A Critical Discourse Analysis on the #StopAsianHate movement: Using Topic Modeling through MDCOR to Identify the Most Salient Themes from the Perspective of Social Representations Theory

Yukun Hu1,a*, Yunxuan Li2,b, Yuhui Zhou3,c, Xiangcun Zhu4,d

1University of California Davis, Davis, 95618, USA
2School of Social Sciences, University of Manchester, Manchester, M13 9PL, UK
3School of Public Affairs, Zhejiang University, Hangzhou, 310058, China
4Suzhou North America High School, Suzhou, 215128, China

Abstract. The Atlanta Spa shooting evoked the Stop Asian Hate movement in the US. This study examines what primary topics were covered in Twitter discourse in #StopAsianHate and which phases of collective symbolic coping emerged from the perspective of social representations theory. 21,474 tweets in #StopAsianHate on Twitter in the first weeks from May to August were analyzed by topic modeling using MDCOR. The results of this research showed that the phases of awareness and divergence emerged in the discourse, while the phases of convergence and normalization remained unidentified. The knowledge generated in this research may contribute to identifying the main topics that appeared in #StopAsianHate on Twitter, and it may also inform the supporters of the Stop Asian Hate movement about the following actions.

1. Introduction

The first COVID-19 case was documented in Wuhan, China, and after that, an exponentially increasing number of COVID-19 cases were documented worldwide [1]. The fears and loathing towards Asian Americans intensified during the COVID-19 pandemic, leading to a soaring rise in hate speeches and violence towards Asian Americans. The terms like "China virus" and "Wuhan virus" used in the pandemic context propagated negative stereotypes and fueled hostility against the Asian American community [2]. A substantial number of anti-Asian statements were disseminated through the media. Moreover, the Stop Asian American Hate organization had received reports of 2,583 anti-Asian incidents since March 2020, while 70.6% of the incidents involved verbal harassment, 8.7% involved physical assault, and 8% involved potential civil rights violations [3]. The Asian American community was, therefore, growingly concerned about their safety and their marginalized status in the US society. According to a report from the Pew Research Centre in 2021, 81% of Asian American adults believed that the violence targeted at Asian communities had witnessed a rise, and 32% of respondents expressed their dread about the potential threats towards themselves and their communities.

On March 16th, 2021, eight women were shot in Atlanta, and six of them were Asians [4]. The Atlanta shooting incident caused significant concerns and outrage within the Asian American community, stimulating a surge of Stop Asian Hate campaigns across the US. Demands such as "Asians must be respected", "please show Asians reverence", and "please cease to hate Asians" were strongly requested by the advocates of the Stop Asian Hate movement in the US. An online social movement named #StopAsianHate was also initiated on Twitter in order to support the Asian community.

2. Literature Review

The theoretical framework that underpins this study is the notion of Collective Symbolic Coping (CSC) developed from Social Representations theory. According to Serge Moscovici [5], who developed the Social Representations theory, social presentations are "a system of values, ideas and practices that function in (a) setting up a social order that allows individuals to organize themselves and control their material as well as social environment; and (b) facilitating community members' communication with each other by a shared code for social exchange and for recognizing and categorizing different facets of their social environment, which involves their personal and collective histories" [6]. Collective Symbolic Coping is a notion that bears upon the Social Representations theory [7]. Collective symbolic coping refers to the process in which the collective social group as a whole "gives meaning to novel situations that threaten the established social order" [8] to integrate the new phenomena into the existing bodies of knowledge [7]. The collective symbolic coping process involves four phases: (1) Awareness; (2) Divergence; (3) Convergence; and (4) Normalization. At the phase of awareness, the event that challenges the established social order is considered important for society by the public. The public is aware of its social
Social media has been utilized by marginalized minority groups to explore their identities, share their experience as marginalized social groups, and organize online or offline social movements [9-11]. There have been several research studying online movements that attempted to support minority groups. For example, Mondragon et al. [8] conducted a lexical analysis of Twitter discourse in #La manada. Mondragon et al. [8] then identified eleven main classes of the discourse and determined the phases of collective symbolic coping that emerged regarding the La Manada case in Spain from the perspective of Social Representations theory. Mondragon’s study motivated our research since there is a lack of research analyzing the main topics that occurred in Twitter discourse in #StopAsianHate from the perspective of Social Representations theory, generating a gap regarding this topic in the academia. For another example, Lee and Jang [12] studied the latent topics emerging in 259,456 English tweets in #StopAsianHate in the first week after the Atlanta shooting through structural topic modeling to identify the trends of the topics. However, Lee and Jang’s study only examined the latent topics of tweets in #StopAsianHate emerging in a 7-day period, while the #StopAsianHate movement lasted much longer than a week on Twitter. Research that studies the discourse of #StopAsianHate on Twitter in a longer time frame may better capture the long-standing topics that emerged. As for another example, Tong et al. [13] studied the main topics that appeared in Twitter discourse in both #BlackLivesMatter and #StopAsianHate to figure out the connections, similarities, and differences between the two movements. Although Tong’s study expanded the horizon of research and examined two movements on Twitter simultaneously, it did not research from the perspective of Social Representations theory, which meant that the nuances in the process in which the public collectively coped with the new challenging social movements were unrecognized. In this case, our research will study the primary topics covered in Twitter discourse in #StopAsianHate from the perspective of Social Representations theory in a four-month time frame to fill in the gap, shedding light on which phases of collective symbolic coping emerged regarding the #StopAsianHate movement on Twitter.

The rationale for our study to research on social media is that social media visualizes the Stop Asian Hate movement and the collective symbolic coping of it. Social media enable users to engage in open, spontaneous, and synchronous discussions to express their opinions regarding a specific event [14]. Hence, social media visualizes the nuances in discourse, actors (especially the most contributing actors), and the network between different actors, which are hard to be observed in the offline world and in the traditional media [15]. It is in this sense that social media decentralizes and formulates a "public sphere", in which one user can influence another, and the knowledge that poses a challenge to the established social order can be circulated and gradually assimilated into the common knowledge in a public setting [6]. In this case, social media allow us to assess the public’s collective symbolic coping process of the Stop Asian Hate movement by visualizing a decentralized public sphere. The rationale for our study to specifically choose Twitter to research on is based on Twitter’s popularity in English-speaking countries, its micro-blogging character, and its text-based character. Twitter is the most famous micro-blogging social media platform [16], and it had 368 million monthly active users in 2022 [17]. The micro-blog is different from the traditional blog, which has relatively shorter texts in each blog and thus has relatively smaller integrated size [18]. The users spend relatively less time writing up a micro-blog, and thus, the micro-blogging character triggers more synchronous discussions that better capture the nuances in the evolving phases of a newly emerging event compared to the traditional blog. Furthermore, Twitter has its strength in orienting the trend of public opinions. Its users engage in political activities when they are participating in discussions about politics or online campaigns initiated on Twitter [19]. There is much valuable data regarding individual involvement in political activities, such as social movements, on Twitter. In general, Twitter has a large number of active users and tweet content that are large enough to formulate a public sphere, and its micro-blogging character allows researchers to capture the nuances in the developing process of an event and study individual engagement in political activities. In this case, this study chooses to research the primary topics discussed in Twitter discourse in #StopAsianHate in order to determine which phases emerged in the collective symbolic coping of the #StopAsianHate movement on Twitter. This research is significant in its contribution regarding the fact that it may highlight the primary topics covered in the public's normalizing process of the Stop Asian Hate movement from the perspective of Social Representations theory. The knowledge produced by this research may have implications in informing the activists and advocates of the anti-racist movements regarding the following actions that need to be taken.

Therefore, the research questions (RQs) of this study are:

RQ1: What are the primary topics covered in #StopAsianHate on Twitter?

RQ2: In light of the Social Representations theory, which phases emerged in the collective symbolic coping of the #StopAsianHate on Twitter?
3. Method

3.1 Data collection method

As for the data collection method, we accessed data from academicwitteR, an R package that is useful for accessing a huge mass of Twitter data, using the #StopAsianHate. The reason why this hashtag was chosen is that it was a prevalent topic in the online movement #StopAsianHate on Twitter, and it was universally used by users from different backgrounds to express their emotions and opinions about the movement.

We used the data to create a Twitter database that contains 91,279 tweets written in English from March 17th, 2021 to August 20th, 2021. For March, we have 38 tweets; for April, we only have 3 tweets; for May, we have 32,710 tweets; for June, we have 16,557 tweets; for July, we have 18,839 tweets; for August, we have 23,132 tweets.

Having considered most of the data are concentrated between May and August, we chose the first week of May, June, July, and August as 4 periods we studied (see appendix). In total, 21,474 tweets about the online movement #StopAsianHate were extracted from the Twitter database we created before. To be specific, the first period from May 1st to May 7th contains 7,045 tweets, the second period from June 1st to June 7th consists of 6,063 tweets, the third period from July 1st to July 7th involves 5,685 tweets, and the last period from August 1st to August 7th is comprised of 2,681 tweets. Moreover, we merged data files of 4 periods into one and created a categorical indicator of time (see appendix) for further analysis of group comparison.

It’s worth noting that we didn’t delete data that were repetitive or irrelevant for two reasons. The first reason is that repetitive data can represent users’ support for specific content that belongs to what we want to study. The second reason is that irrelevant data do exist in this movement and can convey some interesting information.

3.2 Analysis of information

The Machine Driven Classification of Open-ended Responses (MDCOR) method and MDCOR software were used for conducting topic modeling on the data. MDCOR software was developed by Professor Manuel S. González Canché. MDCOR improves the existing topic modeling method by using 4 metrics to identify the optimal number of topics and by adopting a mixed-method approach to conduct topic modeling in order to comprehensively understand the topics [20]. Formal English was not required by machine learning to classify texts in MDCOR, and it is advantageous in processing short texts, such as the text-based posts on social media. MDCOR is, therefore, specifically useful for analyzing tweets [20]. Furthermore, MDCOR is a user-friendly and cost-free software with relatively high efficiency, and it makes the text analyzing process easier and more affordable for researchers [20].

Firstly, the software reads data and starts text cleaning. The frequency is counted and is divided by the number of documents. The results are compared, and the software will determine whether a word needs to be dropped or not. If it is greater than 3/10, this word should be deleted in the following step, or this word will be kept. Secondly, the software provides default values for later calculation of the 4 metrics. The number of topics that correspond to the lowest point of metric CaoJuan 2009 and the highest point of metric Deveaud 2014 will be chosen. MDCOR is then executed to produce the results of text classification and topic identification. Finally, a categorical indicator can be selected to execute a group comparison that measures the relationship between groups and topics as well as between topics.

4. Results

The discourse emerging in the first weeks of May to August 2021 in #StopAsianHate on Twitter were analyzed respectively using MDCOR.

4.1 Period 1: from May 1st to May 7th, 2021

According to MDCOR, the optimal number of topics emerging from the tweets of Period 1 was 9, but 6 of them were repeated in other periods. Therefore, only the 3 unrepeated topics were outlined as the results found in Period 1.

4.1.1 Topic 1: Asian American and Pacific Islander Heritage Month

15.13% of tweets in Period 1 highlighted the significance of Asian American and Pacific Islander Heritage Month in May, and these tweets also showed the recognition of Asian Americans and Pacific Islander American’s contributions to the US society.

4.1.2 Topic 4: Strong dissatisfaction with black racists

In Period 1, 14.33% of tweets indignantly expressed black people were mostly responsible for hate crimes and discrimination against Asians. These tweets also criticized black people for being too preoccupied with their own Black Lives Matter movement and disregarding the rights of other ethnic minorities.

4.1.3 Topic 5: Urging greater participation

Several tweets in this topic called for greater support for the Stop Asian Hate movement, while others expressed gratitude to various different organizations that joined this movement.

4.2 Period 2: from June 1st to June 7th, 2021

4.2.1 Topic 1: Requesting public attention to the Asian victims of hate crimes

In this topic, greater public awareness of the Asian victims of hate crimes was requested, along with calls for money
donations and other forms of help provided to these victims.

4.2.2 Topic 2: Negative attitude towards racist behaviors

Most of the tweets in this topic were not related to the Stop Asian Hate movement. However, there were some tweets criticizing a Serbian volleyball player who made racist gestures towards a Thai player in the FIVB Volleyball Women’s Nations League, showing an attitude of dissatisfaction towards racist behaviors.

4.3 Period 3: from July 1st to July 7th, 2021

4.3.1 Topic 1: Actions to stop Asian hate

21.79% of the tweets in Period 3 described the actions that people had taken for stopping Asian hate, and these tweets also encouraged those who had been harmed by racism to bravely stand up and speak up for themselves.

4.3.2 Topic 2: Stop Asian Hate movement merged with Black Lives Matter

In this topic, a majority of tweets shared black people’s experiences of being brutally abused because of their race, indicating the cooperation between the two anti-racist movements Stop Asian Hate and Black Lives Matter.

4.3.3 Topic 3: Requesting apologies from racist authors

21.81% of the tweets in Period 3 were calling for apologies from the authors who had published racist articles that fabricated facts about southeastern Asians.

4.3.4 Topic 4: Self-promoting advertisements

It was found in this topic that the most representative tweets selected by MDCOR were all not related to the Stop Asian Hate movement. Instead, they were all advertisements of a Youtube channel, which aimed to use the fame of the #StopAsianHate hashtag to increase the followers of that Youtube channel. Advertisements as such occupied 31.08% of the tweets in Period 3.

4.4 Period 4: from August 1st to August 7th, 2021

4.4.1 Topic 1: Harms

The word "make" appeared the most frequently in the tweets in Topic 1 of Period 4. The words with negative meaning were always followed the word "make", such as "hurt", "heart attack", "tortured", and so on. This was a reflection of the wide range of harms suffered by people who had posted the tweets. 22.49% of the tweets in Period 4 mentioned the experiences of being physically or mentally hurt by CIA or the military. Though the original tweets were traced, the context of the tweets was lost, so an accurate interpretation of these tweets could not be made.

4.4.2 Topic 2: Criticisms towards Limeng Yan

18.91% of the tweets in Period 4 were expressing dissatisfaction with Limeng Yan and her statements. Limeng Yan was the former post-doctoral researcher in the University of Hong Kong. She authored several research papers that claimed that the SARS-CoV-2 was created in the Chinese laboratory. She also published these papers on several social media platforms. These rumors created and circulated by Limeng Yan were influential and intensified the discrimination and stigmatization toward Asians, triggering substantial criticisms towards Limeng Yan.

4.4.3 Topic 3: Racist

The word with the highest frequency in Topic 3 was "racist". 31.7% of the tweets in Period 4 mentioned this word. Though it was unable to access the original tweets since the links were not working, this word reflected the rationales of the #StopAsianHate movement on Twitter, that were to share experiences relating to racism, to fight against racist discourse, to support the ethnic minority groups facing racists.

4.4.4 Topic 4: Spreading Stop Asian Hate content in different languages

24.65% of tweets were talking about using different languages to disseminate the content relating to the Stop Asian Hate movement in order to further resist discrimination towards Asians.

4.5 Group comparison

Before showing our research results in this part, we want to emphasize that we did a group comparison before qualitative understanding of the meaning of these codes, as the MDCOR article suggests [20]. However, for the convenience of our readers, we first represent the results of 4 codes and then show readers the outcomes of the group comparison.

In this part, we put 21,474 tweets which combine data of the first week of May, June, July and August together into MDCOR, and they fell into four categories: Code 1 was “people”; Code 2 was “Asian”; Code 3 was “much” and Code 4 was “racist”. Code 1, Code 2 and Code 4 frequently occurred in the first week of each month. The word “racist” reflected the rationales of the #StopAsianHate movement on Twitter to share experiences related to “racist”, to fight against “racist” discourse, and to support the ethnic minority groups facing “racists”. However, Code 3 “much” had never appeared in the previous classification. It might represent that people had made a lot of contributions to the #StopAsianHate movement or that Asian people had been hurt a lot.
4.5.1 Relationships between time and codes

When the independence test was running in MDCOR, the null hypothesis was that the main topics identified in the tweet content in all four periods were consistent, and the selected hypothesis was that the main topics identified in the tweet content in all four periods were inconsistent. As Figure 1 shows, the p-value generated was less than 0.01, so the null hypothesis was rejected, indicating that there was a statistically significant difference in the main topics identified in the tweet content across periods.

In Figure 1, blue represents positive relationships which means the connection between a certain code and tweets of a certain time period is stronger than we imagined. Red indicates negative relationships, which mean the connection between a specific topic and tweets from a particular time period is weaker than we thought. The uncolored boxes don't tell us anything statistically.

Concretely speaking, tweets from May 1st to 7th are more relevant to Code 2 and Code 3, and less relevant to Code 1 and Code 4. Tweets from June 1st to 7th are more likely to be posted on Topic 3 and Topic 4, especially Topic 4, and less likely to post on Topic 1 and Topic 2. Tweets from July 1st to 7th express opinions that are in more connection with Code 1 compared with Code 2, Code 3 and Code 4. Besides, users of this period of time are less likely to provide non-responses. Tweets from August 1st to August 7th are more likely to publish the content of Theme 3 and Theme 4, especially Theme 4, than Theme 1. Based on the analysis above, we can see how codes change over time: in the first week of May, people tended to talk about “Asian” and “much”; in the first week of June, people turned to discuss “racist” and “much”; in the first week of July, people preferred “people” and in the first week of August, people continued their discussion about “much” and “racist”.

![Figure 1](https://example.com/figure1.png)

**Figure 1** Association Between Time Periods and Codes/Non-responses. MDCOR tested the association between respondents in 4 time periods and 4 codes or non-responses and whether participants in a specific period were more likely to express their opinions on a specific topic or gave no responses via an independence test.

4.5.2 Relationships between codes

In addition, according to Figure 2, we can find that Topic 2 and Topic 4 are closely related. Their frequent racist words are "asian" and "racist", which also perfectly echoes the theme of "Stop Asian Hate".

![Figure 2](https://example.com/figure2.png)
5. Conclusion

Three unrepeated topics were found in Period 1, and they were Topic 1: Asian American and Pacific Islander Heritage Month; Topic 4: Strong dissatisfaction with black racists; and Topic 5: Urging greater participation. Two topics were found in Period 2, and they were Topic 1: Requesting public attention to the Asian victims of hate crimes and Topic 2: Negative attitude towards racist behaviors. Four topics were found in Period 3, and they were Topic 1: Actions to stop Asian hate; Topic 2: Stop Asian Hate movement merged with Black Lives Matter; Topic 3: Requesting apologies from racist authors; and Topic 4: Self-promoting advertisements. Four topics were found in Period 4, and they were Topic 1: Harms; Topic 2: Criticisms towards Limeng Yan; Topic 3: Racist; and Topic 4: Spreading Stop Asian Hate content in different languages.

In this study, the collective symbolic coping process of the #StopAsianHate movement was examined from the perspective of Social Representations theory. Twitter, a newly emerging "public sphere", exemplified the collective coping process of the #StopAsianHate movement. Through analyzing the main topics covered in Twitter discourse in #StopAsianHate in the first weeks of May, June, July, and August, the emerging phases of the collective coping towards the Stop Asian Hate movement were identified. It was identified from the results that the collective coping towards the #StopAsianHate movement had arrived at the phase of awareness. Though we had very limited access to all the tweets in #StopAsianHate, there were still 91,279 tweets that could be accessed and analyzed. The relatively large volume of tweets collected in #StopAsianHate indicated that this movement was considered socially significant by a majority of users on Twitter. For example, in Topic 5 of Period 1, some Twitter users were aware of the social relevance of the #StopAsianHate movement and thus called for more support for the movement. Twitter discourse in #StopAsianHate was also involved in the divergence phase. At the phase of divergence, a number of different interpretations of the event, which challenges the current social order, will manifest themselves in the public discourse [8]. The results revealed two conflicting topics. Specifically, Topic 4: Strong dissatisfaction with black racists in Period 1 contradicted with Topic 2: Stop Asian Hate movement merged with Black Lives Matter in Period 3. The majority of tweets in the former topic blamed black people for conducting most of the hate crimes towards Asians and only focusing on their own Black Lives Matter movement, while the latter topic showed cooperation between the Stop Asian Hate and Black Lives Matter movements. The contradictions that occurred in topics indicated that there were various interpretations of the #StopAsianHate movement, thus creating confusion. Hence, the phase of divergence emerged. Furthermore, some Twitter users saw engaging in the Stop Asian Hate movement as calling for more participation, such as Topic 1 of Period 3; while some deemed it fighting back against racism and demolishing the inferior images constructed by the racist individuals, such as Topic 3 of Period 3 and Topic 2 of Period 4.

However, the results did not reveal a consistent and conventional interpretation of the #StopAsianHate movement, nor did the integration of the #StopAsianHate
movement into the existing bodies of common knowledge. In this case, the results showed that the collective coping of the #StopAsianHate movement neither presented itself in the convergence phase nor in the normalization phase. In sum, the collective coping of the #StopAsianHate movement was found to involve the phases of awareness and divergence but not the phases of convergence and normalization.

The implication for this finding is that the #StopAsianHate movement was witnessed to have a good commencement with the public awareness of its social importance and a variety of actions have been initiated to support it, but a shared social representation of the #StopAsianHate movement has not yet been developed, and it was not interwoven into the common knowledge. Hence, continuous practical actions to stop Asian hate and the dissemination of anti-racist knowledge are still required, as the public coping of the #StopAsianHate movement has not reached the normalization phase.

Moreover, in Topic 4 of Period 3, it was found that some people were using the hashtag #StopAsianHate for self-promoting but not for supporting Asians, which deviated from the primary goal of creating this hashtag on Twitter. The content that was unrelated to stopping Asian hate might cover the voices of the people who really needed help, as well as other contributing content. The implication for this is that measures for decreasing the unrelated content in the hashtag, such as commercial advertisements, may be conducted to make the related content more visible.

There were a few imitations recognized in our study. First, the volume of tweets analyzed in this study was limited. Tong et al. [13] identified a peak in tweet volume in #StopAsianHate immediately after the Atlanta shooting event on March 16th; while Lee and Jang [12] extracted 259,456 English tweets in #StopAsianHate in the first-week period after the Atlanta Shooting. However, for this study, we only accessed 38 tweets in #StopAsianHate from March 17th, the day after the Atlanta shooting, to March 31st; and even less, only three tweets could be accessed in April. In addition, the time frame of the data used in this study only covered the first weeks of four months, that was, twenty-eight days; while the #StopAsianHate movement on Twitter lasts much longer than that. As a result, only 21,474 tweets in #StopAsianHate on Twitter were analyzed, and the very limited access to tweets might cause some significant main topics unrecognized in this study. Future studies may analyze a larger volume of data in a longer time frame as to make more main topics covered in Twitter discourse in #StopAsianHate be recognized, giving a better understanding of which phases of collective symbolic coping emerged in #StopAsianHate movement.

Second, the tweets analyzed in this study suffered from a lack of user demographics and context. It was difficult to comprehensively understand the tweet content when ethnicity, social class, income, and other demographical information about the users was lost [13]. The loss of context, such as what influential event happened when the tweet was posted, also prevented us from accurately identifying the main topics discussed in Twitter discourse in #StopAsianHate. Future studies may adopt a mixed-method approach and interview some of the users who posted tweets in #StopAsianHate offline. Offline personal interviews with demographical information about the interviewees might offer analysis with a more detailed context of tweets. Future studies may also cross reference the users’ public accounts (if could be found) on other social media platforms to improve the understanding of user demographics and the context.

Third, only English tweets were studied in this research. It was possible that a majority of tweets in #StopAsianHate were written in Asian languages, and hence, limiting the research scope to English tweets might leave the main topics that emerged in Asian discourse unrecognized. The voices of Asians themselves are of incomparable value for studying the primary topics covered in #StopAsianHate on Twitter. Future studies may therefore study the tweets in Asian languages in #StopAsianHate using software that is specifically apt for conducting topic modeling on texts written in Asian languages.

APPENDIX

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>df = pd.read_csv('StopAsianHate_wh.csv')</td>
<td>Import pandas as pd</td>
</tr>
<tr>
<td>df['Date'] = df['created_at'].str[10]</td>
<td>df['Date'] = # created_at str[10]</td>
</tr>
<tr>
<td>df.head()</td>
<td>df.head()</td>
</tr>
<tr>
<td>author_id</td>
<td>text</td>
</tr>
<tr>
<td>1</td>
<td>2021-08-17</td>
</tr>
<tr>
<td>2</td>
<td>2021-08-17</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
<tr>
<td>df['text'] = df['text'].str[10]</td>
<td>df['text'] = # text str[10]</td>
</tr>
</tbody>
</table>

Figure 3 Codes to Extract Data from the Twitter Database in Python.
We changed the encoding format of the Twitter database to UTF-8 so that we could import it to Python (figure 3). Then we created a column called “Date” and extracted the top 10 elements of “created_at” to normalize the date information. After that, we transferred the data format of “Date” to datetime, filtered out the data of 4 periods and saved them as “May_week1”, “June_week1”, “July_week1” and “August_week1” respectively.

As figure 4 shows, we merged 4 files above into one, created a new column called “Time category”, assigned the 7th character of “Date” which represents months, to “Time_category” and recoded “5” which means May to 1, “6” which means June to 2, “7” which means July to 3 and “8” which means August to 4. At last, we saved this file as “merged”.

In order to fully state the process, we take “merged” for example to explain specific steps we took. In step 1, we clicked “Browse” and uploaded the file we wanted to analyze. Then we chose “author_id” as the ID column and “text” as the Text column in “File upload information” (Figure 5).
In step 2, we clicked “Execute Text Mining” and got “Initial Text Mining Output”, based on which we decided to drop “stopasianhate” and input 1 in the blank of step 3 (Figures 6,7). We clicked “Trim Common Words” and got “Trimming Common Words Output”.

MDCOR dropped 58 responses due to sparsity that evaluates how commonly terms are used in texts. We focused on term frequency results. To identify codes better, we needed to delete terms with a high frequency which might occur in many responses and reduce the distinction between texts. We set a standard to measure how high frequency was based on an example in an article of MDCOR. The article assumed a word was used once among responses that included it, which was the worst situation of commonality of word usage among participants, and considering the prevalence of the word that 3 out of 10 respondents used it was not overwhelmingly popular to merit its removal [20]. Thus, we decided that if the ratio of frequency of a word to the number of remaining responses is greater than three-tenths, the word should be cut off; otherwise, it should be kept. In this case, we deleted “stopasianhate” which was used by about 83.62% (or 21416) respondents.
In step 4, we kept the default values as the article of MDCOR suggested [20]. In step 5, we clicked “Execute Metrics” and got “Metrics Output” (Figures 8,9), based on which we tried 2 numbers that are 2 and 4 in step 6. We clicked “Execute MDCOR” in step 7 and got “Results of 2 Codes” and “Results of 4 Codes”. We compared the results of 2-solution and 4-solution and decided that 4-solution was the most optimal choice on average. We downloaded full MDCOR data and the top 20 most representative cases per code of 4-solution for further interpretation. In step 8, we chose “Time_category” as our categorical indicator and clicked “Execute Group Comparison”. MDCOR represented “Association Between Time Periods and Codes/Non-responses” and “Code Relationships”.

The Y-axis of the upper graph represents the correlation between matrices, the Y-axis of the lower graph represents the dissimilarity between matrices and the X-axis of both graphs represents the optimal number of codes. We wanted to find a number that could minimize correlation and maximize dissimilarity between matrices. However, the results shown by Arun2010 and Griffiths2004 suggested we choose the number of topics as many as possible, which could bring much complexity in code identification. Thus, we determined the optimal number of codes based on the metrics of CaoJuan2009 and Deveaud2014. We chose the lowest point of CaoJuan2009, which means the least correlation and the highest point of Deveaud2014, which means the most dissimilarity. If two points corresponded to the same number of topics, it might be the optimal choice; otherwise, two numbers would be tried. If the greater one could classify texts clearly, we accepted it because we wanted to keep more diversity without losing much distinction. In this case, we got 2-solution and 4-solution.

References


