	that turn bad situat someth memor amaz	negative
I've been trying to watch pretty little liars for so long!!! Thank you so much for this!!!	positive	i've tri watch pretti litl liar long thank much	negative

C. Word cloud

A word cloud consists of two major steps: tokenization and stopword removal. First, as Fig 4 shows, we eliminate emoticons and preprocess the text by removing punctuation marks, capital letters, URLs, etc.

```
df = pd.read_csv('lockdown_cleaned_comments.csv', encoding='utf-8')
def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopword]
    text = " ".join(text)
    text = [snowball_stemmer.stem(word) for word in text.split(' ')]
    text = " ".join(text)
    return text

df['comment_text'] = df['comment_text'].apply(clean)
df
```

Fig. 4. Function that performs text-cleaning operations

Second, since Tiktok has a diverse user base, the comments we collected contain various languages. Therefore, we remove stop words in languages such as Arabic, Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, and Swedish.

D. Natural Language Processing

Natural language processing (NLP) represents a branch of computer science that seeks to analyze, comprehend, and generate human language. Its applications encompass extracting valuable information and knowledge from vast amounts of textual data and employing them in various domains, such as sentiment analysis, text classification, and machine translation. This research endeavors to utilize NLP techniques in scrutinizing TikTok comments pertaining to the pandemic, aiming to explore the emotions and discussions exhibited by users during this period. We assert that this constitutes a significant and challenging research endeavor due to the widespread usage and influence of TikTok as a social media platform, coupled with the global magnitude and societal ramifications of the pandemic. By employing NLP methods, we can delve deeper into the perspectives, emotions, and behaviors of TikTok users, thereby contributing novel insights to the realms of social psychology and communication studies. Additionally, the implementation of word cloud analysis facilitates an examination of the terminology, its significance, and the frequency of its usage among TikTok users. This approach expedites our comprehension of the salient aspects, themes, and trends present within the textual content, enabling the acquisition of invaluable insights.

E. Vader Sentiment Analysis

VADER is a lexicon and rule-based sentiment tool that is specifically attuned to sentiments expressed in social media[9]. The rule-based lexicon analysis uses sentiment libraries or lexicons with a series of rules to evaluate the sentiment of the text[2]. So that we don't need to train or use machine learning models before calculating the sentiment score of posts and comments. VADER not only measures words in sentences but also incorporates numerous lexical features, such as initialisms, emoticons and emojis, to calculate the sentiment score of text. Even the punctuation and capital letters in sentences used to express sentiment can be captured by the VADER lexicon. The sentence structure of comments and posts on tiktok including numerous emojis and emoticons are consistent with the messages texted between friends.Thus, the VADER is suitable for sentiment analysis of posts and comments in TikTok.

The VADER computes four types of sentiment score. The "positive", "negative", "neutral" scores are ratios for proportions of text that fall in each category[9]. The "compound score" is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized[9]. To identify the sentiment of comments on tiktok, we follow the below standardized thresholds of "compound" score and classify comments as either positive, neutral, or negative. A "compound score" greater or equal to 0.05 is positive[2]. A "compound score" lower or equal to 0.05 is negative[2]. Whereas score between -0.05 and 0.05 is neutral[2].

Since TikTok primarily consists of short videos, we are curious whether the emotional tone of TikTok posts affects the sentiment in the comment section. We calculate the mean of compound scores for all the comments on a post and plot it on a scatterplot along with the compound score of the post. We observe that there is no correlation between the two, as shown in Fig 5. Therefore, we can utilize the sentiment of comments to explore users' feelings towards specific hashtags related to covid-19 on TikTok.

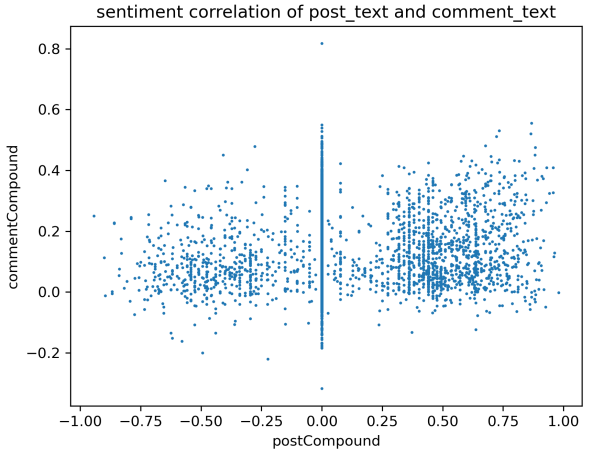

Fig. 5.The scatter plot of the post's compound scores and the mean of comments' compound scores.

To get insight into the sentiment of discussions between users on tiktok during Covid-19 period. For example, whether TikTok users expressed negative comments during Covid-19 period in the relevant hashtags. We compute the sentiment scores of eight tiktok hashtags related to Coronavirus disease during Covid-19 period, from 2019 to 2023. Then we examine the distribution of compound scores and the proportion of positive, negative, and neutral comments under different hashtags. On the other hand, the engagement of discussions on posts related to the pandemic in 2020 is the highest, as shown in Fig 6.

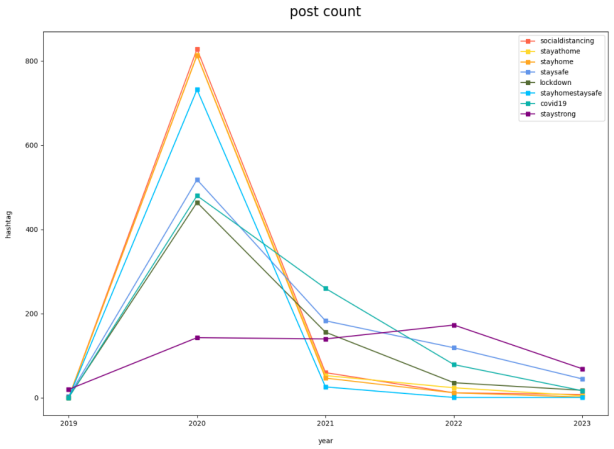

Fig. 6. The line chart of post counts of each hashtag.

V. RESULT

We place particular emphasis on analyzing the comments on TikTok posts. The analysis of the boxplot, as shown in Fig 7, reveals a predominant prevalence of zero values within the range spanning from the 25th to the 50th percentile. Furthermore, the interquartile range, which encompasses the 50th to the 75th percentile, exhibits notable confinement primarily within the interval of 0 to 0.4. Remarkably, the plot signifies a concentrated distribution of extreme values for the variable #covid19, indicating a comparatively diminished level of emotional intensity within the comments.

Upon sorting the data based on the compound score, it yields a representation depicted through the utilization of a pie chart. The charts, as shown in Fig 7, elucidate that a substantial proportion of the comments, constituting the majority, can be characterized as neutral, while approximately 50% of the comments are classified as positive. This suggests that, despite the challenges posed by the pandemic, a significant presence of positive sentiment is evident, with roughly one positive comment observed for every two comments. Surprisingly, the frequency of negative comments is significantly low. Notably, the hashtags #covid19 and #staystrong exhibit the highest proportion of negative comments. Conversely, the hashtags #staynome and #stayhomestaysafe demonstrate the highest proportion of positive comments.

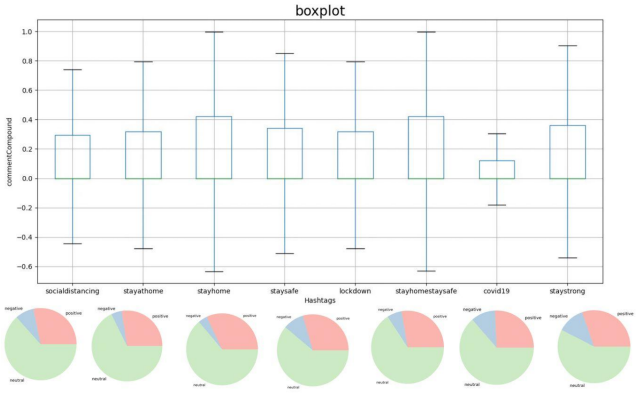

Fig. 7. The boxplot and piechart of 8 hashtags

The barplot, depicted in Fig 8, illustrates the total numbers of neutral, positive, and negative comments across 8 hashtags during the COVID-19 period, using different colors to differentiate comments from various years. It provides a clear visual representation of the overall count of neutral, positive, and negative comments, allowing for easy comparison of their proportions. Furthermore, it is evident from the analysis that the engagement surrounding COVID-19 hashtags has gradually declined from 2020 to 2023.

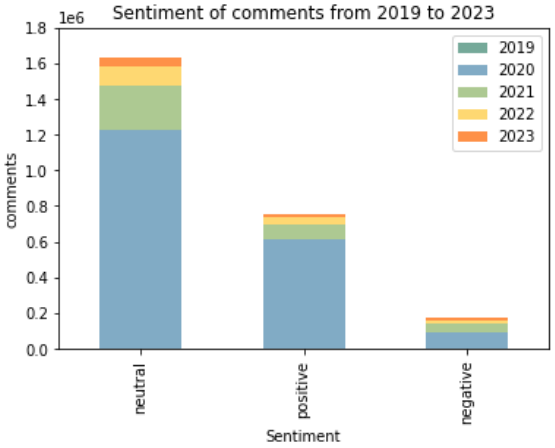

Fig. 8. The barplot of 8 hashtags.

Numerous situations contribute to the prevalence of neutral comments, including the occurrence of consecutive comments and the practice of tagging users without providing additional content. For instance, consecutive comments, as illustrated in the Table V below, consist exclusively of emoticons and emojis, with the intention of creating an image on TikTok. Moreover, it is common for TikTok users to tag '@username' to invite their friends to watch videos. These scenarios are categorized as neutral since the underlying motivation driving users' actions cannot be discerned.

TABLE V
THE EXAMPLES OF COMMENTS CATEGORIZED AS NEUTRAL

comment_text	pos score	neg score	neu score	neu score
D D N'T	0.0	0.0	1.0	neutral
W A T C H	0.0	0.0	1.0	neutral
M Y	0.0	0.0	1.0	neutral
L A S T	0.0	0.0	1.0	neutral
V I D E D	0.0	0.0	1.0	neutral
DON'T WATCH MY LAST VIDEO	0.0	0.0	1.0	neutral
{__/}	0.0	0.0	1.0	neutral
(¬-¬)	0.0	0.0	1.0	neutral
/> i got you buy some more	0.0	0.0	1.0	neutral
@xxmarxaxx	0.0	0.0	1.0	neutral
@monkey.d.uzumaki @sh4d0w_37	0.0	0.0	1.0	neutral
@ambsschultz OMG IM SO EXCITED TO DO THIS	0.373	0.0	0.672	positive

By analyzing the word clouds of comments in 8 related hashtags, as illustrated in the Figure, we gain insights into the discussion content on TikTok during COVID-19. The more frequently used words in TikTok comments include "video," "nice," "super," "like," and "miss," which are represented as larger words. Identifying words related to COVID-19 in the Fig 9 proves to be challenging as they are exclusively found in the negative word cloud. However, a noteworthy finding is that "video" appears as a common word in all word clouds. This observation confirms that TikTok is primarily a social media platform focused on short videos, and the comments largely revolve around video content. This finding aligns with the previous analysis results depicted in the Fig 9 for Vader sentiment analysis. Furthermore, the positive word cloud consists of common words such as "love", "nice", "hahaha", "lol", and "wow." This suggests that the Vader lexicon accurately categorizes positive comments on TikTok.

Fig. 9. The word cloud shows all common words in the comments of 8 hashtags. The positive and negative word cloud individually shows the common words in the comments classified as positive and negative sentiment categories. The size of the word represents the frequency of the word used in the comments.

VI. CONCLUSION

The paper focuses on analyzing comments on TikTok during the COVID-19 period. The aim of the study was understanding through the lens of social media people emotions and feelings about the pandemics. There is a substantial presence of neutral comments, while approximately half of the comments are classified as positive, and the frequency of negative comments is relatively low. The proportion of negative sentiment under #covid19 and #staystrong is highest, while the #staynome and #stayhomestaysafe exhibit a much higher proportion of positive sentiment. The engagement around COVID-19 hashtags gradually declines from 2020 to 2023. Some commonly used words in TikTok comments include “video”, “nice”, “super”, “like”, and “miss”, with COVID-19-related words primarily appearing in the negative word cloud. These findings provide insights into the emotional tendencies and participation trends on TikTok during the COVID-19 period.

VII. FUTURE WORK

Due to limitations of the Tiktok scraper, apify, we only scrape the first thousand posts in the hashtag web page. Thus, the posting time of our data is concentrated in 2020 when the novel coronavirus outbreak has drawn lots of international attention. To have a more comprehensive sentiment analysis in Tiktok, we should figure out another efficient way to scrape the data relating to Covid-19 from 2021 to 2023 in future work. Furthermore, in order to comprehend the Coronavirus related topics that people commonly discuss on TikTok, it is important for us to extract audio-visual materials. Analyzing comments alone can be challenging when attempting to identify popular Covid-19 subjects on TikTok, as comments typically revolve around the video content. Thus, ve should incorporate video analysis in the future to better align with TikTok, a predominantly short-video based social media platform.

VIII. ACKNOWLEDGEMENTS

We sincerely thank Professor Xiang for her exceptional leadership and guidance throughout our research project. Her expertise and extensive experience have provided us with invaluable direction and inspiration, enabling us to progress smoothly in our study.

REFERENCE

[1] Sentiment Analysis of Public Reaction to COVID-19 in Twitter Media using Naïve Bayes Classifier
[2] Senti-COVID19: An Interactive Visual Analytics System for Detecting Public Sentiment and Insights Regarding COVID-19 From Social Media
[3] Deep Learning-Based COVID-19 Twitter Analysis
[4] Sentiment Analysis of Tweets During COVID-19 Pandemic Using BLSTM
[5] Sentiment analysis for news and social media in COVID-19
[6] The impact of Tiktok contents on students’ social perceptions and lifestyles during the COVID-19 pandemic
[7] Carmela Comito, ”Social Media Mining and Analysis to support authorities in COVID-19 pandemic preparedness”, 2022.
[8] Vukosi Marivate, Avashlin Moodley and Athandiwe Saba, “Extracting and categorising the reactions to COVID-19 by the South African public - A social media study”, 2021.
[9] VADER GitHub