Topic Modeling: COVID-19 Early Social Concerns

Haotian Liu

Faculty of Information, University of Toronto

INF 2209: Human-Centered Topic Models

Dr. Shion Guha

November 21, 2022

Abstract

Nearly three years have passed since the outbreak of Covid-19 in early 2020, and the public has been affected to varying degrees. From the early panic about the virus to the present calmness, people's attitude towards the virus has undergone different degrees of change. We were interested in how people's attitudes changed during the pandemic and the public's primary concerns. What were the most important topics of concern during the early period of the spread of the covid-19? Studying this issue can effectively help us understand the essential factors that people consider in a pandemic so that we can better regulate these factors to ease peoples' tension and respond more effectively to a pandemic in the future. In this report, we analyze the tweets data during the early pandemic using topic modeling methods to explore peoples' concerns.

Keywords: Topic Modeling, NLP, Covid-19

Introduction

Covid-19 has had a significant impact on our society nowadays since 2019. As of November 7, 2022, more than 4 million people have been infected in Canada (Government of Canada, 2022). The infected people will suffer from varying degrees of symptoms, such as sneezing, fevers, shortness of breath, etc. After the infection, some people will suffer from different sequelae. Meanwhile, due to epidemic prevention and control, people's daily life has been affected, such as wearing masks in public areas and limiting the maximum number of people gathered. During the early spread of the virus, the lack of awareness led to shortages of essential items such as food and toilet paper. Most of the Covid-19 information is spread from social media and everyday conversations. Among them, Twitter is the most popular microblog social media, and there were around 300 million users in 2020 (Dixon, 2022; Jay, 2022). Posting tweets has become the best way to spread information about Covid-19 early on. Studying these tweets can help us better understand what people are considering during the pandemic and how people's opinion about Covid-19 has changed. In this project, we collect the actual tweets data in 2020 and then apply four topic modeling methods, LDA, NMF, BTM, and CorEx, to retrieve the topic from the text. After that, we study topics from each model and compare them to each other to determine which method can better determine the topic for the text data.

Related Works

There are some studies that have already focused on Covid-19 according to the

tweets data from Twitter. Shahi et al. (2021) studied the misinformation spread on Twitter during the early stage of the pandemic. Kwok et al. (2021) focus on the sentiment of vaccination against Covid-19 in Australia based on Twitter data. Wicke and Bolognesi (2021) studied how people's sentiments and topics changed during the pandemic. Our study will focus more on the people's general concerns during the pandemic and whether these concerns have changed over time. After that, we also include qualitative analysis based on the original tweets text to determine which topic model can conclude the topic best.

Data Description

The tweets data we used was pulled from Twitter and available on Kaggle (Miglani, 2020). It includes over 30,000 data with username, location, Tweet at, original tweets, and label. Among them, the username is represented by a unique ID due to its confidentiality. The "Location" shows where the user comes from, and "Tweet At" shows the time the user posts the tweets. In addition, the label shows the sentiments of the tweets, and the original tweets are the text data we mainly used in the report.

Data Preprocessing

The preprocessing of the data is focused on time and text. We extract the month from January to December and use 1 to 12 to represent them. Then, we categorize 1 to 3 into spring, 4 to 6 into summer, 7 to 9 into autumn, and 10 to 12 into winter. There are additional process techniques used for the text data. First, we set all text to lowercase,

then removed punctuation, and read errors and words containing numbers. After that, we see our text still contains symbols "\n" and "\r\n." Therefore, we remove these two characters from our text. Then based on figure 1, we see some words that may not make sense in our analysis, such as "https," "amp," and "tco." Therefore, we add these words to the stop words list. In addition, words such as "coronavirus" and "covid" are irrelevant to our topic words' results and may affect our analysis. Again, we add these words to the stop word list. After that, we tokenize tweet text, remove the stop words, and then lemmatize words to get each tweet's terms ready for each topic model.

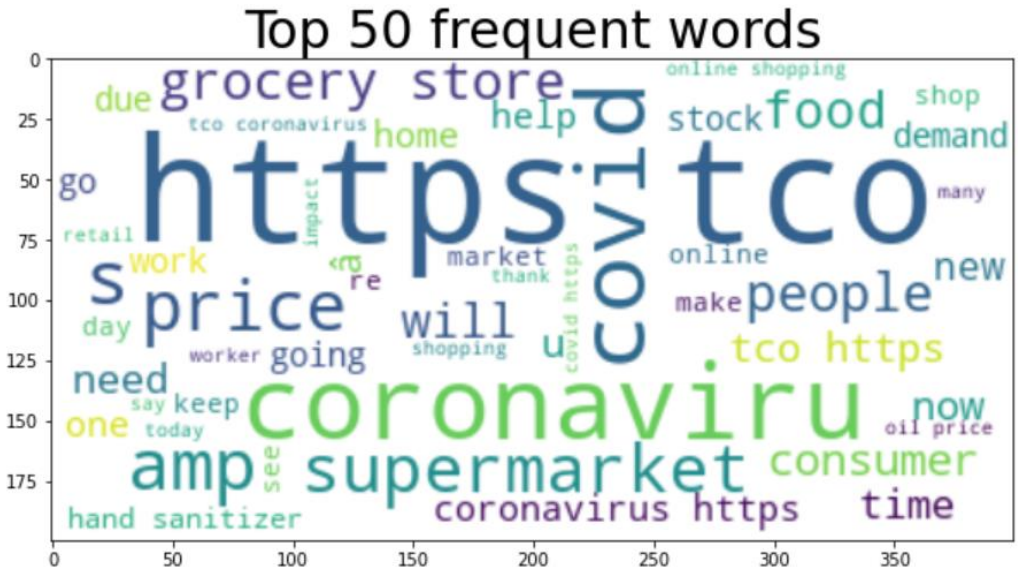

Figure 1. Data Word Cloud

Methods

For each experiment, we apply the following methods:

LDA - Latent Dirichlet Allocation:

LDA is one the famous topic modeling method that can generate three matrices: document with topics, topics with words, and topic weights. We use a document with a

topic matrix to find the top tweets within each topic and then use the topic with words to find topic words within each topic. We manually set topic numbers from 2 to 50 and calculated the perplexity and coherence. Then we determine the optimal topic number according to the intersection of perplexity and coherence and print top words and tweets within each topic.

NMF- Non-negative Matrix factorization

NMF also split the document-term matrix into matrixes. Unlike LDA, NMF only breaks it into two matrices, the topic with words and the topic with documents. We still loop the topic number from 2 to 50 and print the perplexity and coherence score to find the optimal number. After that, we print each topic's top words and tweets for further analysis.

BTM- Biterm Topic Modelling

BTM is a topic modeling technique that is good for short text, and it uses word-word as one term instead of a single word (Cheng et al., 2014). In our experiment, we still loop BTM with the topic number from 2 to 50 to select the topic number and then study the topic words and tweets.

CorEx-Correlation Explanation

CorEx uses the information gain to derive topics. It also allows us to set anchor words when we think a specific topic exists in the documents. We first loop the CorEx model with topics from 2 to 50 with no anchor words and then apply anchor words to find topics. We select the best topic numbers with the highest total correlations to study the topic words and tweets.

We first use four methods for the entire dataset with the original tweets data, and then we analyze the topic according to the seasons.

Result

Entire Tweets Data

Based on our LDA results, we observed an increasing perplexity and coherence according to different topics. Since perplexity measures the overlap and coherence measures the dissimilarity across topics, the ideal topic number should have a minimum perplexity and maximum coherence. However, the perplexity in LDA keeps increasing, and it never converges in our dataset. The short tweets' texts can lead to a sparse matrix and cause a very high perplexity if we keep increasing the topic number. In general, we know that the topic number is around 10 for short text. Therefore, we select the topics with a relatively high coherence of around 10. We see that when topic numbers equal 8,14,15 has around 0.37 coherence score. Therefore, we select the closest number of 8 as the topic number for the LDA model.

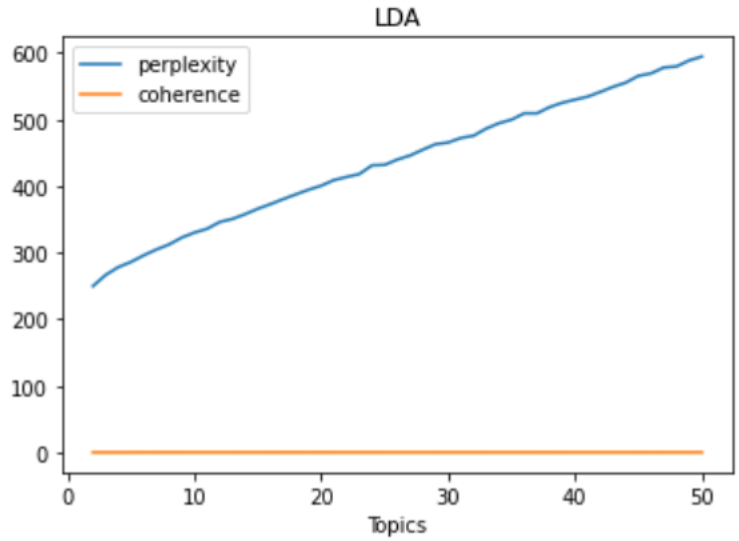

Figure 2. LDA Perplexity and Coherence

```
[(0,
  '0.044*"store" + 0.042*"grocery" + 0.022*"go" + 0.021*"supermarket" + '
  '0.014*"get" + 0.009*"toiletpaper" + 0.009*"socialdistancing" + '
  '0.009*"shopping" + 0.009*"people" + 0.008*"work"'),
 (1,
  '0.019*"online" + 0.013*"shopping" + 0.012*"people" + 0.010*"price" + '
  '0.009*"food" + 0.008*"go" + 0.008*"help" + 0.008*"get" + 0.007*"store" + '
  '0.006*"time"'),
 (2,
  '0.036*"food" + 0.012*"demand" + 0.011*"consumer" + 0.009*"help" + '
  '0.007*"supermarket" + 0.007*"bank" + 0.007*"stock" + 0.006*"people" + '
  '0.006*"need" + 0.005*"essential"'),
 (3,
  '0.036*"price" + 0.016*"food" + 0.011*"supply" + 0.007*"time" + '
  '0.007*"demand" + 0.007*"lockdown" + 0.006*"go" + 0.006*"say" + '
  '0.006*"people" + 0.006*"scam"'),
 (4,
  '0.032*"sanitizer" + 0.031*"hand" + 0.022*"supermarket" + 0.017*"worker" + '
  '0.013*"mask" + 0.012*"people" + 0.009*"make" + 0.008*"use" + 0.007*"store" '
  '+ 0.006*"price"'),
 (5,
  '0.037*"consumer" + 0.013*"business" + 0.010*"easter" + 0.008*"price" + '
  '0.008*"pandemic" + 0.007*"take" + 0.006*"impact" + 0.006*"time" + '
  '0.006*"retail" + 0.005*"new"'),
 (6,
  '0.061*"price" + 0.027*"oil" + 0.010*"low" + 0.009*"fall" + 0.009*"cut" + '
  '0.008*"gas" + 0.008*"food" + 0.007*"market" + 0.007*"pandemic" + '
  '0.006*"year"'),
 (7,
  '0.043*"consumer" + 0.011*"change" + 0.011*"pandemic" + 0.011*"impact" + '
  '0.009*"price" + 0.009*"online" + 0.008*"behavior" + 0.007*"market" + '
  '0.007*"new" + 0.006*"demand"')]
```

Figure 3. LDA Topic Words.

Figure 3 shows the topics with topic words derived from the LDA model with coefficients. Each topic includes 10 words. The topic words show that the first topic focuses on the store and grocery. The second focus is more on online shopping. We see some overlap of topic words across the topics, such as "supermarket," appearing in topic 0, topic 2, and topic 4 with different coefficients. Therefore, we cannot rely only on the keywords to determine the topic's meaning. In addition, after analyzing the top 20 tweets for each topic. We believe that the first topic (Topic 0) should be the "concern of the supplies," and the second topic (Topic 1) should be the "buying options during the pandemic." We also observed some tweets within the same topic may not have similar

topics. Such as, one tweet from topic 1 seems to focus more on the food demand than buying options.

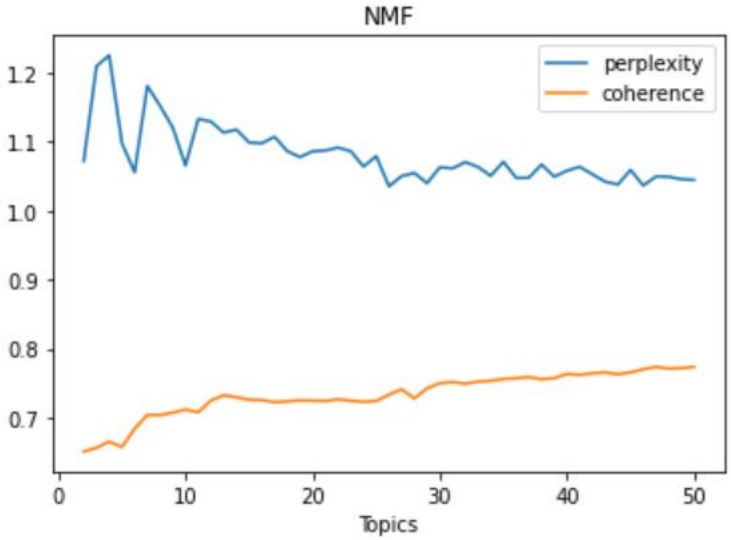

Figure 4. NMF Perplexity and Coherence

The NMF model shows an increasing coherence trend and a decreasing perplexity trend. Still, we believe that our dataset's topic number should be around 10. Therefore, we need to find a topic number around 10 that has low perplexity and high coherence. According to figure 4, there are two options, 6 and 10. Both of them have around 0.7 coherence scores and around 1.05 perplexity. We believe that the topic number is not the large the better because a small number can make each topic more general, so each tweet within the topic can match the topic word better. In addition, after analyzing the topic words and the original tweets, we believe a topic number equal to 6 can be the optimal number of the NMF. Based on the qualitative analysis of topic number equals to 10, we see that topic words are very detailed, and the meaning is very narrow. However, in our opinion, some tweets do not fit the subject topic words very well. According to the 6 topics, most tweets within the same topic can match the topic words

well and derive interpretable meaning compared to topic number 10. Therefore, we select 6 as the optimal number of the NMF model. Figure 5 shows the top 20 words for each topic. And we also observed topic words similar to the LDA model from the results, such as "food demand," "price oil," "hand sanitizer," and "online." After we analyze the top tweets within each topic, we think the result of NMF is more reasonable than the LDA result based on the qualitative analysis of the content of the tweets for each topic.

```
topic 0:  store, grocery, go, worker, work, get, people, retail, employee, time, online, shopping, line, thank, close, day, say, home, see, essential
topic 1:  price, oil, low, market, pandemic, due, increase, gas, fall, high, demand, drop, time, see, take, supply, say, crisis, global, go
topic 2:  food, panic, stock, demand, buy, supply, need, people, bank, help, buying, get, increase, due, chain, say, shortage, pandemic, stop, keep
topic 3:  supermarket, people, go, get, work, worker, shop, shelf, time, shopping, home, need, buy, day, staff, online, see, delivery, local, take
topic 4:  consumer, online, business, pandemic, shopping, change, impact, help, behavior, new, time, crisis, report, good, demand, due, see, take, product, economy
topic 5:  hand, sanitizer, make, use, mask, wash, soap, help, alcohol, stay, glove, home, face, need, safe, water, keep, bottle, get, people
```

Figure 5. NMF Topic Keywords.

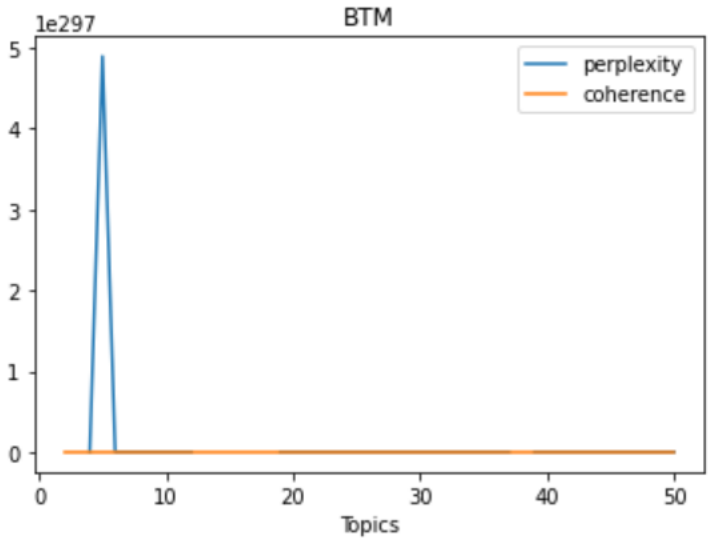

Figure 6. BTM Perplexity and Coherence

```
Topic 0: consumer, online, shopping, price, store, business, help, time, pandemic, retail
Topic 1: supermarket, go, people, store, hand, grocery, sanitizer, get, mask, buy
Topic 2: store, grocery, supermarket, worker, people, go, work, get, food, delivery
Topic 3: food, panic, people, price, buy, need, supermarket, store, stock, go
Topic 4: food, price, consumer, hand, supply, supermarket, sanitizer, store, people, demand
Topic 5: price, consumer, oil, demand, pandemic, market, due, impact, low, food
```

Figure 7. BTM Keywords.

In the BTM model, we still use the low perplexity with the high coherence to get the topic number. Figure 6 shows the change of perplexity and coherence according to the topic number. We observe the perplexity and coherence scores are very unstable,

whereas the perplexity is considerable, and the coherence is always negative. Therefore, we run the model multiple times to find the optimal topic number with low perplexity and high coherence. After that, we decided to use 6 as the topic number for the BTM model. This number is the same as the topic number of the NMF model. Figure 7 shows the topic words for each topic, and we see similar topics. Both have a topic containing "Store" and "grocery," "food" and "panic," "consumer" and "online." According to the observation, most topics overlapped for the BTM and NMF models. However, BTM has a topic that contains "food" and "price," which is not shown in the NMF. Although the two models have the same topics, the result is not precisely the same. We will explore these two models further in the discussion.

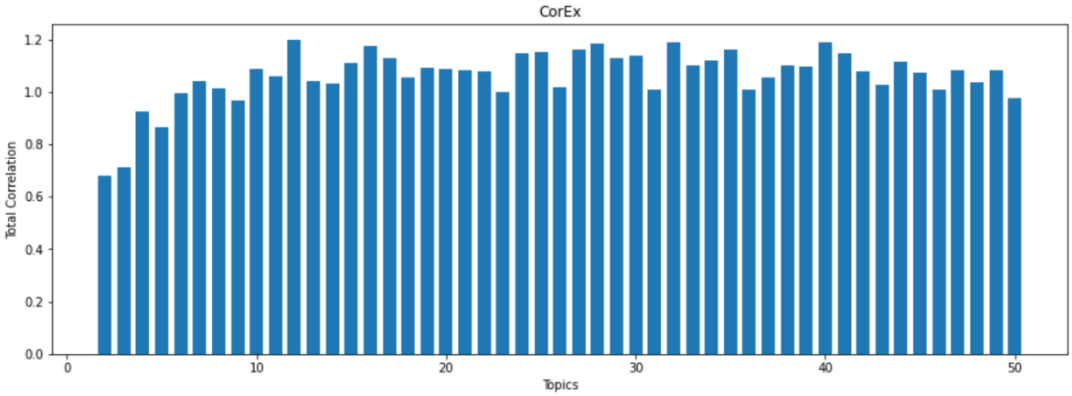

Figure 8. CorEx Total Correlation

```
Topic 0: store, grocery, ~consumer, retail, worker
Topic 1: price, oil, ~supermarket, low, gas
Topic 2: hand, sanitizer, mask, glove, wear
Topic 3: online, shopping, shop, delivery, order
Topic 4: food, demand, supply, stock, chain
Topic 5: toilet, paper, toiletpaper, roll, find
Topic 6: stay, social, home, safe, distancing
Topic 7: go, people, get, shelf, empty
Topic 8: panic, buying, buy, stop, fear
Topic 9: support, help, community, business, small
Topic 10: week, last, month, next, year
Topic 11: change, behavior, shift, new, pandemic
```

Figure 9. CorEx Topic Words

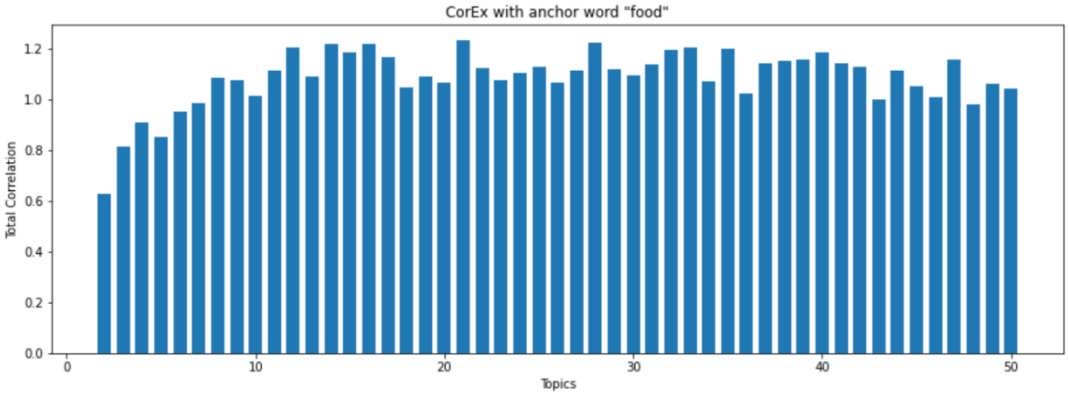

Figure 10. CorEx with anchor word "Food" Topic words Total Correlation

```
Topic 0: food, panic, stock, buy, buying
Topic 1: make, sure
Topic 2: consumer, ~supermarket, demand, impact, business
Topic 3: people, get, work, time, itâ
Topic 4: online, shopping, shop, delivery, order
Topic 5: toilet, paper, toiletpaper, roll, find
Topic 6: price, oil, fall, market, high
Topic 7: stay, social, safe, distancing, home
Topic 8: low, year, gas, old, drop
Topic 9: store, grocery, retail, employee, go
Topic 10: hand, sanitizer, mask, glove, wear
Topic 11: worker, driver, health, staff, thank
```

Figure 11. CorEx with anchor word "Food" Topic words

The CorEx model uses the total correlation to determine whether the model is good from the quantitative perspective. We have tested two CorEx models with topic numbers from 2 to 50. One is the original model with no anchor words. The other uses the anchor word "food." The original CorEx shows that there will be 12 topics for the entire tweet, and we observe similar words such as "price oil," "hand sanitizer," and "online shopping." There are also "stay social," "toilet paper," and "change behavior" that have not been shown in the previous models. After that, we set the anchor word "food" because we believe there should be topics related to the food supply or food demands during the pandemic. It has the same topic number as the original CorEx model, 12. We see a topic with "food panic" after we set the anchor word, and there is

a topic with the words "make sure," which has no meaning. After reviewing the original tweets for the "make sure" topic, we believe it is a general cluster of tweets with many topics. We think that is because the anchor words we choose is too general, and we may need more anchor word for that. Overall, we believe the CorEx is still an excellent model to analyze the tweets data because it can derive topics that are never shown from the previous model after we choose the optimal topic numbers.

Experiment 1 (Over season)

One important topic modeling analysis is about finding topic changes over time. In our project, we split time into seasons: spring, summer, autumn, and winter. The proportion of tweets is shown in figure 12, and each model's topic number per season is also different. We use the optimal topic number from the previous experiment for each model to compare whether there are topic differences between each season. The topic number we used for comparison is 8,6,6,12 for LDA, NMF, BTM, and CorEx.

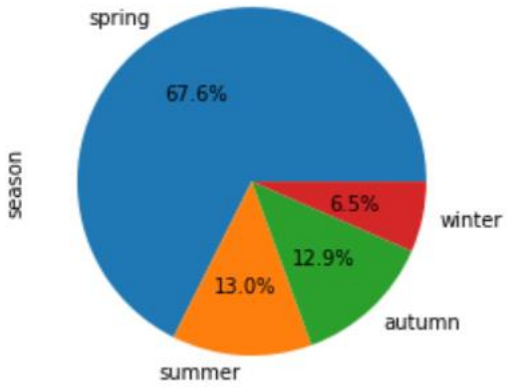

Figure 12. Data Proportion - Season

```
spring
Topic 0: store, grocery, go, people, stay
Topic 1: food, people, panic, lockdown, waste
Topic 2: price, oil, low, consumer, economy
Topic 3: supermarket, store, crash, ju, teeter
Topic 4: food, consumer, demand, dog, say
Topic 5: supermarket, price, wealth, norwegian, slum
Topic 6: prior, batch, sovereign, price, today
Topic 7: price, chain, coffee, food, supply

summer
Topic 0: price, supermarket, store, due, say
Topic 1: supermarket, people, grocery, store, shopping
Topic 2: price, food, supermarket, worker, pandemic
Topic 3: food, price, grocery, store, market
Topic 4: price, essential, hand, sanitizer, people
Topic 5: food, consumer, demand, price, store
Topic 6: price, go, shopping, supermarket, online
Topic 7: grocery, store, supermarket, price, shopping

autumn
Topic 0: store, grocery, sanitizer, price, get
Topic 1: grocery, store, shopping, online, consumer
Topic 2: consumer, price, crisis, pandemic, take
Topic 3: hand, store, sanitizer, food, price
Topic 4: consumer, pandemic, grocery, people, late
Topic 5: consumer, price, pandemic, hand, help
Topic 6: food, store, grocery, consumer, help
Topic 7: consumer, food, supermarket, online, pandemic

winter
Topic 0: supermarket, price, food, consumer, stay
Topic 1: sanitizer, hand, supermarket, food, people
Topic 2: price, consumer, store, grocery, people
Topic 3: toiletpaper, get, shopping, hand, sanitizer
Topic 4: sanitizer, consumer, buy, price, pandemic
Topic 5: price, food, home, online, stay
Topic 6: food, price, mask, time, business
Topic 7: oil, price, supermarket, come, get
```

Figure 13. LDA Topic Words - Season

In LDA, we still observe an increase in perplexity in all seasons. Figure 13 shows the topic words, and based on our analysis, we think the topic has shifted from the food supply to the sanitizing demand over seasons. I think this matches the reality that when the pandemic started, people were more concerned about the price of supplies such as food and oil. After that, people become aware of wearing masks and sanitizing. That's why words such as "sanitizer" do not appear in the spring topics but start appearing in

the summer.

From the NMF, we observed a similar trend of sanitizing through time. Figure 14 shows the topic words per season. We see the words are listed according to the coefficients, and we see that "sanitizing" appear in the spring with a very low coefficient, which means people are not aware of the importance of sanitizing. However, starting in the summer, "hand sanitizer" has a very high coefficient within the topic. BTM demonstrates the same conclusion of the trend of sanitizing by displaying the "sanitizer" in the summer with high coefficients.

```
spring
topic 0:  store, grocery, worker, work, retail, employee, close, time, thank, get, line, hour, open, say, day, pandemic, essential, go, home, make
topic 1:  food, panic, stock, buy, supply, demand, need, buying, bank, help, get, say, due, increase, shortage, stop, chain, enough, shelf, pandemic
topic 2:  price, oil, low, market, increase, gas, demand, pandemic, fall, due, high, drop, time, hand, mask, take, sanitizer, sell, see, say
topic 3:  supermarket, worker, shelf, work, staff, get, local, empty, time, delivery, shop, day, see, thank, keep, need, hour, home, social, week
topic 4:  consumer, online, shopping, business, pandemic, help, impact, time, change, new, home, good, crisis, behavior, due, report, shop, demand, make, service
topic 5:  people, go, get, buy, need, online, hand, home, shop, shopping, sanitizer, make, stay, think, work, panic, time, stop, see, thing
summer
topic 0:  store, grocery, go, worker, mask, employee, work, wear, day, say, get, die, line, retail, essential, people, time, week, keep, thing
topic 1:  price, oil, low, pandemic, market, due, cut, high, fall, demand, gas, global, impact, people, increase, time, drop, supply, see, crisis
topic 2:  food, demand, supply, stock, need, help, panic, increase, bank, chain, buy, due, delivery, shortage, people, service, pandemic, farmer, global, say
topic 3:  supermarket, go, people, worker, get, work, home, stay, time, keep, staff, need, day, thank, stayhomesavelive, week, many, still, shop, take
topic 4:  online, consumer, shopping, shop, pandemic, new, time, delivery, business, change, home, crisis, order, stay, get, good, see, due, impact, take
topic 5:  hand, sanitizer, make, mask, use, glove, wash, help, alcohol, time, clean, paper, keep, donate, soap, fight, need, face, first, pandemic
autumn
topic 0:  store, grocery, worker, go, get, employee, line, work, die, mask, people, pandemic, day, front, wear, socialdistancing, customer, thank, retail, see
topic 1:  price, oil, market, pandemic, due, low, high, drop, global, fall, rise, see, say, year, cut, impact, gas, supply, increase, home
topic 2:  consumer, business, change, behavior, pandemic, impact, new, report, brand, time, online, crisis, help, behaviour, see, trend, insight, read, datum, update
topic 3:  hand, sanitizer, use, make, mask, wash, help, soap, alcohol, handsanitizer, get, pandemic, face, glove, stay, donate, base, wear, safe, free
topic 4:  food, demand, help, stock, increase, supply, bank, need, pandemic, crisis, keep, due, people, panic, see, chain, lockdown, service, delivery, meet
topic 5:  supermarket, people, go, online, get, shopping, worker, work, shop, time, delivery, home, make, stayhomesavelive, day, take, keep, essential, socialdistancing, need
winter
topic 0:  store, grocery, online, shopping, socialdistancing, get, people, mask, go, shop, wear, see, pandemic, way, worker, stay, work, retail, delivery, good
topic 1:  consumer, business, change, pandemic, behavior, impact, new, product, online, time, insight, help, trend, need, see, habit, crisis, retail, economy, marketing
topic 2:  sanitizer, hand, mask, use, wash, make, face, soap, alcohol, runwalofficial, help, glove, handsanitizer, fight, get, water, frequently, product, spread, stay
topic 3:  food, demand, supply, bank, help, stock, chain, increase, need, meet, work, panic, people, day, milk, many, due, go, pandemic, crisis
topic 4:  price, oil, cut, low, good, due, fall, market, supply, deal, economic, take, economy, production, high, country, pandemic, get, global, call
topic 5:  supermarket, people, go, queue, get, day, time, socialdistancing, stayhomesavelive, staff, spread, local, need, see, cough, home, back, mask, say, shop
```

Figure 14. NMF Topic Words – Season

```
spring
Topic 0: consumer, online, price, help, business
Topic 1: store, grocery, supermarket, food, people
Topic 2: store, supermarket, grocery, people, go
Topic 3: food, store, supermarket, people, go
Topic 4: food, go, hand, supermarket, people
Topic 5: price, consumer, oil, demand, food
summer
Topic 0: price, hand, sanitizer, people, food
Topic 1: store, grocery, supermarket, go, worker
Topic 2: food, online, shopping, store, price
Topic 3: price, food, supermarket, go, store
Topic 4: price, oil, consumer, pandemic, demand
Topic 5: food, price, online, help, time
autumn
Topic 0: food, online, store, supermarket, help
Topic 1: consumer, price, business, change, help
Topic 2: price, oil, demand, market, pandemic
Topic 3: store, grocery, hand, worker, sanitizer
Topic 4: consumer, price, pandemic, new, behavior
Topic 5: store, grocery, supermarket, go, worker
winter
Topic 0: supermarket, store, people, online, get
Topic 1: hand, sanitizer, store, mask, grocery
Topic 2: consumer, price, oil, food, demand
Topic 3: supermarket, store, food, sanitizer, people
Topic 4: consumer, pandemic, food, price, business
Topic 5: food, demand, consumer, price, supply
```

Figure 15. BTM Topic Words - Season

From CorEx, we changed the anchor word to "oil" because the oil price is also an important term based on the previous result. The result shows no precise observation of the topic related to "oil" in the spring. However, it started showing "oil low" in the summer. This reminds me of the oil price change during the pandemic, and I remember the price is very low at that time, although the oil price is very high nowadays. In addition, we do not observe a trend of sanitizing in the CorEx model, and the CorEx model also reports the term "social distancing," which we rarely observe in other model results.

```
spring
Topic 0: store, grocery, ~price, retail, ~consumer
Topic 1: health, public, emergency, state, government
Topic 2: help, support, small, vulnerable, business
Topic 3: people, think, bad, know, need
Topic 4: social, distancing, follow, post, measure
Topic 5: online, shopping, shop, order, deliver
Topic 6: demand, ~supermarket, increase, impact, pandemic
Topic 7: sanitizer, hand, mask, glove, use
Topic 8: panic, food, buy, buying, stock
Topic 9: toilet, paper, roll, toiletpaper, find
Topic 10: worker, home, work, stay, driver
Topic 11: go, get, week, last, day
summer
Topic 0: price, oil, ~supermarket, low, ~consumer
Topic 1: mask, wear, glove, face, stop
Topic 2: go, week, last, back, thing
Topic 3: stay, home, safe, delivery, keep
Topic 4: hand, sanitizer, use, clean, fight
Topic 5: test, happen, positive, itâ, shelf
Topic 6: toiletpaper, paper, ~company, ~many, toilet
Topic 7: store, grocery, worker, retail, employee
Topic 8: socialdistancing, line, try, ~state, ~economy
Topic 9: global, due, impact, pandemic, rise
Topic 10: food, demand, supply, social, chain
Topic 11: people, staff, tell, lot, man
autumn
Topic 0: price, oil, ~supermarket, market, low
Topic 1: demand, supply, bank, production, meet
Topic 2: time, important, information, ever, close
Topic 3: stay, safe, get, people, home
Topic 4: increase, lead, result, see, current
Topic 5: go, week, last, back, still
Topic 6: consumer, behavior, change, behaviour, brand
Topic 7: sanitizer, hand, mask, glove, wear
Topic 8: grocery, store, worker, employee, die
Topic 9: food, stock, delivery, need, essential
Topic 10: health, test, positive, care, first
Topic 11: online, shopping, shop, paper, toilet
winter
Topic 0: price, oil, cut, low, fall
Topic 1: paper, toilet, toiletpaper, shelf, run
Topic 2: food, bank, stock, panic, milk
Topic 3: store, grocery, online, shopping, socialdistancing
Topic 4: supply, chain, hit, full, much
Topic 5: change, new, live, future, world
Topic 6: demand, ~virus, ~spread, ~start, ~know
Topic 7: time, thing, customer, need, take
Topic 8: social, distancing, tip, system, health
Topic 9: consumer, behavior, impact, spending, habit
Topic 10: supermarket, go, queue, cough, people
Topic 11: sanitizer, hand, mask, glove, wear
```

Figure 16. CorEx Topic Words – Season

According to the topic words analysis of these models, there are common words such as "oil," "price," "food," "demand," "sanitizer," and "grocery" in our topics, which means every model can fit our tweets' text. These topic words illustrate what people are concerned about during the pandemic. However, each model has different topic results

by having topic words with different coefficients, and some words may not be retrieved in some models, such as "social" and "distancing." In order to evaluate these models further, we will apply the qualitative analysis based on the top 20 tweets within each topic from models to determine which model can derive a meaningful topic.

Discussion

In the project, we apply four topic modeling techniques, LDA, NMF, BTM, and CorEx, to the tweets data to find a topic during the early stage of Covid-19. Each model can derive reasonable topic words, but the results are slightly different. We observed that each model would generate different numbers of topics based on the perplexity and coherence score. For the LDA model, we observe an increasing perplexity. Meanwhile, our BTM always outputs a very large perplexity from infinite to around 200. In addition, we observe a negative coherence score in BTM but positive coherence in LDA and NMF. Therefore, our topic number selection does not follow a unique standard for each model. And thus may affect our final results.

Other than that, our dataset contains non-English characters. In my opinion, we do not have to drop these characters because tweets use informal language because there will be some special words and phrases for online communications. We cannot simply remove them when we preprocess them. In addition, the amount of these non-English characters does not account for many original tweets. The frequency-based topic model can ignore them if these characters are not frequent in the dataset, so our result will not be affected if we include them. For the special character that forms a topic, we need to

study further to understand whether it is useful, and then decide to remove it in the future.

In addition, we observed some common words for the topic across different models. Among them, words such as "online," "store," "supermarket," "grocery," "oil," "price," and "low" frequently appear in our results. However, a model such as NMF and LDA that breaks sentence grammar cannot get reason meaning only based on the words. One observation we found for a topic that contains "oil," "price," and "low" can conclude that the oil price has decreased to reach a very low level. However, after the qualitative analysis, we see most tweets within the topic talk about the price of supplies such as food and materials changed during the pandemic. The word "high" also appears frequently in the original tweets within the topic because it talks about some low prices and some high ones. Therefore, we cannot simply conclude that the price has decreased during the pandemic. The conclusion may differ through the quantitative and qualitative analysis according to the NMF and LDA.

Compared to the NMF and LDA, CorEx makes more sense when we analyze the original tweets with topic modeling results. For example, one topic which contains "price," "oil," "low," and "gas" can conclude that the oil and gas price is very low. The qualitative analysis shows that most of the tweets on this topic talk about the price of gas and oil decreasing significantly during the pandemic, which is the same conclusion that the CorEx model summarizes.

BTM had the worst performance in these four methods. Based on the analysis of a topic that contains the words "consumer," "online," and "shopping," we can derive the

conclusion that people's shopping behavior may be changed to online shopping during the pandemic. Although one tweet talks about online shopping within the topic, most tweets mention the consumer and grocery store individually. It is hard for us to derive the same conclusion as the BTM model result according to the qualitative analysis. Also, the perplexity and coherence in BTM are not that good for choosing the topic numbers. We may consider studying the BTM further to see why its performance is bad in our project. Overall, we believe that the CorEx model has the best performance in analyzing topics for our tweets data, and its output topics are the same as the qualitative analysis based on the original tweets.

Thematic Analysis using CorEx

Based on our analysis from quantitative analysis, we noticed that the CorEx model is superior to the other model. LDA and NMF break sentence grammar, so the topic word cannot accurately reflect the real meaning of the original tweets. In addition, BTM shows even worse results, and the relationship between the original tweets and the topic words is not apparent. Therefore, we select CorEx for our further qualitative analysis. According to the total correlation of each of our models, we see that topic equal to 12 has the highest correlation among others. Therefore, our qualitative analysis focus on these 12 topics. First, we will print out the first 10 words for each topic and then study the top 20 tweets belonging to each topic to find the real meaning. However, we find that our tweets usually have very short text, so 10 words analysis may be too much for each topic. Therefore, we will also analyze the meaning based on the first five words and see whether top-5 or top-10 words can better reflect the real meaning of the tweets.

Topic 0: store, grocery, ~consumer, retail, worker, employee, close, line, open, hour

Top-5 words initial meaning: worker of the retail store
Top-10 words initial meaning: working hours of the employee in a retail store

Meaning based on the Top 20 tweets:
People thanks the workers who still work during the pandemic in grocery stores and retail stores. In addition, people care about their pay and health

Qualitative analysis meaning of the topic:
Care and appreciation for stores' workers.

Topic 1: price, oil, ~supermarket, low, gas, market, fall, drop, global, high

Top-5 words initial meaning: energy price decrease
Top-10 words initial meaning: The global energy market has decreased

Meaning based on the Top 20 tweets:
Oil and gas prices have decreased significantly during the pandemic. The lowest price of oil and gas in different countries. The oil war between Saudi Arabia and Russia is a reason that leads to the decreasing oil price

Qualitative analysis meaning of the topic:
Energy price decrease

Topic 2: hand, sanitizer, mask, glove, wear, use, face, make, fight, protect

Top-5 words initial meaning: Epidemic prevention supplies
Top-10 words initial meaning: People need these epidemic prevention supplies to protect themselves.

Meaning based on the Top 20 tweets:
Call for people to use sanitizer and face masks during the pandemic to prevent Covid-19. It also shows the shortage of sanitizer and face masks. People are becoming aware of protection.

Qualitative analysis meaning of the topic:
Using sanitizer and face masks.

Topic 3: online, shopping, shop, delivery, order, deliver, retailer, offer, list, sale
Top-5 words initial meaning: Online shopping/delivering
Top-10 words initial meaning: Online shopping/delivering with discounts

Meaning based on the Top 20 tweets:
Stores offer online shopping to customers. More and more people were starting online shopping. People were getting used to online delivery and not shopping at local stores. Discounts and special offers for online shopping. People are encouraged to shop online.

Qualitative analysis meaning of the topic:
Online shopping and delivering.

Topic 4: food, demand, supply, stock, chain, bank, shortage, meet, enough, item
Top-5 words initial meaning: Supplies of food
Top-10 words initial meaning: Shortage of food because the supply chain has a problem

Meaning based on the Top 20 tweets:
People care about the shortage of food. The demand for food has increased, and the food supply chain has problems. Food hoarding leads to food supply chain tensions.

Qualitative analysis meaning of the topic:
Food hoarding.

Topic 5: toilet, paper, toiletpaper, roll, find, ~big
Top-5 words initial meaning: toilet paper
Top-10 words initial meaning: toilet paper

Meaning based on the Top 20 tweets:
People are concerned about the toilet paper price and face a toilet paper shortage. The demand for toilet paper has increased.

Qualitative analysis meaning of the topic:
Toilet paper is very expensive

Topic 6: stay, social, home, safe, distancing, work, keep, care, health, driver
Top-5 words initial meaning: keep social distance and stay home
Top-10 words initial meaning: stay home and keep social distancing to take care of health
Meaning based on the Top 20 tweets:
People are suggested to stay at home and keep social distance.

Qualitative analysis meaning of the topic:
Stay home and keep social distance

Topic 7: go, people, get, shelf, empty, think, back, thing, day, itâ
Top-5 words initial meaning: Shelf empty
Top-10 words initial meaning: People find out the shelf is empty of supplies

Meaning based on the Top 20 tweets:
People start to panic when they see the empty grocery store shelf. People hoard food due to panic.

Qualitative analysis meaning of the topic:
Empty shelf in grocery store.

Topic 8: panic, buying, buy, stop, fear, ~check
Top-5 words initial meaning: People panic buying due to the pandemic.
Top-10 words initial meaning: People panic buying due to the pandemic.

Meaning based on the Top 20 tweets:
The fact of people panic shopping, and they were suggested to stop panic shopping and overbuying. People complain about overbuying, and government calls for no overbuying.

Qualitative analysis meaning of the topic:
Panic buying/Overbuying

Topic 9: support, help, community, business, small, local, time, way, provide, call
Top-5 words initial meaning: Community support small business
Top-10 words initial meaning: Local community help and support small business.

Meaning based on the Top 20 tweets:
Small businesses closed during the pandemic, and people want to support these small businesses. People ask local communities to support their small businesses. Government supports these small businesses by changing the electricity price. Support small businesses online. Some repeat advertisements provide financial support to small businesses.

Qualitative analysis meaning of the topic:
Supporting small businesses

Topic 10: week, last, month, next, year, start, say, happen, already, end
Top-5 words initial meaning: Time change
Top-10 words initial meaning: Something happened during the time

Meaning based on the Top 20 tweets:
This is a general topic. It contains the house price changes through time, what people do in their morning and night, and the expected effect of Covid-19 in the future. It also contains topics about food and toilet paper supply.

Qualitative analysis meaning of the topic:
General concern.

Topic 11: change, behavior, shift, new, pandemic, read, post, learn, affect, ~die
Top-5 words initial meaning: People change their behavior during the pandemic
Top-10 words initial meaning: People read and learn about the pandemic and affect their behavior

Meaning based on the Top 20 tweets:
People are getting used to online shopping, and online retailers have become a new trend. Food delivery is also becoming more acceptable for people who work at home. People start preferring nail kits rather than makeup during the pandemic.

Qualitative analysis meaning of the topic:
Customer behavior has changed

Thematic Analysis Discussion

Through the analysis of the topic words and qualitative analysis, we see that the topic words can cover the real meaning of the topic, but it lacks details. With top-5 topic words analysis, we see it can almost represent the qualitative analysis in our research. For the top-10 topic word analysis, it covers more details, but the meaning may be too narrow to represent the real meaning of the topic, and it may confuse me about the results. Therefore, we believe that for short text data, such as tweets text, it is best to use the top 5 topic words for preliminary analysis rather than the top 10 topic words.

However, the quantitative analysis is not enough to retrieve the true meaning of the topic. For example, topic 5 shows the topic is toilet paper, but toilet paper is too general to represent a topic. After we read the original tweets of topic 5, we find that most tweets complain about toilet paper's price. These complaints use many similes and metaphors, so it may be difficult for CorEx to determine the keywords, such as price and high.

In addition, we see topic 0 talks about the workers and employees. After analyzing the original tweets, we find that people are concerned about these workers' and employees' health when they are working in the store, and people are thankful for their

work during the pandemic. It is hard to say the topic retrieved from the quantitative analysis is wrong, but we see that it lacks details to retrieve the more accurate meaning. Therefore, qualitative analysis of the original tweets is still essential to find out the real meaning of the topic.

Figure 17. Thematic Analysis Diagram

According to the thematic analysis, we observed that public concerns during the early stage of Covid-19 could be divided into 6 categories: Social Behavioral Change, Social Supports, Economic Crisis, Shortage of Supplies, Awareness of Protection, and general concerns. Among them, general concern is a very large topic that contains living concerns, stock price expectations, and supply prices through time. There is no clear way to conclude these general concerns, so we just categorize them as an individual topic.

Also, we find that there is a social behavioral change according to the 3 topics we retrieved from the tweets. We first see that people are encouraged to stay home and

keep social distancing in topic 6, so people need to find a new way to get food and other supplies. Therefore, we see a topic relates to online shopping and delivery in topic 3. Furthermore, we observe that people are getting used to shopping online rather than shopping in-store, as shown in topic 11 in figure 17. In conclusion, we believe that there was a behavioral change in the shopping style during the early stage of Covid-19 because they were staying at home, and the items people were buying were also changing, such as nail kits, face masks, and hand sanitizer.

We also see a there was social support during the pandemic, where people are taken care of the employees from the retail stores and supported small businesses. In addition, we see the price of energy, such as oil and gas, decreased significantly at that time, which led to the economic crisis and recession concerns. Other than that, the shortage of supplies was also a problem faced by the people at that time. According to our analysis, one perspective is that toilet paper price at that time was extremely high, and there is no clear evidence showing why it is expensive, according to the tweets.

Also, people were facing a food shortage, and grocery store shelves were empty. The reason for that relates to the other two topics, food hoarding and panic buying/Over buying. These two topics are similar because people at that time were so worried about the food shortage so that they started overbuying and stocking food and supplies. The huge demand for the foods breaks the traditional supply chain and causes empty shelves in grocery stores. Thus, people were panicking about that and called to stop overbuying. Those can be categorized as the shortage of supplies, which shows the people's concern about the essentials.

Another thing people considered is using sanitizer and face masks, and they also urged people to wear masks to prevent the spread of the virus and use sanitizer to protect their health. Also, some tweets suggest wearing gloves to prevent the virus. But overall, we believe that people have started to think about using anti-epidemic products to protect themselves from Covid-19. Through the thematic analysis, we illustrate people's concerns about Social Behavioral Change, Social Support, Economic Crisis, Shortage of Supplies, and Awareness of Protection during the early stage of Covid-19. The thematic analysis provides a good compliment to the CorEx result and makes topics more meaningful. However, we do not know whether people's concerns changed during the different periods of the pandemic. Therefore, we will continue to study the time-related topic's change in the future.

Conclusion

In conclusion, we have analyzed the topic of tweets during the early stage of the pandemic by using four topic modeling methods. We find that all of them can generate reasonable topic words through quantitative analysis. Among them, we believe that the CorEx method is superior to the others because it can generate the same result as we use qualitative analysis. Through the results, we find that during the pandemic, people are concerned about the shortage of supplies such as food and toilet paper, and oil prices have decreased significantly, whereas food prices have increased a little bit. Meanwhile, shopping behavior may also change because some topics contain the words "online" and "shopping." Also, people's concerns changed during the pandemic. We see people

start becoming aware of sanitizing over time because we observe these words in the summer, autumn, and winter but not spring. Still, our project is limited by the raw data and the model quality. A model such as BTM does not fit well with the original tweets, and we do not deeply study hashtags and web links. In the future study, we plan to optimize the BTM model and other features with hashtags to derive a more accurate topic for our dataset.

References

Cheng, X., Yan, X., Lan, Y., & Guo, J. (March 26, 2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Dixon, S. (July 27, 2022). *Twitter: number of worldwide users 2019-2024.* Statista. https://www.statista.com/statistics/303681/twitter-users-worldwide/

Government of Canada. (November 14, 2022). *Coronavirus disease (COVID-19).* https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19.html

Jay, A. (November 6, 2022). *Number of Twitter Users 2022/2023: Demographics, Breakdowns & Predictions.* Finances Online. https://financesonline.com/number-of-twitter-users/

Kwok, S. W. H., Vadde, S. K., & Wang, G. (May 19, 2021). Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. *Journal of Medical Internet Research*, *23*(5), e26953–e26953. https://doi.org/10.2196/26953

Miglani, A. (September 8, 2020). *Coronavirus tweets NLP – Text Classification.* Kaggle. https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (March 9, 2021). An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media*, *22*, 100104–100104. https://doi.org/10.1016/j.osnem.2020.100104

Wicke, P., & Bolognesi, M. M. (March 16, 2021). Covid-19 Discourse on Twitter: How

the Topics, Sentiments, Subjectivity, and Figurative Frames Changed Over Time. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.651997