COVID-19 Information Spread around Political Engagement on Facebook

POOREUMOE KIM
Department of Informatics
Technical University of Munich
München, 80333 Germany
prm.kim@tum.de

Tobias Eder
Department of Informatics
Technical University of Munich
München, 80333 Germany
tobias.eder@in.tum.de

Abstract

To observe how Facebook users behave during the pandemic, this paper investigates how they spread the information. First, we crawl Facebook posts regarding COVID-19 in 2020 by Crowdtangle, a social networks analysis tool. Then we analyze the data with four steps. 1) We explore general post-sharing patterns. They have skewed distributions, but information spreading speeds differ by posts and accounts. 2) To understand the spread through linked posts, we utilize network mapping. 3) We also apply topic modelings and figure out which topic surges by dates and information sources. 4) Finally, we make a prediction model that can predict a post's future sharing patterns with its initial behavior and internal features. One of the briefest highlights of this article is that different account types, show different information spreading pattern.

0 Introduction

Even though Facebook has more than 2 billion user accounts, there has been little detailed study of how COVID-19 information is spread in the massive social network. Still, it is crucial to characterize how the vast social network affects public information, given the pandemic changes everyone's daily life.

A post's share count is our primary metric to measure the information diffusion. To collect and analyze posts, we use Crowdtangle, a social network crawling and analyzing tool developed by Facebook. Besides the number of shares, Facebook posts can be characterized by total interactions, emotional reactions, their messages, topics, authors' accounts, and linked URLs. We would utilize the features to understand the spread.

Also, we observe political engagement on Facebook by collecting posts from the pages of CNN and Fox News, Biden, and Trump, as well as opinions from lockdown advocates and haters.

0.1 Research Questions

This paper studies the four research questions:

1. How does a general sharing pattern of COVID-19 posts appears?
2. How does COVID-19 information spread through linked posts? Especially from official sources to unofficial on Facebook?
3. What are the main COVID-19 topics and trends in 2020? How did users react to them?
4. Can we predict a post's long-term future share counts with its initial behavior? If then, which feature plays a crucial role?

The questions would be answered by the corresponding section number. For example, the first section, Sharing Pattern Exploration, will answer the first question, while the second section, Network mapping, would do the second. Likewise, Topic modeling answers the third. Finally, we use a panel data forecasting in machine learning approaches to answer the fourth in Prediction modeling.

0.2 Related works

We introduced related works and maintain that our analysis is unique compared to them. One of the primary references is the exploratory study of Ordun et al. about COVID-19 Tweets [1]. They explored specific topics about COVID-19, and analyzed retweet network, as well as surveyed related-exploratory works on Twitter. Including the article and the surveyed studies, others also used the topic modeling and network analysis. However, our analysis focuses on how the COVID-19 information spreads, and its changing pattern by Facebook accounts and posts features.

We also introduce articles studying information spread on Facebook but having different focuses.

Papakyriakopoulos et al. show the diffusion of misinformation related to the origin of COVID-19 on four social media platforms [2]. They tracked and 11,023 unique URLs of information sources and figured out 1) traditional media contribute overall more to a conspiracy diffusion than the alternative, and 2) content moderation practices can mitigate its spread. Bruns et al. also studed the spread, but focusing on a specific rumor that the pandemic outbreak was somehow related to the rollout of 5G mobile telephony technology [3]. On the other hand, Boberg et al. suggested a different argument considering the misinformation spread. Their analysis argued that the alternative news media do not spread apparent lies; they are predominantly sharing overly critical, even anti-systemic messages, opposing the view of the mainstream news media and the political establishment [4].

Next, we briefly introduce references about general information spread, not specialized in COVID-19. Romero et al. quantitatively evaluated how different types of information spread on Twitter by analyzing hashtags [5]. Lerman and Ghosh tracked the information spread on Digg and Twitter and recognize that network structure affects dynamics of information flow [6]. Zhao et al. analyzed contents on Twitter, comparing traditional media using a a topic model, Latent Dirichlet Allocation (LDA) [7].

Finally, we list up other prediction research for Facebook posts. One difference between our prediction and others is distinct feature usage like account properties. Also, we target to understand which features play a crucial role in forecasting information spread over time, not to develop an advanced neural network model. For example, Krebs et al. predicted the reaction distribution on Facebook posts. The authors used neural network architectures (convolutional and recurrent neural networks) using pretrained word embeddings [8]. Straton et al researched a methodology to predict user engagement based on eleven characteristics of a post: post type, hour span, Facebook Wall category, level, country, isHoliday, season, created year, month, day of the week, time of the day. They primarily use neural networks [9]. The most closest work is the study of Kamaljot Singh. The author predicted comment volume on Facebook posts in next few hours, targeting Facebook pages. The main focus of this study is to experiment with variety of regressive modeling techniques such as meta-learning and neural networks [10].

1 Sharing Pattern Exploration

1.1 Data collection

We crawled COVID-19 related posts written in 2020 by Crowdtangle's API. It allows one to access Facebook posts, which are 1) made by a public page, group, or verified public person, who has ever (since 2014) had more than 110,000 likes, and 2) are without the poster aiming at a particular audience using Facebook targeting and gating tools [11].

In particular, we use four types of Crowdtangle's API: 1) GET/posts, 2) GET/post, 3) GET/links, 4) GET/posts/search. The first API is used to retrieve a set of posts for given parameters like time range. The second one is used to acquire a specific post with its ID [12]. The acquired data can include historical time-series format. Next, the link API can track which Facebook posts share a specific URL. Finally, the search API provides a list of posts with a query such as 'COVID AND lockdown.' In the following sections, we would crawl and process data by the APIs.

1.2 Data exploration

1.2.1 Pfizer's vaccine and Trump's argument

We start the exploration with two examples: Pfizer's vaccine announcement and Trump's argument. First, Pfizer's vaccine was announced on 9.Nov.2020.[1] They pronounced the news on their official website, and it triggers Facebook users to share. We tracked the 430 sharing accounts by Crowdtangle link API. In Figure 1, the 406 red dots mean hourly share counts by unverified accounts, and the 24 blue do the share by verified users. Facebook users have shared this even until 29.Jan.2021.

The second example is Donald Trump's post on 27.Oct.2020 arguing that 'ALL THE FAKE NEWS MEDIA WANTS TO TALK ABOUT IS COVID, COVID, COVID.'[2] We could track 180 unverified accounts share this post, and two verified ones do. Plus, users share them only within five days.

The two cases show that 1) both have skewed distributions. However, 2) the proportion of verified posts are significantly different (24/430 and

[1] https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-announce-vaccine-candidate-against

[2] https://www.facebook.com/DonaldTrump/posts/10165703998545725

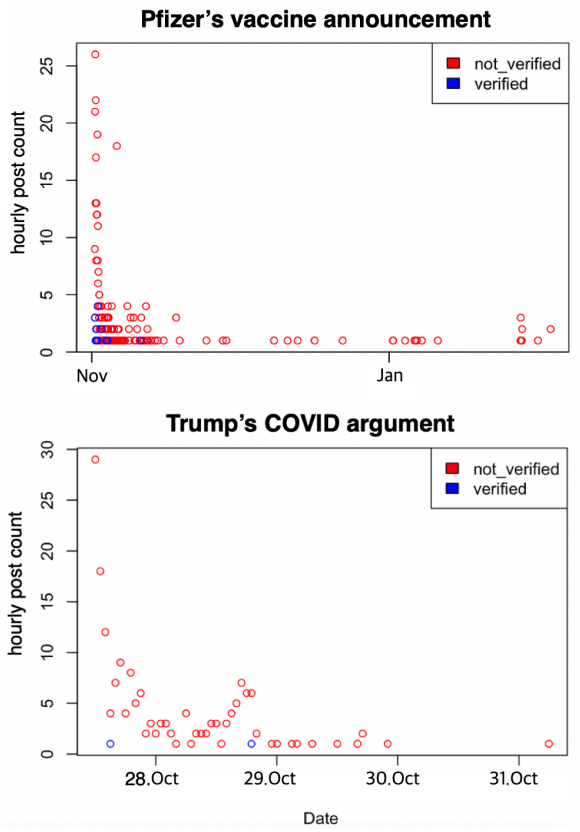


Figure 1: Hourly share counts of two information sources: Pfizer's announcement and Trump's post.

2/182). Also, 3) the sharing periods differ(81 and 5 days).

1.2.2 CNN, Fox News, Trump, and Biden

Next, we explored cumulative sharing patterns from the following four pages: CNN, Fox News, Donald Trump, Joe Biden. The Figure 2 shows the posts with the five most total interactions from each page. The total interaction includes the total number of reactions, shares, and comments on a Facebook post. The y-axis means scaled share counts from 0 to 1, where the x-axis is timestep, a logarithm of a post's age.

We would name a post as 'breaking news' if its scaled share reached 0.8 before its timesteps become 10. CNN's second, third, and Biden's fifth plots show such rapidly increasing patterns. They are highlighted in red rectangular. Considering the 100 most interacted posts from each page in 2020, we observed CNN page has 16, Fox News has 2, Donald J. Trump has 3, and Joe Biden has 4 breaking news posts. In summary, a post's spread depends on features like the author's account, as well as its content.

2 Network mapping

We want to understand how COVID-19 information spread through linked posts; thus, we adopt network mapping in this section. We first analyze how lockdown haters and supporters behave in the social network. Next, to see the information diffusion from official sources to unofficial, we will study the linked posts from CNN and Fox News.

2.1 Lockdown haters vs supporters

First, we searched the 100 most interacted posts in 2020, which support lockdown, by the query, 'lockdown AND (positive OR lucky OR play OR safe).' The resulting posts say the authors try to be positive while standing the lockdown policies. From the 100 posts, we could track the linked 70 URLs, including external websites. In the end, we traced 2809 posts sharing the URLs by the link API.

On the other hand, lockdown haters' posts could be crawled with the query, 'lockdown AND (liberty OR unconstitutional).' They maintain the lockdown violates their freedom, therefore illegal. From the initial 100 hating posts, we acquired trackable 69 websites. Finally, we crawled 1496 posts sharing the sites.

Then we created the network where a node is either a Facebook account or a shared URL, and a directed edge means that an account shares a specific web address. An account has three types: Facebook page, group, and profile. Profile type has less than 1% of proportion in our analysis.

Besides, we labeled each node as lockdown supporters if the node is linked to the 70 lockdown supporting web sources—Vice versa for the haters' case. If an account shares both lockdown supporting and hating websites, it is named both sider. In the network, supporting nodes occupy 64.69%, while haters do 32.22%, while both siders are 3.09%. The network Figure 3 illustrates the both-siders(green dots) are located between the supporters and haters.

The node with the biggest in-degree on the supporting side is an announcement of the Pakistan prime minister assuring donation for COVID-19[3]. On the other hand, such a node on haters side maintains COVID-19 is a great lie[4]. Thus, we could understand that the conspiracy belief has been spread among the haters.

[3] https://www.facebook.com/NHSRCOfficial/videos/935719573533578/

[4] https://www.facebook.com/jordanbelfort/videos/262201088124997/

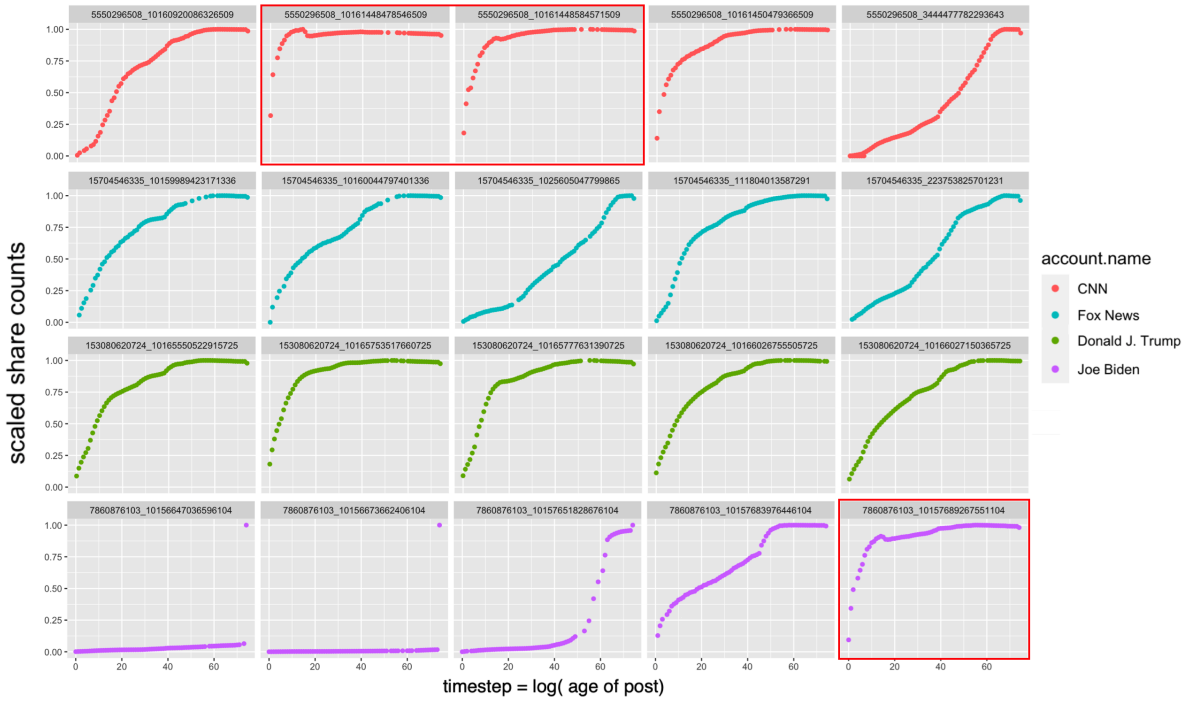
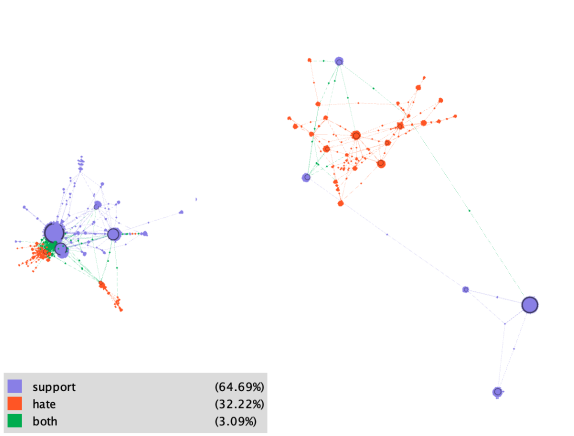


Figure 2: Cumulative sharing patterns



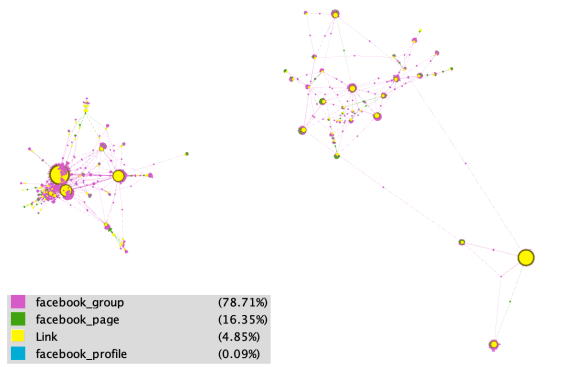


Figure 3: Facebook users' engagement around lockdown

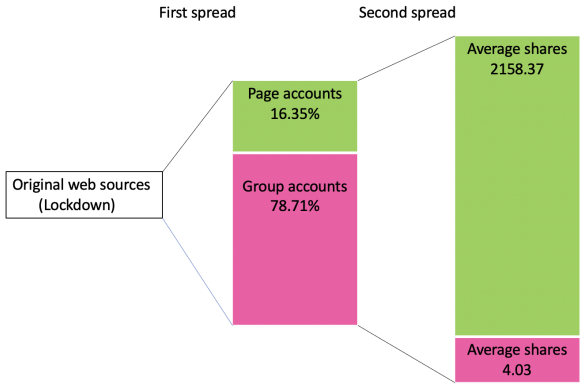


Figure 4: Share distribution of lockdown posts

Next, we want to see how two types of Facebook accounts(page or group) perform differently in the network. The group accounts occupy 78.71% of the total nodes, 16.35% nodes are page accounts, while 4.85% are linked web sources and Facebook profiles. The remaining 0.09% nodes are Facebook profiles. However, the posts' average number of shares is 2158.37, when the authors are page accounts. The average for the group accounts is just 4.03.

The Figure 4 explicitly explains this. Group accounts play a major role as information intermediators in the first shares. Then in the second information spread, page accounts does a more significant role. To summarize, page accounts spread

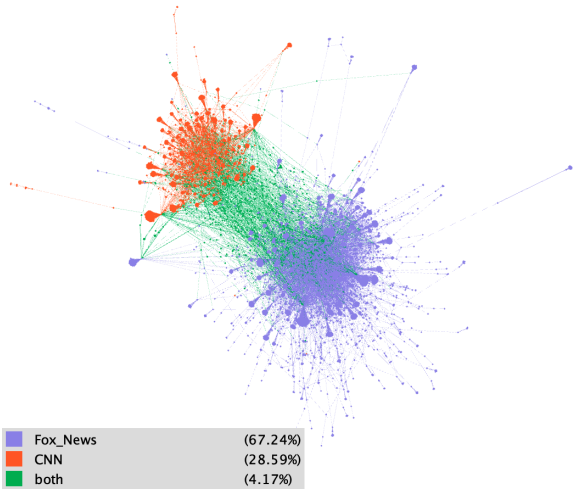


Figure 5: COVID-19 news sharing network from CNN and Fox News posts

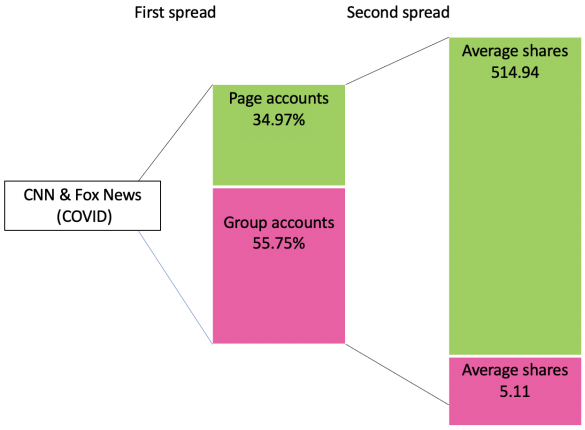


Figure 6: Share distribution of CNN and Fox News' COVID-19 information

lockdown information 535 times more actively than group accounts, even they are only 16.35% of the initial sharers of the original lockdown information.

Regarding account verification, group accounts always remain unverified state by Facebook. On the other hand, among the page accounts, only 13.68% are verified. It means any unverified account can also be a substantial information diffuser.

2.2 CNN vs Fox News

To understand how information spreads from traditional media to the public via Facebook, we make a network from CNN and Fox news posts. Like the network mapping about lockdown, we searched the 100 most shared posts in 2020 from each CNN and Fox News, with the query 'COVID.' Then, we tracked which posts shared the 200 information sources. If an account shares CNN's news, we label them as CNN vice versa for Fox News' case.' Both siders stand for accounts that share both news sources.

The Figure 5 visualizes this network. The light purple cluster means the Fox News' information sources and Facebook accounts sharing them. Orange ones stand for the CNN sides. Unlike the lockdown case, two clusters are conflicting and facing.

Similar to the lockdown network, Figure 6 explains that posts on the page account spread 100.77 times more. In particular, the average share of page accounts is 514.94, and the average of the group is only 5.11. On the other hand, 55.79% of network nodes are group accounts, and 35.00% are pages. Connected web sources make up 8.94%, and the remaining 0.27% nodes are the Facebook profile type.

At last, CNN's COVID-19 information spreads 1.58 times more than Fox News' one. Precisely, we summed up the shares of 1,000 COVID-19 posts from CNN and Fox News that were the most shared in 2020. In total, CNN posts were shared 3,162,516 times, and Fox News posts 2,001,144. Thus, even though the accounts initially sharing Fox News occupy 67.24% of the network, CNN's information spreads more broadly.

3 Topic modeling

In this section, we do topic modeling to understand what kind of topics spread during 2020 from official information sources to the public. Specifically, we collected 3,000 most shared COVID-19 posts on CNN and Fox News's Facebook pages in 2020. As a result, we acquired 2146 posts from CNN and 854 from Fox News. With this data, we experiment with three topic modeling methods: Latent Dirichlet Allocation(LDA), Cluwords, and Contextualized Topic Modeling(CTM). Besides, we would analyze the topics in a time series format as well as by sentiment analysis.

3.1 Preprocessing

Before actual modeling, we process data as follows: 1) remove stopwords by NLTK library, 2) further remove COVID-19, coronavirus, Fox News, CNN, and web protocol terms such as HTTP, www, .com, 3) do lower-casing. 4) remove small words whose lengths are shorter than three, 5) do tokenization.

Topic	Component words
1	trump, president, positive, tested, donald
2	social, source, content, medium, fbcnn
3	state, pandemic, york, city, states
4	vaccine, pandemic, first, people, says
5	vaccine, health, said, fauci, people
6	news, patients, health, said, hospital
7	trump, control, disease, president, said
8	questions, kids, gupta, sanjay, street
9	said, nation, home, sunday, people
10	mask, says, mental, infection, hospital

Table 1: Ten topics made by LDA

Topic	Component words
1	richardson, myers, jeffrey, hancock, hayden
2	children, families, homes, residents, kids
3	president, election, government, council, office
4	reported, found, claimed, sent, discovered
5	health, medical, disease, research, efficacy
6	full, numerous, first, certain, many
7	thing, really, actually, certainly, obviously
8	cases, year, total, times, instances
9	take, bring, develop, come, seek
10	keeping, putting, creating, bringing, moving

Table 2: Ten topics made by Cluwords

3.2 Latent Dirichlet Allocation (LDA)

LDA is a popular and primary method for topic modeling, which we use to compare the others' results. Table 1 shows ten topics ordered in marginal distributions. The most important topic says Trump's positive confirmation of the COVID-19 diagnosis.

3.3 Cluwords

The next algorithm is named Cluwords, using a modified TF-IDF matrix. In detail, they calculate an adjusted TF-IDF matrix defined according to the equation.

$$C_{TF-IDF} = C_{TF} \times idf(C) \tag{1}$$

where $C_{TF} = T \times C$, T is term-document matrix. Especially, the algorithm only leaves similar words when creating the C matrix. Then the method generates topic vectors by non-negative matrix factorization [13].

Table 2 suggests the modeling result. It seems the topics are clusters of similar words, technically because the algorithm discards non-similar words when creating its TF-IDF matrix. Our experiment reveals the method shows the lowest performance score.

3.4 Contextualized Topic Modeling

Next, we apply a model suggested by Federico Bianchi et al., named Contextualized Topic Models(CTM) [14]. The model is designed to deliver a more interpretable and coherent result than the standard LDA or existing neural topic models. Besides, they provide code for this on their Github. Table 3 shows the topic modeling result, ordered in marginal distributions.

Topic	Component words
1	provides, holds, andrew, looking, governor
2	study, likely, disease, blood, infection
3	died, test, police, year, three
4	**trump, president, donald, biden, white**
5	workers, kids, parents, back, care
6	vaccine, pfizer, vaccines, administration, biontech
7	relief, bill, senate, economic, package
8	social, health, content, source, medium
9	social, content, source, fbcnn, medium
10	reported, states, number, infections, cases

Table 3: Ten topics made by CTM

3.5 Model evaluation

To compare the three modeling results, we measure UCI, Umass, and NPMI coherences, refering to the paper [15]. The UCI coherence measures pointwise mutual information(PMI), calculated as follows:

$$C_{UCI}=\frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} PMI(w_i,w_j) \tag{2}$$

$$PMI(w_i,w_j)=\log\frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)} \tag{3}$$

The method only considers not word order but only co-occurrence.

The next measure is UMass coherence proposed in the article [16]. The summation of UMass coherence accounts for the ordering among the top words of a topic. The formula is as follows:

$$C_{UMass}=\frac{2}{N(N-1)}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\log\frac{P(w_i,w_j)+\epsilon}{P(w_i)} \tag{4}$$

The last coherence metric is C_{UCI} with Normalized PMI(NPMI). According to the articles [17] and [18], C_{UCI} performs better if the PMI is replaced by the NPMI. The NPMI formula is as follows:

$$NPMI(w_i,w_j)^{\gamma}=\left(\frac{PMI(w_i,w_j)}{\log P(w_i,w_j)}\right)^{\gamma} \tag{5}$$

where γ is a hyperparameter.

We measured the three metrics by Gensim library in python given default parameters and listed the results on the Table 4. Regarding C_{UMass} and C_{UCI}, LDA outperforms, while CTM does concerning NPMI. We chose the CTM's result as the

	LDA	Cluwords	CTM
C_{UCI}	-0.34	-10.83	-0.42
C_{UMass}	-2.85	-15.39	-3.58
C_{NPMI}	0.12	-0.38	0.21

Table 4: Topic modeling evaluation

final topics, as NPMI achieved the most considerable correlation to human topic coherence ratings in G. Bouma's experiment [19].

3.6 Time-series exploration

Then, we analyze when and which topics surged in 2020, referring to the paper [1]. The Figure 7 describes the ten topics' trends. The topic five rapidly increases in October, peaking on 3.Oct. The topic's three most shared posts on the day show an event: positive COVID-19 diagnosis of Trump and Kellyanne Conway. Thus, the social event was a trigger of information spread in the social network.

3.7 Sentiment analysis

Next, we analyze the emotional reactions by topics. Counting the most dominant emotion, Sad has four, Angry does two, Love does one, and Haha does one as described in the left plots of Figure 9. The result is highly different when we observe posts on the general Facebook group in the following subsection.

Especially, Topic 4 has Haha as the most count followed by Love. If we separate the two information sources, we could see a clear difference: Angry is the second dominant in CNN, while Love is the one in Fox News. It shows a different sentiment distribution depending on users' political stances.

3.8 Topics on Facebook groups

We applied the same preprocessing pipeline and CTM model to posts from the Facebook groups by querying the keyword COVID. Table 5 shows the ten topics. Interestingly, topic G is about religious pray that did not show up from the CNN and Fox News case. Also, the sentiment distribution is significantly different. In group posts, Love is dominant (Figure 9).

Topic	Component words
A	changing, sound, wore, thanksgiving, copied
B	mexican, advance, cheese, double, posting
C	pastor, testimony, soon, shame, wore
D	test, patients, hospital, positive, tested
E	people, virus, like, home, even
F	public, must, health, community, government
G	**jesus, pray, lord, amen, heart**
H	world, trump, country, president, lockdown
I	million, name, share, fraud, organize
J	please, family, would, home, keep

Table 5: Ten topics from group posts

4 Prediction modeling

In this section, we predict posts' future share counts by initial information. The goal is to model multiple posts' information spreads and observe which features play a crucial role in the model. Initially, we crawled 100 most interacted posts for each following queries: 1) COVID and mask, 2) COVID and lockdown, 3) COVID and vaccine. From the 300 posts, we acquired 254 trackable posts.

In the following subsections, we explore and preprocess the data and apply different modeling approaches. The problem is that Crowdtangle limits the API access time; thus, the size of the dataset is not sufficient despite days of crawling. We would explain the details, including our solution.

4.1 Preprocessing

We need to process the data because 1) each post has a different spreading period and 2) missing datapoints. First of all, when we explore the 300 posts, the number of shared posts drops down when the timestep becomes bigger than 47.

Specifically, Figure 10 illustrates this where the x-axis is timestep, and the y-axis is the number of posts existing on a specific timestep. Therefore, we fix the 47th timestep as the maximum lifetime. Also, the plot only explains the 300 most popular posts, and if we crawl additional datasets, such as 1200 posts in 2020, then the sharing period would typically decrease. Therefore, we trim a post history after the 47th timestep and impute if it does not last until the point.

Regarding the imputation method, we do it by linear average and regression. For example, if historical data is missed at the 15th timestep but not at the 16th and 17th, we complement it as an arithmetic average. Also, if a post is no more shared after the 45th timestep, we complement data between 46th and 47th by linear regression.

4.2 Feature engineering

We make the posts have both historical data and individual features. The historical features include shares, likes, and the number of reactions from timesteps within the initial period. The period de-

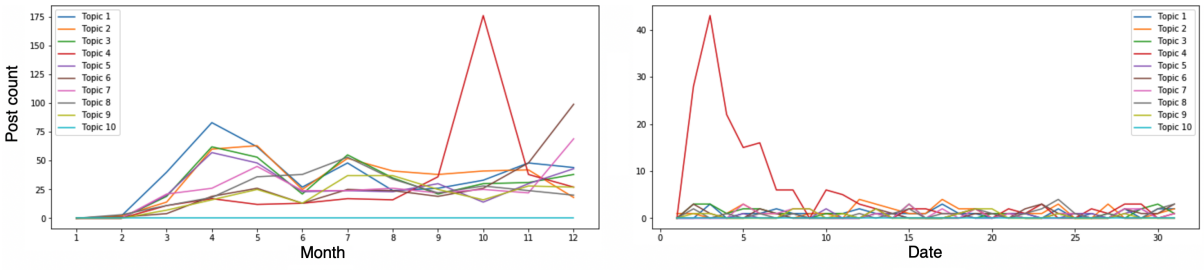


Figure 7: Monthly topic distribution in 2020(left), and daily topic distribution in Oct.2020(right)

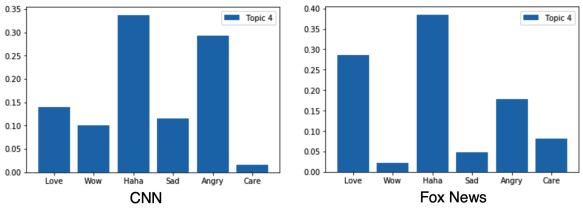


Figure 8: Sentiment distribution of Topic 4 by the information sources

pends on the dataset, but at first, we set it between timesteps one and ten. The individual ones do account verification, page/group, topics such as lockdown or mask, subscriber count, and embedded message. Especially, we applied a pre-trained Bert encoder to embed the post messages and applied PCA to reduce its dimension to 100. Besides, the target variables consist of share counts in the future that we initially set timesteps between 11th and 47th.

4.3 Machine learning for panel data

Our data is panel data, but we implement machine learning methods instead of conventional panel data modeling. Specifically, panel data means time-series data that has not only multi-variates but also multi-subjects. Statisticians mainly take two methods to model such panel data: fixed-effect model and random effect model. The first model considers the personal effect as constant varying by a subject. The second one does it as a random error varying by a subject. However, we have specific data for each post and account. To utilize it, we train machine learning models.

4.4 First dataset

To train machine learning models, we crawled more data. We collected 100 most shared posts per month by the same queries; as a result, we acquired 3600 posts, and 2492 of them were trackable. It took more than 36,000 sec to crawl because Crowdtangle allows only six API accesses within

	Data1	Data2	**Data3**
size	2492	4190	50280
μ	7200.24	7765.38	5267.28
σ	3325.89	2260.11	**790.26**

Table 6: RMSEs comparison. μ means average RMSE and σ is standard deviation.

1 minute. Including data processing time, it takes less than a day in practice.

Unfortunately, Root Mean Square Error(RMSE) highly fluctuated regardless of three model types: linear regression, XGBoost, a neural network with one hidden layer. The fluctuation is because the dataset is too small, so train and validation sets' distributions keep changing whenever sampling. The Figure 11 explains it.

4.5 Second dataset

The straightforward solution for the variation is just crawling more data. Thus we crawled posts with additional three keywords: Trump, Biden, and Jesus, and obtained 4,190 trackable posts in total. The words were chosen based on the topic modeling analysis. This time, we experimented with linear regression for 10-fold cross-validation. The result says that the standard deviation of errors decreased from 3325.89 to 2260.11, illustrated in Table 6. The crawling managed to decrease the error, but still not satisfactory as well as practical in that it took more than 72,000 secs in total.

4.6 Third dataset

To generate more data without further time-consuming crawling, we sliced the existing data by sliding a time window whose size is six. For example, we acquired the first train record from time step first to sixth timesteps, and the second one from second to seventh et cetera. Also, the target variables became the future share on the next 30 steps. Finally, we obtained 50,280 records. The standard deviation of the cross-validation errors de-

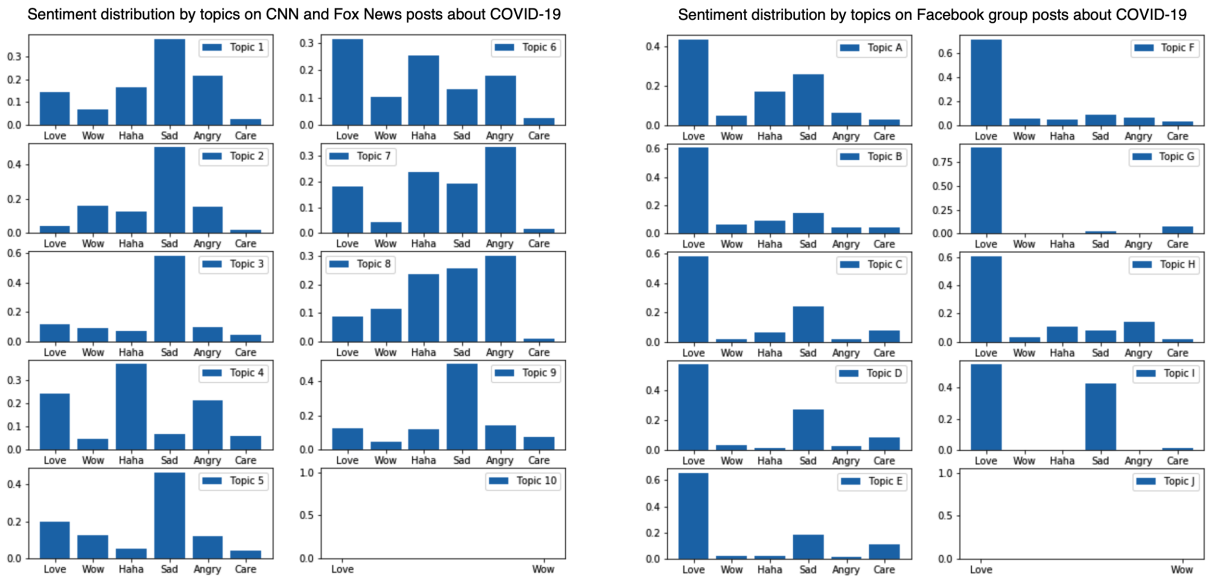


Figure 9: Sentiment distributions by topic and post types. The last topics have zero records when counting the most dominant topic of each post

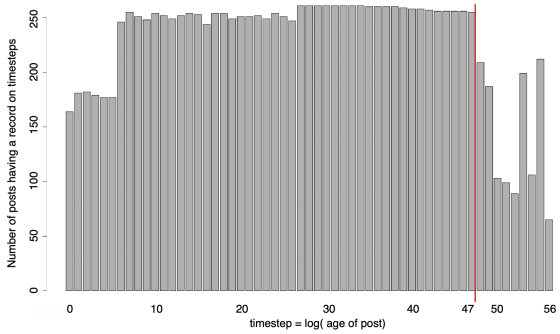


Figure 10: Distribution of post ages

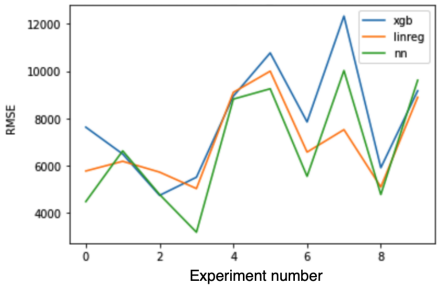


Figure 11: Modeling issue: fluctuating errors by the first dataset

	Validation RMSE	Configuration
Linear regression	6111.94	default
XGBosst	3143.16	default
Neural network	**1053.54**	2 hidden layers 2000 epoch

Table 7: Predicting models evaluation

creased to 790.2588. Plus, the average error was also the lowest. Table 7 describes the detail.

4.7 Final model

With the third dataset, we tested models including linear regression, XGBoost, and neural networks. We separated this into train, validation, and test datasets in a ratio of 80%, 10%, 10%. Then, we measure the validation RMSEs by the models, using Scikit-learn and Tensorflow. As summarized in Table 7, the regression and XGBoost were configured as default. The neural network with two hidden layers trained on 2,000 epoch shows the best result whose RMSE is 1053.54.

The plots in Figure 12 demonstrate the prediction result on the test dataset whose x-axis is predicted target variables(30 timesteps in the future), while the y-axis is the number of shares. The plot (a) is an overall average of all records. Next, plot (b), (c) are test cases whose shares are above 5,000. The model captures the general trend. At last, plots (d) seems to have a more significant gap between predictions and actual values. However, the gap is smaller than 150, while the model captures their

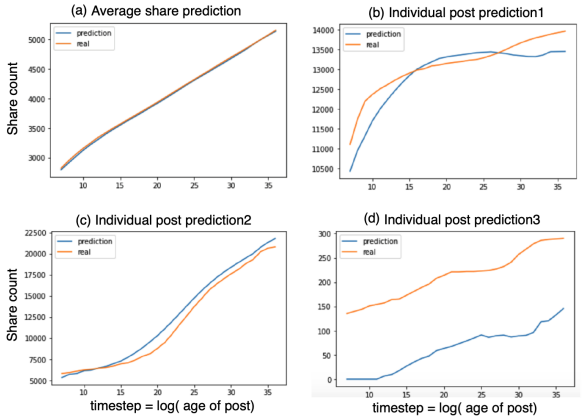


Figure 12: Prediction result

Feature	Importance
Share at timestep 6	1134.82
Share at timestep 5	667.86
group account	382.08
page account	375.12
...	...
Love at timestep 2	25.10
Haha at timestep 4	23.50
Angry at timestep 4	21.06
Haha at timestep 3	18.54

Table 8: Feature importance

trend.

4.8 Feature importance

We recognize essential features by the variable perturbance. The method adds random noises to the input features and measures how much it affects the final result. Table 8 summarizes the importance. The two most important features are share count at the sixth and fifth-time steps. Interestingly, the third and fourth important ones are group and pages. It means that an author's account type is a crucial barometer to predict a post's information spread. It is reasonable according to the network mapping analysis: page accounts are major information diffusers. On the other hand, emotional reactions are graded as not significant. They are even less meaningful than post-embedding's elements.

5 Conclusion

We hope this paper delivers meaningful results to understand the enormous social network's role in COVID-19 information spread. We finalize the article by suggesting a summary and possible future tasks on top of it.

5.1 Summary

So far, we collected and analyzed Facebook posts to answer the four research questions mentioned below.

1. How does a general sharing pattern of COVID-19 posts appear?

   According to our sharing pattern exploration, it has a skewed distribution changing by post features. In detail, Pfizer's vaccine announcement and Trump's argument have distinct the shared periods. Their proportions of verified sharing accounts also differ.

   Analyzing the 100 most shared posts from the CNN, Fox News, Biden, Trump's pages, we observed that the accounts have different breaking information ratios.

2. How does COVID-19 information spread through linked posts? Especially from official sources to unofficial on Facebook?

   In a nutshell, page accounts play a major role in the diffusion. Especially, CNN's information spread 1.58 times more than Fox News one.

   To recognize it, we created two network through the linked posts: the one from lockdown posts, the other from CNN and Fox News posts.

   In both network, page accounts play a significant role in the COVID-19 information diffusion, which we visualized as two steps of spreading.

   In the lockdown network, we recognized the conspiracy had been the main information source among the lockdown haters' accounts.

   The second network explains the information dissemination from official sources to unofficial. While the accounts initially sharing Fox News occupy 67.24% of the network, CNN's information spread 1.58 times more because of the larger second spread.

3. What are the main COVID-19 topics and trends in 2020? How did users react to them?

   Our topic modeling with sentiment analysis figured out that Trump's coronavirus confirmation made an significant impact on the topic trends, and the emotion towards this event differs by users' political stances.

To analyze this, we crawled the 3,000 most shared posts from CNN and Fox News pages and tested the three models: LDA, Cluwords, CTM. Evaluating them with the three coherence metrics, we chose CTM as our final topic model with the most NPMI.

In addition, groups' posts have distinct topics and corresponding emotion distributions. One popular topic is about religious pray, not found in the topics from the official news sources. Plus, the dominant emotional reaction is Love regardless of the topics.

4. Can we predict a post's long-term future share counts with its initial behavior? If then, which feature plays a crucial role?

   We could predict the number of shares in the next 30 timesteps using the initial 6 timesteps information. Note that the timestep is log of a post's age.

   The share counts at timestep fifth and sixth are the two most critical features, followed by account types: page/group. Emotional reactions are less crucial than embedded message.

   According to our experiment with three datasets, it is more efficient and practical to slice time-series data with time window to decrease time-consuming crawling.

   Finally, we chose the neural network model whose RMSE is the lowest. Its validation error was lower by 6 times than linear regression's.

5.2 Future works

We suggest two points of complementing in the future.

First point is the relationship between the information spread and network statistics such as density and modularity. Higher density in a network from linked posts means the information is more actively shared. Also, high modularity has active interactions among accounts within modules but fewer connections to other modules. It would be meaningful to explore and compare such graph metrics considering the spread.

Second, one can utilize more advanced model as the article [10] suggested. We use the simple neural network with two hidden layers, because our modeling goal is to analyze feature importance on top of a satisfactory result.

References

[1] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs," *arXiv preprint arXiv:2005.03082*, 2020.

[2] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "The spread of covid-19 conspiracy theories on social media and the effect of content moderation," *The Harvard Kennedy School (HKS) Misinformation Review*, vol. 1, 2020.

[3] A. Bruns, S. Harrington, and E. Hurcombe, "<? covid19?>'corona? 5g? or both?': the dynamics of covid-19/5g conspiracy theories on facebook," *Media International Australia*, vol. 177, no. 1, pp. 12–29, 2020.

[4] S. Boberg, T. Quandt, T. Schatto-Eckrodt, and L. Frischlich, "Pandemic populism: Facebook pages of alternative news media and the corona crisis–a computational content analysis," *arXiv preprint arXiv:2004.02566*, 2020.

[5] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proceedings of the 20th international conference on World wide web*, pp. 695–704, 2011.

[6] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, 2010.

[7] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*, pp. 338–349, Springer, 2011.

[8] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social emotion mining techniques for facebook posts reaction prediction," *arXiv preprint arXiv:1712.03249*, 2017.

[9] N. Straton, R. R. Mukkamala, and R. Vatrapu, "Big social data analytics for public health: Predicting facebook post performance using artificial neural networks and deep learning," in *2017 IEEE International Congress on Big Data (BigData Congress)*, pp. 89–96, IEEE, 2017.

[10] K. Singh, "Facebook comment volume prediction," *International Journal of Simulation: Systems, Science and Technologies*, vol. 16, no. 5, pp. 16–1, 2015.

[11] N. Shiffman, "Crowdtangle cookbook," 2019.

[12] N. Shiffman, "Crowdtangle api wiki," 2020.

[13] F. Viegas, S. Canuto, C. Gomes, W. Luiz, T. Rosa, S. Ribas, L. Rocha, and M. A. Gonçalves, "Cluwords: exploiting semantic word clustering representation for enhanced topic modeling," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 753–761, 2019.

[14] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," *arXiv preprint arXiv:2004.03974*, 2020.

[15] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408, 2015.

[16] L. AlSumait, D. Barbará, J. Gentle, and C. Domeniconi, "Topic significance ranking of lda generative models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 67–82, Springer, 2009.

[17] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pp. 13–22, 2013.

[18] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539, 2014.

[19] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of GSCL*, pp. 31–40, 2009.