COMP 598 Final Project: COVID-19 Discussion Topics in Social Media

Zichuan Guan 260870895*, Ade Thornhill 260896212 *, Stanley Wu 260909721*,

[1]McGill School of Computer Science
McConnell Engineering Bldg. Room 318
3480 University St., Montréal, Québec, Canada H3A 0E9
zi.guan@mail.mcgill.ca, ade.thornhill@mail.mcgill.ca, stanley.wu@mail.mcgill.ca

Introduction

Since the start of the pandemic almost 2 years ago, COVID-19 has been consistently under the spotlight by world leaders, major news source, and ultimately the general population. Several tools serve as a way of communicating any COVID-related information, with the most easily accessible and widely used being social media. Therefore, it is of great interest to analyze social media platforms to determine the current discussion space of COVID on social media, which is precisely what we have performed in this study. More specifically, we fetched 1000 COVID-related tweets, developed the most salient topics based on these tweets, computed the relative engagement for each topic, and determined the overall sentiment to the pandemic/vaccination. The developed topics were Policies and Health/Sanitary Measures, Daily Life Impact, Economic and Social Impacts, Breakthroughs, Pandemic Severity, and Vaccination. We found that engagement tends to involve personal matters, notably for the most common topics, which are Policies and Health/Sanitary Measures, Daily Life Impact, and Vaccination. More importantly, the overall response towards the pandemic and vaccination is more negative than positive.

Data

We collected 1200 tweets using the Twitter API. The tweets spanned evenly within a three day window, from November 23 to November 25, 2021. To analyze COVID discussions, we filtered the tweets with the API query such that they need to

- have an English language tag;
- contain at least one of the following keywords: *covid*, *covid19*, *coronavirus*, *vaccination*, *pfizer*, *moderna*, and *astrazeneca*; and
- be original tweets (i.e. not a retweet, reply or quote).

The Twitter API is case insensitive and searches not only the tweet's text but also the contents in the link if it has one. We decided to only collect original tweets to eliminate duplicates in order to focus more on the centre of discussion. We also included public metrics such as retweet count and like count in our query.

The tweets are then manually annotated on their topic and sentiment. Invalid tweets (e.g. incomprehensible tweets, tweets with only a broken link) and irrelevant tweets (e.g. spams, tweets with multiple unrelated hashtags to gain exposure) are filtered out during the process to have 1000 annotated tweets.

Methods

Data Collection

Fetching the Data We have created a script in Python, relying on the Twitter API, as mentioned above. Specifically, the script sends GET requests to the API endpoint representing the most recent tweets. In our GET request, we send our authorization token wrapped in the headers of the request, and the keywords, start/end time, and desired fields (such as public metrics and context annotations) contained in the parameters in the request. The GET request is located in a loop, and in each loop, the start and end times increment by 1 day (in the 3 day window). This is repeated until 1200 tweets are fetched, before being processed into a TSV format, containing the id, metrics (retweet count, reply count, like count, and quote count), tweet message, coding, and sentiment as headers.

Filtering In addition to filtering by passing the keywords to the API query, to filter the invalid and irrelevant tweets (defined in the data section), we manually inspected and replaced them with a randomly sampled tweet from the remaining 200 tweets out of 1000. We were careful to not remove any tweet that was almost incomprehensible or almost unrelated, in order to avoid modifying the discussion space to fit the data to our developed typology. We have also considered removing any tweet containing a link (as many invalid or irrelevant tweets contain links). However, we decided to proceed with manual filtering as we believe that the automatic removal of links will filter out many tweets potentially belonging to a topic.

Data Annotation

Topic Design Our open coding phase was executed in a trial and error fashion. Initially, we skimmed through the 200 tweets and hypothesized potential topics based on the contents of the tweets and based on previous studies on the most common COVID topics on Twitter. After a few rounds

*These authors contributed equally.

of coding, we eventually agreed on a set of topics that attempts to maximize comprehensiveness while minimizing subjectivity. To ensure that our topic definitions were well-defined and to avoid misinterpretation, we have created a code book containing the definitions of each topic along with some common examples.

Annotations We have split the collected tweets into 3 equal sections, so each member annotates 333 tweets. During our annotation process for the topics and for the sentiments, in addition to referring to the code book to maintain consistency, edge case tweets were shared and reviewed by all 3 members.

In the process of categorizing the tweets, paired with the message of a tweet, we decided to factor in other information that can help to infer/confirm the context of the tweet, which includes information such as the twitterer and his/her pasts tweets, the contents of any potential links and photos, and the framing of the tweet itself, corresponding to the type of language used (for example, medical language or advertising language).

For the annotation of the sentiments, in general, we have decided to primarily evaluate a tweet based its tone, such as the usage of exclamation marks, emojis, and interjections, rather than the content of the tweet itself. This is because the subject of the discussion space of the tweets, corresponding to the pandemic, is negative by default, and by this argument, tweets reporting a statistic about the pandemic will be categorized as negative, when we believe that citing numbers are objective, and therefore its sentiment should be neutral.

Data Analysis

Before computing the tf-idf scores, $\textit{tf-idf}(t, d, D) = \textit{tf}(t, f) \cdot \textit{idf}(t, D)$, where t denotes the word, d denotes the the topic, and D denotes all 1000 tweets, we preprocessed the data. Specifically, we removed common stopwords and punctuations, and we also performed lemmatization (using the NLTK library). This avoids having the singular and plural versions of the same word (e.g., ”vaccines” vs ”vaccine”) listed in the same top 10 words by tf-idf score output of a topic, and overall allows for a deeper insight on the characteristics of the topic.

We also analyzed public metrics to better understand the preference and popularity of each topic ideas. Total number of likes and retweets are calculated for each topic and by sentiment. This allows a clear view on the distribution of such metrics over our point of interests.

Results

Salient Topics Found

We have developed 6 topics making up our typology as a result of the open coding. These 6 topics are (1) policies and health/sanitary measures, (2) daily life impacts, (3) economic and social impacts, (4) breakthroughs, (5) pandemic severity, and (6) vaccination.

1. Policies and Health/Sanitary Measures Definition This topic includes any discussion (which can be opinionated) relating to any policy on limiting the spread of COVID-19, and any health/sanitary measure implementing these policies, such as social distancing, travel bans, and lock-downs. Categorizing a tweet to this topic is independent of the twitterer, but dependent on the generalization of the meaning of the tweet. More specifically, the context of tweets in these topics are societal, which in general corresponds to any group of people defined by an administrative division, while any reference or hint to a personal impact caused by a COVID policy will automatically categorize the tweet to the next topic, Daily Life Impacts. Note, however, that this topic strictly excludes vaccination, since we believe that defining vaccination as a separate topic from the topic of policies allows for a more explicit and direct evaluation of the sentiment of the response to vaccination in general.

2. Daily Life Impact Definition This topic regroups any tweet that refers to an individual or small-scale impact or effect (hence the term ”Daily Life”) caused by anything related to COVID, including (more commonly) the reaction to the impacts of COVID. Common examples of causes of the impacts could be the virus itself, a health measure, a political person linked to the pandemic, or an economic consequence of COVID. The scale of the individual impact ranges from 1 person to a small community or group of people, such as a school, hospital, or party. Common examples of tweets in this category include rants on the pandemic, opinions on political figures (e.g., insults, wishes, questions) and re-opening of events. Moreover, similarly to the previous topic, any mention of vaccinations, even if referencing personal concerns or facts, will be categorized in the vaccination topic instead.

3. Economic and Social Impacts Definition This topic categorizes all tweets mentioning any impact of COVID on markets, industries, or entire economies. These include factors and metrics such as stock market prices, inflation rates, prices of commodity (such as oil), and unemployment rate. This topic also includes social impacts, corresponding to effects on people and communities at a large scale due to the pandemic, such as inequality, financial hardship, or national holidays during the pandemic.

4. Breakthroughs Any tweet referencing a novelty will be classified as Breakthrough. Specifically, breakthroughs primarily include the development or use of new technologies, any new scientific discoveries relating to COVID (such as the development of a new vaccine), any new virus variants, and any new studies relating to COVID. In general, tweets in this topic should be objective.

5. Pandemic Severity This topic consists of tweets primarily mentioning statistics and facts about the pandemic, notably the death rate, infection rate, forecasts, warnings, and vaccination rates, which, in this case, do not belong to the topic Vaccination. Similarly to the topic above, tweets in this topic should also be objective.

6. Vaccination This topic regroups any tweet mentioning anything about vaccines, factual or opinionated, excluding those that state objectively the vaccination rate, which belongs instead to Pandemic Severity. Unlike the first two topics, the context of the references to vaccinations can be individual to societal. Examples include vaccination opinions, concerns, effectiveness, and booster shots.

Topic Characterization

Table 1 lists the top 10 words by tf-idf score for each topic, following the data analysis methodology explained above.

From table 1, for Policies and Health/Sanitary Measures, as expected, words referring health regulations, such as ”passport”, ”restaurants” (potentially indicating the health regulations, such as vaccine passports, implemented in restaurants), ”masks”, and ”measure” are among the highest tf-idf score words. Most of the words in the list denote objects and places, and are therefore general and objective by itself, with the exception of ”gop”, most likely standing for the Republican Party in the U.S.

For Daily Life Impact, the words are more specific to events, highlighted by ”toe”, ”rodgers”, and ”aaron”, referring to the NFL player Aaron Rodgers’ toe injury and its connection his positive COVID test recently. Moreover, when interpreting the words alone, there are more words that belong to an opinion-based context, such as ”fuck” or ”positive”, unlike topic 1. The role of opinions in Daily Life Impacts also explains the incoherent words in the rest of the list.

The list of words obtained in the topics Economic and Social Impacts are coherent to its topic definition, as most of the words are directly related to the economy, markets, or society, such as ”price”, ”stock”, ”job”, ”economy”, and ”program”. All of these words, when interpreted alone, are also neutral, which aligns with the topic definition. This is also the case for the topic Breakthroughs, where we have words related to research and novelty, such as ”mutated”, ”variant”, ”research”, and ”scientist”.

For Pandemic Severity, a surprise is that terms referring to data science, such as ”analytics” or ”datavisualization” have a high tf-idf score, when we were expecting reporting terms, such as ”reports” or ”cases”, to be on the top of the list.

Finally, it is expected that ”vaccine” sits at the top of the list of the topic Vaccination. The occurrence of ”baby” and ”age” suggests a discussion trend about vaccination for different age groups. However, the relation between vaccination and ”musician” is unexpected.

Topic Engagement

Table 2 shows the distribution of sentiments for each topic. We can observe a higher number of posts in topic 2, followed by 6, 1, and 5, with topics 3 and 4 being the least discussed. From figure 1, we can observe that topic 2 is highly liked and retweeted among all the topics. Although like and retweet counts primarily depend on the popularity of the original poster, they correlate with the observed distribution of posts.

On the sentimental aspect, an observed pattern shared by all 6 topics is that each topic has more neutral tweets than negative tweets and more negative tweets than positive tweets, and that the overall sentiments are skewed towards the negative side, but the difference between the sentiments vary. Specifically, for Policies and Health/Sanitary Measures, there are 38 more tweets that are negative than positive, and this difference is the largest among all topics. Moreover, based on the tf-idf score list, the most referenced topics in terms of health regulations are vaccine passports and masks. For Daily Life Impacts, most individual opinions referred to Aaron Rodger’s COVID toe case. For Economic and Social Impacts, twitterers seem more engaged at economic factors, rather than social factors, illustrated by ”price” and ”stock” having a higher tf-idf score than ”support”, ”job”, or ”program”. For Breakthroughs, engagement is more towards new COVID variants, demonstrated by the words ”raise” (concerns), ”heavily”, ”mutated”, and ”variant”. This also aligns with a higher negative tweet count than positive tweet count. For Pandemic Severity, on the surface, attention seems to be more directed at statistics of the pandemic obtained from a data analytics method. Finally, for Vaccination, vaccines, booster shots, and in some ways, babies seem to be the main focus for this topic.

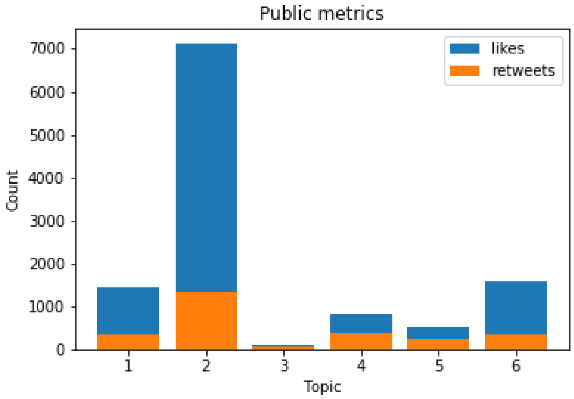

Figure 1: Like and retweet count per topic

Figure 2 reveals further trends. While most topics have likes and retweets skewed towards the negative side, topics 2 and 6 show the opposite. Although topics 2 and 6 have both less positive tweets than negative or neutral tweets, their like counts dominate slightly.

Discussion

General Overview

The overall skew towards a negative sentiment for all 6 topics suggests a negative response to the pandemic and vaccination on social media. This is also supported by the public metrics, with the exception of topics 2 and 3 which will be detailed below. Moreover, the discussions seem to be largely personal. This is expected for social media posts. In fact, people seem to care more about vaccination, various sanitary measures put in place to counter the pandemic, as well as the severity of the pandemic. These topics have direct implications on the general public, while the broader topics - social-economical impact and COVID-related scientific breakthroughs - are the least discussed. This polarity suggests that social media posts tend to involve personal interest.

topic	word	tf-idf score	topic	word	tf-idf score	topic	word	tf-idf score
1	passport	8.9588	2	toe	34.0434	3	price	4.3944
	mask	8.1093		rodgers	26.8764		stock	3.2958
	restaurant	4.3944		aaron	23.2929		support	2.7726
	measure	4.3944		fuck	17.9176		foundation	2.1972
	worst	4.1589		positive	12.0847		economy	2.0794
	school	4.0547		initially	10.7506		celebrate	2.0794
	vaccine	4.0111		denied	10.7506		job	2.0794
	variant	3.4641		home	9.7041		china	2.0273
	five	3.2958		lost	8.9588		pandemic	2.0055
	gop	3.2958		wa	6.7459		program	1.3863
4	raise	17.9176	5	analytics	32.2517	6	vaccine	20.7847
	heavily	16.1258		usafacts	16.1258		baby	8.9588
	mutated	16.1258		datavisualization	16.1258		booster	6.3813
	variant	7.8398		datascience	16.1258		vaccinated	4.558
	scientist	4.4601		columbia	10.7506		dead	4.3944
	evade	4.3944		infection	6.3813		musician	4.3944
	horrific	4.3944		positive	5.4931		refusing	4.3944
	nu	4.3944		warns	4.3944		age	4.3944
	research	3.2958		recorded	4.3944		available	4.3944
	south	2.9171		yesterday	4.3944		forced	4.3944

Table 1: Top 10 words by tf-idf score for each topic

topic	positive	neutral	negative	total
1	13	128	51	192
2	51	160	82	293
3	9	36	12	57
4	2	54	14	70
5	5	149	22	176
6	38	120	54	212
total	118	647	235	1000

Table 2: Distribution of tweets by sentiment and topic

Looking at figure 2, we can approximate 3 different distributions based on the like and retweet activity. The first distribution has more noticeable negative like/retweet counts than positive counts, with neutral counts being the highest. This corresponds to topics 1 and 3. The second distribution seems bi-modal, where we observe higher like/retweet counts for the positive and negative tweets than counts from neutral tweets, which is the case for topics 2 and 6. The third type of distribution resembles a normal distribution, fitting topics 4 and 5. Although we cannot conclude anything with certainty, since the sample sizes are small (especially for topic 3), we hypothesize that a similar like/retweet activity distribution could imply similar characteristics between topics. A possible explanation is that topics 1 and 3 regroup discussions concerning the general population (e.g., health regulations and policies are societal), topics 2 and 6 contain many opinionated and individualized tweets possibly explaining the emphasis of the like/retweet counts on the positive/negative sentiments, and topics 4 and 5 are generally more based on facts (objective), which is reflected by the higher like/retweet activity on neutral tweets than on positive/negative tweets.

Interpretation on Individual Topics

1. Policies and Health/Sanitary Measures The list of the top 10 tf-idf scores and the sentiment results indicate a negative response towards vaccine passports, mask mandates, the GOP, and health measures in general. Overall, the heavier weighting towards a negative response suggests that COVID-19 related health policies are controversial. However, based on our method of sentiment annotation, we cannot conclude that the health regulations of masks and vaccine passports are unpopular by the general twitter discussion space, since tweets can also react negatively to any opposition and violation to these policies.

2. Daily Life Impact The tf-idf scores highlighting the prevalence of the COVID toe symptom of Aaron Rodgers indicates that a majority of individual opinions and reporting on any COVID-19 related impacts on social media are centred around trending topics, while the rest of the words in the tf-idf scores relate to conveying personal opinions on the pandemic, such as ”fuck”, or personal experiences with COVID, such as ”positive”, meaning ”tested positive” when referring back to the dataset.

In fact, the high frequency of tweets referring to the COVID toe of the NFL player demonstrates how time-sensitive our design of Daily Life Impacts is to short-term trends, since the spread of the information of this incident happened around November 24. This indicates that if we were to slide the 3 day window to a different period, we would have most likely observed another COVID-19 related trend (if any) that does not involve health regulations, the virus, or vaccinations, which would have been reflected in the top tf-idf scores. For the scores list to be less sensitive to trends, we would have had to increase the number of days in the window.

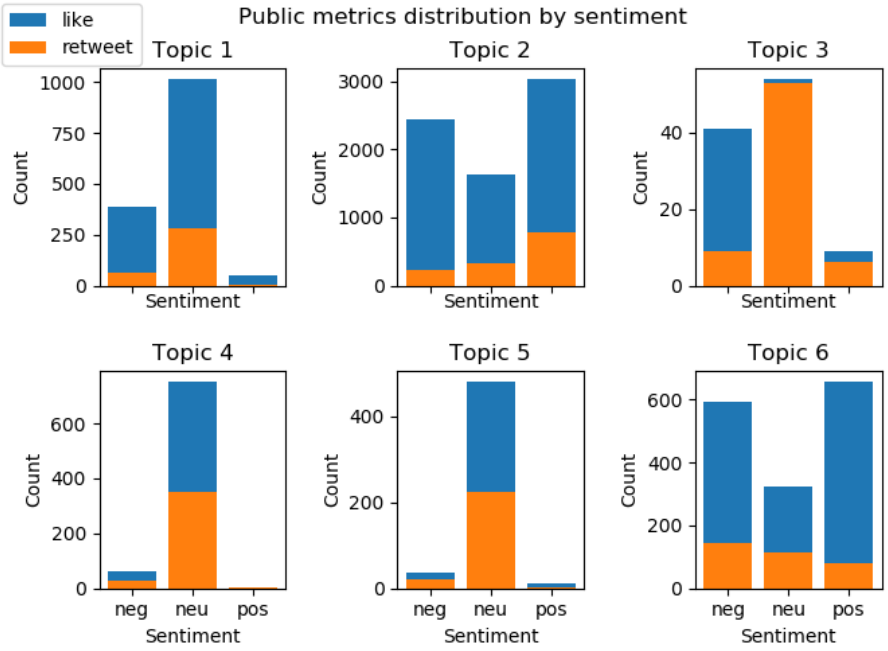

Figure 2: Distribution of likes and retweets by sentiment for each topic. The sentiment labels ”neg”, ”neu” and ”pos” refer to ”negative”, ”neutral” and ”positive” respectively.

An interesting observation is the high scores for ”initially” and ”denied”, both of which refers to the message ”COVID-19: Fully vaccinated B.C. man, initially denied COVID vaccine card, gets his card”. Based on the dataset, this message appears 4 times, each with a different tweet ID, and 3 with the same news link article, but with a different original twitter account. We believe that this corresponds to a type of spam that we have not taken into account when designing our manual filtering process. Nevertheless, this highlights the pervasiveness of spam messages in the COVID discussion space on Twitter.

Furthermore, the higher like count and retweet count for positively annotated tweets in this topic is worth investigating. We found that the most-liked tweet is classified as positive. It is also the most retweeted post with 2210 likes and 689 retweets, which dominates over all other tweets. The tweet promotes a new book by Scott W. Atlas: ”My new book is out! ”A Plague Upon Our House: My Fight at the Trump White House to Stop COVID From Destroying America””. Other positive tweets mostly refer to appreciation to help received and to supporting decisions made to combat the pandemic.

3. Economic and Social Impacts The top 10 tf-idf scores list of this topic are more uniform, unlike the other topics. This could be explained by the relatively small sample size, but also by the topic design. In fact, Economic and Social Impacts is a broad topic possessing many keywords, similar to the words on the list, which could potentially contribute to the uniformity of the scores.

Furthermore, the number of positive tweets (9) is almost equal to the number of negative tweets (12). Looking at the original dataset, positive tweets referred to the re-opening of social events, such as holiday shopping, call for actions for a societal change in a positive tone, and positive news on stocks and sales. Negative tweets commonly evoke the negative impacts of COVID on the economy, jobs commonly referring to unemployment, and workers, such as nurses, or employers being overworked in general, all in a negative tone. Therefore, although positive and negative counts are approximately balanced, the reason for this observation is not purely random, as the sentiment of a tweet is not only determined by the language used but also influenced by the subject it is referring to.

We also observe the highest retweet to like rate for neutral and positive sentiments. This may suggest that people are more likely to share the tweet if they also like it. However, the data size is not sufficient to make any significant conclusions.

4. Breakthroughs For this topic, we have originally envisaged a primarily neutral sentiment. Nonetheless, the results indicate 14 negative tweets out of 70. However, in our annotation, negative words, notably ”heavily” and ”horrific, were usually placed in an objective context, such as ”New Botswana variant with 32 ’horrific’ mutations”, and therefore were annotated neutral. Referring to our original dataset, we found that there are still a handful of the tweets on new variants, vaccines, and research related to the pandemic that have been reported in a personal context, with the use of exclamation marks and emojis, for example. This demonstrates that any message on scientific breakthroughs on social media do not always come from experts, such as researchers and medical doctors, and indicates individual engagement and viewpoints on new discoveries.

5. Pandemic Severity Although we were expecting terms related to reporting and numbers to be at the top of the list, table 1 indicates that terms related to Data Science have the highest tf-idf scores instead. One possible explanation could be that in the tf-idf formula, $idf(t, D)$ rewards more unique words and penalizes more common words. This suggests that data science terms are more unique to this topic, while terms on reporting, surges, deaths, and case numbers are also common in other topics, which in turn decreases its score. When inspecting the dataset, this seems to be the case: for example, while the term ”death” is frequently used in tweets belonging to this category, this term is also common in topics 1, 2, and 6, thereby lowering its tf-idf score.

As expected, the majority of tweets in this topic are neutral. From the dataset, for tweets that were annotated as positive or negative, similar to the reasoning in topic 4, the original twitterers did not correspond to news source accounts, but rather individual users who added their opinions when reporting some statistics or when linking to a news article.

6. Vaccination The list of the top 10 tf-idf scores contain both coherent (such as ”vaccine”, ”booster”, ”vaccinated”) and incoherent (namely ”musician” and ”baby”) terms. We also observe terms that possess a more negative sentiment by default, such as ”dead” and ”forced”, which most likely contributes to a stronger negative sentiment for the topic.

Interestingly, the higher difference between negative and positive counts in topic 1 vs topic 6 could imply that mask and vaccine passport mandates are more controversial than vaccination mandates. However, based on our topic design, this could also be due to the individual aspect of vaccination, such as the type of the vaccine received and its feeling, placing topic 6 on both the individual and societal level, while passports and mask mandates are (generally) more societal. In other words, giving an individual opinion or fact (which vaccine did I receive?) on vaccination is more common than communicating some individual message on other health regulations.

This topic is the other aberrant case based on figure 2, with higher like counts on positive tweets and almost equal retweet counts for all sentiments. However, it is a different case than topic 2, with no clearly dominating tweets. Several of the top-liked tweets are positive and have similar like counts. All of them are about people taking booster shots and calling for actions to increase vaccination rates. This indicates that even though the tweets are mostly negative and neutral, the preference is still slightly positive towards vaccination.

For the incoherent terms, referring back to the original 1000 tweets, the reason for the high tf-idf score of the word ”musician” is because of the 5 tweets referring to the death of a pianist partially caused by a vaccine jab being categorized as Vaccination. The high tf-idf score for ”baby” relates to a raise in concern about the effects of vaccination in pregnant women to the newborn babies. After investigating the original tweets, people seem to have a negative sentiment over this subject matter. In fact, several posts relate to a higher rate of stillborn babies having vaccinated mothers.

Group Member Contributions

- Zichuan Guan contributed to the data collection, data annotation, data analysis, interpretation, and the writing of this report.
- Ade Thornhill contributed to the data collection, data annotation, data analysis, interpretation, and the writing of this report.
- Stanley Wu contributed to the data collection, data annotation, data analysis, interpretation, and the writing of this report.