The first year of the Covid-19 pandemic through the lens of r/Coronavirus subreddit: An exploratory study

Zachary TAN, Anwitaman DATTA
School of Computer Science and Engineering
Nanyang Technological University, Singapore
E-mail: zacharytan.cs@gmail.com, anwitaman@ntu.edu.sg

Abstract This study looks at the content on Reddit's COVID-19 community, r/Coronavirus, to capture and understand the main themes and discussions around the global pandemic, and their evolution over the first year of the pandemic. It studies 356,690 submissions (posts) and 9,413,331 comments associated with the submissions, corresponding to the period of 20th January 2020 and 31st January 2021. On each of these datasets we carried out analysis based on lexical sentiment and topics generated from unsupervised topic modelling. The study found that negative sentiments show higher ratio in submissions while negative sentiments were of the same ratio as positive ones in the comments. Terms associated more positively or negatively were identified. Upon assessment of the upvotes and downvotes, this study also uncovered contentious topics, particularly "fake" or misleading news. Through topic modelling, 9 distinct topics were identified from submissions while 20 were identified from comments. Overall, this study provides a clear overview on the dominating topics and popular sentiments pertaining to the pandemic during the first year. Our methodology provides an invaluable tool for governments and health authorities to obtain a deeper understanding of the dominant public concerns.

Keywords COVID-19, Reddit, Sentiment analysis, Topic modelling

Code and Data Availability: https://doi.org/10.21979/N9/0LGZYN

1 Introduction

The outbreak of the 2019 novel coronavirus (COVID-19) was an event that dominated global news since 2020 and has continued to do so as of the writing of this paper, well past mid-2021. With its first reported cases in Wuhan, China, the highly infectious virus spread rapidly and was declared a global pandemic on 11 March 2020 (Cucinotta and Vanelli, 2020). To 'flatten the curve' and reduce the burden on overwhelmed healthcare systems, many countries worldwide implemented a series of measures, such as travel bans, temporary closures of public spaces and businesses. In a time of reduced in-person social interactions, social media became the best way to stay connected with other people. It also provided an outlet for people to express their frustrations and anxieties caused by the virus and the various impacts on individuals as well as the society at large due to the restrictive measures.

Apart from the social element, social media has also become a popular way to get the latest news. One such platform is Reddit, which provides a platform for both sharing news (links/content) and carry out lengthy discussions unhindered by practical limits to express oneself or discuss and debate topics in great details, since comments on Reddit have a 10000 characters limit (as opposed to 280 characters limit on Twitter) and Reddit allows pseudonymous participation. Home to more than 400 million monthly users (Kastrenakes, 2020), Reddit is a social news aggregator that features a collection of news articles, text posts and visual content submitted by users. Users then curate the submissions and their comments with upvotes and downvotes. The forum is divided into communities called subreddits, each revolving around a central theme or topic. This research focuses on one such subreddit, r/Coronavirus which discusses issues related to COVID-19. Some distinctive aspects of Reddit from conventional social media platforms such as Facebook is its nature of user anonymity and user curation. These distinctions have fostered an environment where truthful, unfiltered sentiments might be more easily shared, and popular ones are upvoted and showcased.

As more people use social media to find and share news (Newman, 2020), social media content has become invaluable data sources for mining trends and sentiments. There have been recent works focused on extracting and analysing online sentiments with respect to COVID-19. For instance, H. Yin et al. (2020) proposed a framework to analyse topic and sentiment dynamics due to COVID-19 based on tweets from Twitter over a span of two weeks. Another study by D. Low et al. (2020) utilised natural language processing (NLP) techniques to characterise changes in fifteen subreddits focused on mental health before and during the pandemic. However, more research is necessary to analyse the shifts in popular topics and sentiments over a longer timespan to identify patterns and trends.

This study is exploratory in nature, aimed to gain insight on the public's sentiments and attitudes towards COVID-19, for general understanding but also since it may help policymakers and health officials understand public opinions and perceptions, and respond more effectively to the people's concerns. We do so by identifying popular sentiments and topics on the subreddit r/Coronavirus and evaluating how these elements have evolved in the span of a year, beginning from the virus outbreak to the rollout of vaccines.

Data comprising submissions and comments was first crawled from r/Coronavirus and preprocessed. Subsequently, the Valence Aware Dictionary and Sentiment Reasoner (VADER) (Hutto and Gilbert, 2014) was employed to acquire and understand sentiments from the data. Series of topics were also mined using the latent Dirichlet allocation (LDA) model for further analysis. These methods would help to achieve a more comprehensive overview of the various sentiments and trends in the one year after COVID-19. Our findings revealed that while there are 9 distinct themes in Reddit submissions, the comments in the subreddit are more diverse with 20 distinct topics. Sentiment of submissions are generally more negative than positive due to the nature of COVID-19 news, while the ratio of negative and positive sentiment in comments is fairly equal. We also observed through upvoted and downvoted comments that the r/Coronavirus community generally disapproves of comments that do not treat the COVID-19 virus seriously.

Overall, the contribution of this work is exploratory in nature, and its value is in (i) the establishment of the summary insights from r/Coronavirus subreddit data, in the process, (ii) creating a curated dataset (made available at https://doi.org/10.21979/N9/0LGZYN) which would serve as a valuable resource for any future studies by the research community, and (iii) the accompanying code (also available at the same link) and methodology establishes a base-line approach, which can be reused and extended upon, to continue to gather similar insight over time, as and if the pandemic continues to persist.

Next, we review past studies of COVID-19 on social media, followed by this study's methods of data collection and preparation, and then we delve deeper with the analysis and evaluation of our findings.

2 Literature Review

In the early stages of COVID-19, there were several studies on the sentiments present on social media. Twitter was amongst the social media platforms for which such research was carried out. One study analysed the topic and sentiment dynamics surrounding COVID-19 based on a compilation of 13 million tweets over two weeks (Yin et al., 2020). Another study mined a collection of 107,990 tweets related to COVID-19 in the first three months of the outbreak and identified topics using the LDA model (Boon-Itt and Skunkan, 2020). Xue et al. (2020) also utilised the LDA model to analyse 1.9 million tweets and discovered 11 topics related to COVID-19.

Despite Twitter's popularity, it is greatly limited in the length of its average content. Given that the most common length of a tweet is 33 characters (Perez, 2018), the data may be limited, which may consequently affect the comprehensiveness of topics and evaluation. The studies also focused on data within a maximum span of three months, which is a relatively short duration compared to the length of the pandemic. Therefore, there is a need to collect longer text data over an extended period for a more complete analysis. To address this gap, we have chosen a period of approximately a year, spanning the period from the public knowledge of the pandemic till the time point when the first vaccine rolls out, which in retrospect can be viewed as the early (pre-vaccines) stage of the pandemic. While it is still an ongoing event, and vaccination is happening across the globe in a very heterogeneous manner, we believe that in future, understanding the social discourse on COVID -19 may be decomposed in three logical phases – pre-vaccination, during the period of initial vaccination phase, and eventually (and hopefully) when society can view COVID -19 endemic but its impact reasonably under control. In that context, the current study captures the first (pre-vaccines) phase.

3 Methods

3.1 Reddit Dataset

3.1.1 Data Collection

Reddit is organized as theme or topic centric subreddits. Users post submissions within any given subreddit, and the submission comprises a title, often along with an URL link referencing some online content, and possibly accompanied with a further body of text. Users then post comments within a submission thread, as response to the original post or as response to other comments. Users can also up or downvote the original submission, as well as individual comments.

Data was retrieved from the subreddit r/Coronavirus using Pushshift API. 356,690 submissions and 9,413,331 comments between 20th January 2020 (the beginning of the subreddit) and 31st January 2021 were extracted. The features of the data are shown in Table 1.

Data	Attributes
Submission	Id, title, author, created_utc, domain, full_link, num_comments, score, total_awards_received, is_self, subreddit_subscribers
Comment	Id, author, body, score, total_awards_received, created_utc, parent_id, permalink

Table 1 Attributes available with the data collected. The aggregate votes are represented by the 'score' attribute.

3.1.2 Data Preprocessing

In the submission dataset, discussion threads submitted by the subreddit's moderators were removed. Submissions with titles that were not in English were also removed using Python's *langdetect* library. In the comment dataset, deleted, removed, duplicated and bots' comments were removed. Subsequently, URLs,

email addresses and special characters[1] were stripped from the remaining submissions and comments as they did not contribute meaning to the text. This would improve the quality of tokens extracted in the later stages for sentiment analysis and topic modelling. Data that returned empty strings as a result of text cleaning were further removed. Fig. **1** shows the composition of comments that were filtered out, and that the final cleaned data comprised 90.3% of the comments originally retrieved.

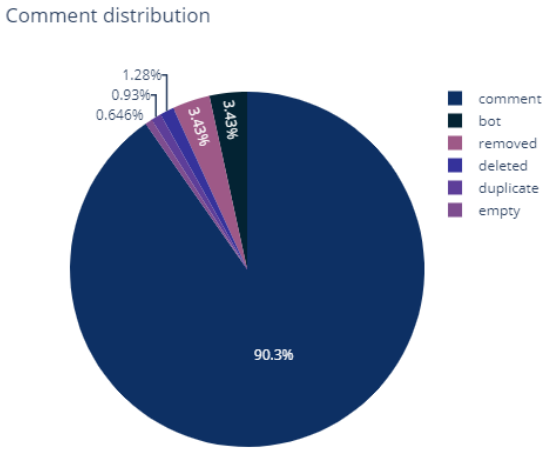

Fig. 1 Distribution of comment dataset

In preparation for the topic modelling phase, the submission and comments datasets were sanitised further. Both datasets were pre-processed by the lemmatisation of tokens, removal of stop words and phrase extraction. Terms that were redundant in submission texts, such as 'coronavirus' and 'covid' were also removed from the submission dataset.

This was followed by the filtering of submissions with less than 5 tokens and comments with less than 7 tokens as these did not contribute substantially useful information. The difference in minimum token count was due to the headline-like nature of submissions (and thus submission texts being inherently shorter) and the conversational nature of comments.

[1] Punctuation marks such as '!' were retained for sentiment analysis.

268,109 submissions constitute the submission dataset for sentiment analysis and topic modelling; 8,499,244 comments (short comments preserved) constitute the comment dataset for sentiment analysis while a smaller subset of 5,243,419 comments were used for topic modelling.

3.2 Sentiment Analysis

3.2.1 Comments and Submissions with VADER

Sentiment analysis involves the contextual mining of text to extract its sentiment polarity. In this study, VADER (Valence Aware Dictionary and Sentiment Reasoner, (Hutto and Gilbert, 2014)) is utilised to identify if each submission or comment is positive, negative, or neutral. VADER is a lexicon and rule-based sentiment analysis tool that is specifically adapted to sentiments expressed on social media. It takes into consideration negation, punctuation, sentiment-related initialisms and commonly used slangs (Hutto and Gilbert, 2014), making it more applicable for sentiment analysis of informal content on Reddit than other popular tools such as Textblob and NLTK (Bonta et al., 2019).

VADER's sentiment analyser was employed on each submission and comment individually. Consequently, each submission or comment is given a set of scores, of which the most important is the compound score that lies between 1 and -1. The score represents the overall sentiment of the given text, where -1 is the extreme negative and 1 is the extreme positive. A text document with a score between 0.05 and -0.05 is classified as neutral; a score greater than or equal to 0.05 was given a "positive" classification while a score less than or equal to -0.05 was attributed a "negative" classification.

On a further note, a phrase such as "tested positive" is semantically positive due to the lexicon "positive" but is in fact negative in context. Therefore, the terms "positive" and "negative" were swapped prior to sentiment analysis to improve the precision of sentiment scores and restored afterwards. An example of the impact of the swapping is illustrated in **Table 2**.

Text		Compound score
Aide to Vice President Mike Pence tests positive for coronavirus	Before swapping	0.5574
	After swapping	-0.5719

Table 2 Example of sentiment analysis with the term 'positive'

To capture the differences between the submissions with positive sentiments and those with negative sentiments, word clouds were generated using the *wordcloud* library in Python. Treating them as two separate documents, positive and negative submissions were subsequently converted to the bag-of-words model using Scikit-Learn's CountVectorizer with *"ngram_range = (1,3)"* for bigram and trigram generation. The result is a 2 x 2151767 sparse matrix, excerpt of which is shown in **Table 3**.

To better capture the differences in submissions of opposing polarity, log odds ratio was calculated for each token in the matrix. This method helps to identify tokens that are more common in the positive submissions than in the negative submissions, and vice versa. The formula below is applied to each value in the matrix:

$$\log_2\left[\frac{\left(\frac{1+\text{frequency of X in A}}{1+\text{total word count in A}}\right)}{\left(\frac{1+\text{frequency of X in B}}{1+\text{total word count in B}}\right)}\right] \quad (1)$$

where X is a token, and A and B are different corpuses. This transforms the matrix to the one, excerpt of which is shown in **Table 4**. Subsequently, word clouds were generated using both matrices to visually identify and interpret the most prominent words and corresponding themes.

	cases	new	help	vaccine	people	says	health	china	care	masks	...
negative	9029	10062	747	2982	6980	5804	4324	4837	692	2269	...
positive	6396	5387	4216	3910	3817	3318	3135	2600	2415	2392	...

2 rows × 2151767 columns

Table 3 Bag-of-words model

	cases dead	total cases dead	new deaths raising	daily death toll	uk death	uk death toll	positive trump	daily death	worst hit	death toll passes	...
negative	7.838227	7.571834	7.371965	7.004420	6.760691	6.530904	6.467245	6.330826	6.318853	6.294606	...
positive	-7.838229	-7.571835	-7.371966	-7.004421	-6.760692	-6.530905	-6.467246	-6.330827	-6.318855	-6.294607	...

2 rows × 2151767 columns

Table 4 Matrix with log odds ratio

3.2.2 Score-weighted Sentiment Score

To reflect the sentiment of the subreddit's community more accurately, the "score" attribute from each submission was used to compute a score-weighted sentiment. The score indicates the net popularity of the content, therefore taking into consideration may allow clearer comparisons of the popularity of negative and positive content on the subreddit.

Table 5 describes the score in the data which ranges from 0 to 77296. The minimum score is zero as submissions can only be downvoted to zero irrespective of the true number of downvotes it received. Nonetheless, the score still reveals the popular submissions on the subreddit. Reddit's nature of making popular content more visible justifies the greatly right-skewed distribution observed in **Fig. 2**. By nature, submissions have a score of 1 at the time of creation. Since 85.9% of submissions in the dataset has a score of 1, this suggests that they were likely not upvoted or seen by the community. To reduce the skew, log transformation is applied to the score and the resulting value is used as a multiplier. The range of the multiplier is shown in Fig. 2. The score-weighted sentiment is then derived using the following equation:

$$\text{Score-weighted sentiment} = \log_{10}(score + 1) \times \text{compound score} \quad (2)$$

Mean	19.93
Std	619.71
Min	0
25%	1
50%	1
75%	1
Max	77,296

Table 5 Statistics of submission score

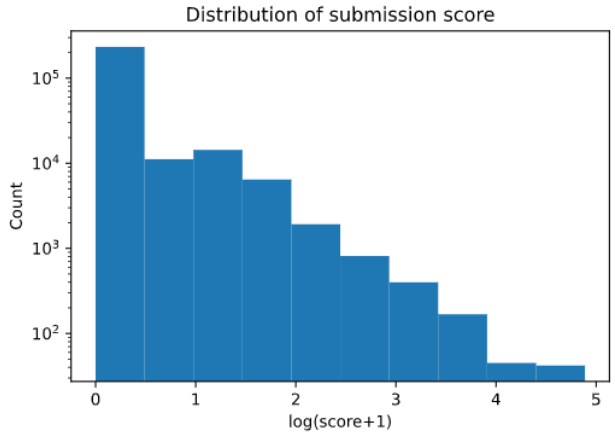

Fig. 2 Multiplier values after log transformation of score

3.2.3 Comment Score

The score of a comment indicates the net quantity of upvotes or downvotes it has received from the community. **Table 6** shows the range of the score in the comment dataset. As each comment is given a default score of 1, it is unsurprising that 67% of comments has a score of 1. Comments with a negative score are interpreted as unpopular in the subreddit, while those with a positive score may be viewed as popular. The Pearson correlation coefficient between the sentiment score of a comment and its net score (votes) is 0.000326, indicating that there is no relationship between these two variables. This can be observed in the scatter plot displayed in Fig. 3.

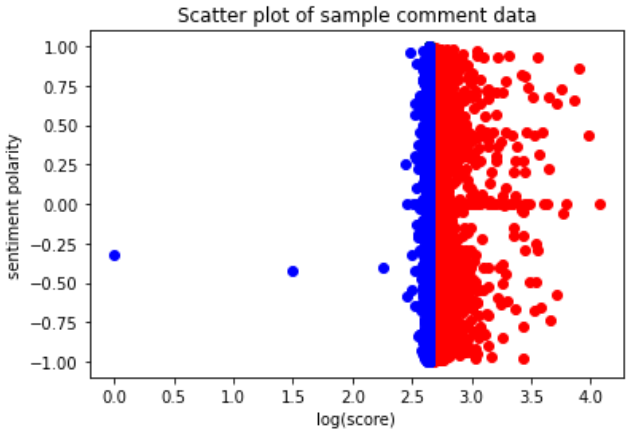

Fig. 3 Scatter plot with sentiment score on the y-axis and log-transformed score in the x-axis. The blue dots represent downvoted comments while the red dots represent upvoted comments.

Mean	2.75
Std	29.6
Min	-642
25%	1
50%	1
75%	2
Max	11,247

Table 6 Statistics of comment score

	Upvoted comments	Downvoted comments
Min	2	-642
25%	2	-2
50%	3	0
75%	6	0
Max	11,247	0

Table 7 Score statistics of upvoted and downvoted comments

Highly upvoted or downvoted comments are more representative of the subreddit's opinions. Hence based on the statistics shown in **Table 7**, upvoted comments with a score in the upper 25% and downvoted comments with a score in the lower 25% were selected to provide insights on the popular and unpopular opinions on r/Coronavirus. These two sets of data were subsequently used to build two corresponding sets of corpora with unigrams and bigram. The bag-of-words model and log odds ratio method were applied, followed by the generation of word clouds.

3.3 Topic Modelling

Python's *gensim* library was used to create LDA (Latent Dirichlet Allocation) models on each submission and comment datasets. LDA is an unsupervised, generative statistical model that assumes each document – in this case, a submission or comment – to be a mixture of topics and each topic a mixture of words. To find the topic representation of each document and the words that

contribute to each topic, LDA goes through each document and randomly assigns each word to one of the *K* topics. It subsequently repeats a generative process multiple times to improve on the assignments to produce a final set of topics and their associated words.

To achieve the optimal number of topics for each model, multiple models with varying numbers of topics were generated. The *gensim* topic coherence pipeline (Röder et al., 2015) was implemented on each model to calculate its coherence value 'c_v' to measure the extent of semantic similarity between high scoring words of topics. The model that gave the highest coherence value was selected for further study.

3.3.1 Submissions

268,109 pre-processed submissions were used to create a dictionary of 16,069 unique tokens and bigrams, for each token which has appeared in more than 5 submissions. The dictionary was then applied to the submission dataset to create a bag-of-words corpus which was used to generate six LDA models with 3, 5, 7, 9, 11 and 13 topics. Evaluation of their coherence is shown in Fig. 4, where the model with 9 topics had the highest coherence value. The final LDA model with 9 topics was regenerated with 20 passes.

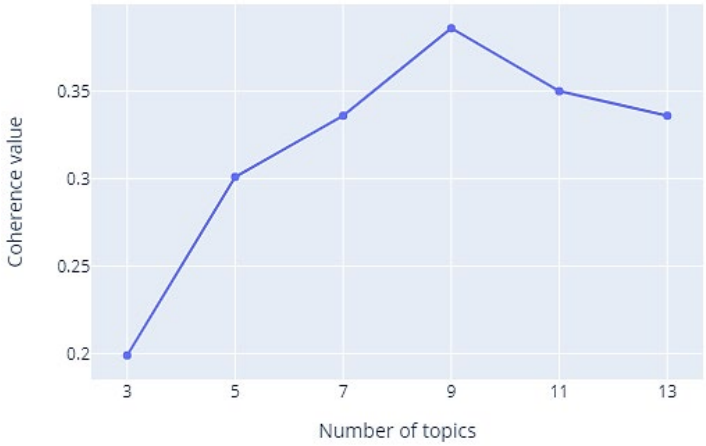

Fig. 4 Coherence values of LDA models for submissions

3.3.2 Comments

5,243,419 pre-processed comments were used to create a filtered dictionary of 62,017 unique tokens and bigrams. The dictionary was then applied to the comment dataset to create a bag-of-words corpus that generated six separate LDA models with 4, 8, 12, 16, 20 and 24 topics. As illustrated in Fig. 5, the model with 20 topics produced the highest coherence value, hence it was selected for topic identification and analysis.

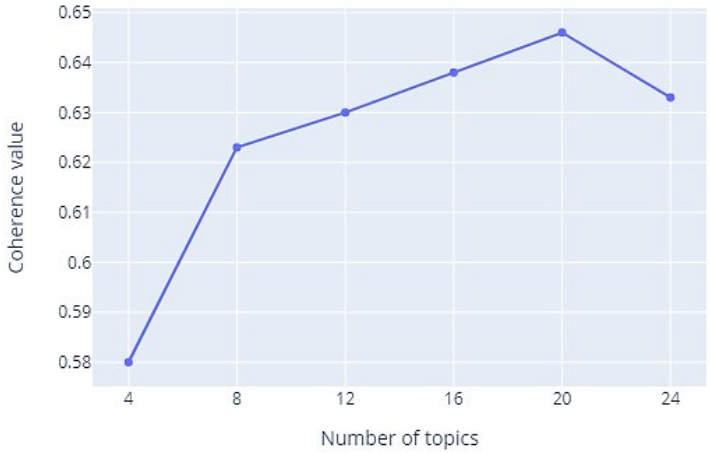

Fig. 5 Coherence values of LDA models for comments

4 RESULTS AND DISCUSSION

4.1 Submissions

4.1.1 Sentiment Analysis

Submissions with positive, negative, and neutral sentiments for each week were tallied and plotted over the whole year of data, as shown in Fig. 7. For all this period, the number of negative submissions was consistently higher than the number of positive submissions. This reflects that the nature of news tended to be negative.

Submissions on r/Coronavirus began from the 3rd week of 2020, when the first confirmed cases of the novel virus were reported outside China (Taylor, 2021). The subreddit gradually received more submissions during the month of February (weeks 5 to 8) when outbreaks occurred in several countries such as Italy, South Korea and first cases surfaced in numerous other countries, as shown in Fig. 6. From week 8 onwards, the number of submissions increased rapidly and reached its peak in week 10 when the WHO (World Health Organization) declared the COVID-19 outbreak a pandemic on 11th March (Cucinotta and Vanelli, 2020). Following this, the number of submissions fell quickly back to pre-March levels despite the continuous rise in new cases. This indicates a possibility that March was the point in time when the virus was taken more seriously due to the influx of negative news, subsequent to which, a form of ‘normalization’ set in. This influx also resulted in the biggest gap between positive and negative submissions in March.

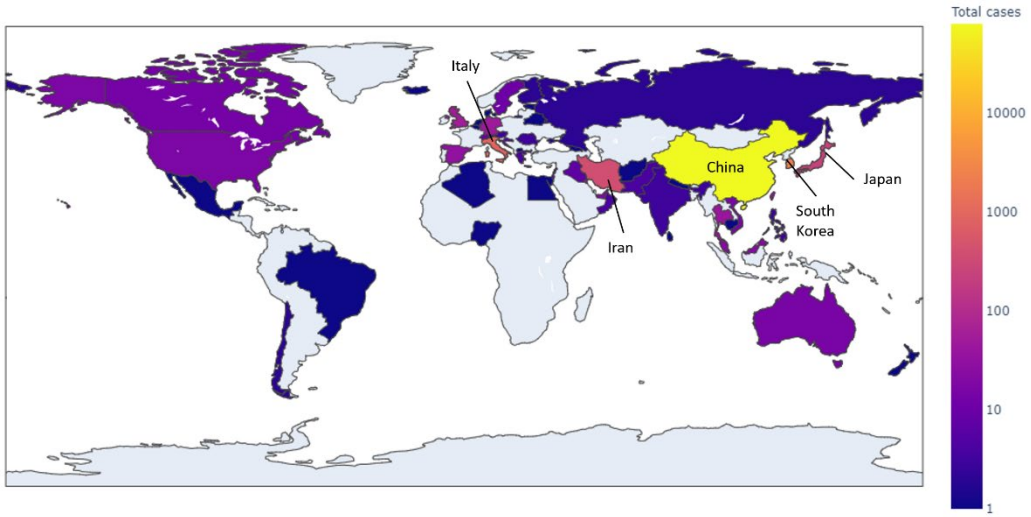

Fig. 6 Countries, territories, or areas with reported confirmed cases of COVID-19, 28 February 2020. Italy, China, South Korea, and Japan were amongst the first few countries to report high number of COVID-19 cases.

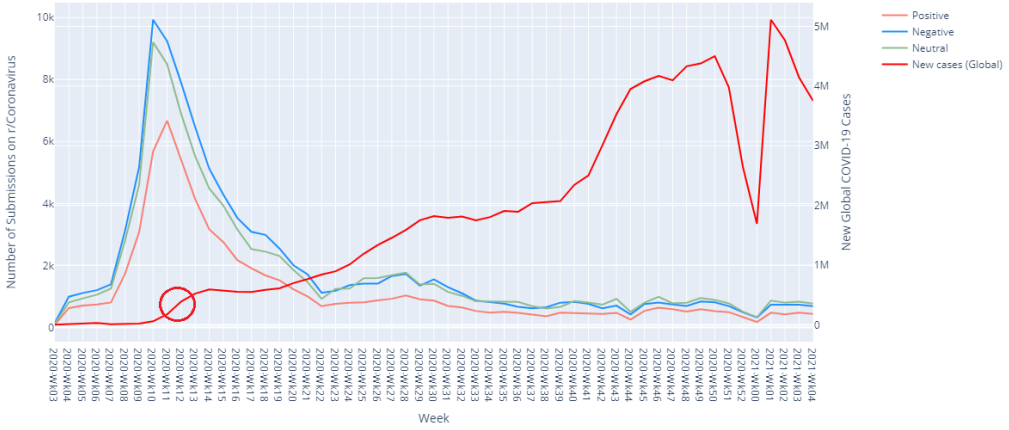

Fig. 7 Weekly count of submissions of varying sentiments. The orange, blue, and green lines represent the trend of positive, negative, and neutral submissions across the year. The red line represents the trend in the number of new reported COVID-19 cases worldwide (Hasell et al., 2020). The peak in submissions aligns with the sudden uptick (red circle) in new reported COVID-19 cases globally, which led to a declaration of a global pandemic in March 2020 by WHO.

To evaluate the extent of the positivity and negativity of the subreddit, the score-weighted sentiment as described in Sect. 3.2.2 is employed. As the votes received by each submission signifies its popularity on the subreddit, the score-weighted sentiment allows us to give submissions with higher number of votes greater significance in its sentiment polarity.

Fig. **8** shows the score-weighted sentiment scores of each submission in every month. It can be observed that in most months, there are more submissions with more extreme negative scores than positive ones. A sample of submissions are shown in Table 8.

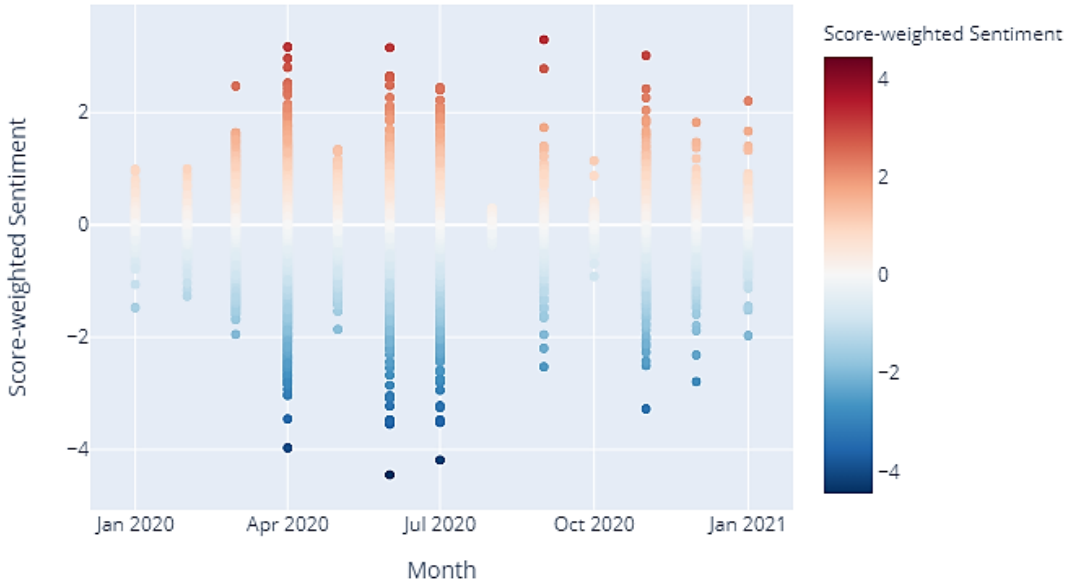

Fig. 8 Score-weighted sentiment score of submissions

Score	Sentiment	Score-weighted sentiment	Title	Month
56125	**Negative**	-4.45	By the end of this weekend more Americans will have died from COVID than were killed in combat during World War One	Jun 2020
58721	**Negative**	-4.19	A COVID positive California woman was instructed to	Jul 2020

			quarantine isolate and inform the people she had been in contact with. Instead she went to work while contagious didn't tell her roommates and infected all people in her household	
25167	**Negative**	-3.97	Charlotte woman hadn't left her house in weeks but tested positive for COVID. Merely touched package delivered by someone who also tested positive	Apr 2020
10796	**Positive**	2.96	Boris Johnson praises immigrant nurses who saved his life	Apr 2020
63562	**Positive**	3.16	The Sikhs Know How to Feed Crowds in a Protest or Pandemic. The Sikh Center of New York has served more than free meals in the last months as part of their faith tradition of feeding anyone in need	Jun 2020
5905	**Positive**	3.17	PlayStation Announces Play at Home Free Games will be given during pandemic worldwide to encourage self-quarantine	Apr 2020

Table 8 Sample of positive and negative submissions with high score-weighted sentiment score.

The word clouds in Fig. 9 and Fig. 10 allow comparisons of positive and negative submissions based on word frequency. In Fig. 9, words such as ‘help’ and ‘vaccine’ are in reference to more resources and an expectation of end to the pandemic or relief from it, hence associated with positive sentiments. In Fig. 10, ‘death’ and ‘positive’ mainly refers to positive COVID-19 cases and COVID-19-related deaths.

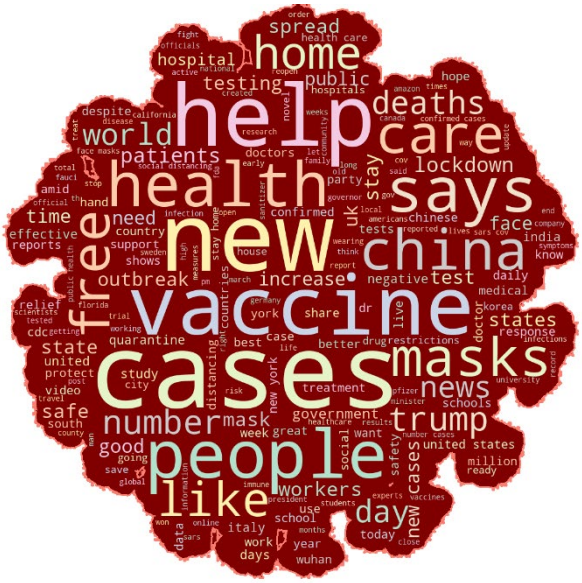

Fig. 9 Word cloud of positive submissions (word frequency)

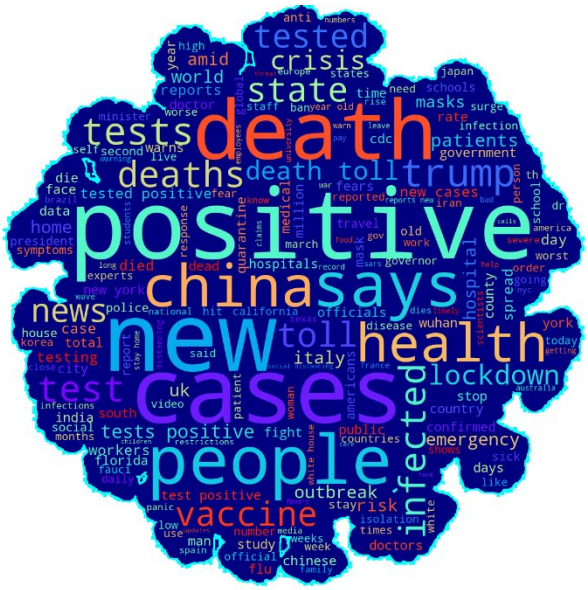

Fig. 10 Word cloud of negative submissions (word frequency)

The differences between positive and negative submissions become more distinct when comparing Fig. 11 and Fig. 12, which are word clouds generated using log odds ratio of tokens in positive and negative submissions, respectively. The phrases “health authority update” and “negative case” being associated positively suggests that they are often used in the context of positive headlines such as negative test results for COVID-19, as well as “safe effective” which refers to vaccines. On the other hand, the distinctive phrases in negative submissions are “total cases dead”, “daily death toll” and “UK death toll”. This strongly indicates that negative news on r/Coronavirus is predominantly on COVID-19 deaths.

Fig. 11 Word cloud of positive submissions (log odds ratio)

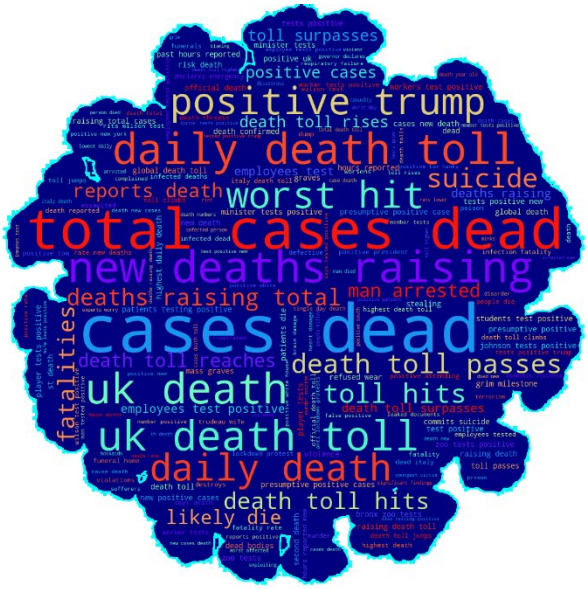

Fig. 12 Word cloud of negative submissions (log odds ratio)

The Kendall Tau correlation scores were also computed between the top *N* most frequently appearing tokens in positive and negative submissions, as shown in Table 9. The low Tau scores indicate that there is little monotonous relation between the ranks of tokens in the positive and negative submissions. Some tokens appeared frequently in both positive and negative submissions (as shown in Fig. 9 and Fig. 10), but as N increases, we observe that Tau score decreases, indicating that there is little correlation between tokens in negative submissions and those in positive submissions.

N	Kendall Tau score
10	0.16735
20	0.028283
30	0.029083
50	0.10908

Table 9 Kendall Tau score of token ranks in positive and negative submissions

4.1.2 Topic Model

The objective of topic modelling was to examine the emergent themes and discourse surrounding COVID-19. Presented in Table 10, 9 topics were identified

based on the model with the highest topic coherence, along with each of their most frequent words. The words were subsequently used to manually label each topic's theme.

Topic	10 most frequent words	**Theme**
1	Case, new, death, report, day, record, high, surge, rise, number	Infection rate and death toll
2	Mask, face, wear, make, governor, people, go, help, get, need	Face masks and other preventive measures
3	News, health, world, live, warn, Fauci, care, country, say, global	Global news
4	State, school, order, public, close, health, county, California, home, open	Government Stay-At-Home orders
5	Hospital, patient, die, doctor, risk, people, child, nurse, drug, man	COVID-19 patients, risks of infection, and drugs
6	Vaccine, study, say, show, CDC, find, scientist, spread, vaccination, SARS-CoV	Spread of COVID-19
7	Test, positive, variant, student, testing, day, school, flight, staff, July	Testing and positive cases
8	Trump, lockdown, reopen, say, travel, quarantine, rule, restriction, White House, president	Lockdown and travel restrictions
9	Vaccine, worker, dose, Pfizer, trial, say, Moderna, medical, receive, get	Medical supplies and vaccines

Table 10 Submission topics identified using LDA model

Python's *pyLDAvis* library also enables the observation of topic segregation through an inter-topic distance map shown in Fig. 13 to examine any possible overlaps. The final model generated has mostly distinctive topics that contain different tokens at various frequencies.

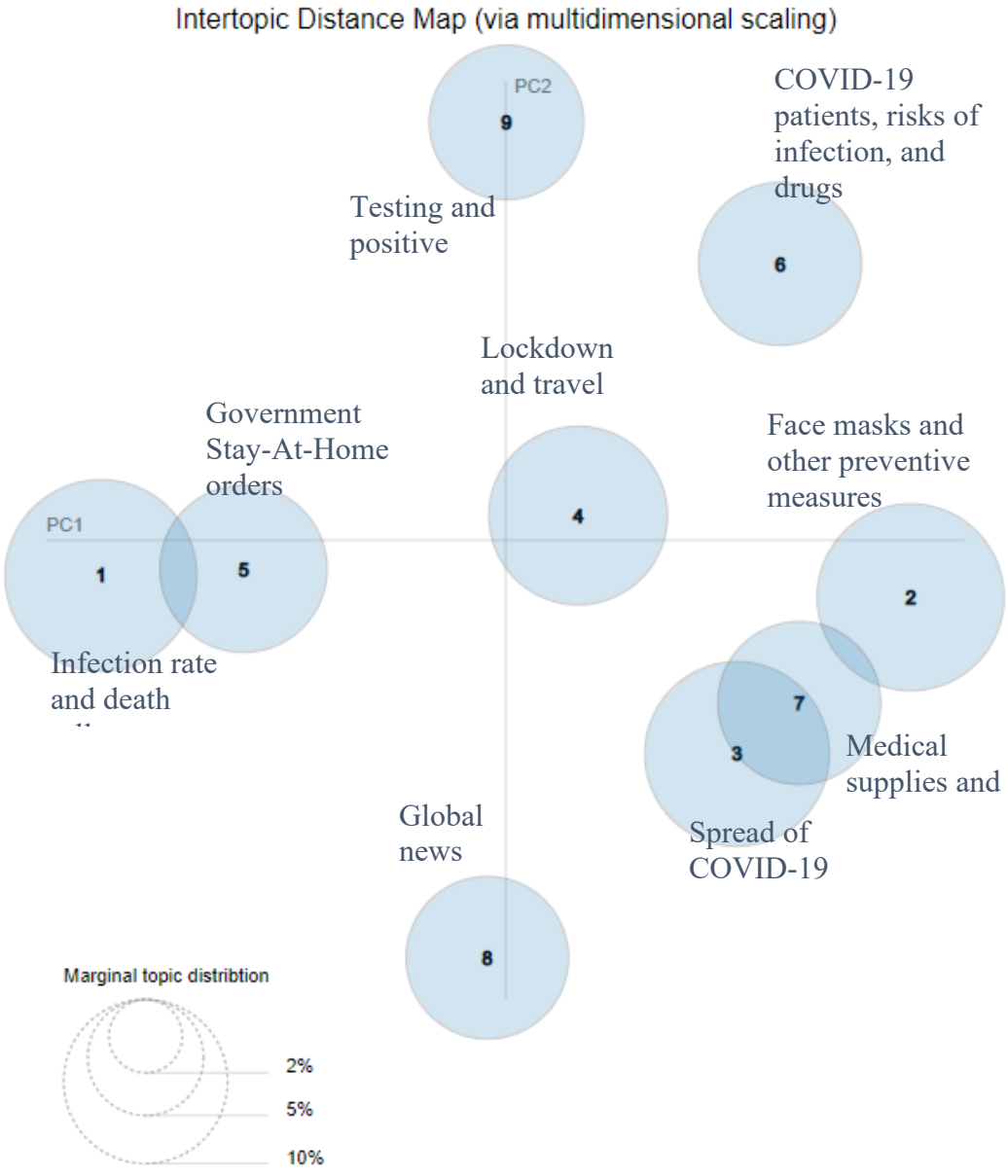

Fig. 13 Visualising the fit of the submissions' LDA model to the submission corpus. Each of the 9 circles represent one topic, whose area is proportional to the proportions of the topics across the number of tokens in the corpus. The circles are labelled in descending order of their areas. The centres of the topic circles are laid out in two dimensions according to a multidimensional scaling algorithm that

is run on the inter-topic distance matrix. The distances between topics are computed using Jensen-Shannon (Dagan et al., 1997) divergence.

The themes identified in our model share some similarities to the 10 themes discovered by Xue et al. (2020) using COVID-19-related tweets published between 23 January 2020 and 7 March 2020. The common themes are highlighted by the same-coloured cells in Table 11.

Current study	Xue et al.
Infection rate and death toll	Updates about the number of COVID-19 cases
Face masks and other preventive measures	COVID-19 related death
Global news	Cases outside China
Government Stay-At-Home orders	Outbreak in South Korea
COVID-19 patients, risks of infection, and drugs	Early signs of the outbreak in New York city
Spread of COVID-19	Diamond princess cruise
Testing and positive cases	Economic impact
Lockdown and travel restrictions	Preventive measures
Medical supplies and vaccines	Authorities
	Supply chain

Table 11 Comparison of themes with the study by Xue et al. (2020)

Despite the difference in volume of data (356,690 submissions on r/Coronavirus compared to 1.9M tweets), it can be observed that the submissions encompass a more diverse set of themes as the dataset was collected over the duration of a year. The informative nature of submissions on the subreddit also meant that there was more content associated with the science behind COVID-19 and its vaccines.

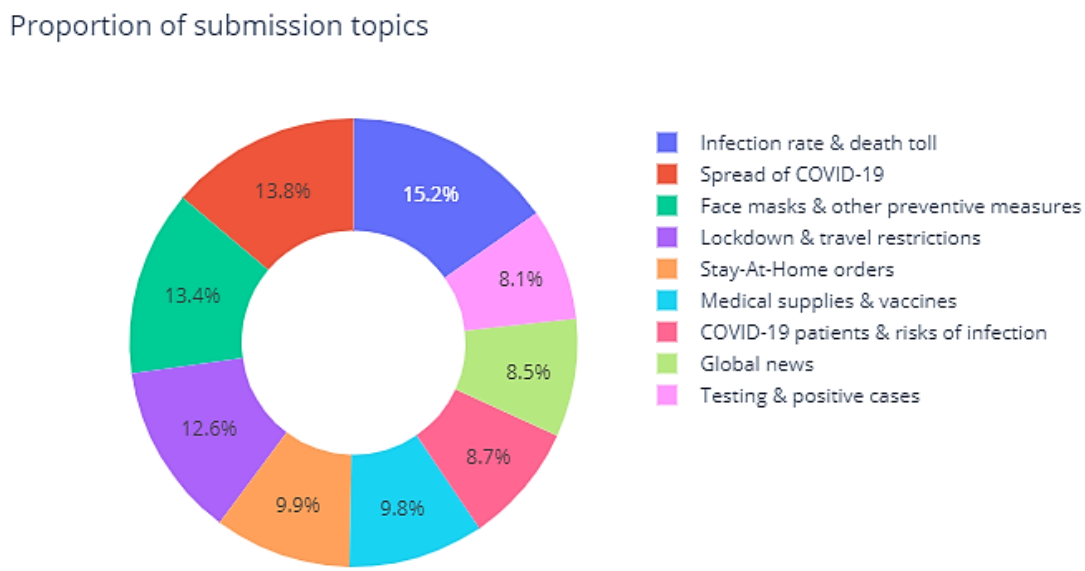

Fig. 14 Proportion of submission topics in a year

The proportions of topics in the submissions are presented in Fig. 14. It reveals that the most dominant theme of all submissions is "infection rate and death toll of COVID-19". An example of a submission of this topic is "New York state reports 3,832 new cases of coronavirus and 58 new deaths, adding to the 2,461 new cases and 67 new deaths reported 1 hour ago". This suggests that the r/Coronavirus subreddit as a community monitoring regularly the latest news on new cases and deaths. The second most dominant theme is "spread of COVID-19", for which a representative example is "University of Hong Kong study finds eyes are 'important route' for coronavirus, up to 100 times more infectious than SARS". This topic relates to the new studies and discoveries on COVID-19 with regards to its transmission, occasionally in comparison to other flus such as SARS. The third most dominant topic amongst submissions is "face masks and other preventive measures". This topic may be popular as the wearing of face masks was at the centre of political divide in the United States (McKelvey, 2020), therefore, many submissions were intended to educate and inform the community. For example, a submission was titled "Masks don't merely help to protect the wearer, they limit the amount of virus exposure, which in turn limits the severity of symptoms should the wearer get sick".

At the start of 2020 when COVID-19 first emerged, most submissions were related to "face masks & other preventive measures", as illustrated in Fig.

15. This suggests that a majority of submissions in the early phase of COVID-19 was about staying safe as individuals. Submissions of this category decreased in ratio as the situation evolved over time.

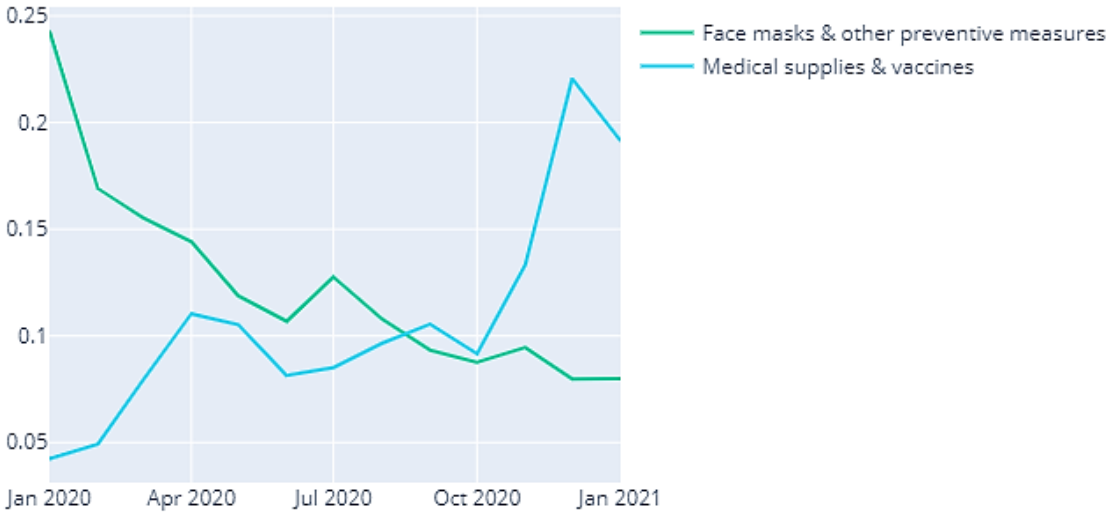

Fig. 15 Fraction of submissions related to "Face masks & other preventive measures" and “Medical supplies & vaccines”

The development of vaccines, along with the availability of medical supplies, was a prominent topic in the submission dataset as well. Presented in Fig. 15, it rose quickly in the last quarter of 2020 to become the most dominant topic on the subreddit. This trend coincides with the moment when the first COVID-19 vaccine was approved by the U.S. Food and Drug Administration on 14 December 2020 (FDA, 2020). It can therefore be inferred that the dominant subject in submissions evolved from preventive measures to vaccinations in the span of one year.

The full graphs on each topic’s distribution over time is shown in Fig. 26 in the appendix.

4.2 Comments

4.2.1 Sentiment Analysis with VADER

Fig. **16** presents the overall sentiment distribution of the subreddit's comments in the one-year period. The number of comments follows a trend similar to the number of submissions as shown in Fig. 7, likely due to the chain of events described in Sect. 4.1.1. Similarly, comment activity was the highest in the month of March when COVID-19 was officially classified a pandemic. Nearly one-third of comments in the dataset were from March.

However, unlike the sentiment trend seen in submissions, the level of negative and positive comments remained consistently balanced. The number of positive comments actually exceeded that of negative comments in March, while the opposite was observed in the submission dataset. This indicates that even in the shadow of negative news on COVID-19, the community remained positive. This is consistent with the findings in other studies (Bhat et al., 2020; Yin et al., 2020) where it was found that people remained hopeful that government restrictions and proper personal hygiene measures would end the pandemic.

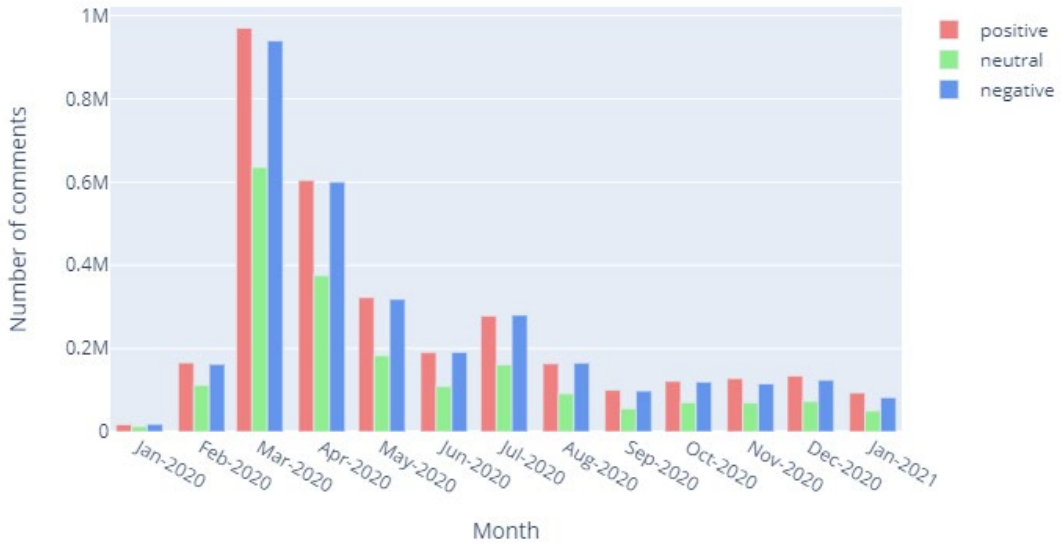

Fig. 16 Sentiment polarity of comments on r/Coronavirus

To get an overview of the sentiments and their corresponding words, a word cloud was generated for each sentiment. The size of the words in each word

cloud is proportional to its log odds ratio, which indicates their regularity of appearance in positive comments relative to negative ones.

Fig. **17** is a word cloud containing words with a high log odds ratio. Many of the tokens that stood out are related to the mild symptoms of COVID-19, such as coughing and sneezing in “cough_sneeze”, as well as previous viruses such as the Spanish flu and the swine flu. On further inspection, it suggests that the comments containing “spanish flu” and “swine flu” were often assuring other people that there were pandemics in the past, and that the situation would return to normal eventually. Users in the community reassured one another that showing mild symptoms such as coughs were not conclusive of being infected. They also encouraged others to wear masks to cover their mouth and nose. The phrase “icu bed” is also associated positively when referring to sufficient medical resources. On the other hand, comments containing “spring break” are classified positive as they mostly described fun activities common during spring break, yet in reality, they were viewed negatively by the commenters. Some representative comments are shown in Table 12.

Fig. 17 Word cloud for comments with positive sentiments

Token	Comment
‘icu_bed’	“…They have changed that now and are sending the mild cases into the arenas where beds have been added while trying to keep the **icu beds** only for critical patients”
‘spring_break’	“I'd imagine all the bars are open no problem with groups of people gathering in the parks citizens relaxing and enjoying their lattes while browsing the morning news the beaches full of youngsters enjoying **spring break**”
‘spanish_flu	“Slowing is good. You don't need a vaccine. All you need is better treatment options. And again the goal isn't to eradicate the virus. It's to eliminate the spread locally. This was all done with the **spanish flu** as well. This is nothing new. Mankind has fought against epidemics for centuries”
‘swine_flu’	“…it could get better with wide containment south korea style or china style where the cases are known and we can reopen things again. It could get better like **swine flu** did where the virus mutates to be less lethal, it could get better with herd immunity, it could get better by us identifying good treatment options…”
‘cough_sneeze’	“…please start from ourselves by giving more understanding to people around us when we are out if they **sneeze** or **cough** abit don't panic and don't be aggressive towards them. A time like this people are easily anxious and it turns into frustrations very quickly. They have those behaviours doesn't mean they have the virus…”

Table 12 Sample of positive comments containing specified tokens

Fig. **18** visualises the word cloud of words that appear more frequently in comments with a negative sentiment score. The biggest token is “wash_hand”, which is commonly encouraged as a way to prevent being infected with the virus. The phrase “wild animal” is commonly associated in negative comments as the source of COVID-19, while pre-existing conditions (“pre_exist) are generally

linked to COVID-19-related deaths. Lastly, “dry_cough” which to the COVID-19 symptom of dry cough is also viewed negatively. Contrasting this with “cough_sneeze” in Fig. 17 which is associated more positively, dry cough is commonly mentioned as a more severe effect of COVID-19. It is also often associated with other serious symptoms such as pneumonia, therefore it is seen in a highly negative light. Some representative comments are shown in Table 13.

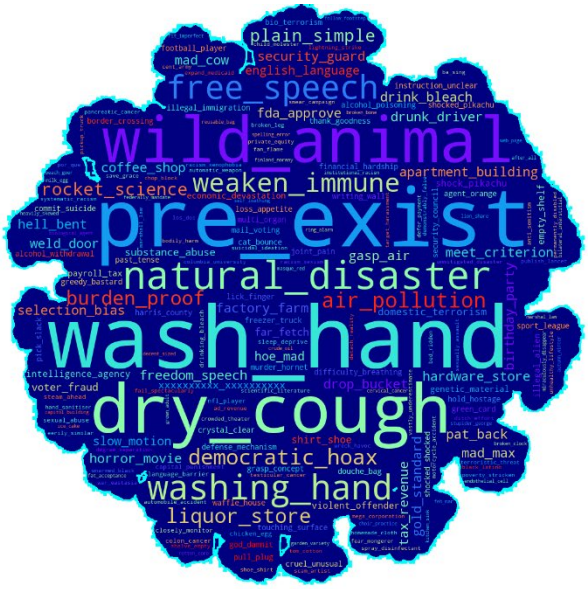

Fig. 18 Word cloud for comments with negative sentiments

Token	Comment
‘wash_hand’	“The amount of people saying this is just fear mongering a new virus every year to scare people and distract them and now to blame Trump is scary… I posted basically telling friends and family to be aware of what’s going on to buy sanitizer and apparently TP and to **wash hands** and not touch their face…”
‘pre_exist’	“…The disease infects a ton of people but most of the diseased are old or frail and with **pre existing** conditions. How is that a conspiracy it's true. If young people die that doesn't mean the virus is completely deadly as many young people also die due to many other diseases…”

'dry_cough'	"I was one of those people I had a severe respiratory illness that was textbook covid I have never had pneumonia in my life or had that much difficulty breathing or feel that kind of truck run over me. Never lost sense of smell or taste. But those severe symptoms lasted a few days It was the **dry cough** that killed me for weeks"
'wild_animal'	"…COVID most likely originated from bad hygiene practices at markets where they treat the poor **wild animals** abominably and put them next to domesticated animals in dirty horrid conditions with no biosecurity also Pangolins the most trafficked animal in the world…"
'natural_disaster'	"…That doesn't mean this isn't a natural disaster. Viruses are natural. The disaster caused by it is a **natural disaster** regardless of if it could have been mitigated"

Table 13 Sample of negative comments containing specified tokens

The Kendall Tau scores were also computed between the top *N* most frequently appearing tokens in positive and negative comments, as shown in Table 14. The relatively low scores indicate that distinctiveness of the phrases carrying positive or negative connotations, even as there is a lot of shared vocabulary across such posts (which is why the score is not too low).

N	Kendall Tau score
10	0.49388
20	0.46061
30	0.62058
50	0.58233

Table 14 Kendall Tau score of token ranks in positive and negative comments

4.2.2 Analysis of Upvoted and Downvoted Comments

650,753 upvoted comments and 151,199 downvoted comments were analysed for their topics. This subset of comments was selected as their score were deemed more representative of the subreddit.

Fig. 19 illustrates the word cloud based on the log odds ratio of tokens in upvoted comments. Comments that contained “spring break”, “shake hand” and derided people for flouting social distancing rules were widely upvoted by the community. The token “intensive_care” generally appears in comments discussing the appropriate allocation of medical resources, while “false_negative” appears in discussions on COVID-19 tests that gave false negative results, for example: “It is unlikely a new infection would show symptoms that quickly. That strongly suggests this is an existing infection that lingered and showed a false negative.” Among comments that contain the bigram “false negatives”, there were 50% more comments with negative sentiments than positive ones. Given that COVID-19 tests showed risks of producing false negatives (Kanji et al., 2021), this suggests that false negative cases were a topic of concern on the subreddit.

Fig. 19 Word cloud for upvoted comments

On the other hand, Fig. 20 presents the word cloud for downvoted comments. The most prominent tokens are “long_term” and “fake_news”. The token “long_term” appears in many downvoted comments that were unsupportive of COVID-19 restrictions, citing reasons such as “long term unemployment” or debating that the restrictions do not work “long term”. Similarly, comments that mentioned “mental health” as a reason to disregard restrictions were also downvoted by the community. Additionally, downvoted comments that included “nursing home” consisted of comments minimising COVID-19-related deaths in nursing homes in the United States, some of which accused New York State Governor Andrew Cuomo of “sending COVID patients into nursing homes”. This was a point of controversy as, at some time point, one-third of COVID-19-related deaths in the United States were nursing home employees or residents (Yourish et al., 2020).

Many comments that claimed information as “fake news” were also unpopular, many of which were mistrusting towards science or the media.

Fig. 20 Word cloud for downvoted comments

Comparing the attitudes adopted in upvoted and downvoted comments, it can be observed that majority of the r/Coronavirus community treated COVID-19 restrictions and recommendations with great importance, even being cautious of false negative results that might hint at undetected cases. The community largely did not tolerate the underestimation of COVID-19’s impacts and the undermining of government orders, as well as the outright dismissal of information using phrases such as “fake news”.

4.2.3 Topic Model

The comment dataset was used to train the LDA model. 20 topics were identified based on the model with the highest topic coherence and these are presented in Table 15, along with the most frequent associated words. The words were subsequently used to manually determine and label each topic's theme. Examples of comments are shown in Table 16 in the appendix.

Topic	10 most frequent words	**Theme**
1	Pfizer, food, buy, restaurant, order, supply, eat, store, go, car	Food and supplies
2	People, f***, say, trump, s***, think, want, get, just, right	Angry and hateful comments
3	Virus, vaccinate, people, flu, spread, population, get, year, variant, infection	Comparison of COVID-19 with other strains of flu/viruses
4	Mask, wear, people, hand, face, store, grocery, work, air, use	Personal preventive measures
5	Government, country, china, world, American, America, federal, pandemic, lie, blame	Handling of COVID-19 by US and China
6	People, thing, think, good, go, will, bad, know, well, feel	Negative outlook towards the situation (fear, pessimism, etc)
7	State, public, health, say, new, follow, Florida, law, pandemic, official	Government response
8	Say, article, read, news, post, datum, study, comment, know, source	Scientific studies and sources
9	Case, death, number, day, rate, week, high, report, new, march	Infection and death rates
10	Make, sense, pretty sure, play, know, see, big, hard, difference	Miscellaneous topics

11	People, risk, take, die, go, get, want, life, live, think	Putting lives at risk
12	Home, school, kid, work, family, go, stay, friend, get, year	Staying at home
13	Human, virus, cell, body, transmission, animal, change, host, nature, protein	Virus transmission from animals
14	Test, positive, symptom, result, get, question, testing, negative, people, answer	Accuracy of tests
15	Country, lockdown, travel, state, restriction, city, area, place, people, population	Travel restrictions and lockdown
16	Hospital, patient, medical, care, doctor, nurse, worker, healthcare, staff, capacity	Shortage of medical resources
17	Old, age, young, year, term, long, dose, cause, effect, people	Long-term effects
18	Day, month, week, time, dose, get, go, year, long, take	Measure of time
19	Pay, work, money, job, company, business, worker, government, cost, need	Economy, jobs, and income
20	Vaccine, trial, vaccination, effective, work, approve, new, efficacy, use, available	Effectiveness and availability of vaccines

Table 15 Comment topics identified using LDA model

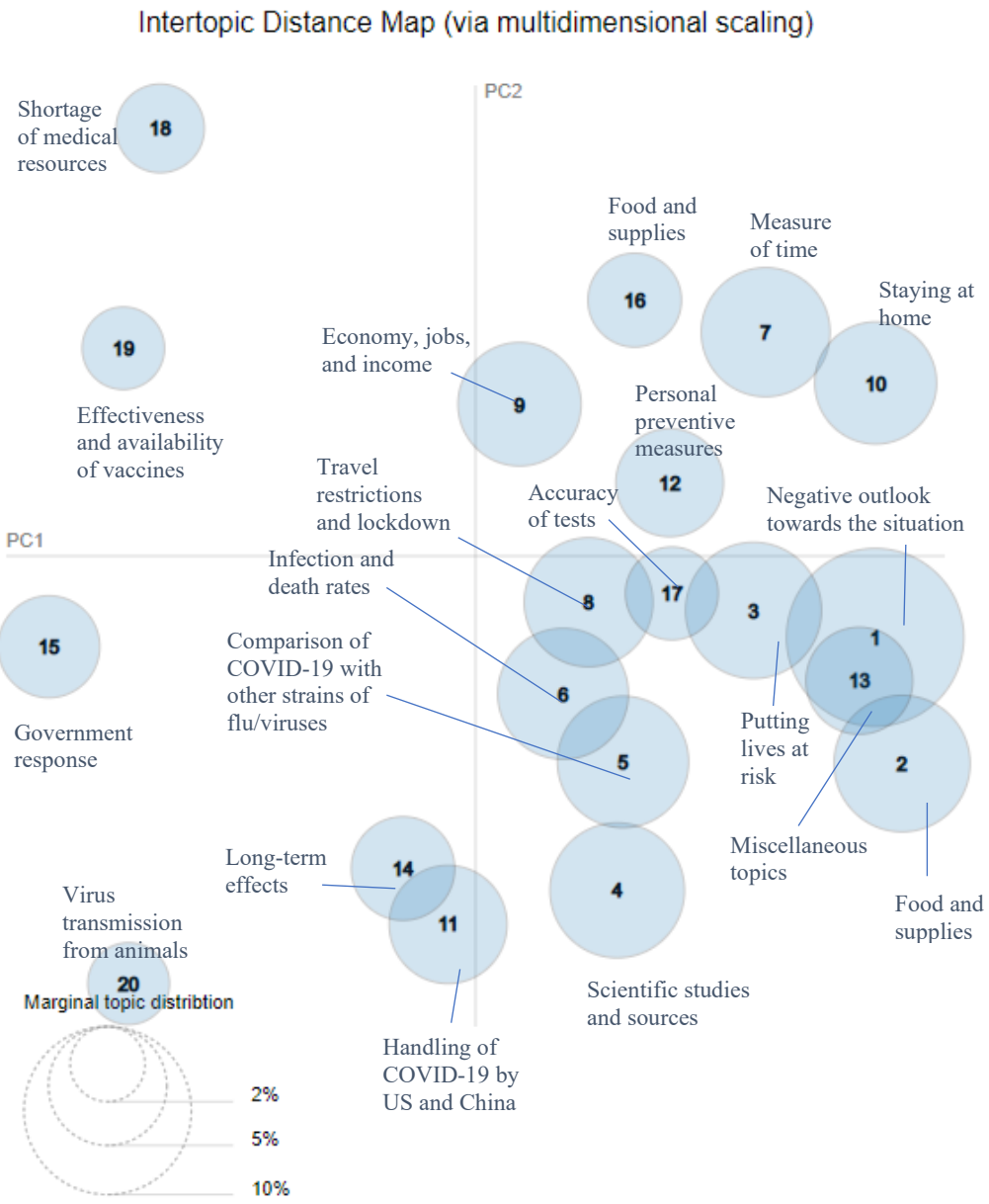

Fig. 21 Visualising the fit of the comments' LDA model to the comment corpus. Each of the 20 circles represent one topic, whose areas are proportional to the proportions of the topics across the number of tokens in the corpus. The circles are labelled in descending order of their areas. The centres of the topic circles are laid out in two dimensions according to a multidimensional scaling algorithm that is run on the inter-topic distance matrix. The distances between topics are computed using Jensen-Shannon (Dagan et al. 1997) divergence, a measure based on (dis)similarity between probability distributions.

Fig. **21** illustrates the inter-topic distance of the LDA model containing 20 topics and it can be observed that the topics are generally distinct from one another.

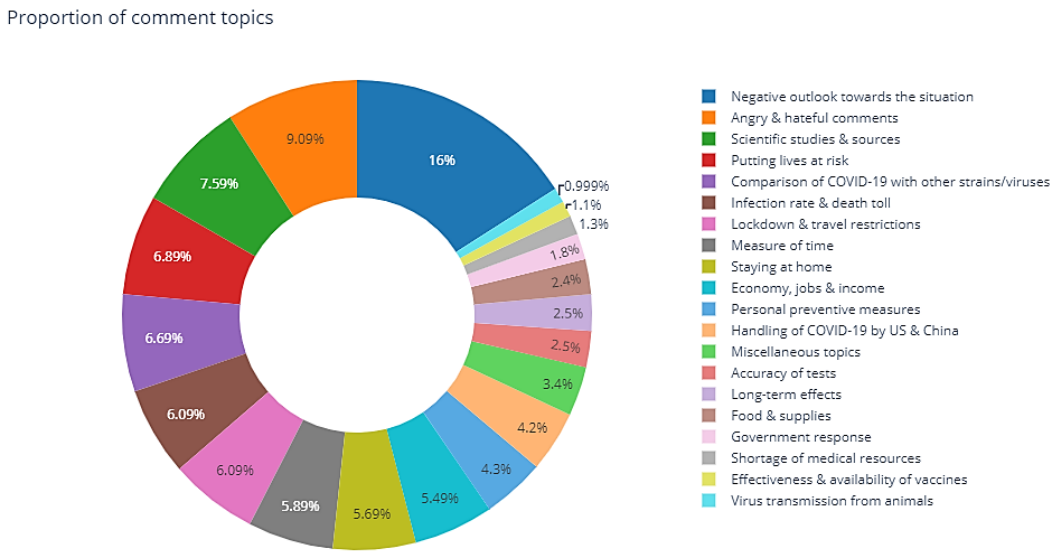

Fig. 22 Proportion of comment topics across the year

Fig. 22 Proportion of comment topics across the year charts the percentage of occurrence of each topic in the comments on r/Coronavirus. Through the LDA model, the topics have provided insight into the discussion themes occurring during the span of the year.

The most notable themes were as follows: (6) Negative outlook towards the situation, (2) Angry and hateful comments, (8) Scientific studies and sources. The first one largely reflects people's worry and scepticism towards the pandemic, commonly believing that the situation would become worse. This can be inferred from the tokens shown in Fig. 23.

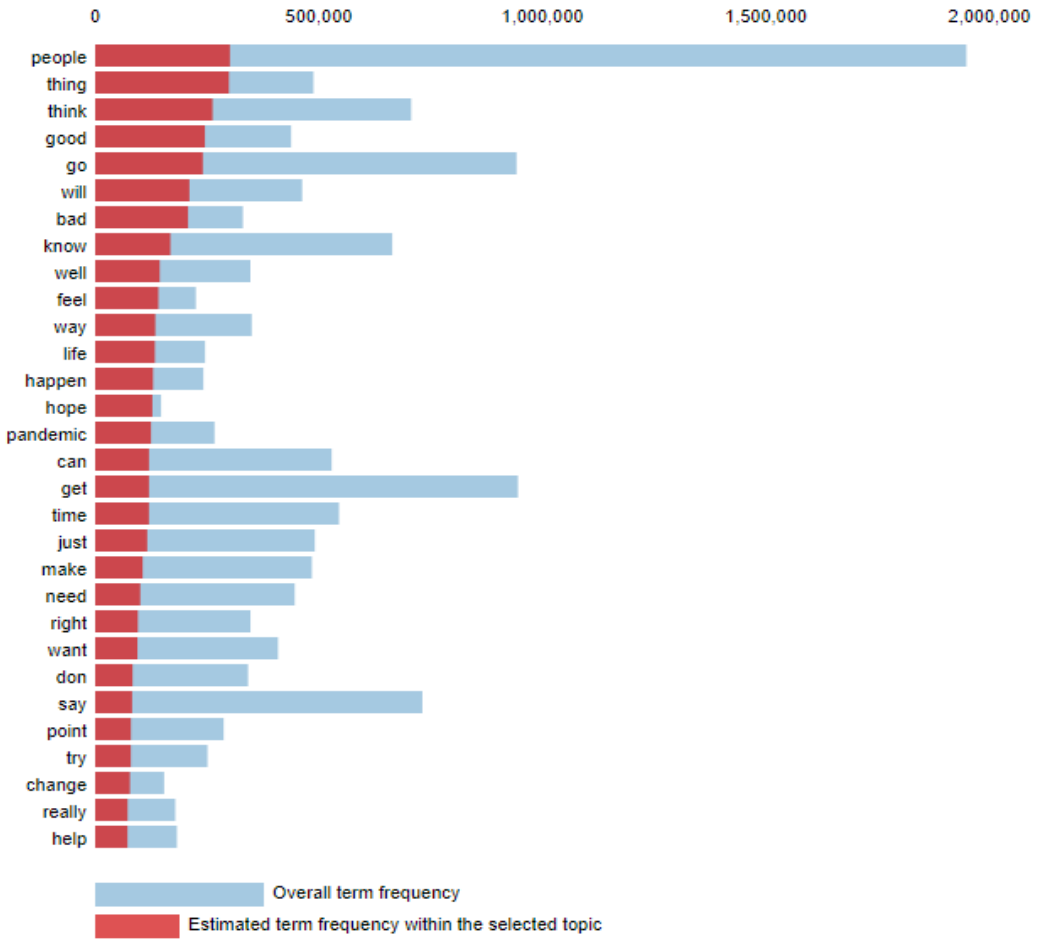

Fig. 23 Top 30 most relevant terms for Topic (6) Negative outlook towards the situation

The second most dominant theme falls under “Angry and hateful comments”, which is characterised by tokens that include “hate” and “dumb”, amongst many other expletives. These were intentionally left in the text as they represented an outburst in human emotions such as frustration, anger, and contempt. These terms are also unique to the comment dataset due to its conversational nature. The high prevalence of this theme suggests that COVID-19 was a divisive subject that invited uncivilised discourse online, and amplified political and other societal divides.

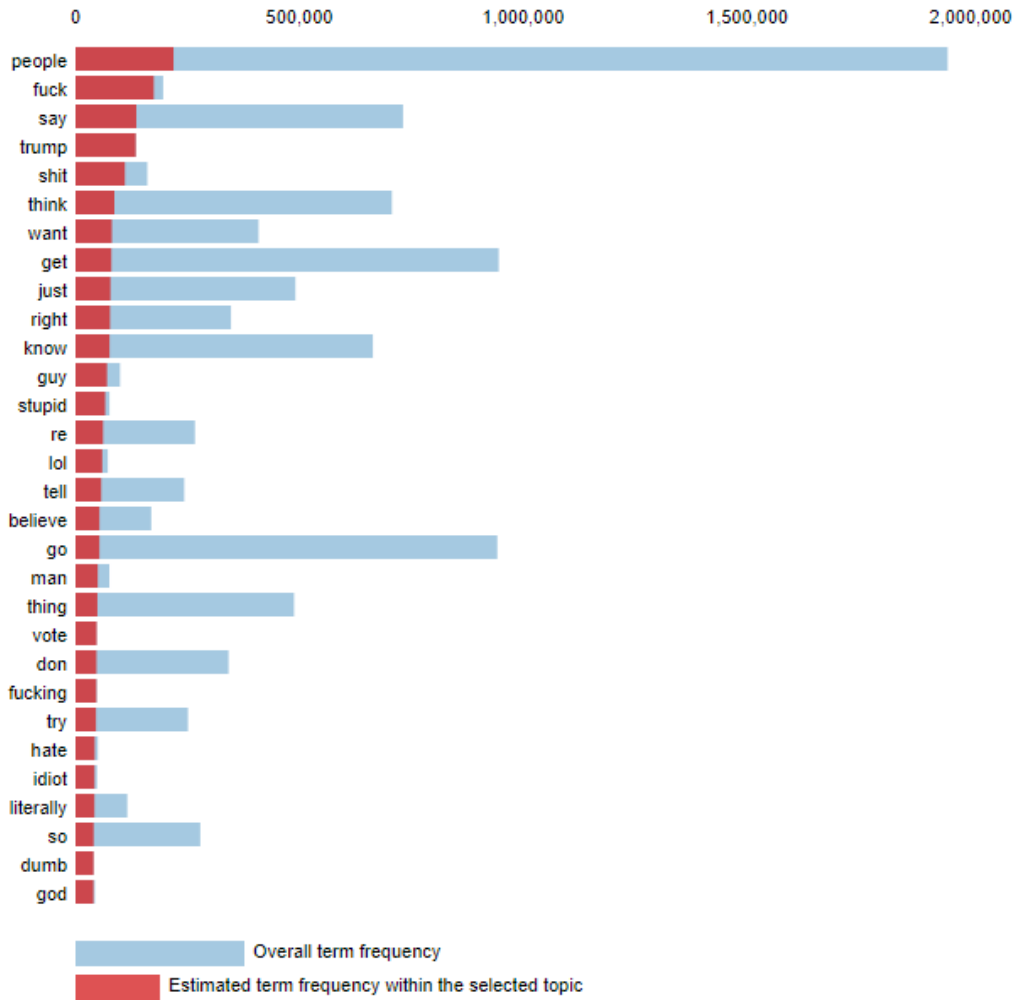

Fig. 24 Top 30 most relevant terms for Topic (2) Angry and hateful comments

The third largest theme is “Scientific studies and sources”. Fig. **25** shows that the relevant terms are “article”, “news”, “datum”, “study”, etc, which clearly implies that the sources of information were a widely discussed topic on r/Coronavirus and is an important factor as to whether a piece of information should be trusted. This also suggests that the community is cautious and critical towards the accuracy of any claims or posts that appeared on the subreddit.

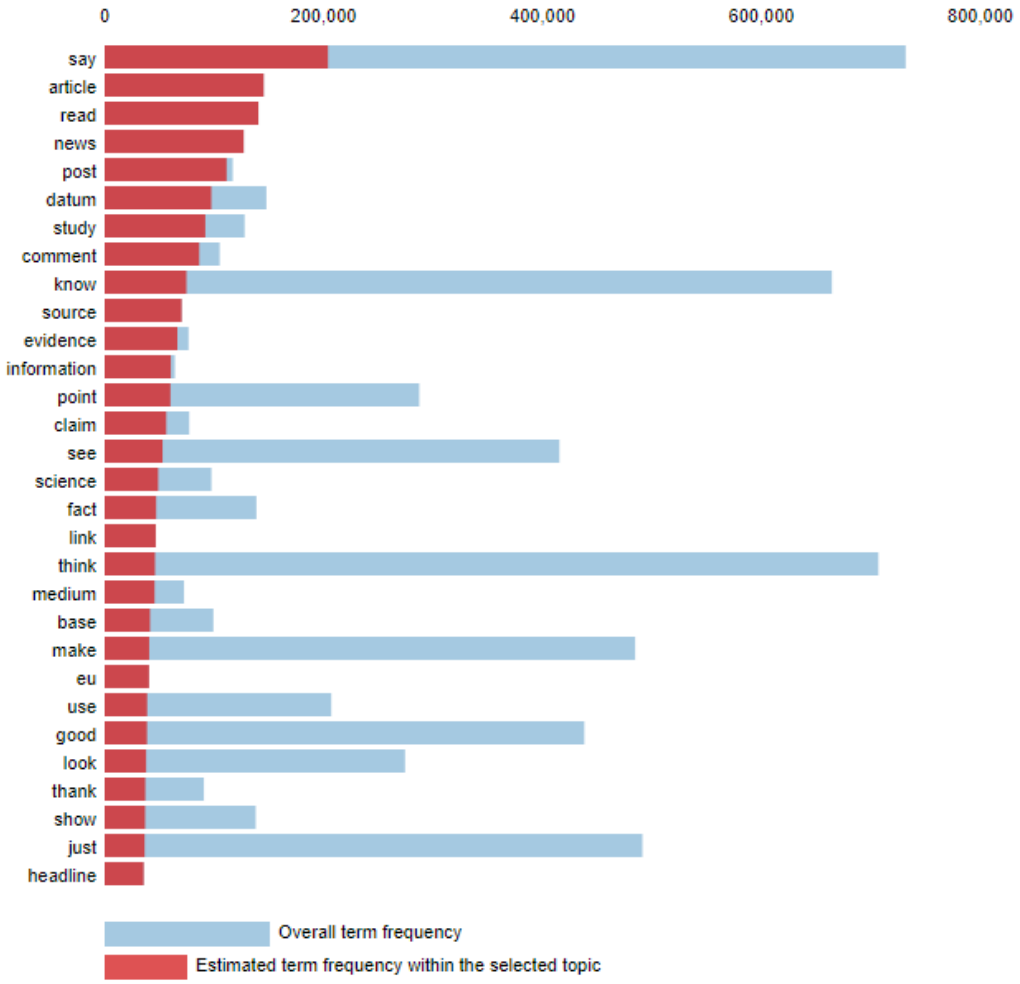

Fig. 25 Top 30 most relevant terms for Topic (8) Scientific studies and sources

Topic distribution was assessed at each quarter of the year to capture the changes in topic prevalence. Inspection of the topic distribution across all timeframes indicated that there was no significant shift in any topics that aligned with significant events revolving around COVID-19. The graphs on each topic's distribution over time is shown in Fig. 27 in the appendix.

4.3 Comparison of Submissions and Comments

4.3.1 Sentiments

The sentiment polarity of each submission is greatly dependent on the nature of its subject matter. For example, news about vaccines and medical aid are often positive while news about positive COVID-19 cases and COVID-19 death toll are negative. As the sources of the submissions are generally from news media and healthcare authorities, the subjects become quite one-dimensional. In contrast, Reddit comments are discussion-based and conversational in nature. Hence it is unsurprising to see that the topics found in comments with positive or negative sentiment are more diverse, ranging from flu symptoms to pre-existing medical conditions and wild animals. The vote feature on Reddit also allows the identification of popular and unpopular comments, and consequently, the popularity of the associated themes. Evaluation of these polarised comments revealed that both popular and unpopular comments carried negative sentiments towards a diverse set of topics, such as spring break (popular) and fake news (unpopular). It is evident that comments offer deeper insights into people's perceptions and sentiments towards topics that are more societal in nature, while submissions tend to be confined to a smaller set of themes, which are often more scientific and authoritative.

4.3.2 Topics

The sets of topics generated from the submission dataset and the comment dataset share a small degree of overlap, such as "face masks and other preventive measures", "infection rate and death toll", and "travel restrictions and lockdown". However, as noted previously, comments are discussion-based and tend to relate to personal concerns and experiences, therefore the topics found in comments are largely different and much more diverse than those found in submissions. For example, the most dominant topic in comments is about people's outlook towards the COVID-19 situation. This is followed by emotional debates between users on the subreddit, while the third is associated with comments discerning news sources. Given these observations, it can be noted that submission topics represent

the category of news related to COVID-19 across the year, while comment topics correspond to the list of concerns that the public had.

5 CONCLUSION

5.1 Principal Findings and Practical Implications

This study examined the themes that have emerged around the central topic of COVID-19 through the examination of sentiments in submissions and comments, as well as the popular and controversial views that emerged in the form of upvotes and downvotes. From the submissions, it was noted that death tolls dominated the negative news and authority updates dominated the positive ones as they were perceived as progress, or hope thereof. The topics extracted from the submission data also clearly shows that information on the transmission of COVID-19 and the proper preventive measures were highly reported on r/Coronavirus. Vaccine-related news became the most prevalent topic at the end of 2020. These reflected the dominant focus and narratives in the mainstream media.

On the other hand, the comments presented a more vivid understanding of the sentiments and concerns of the populace at large, and our analysis captures a varied picture, in which the data better revealed a divide in attitudes towards COVID-19. Analysis of the sentiments showed that there were two main groups of people on the subreddit: one that was cautious and vigilant of COVID-19, contrasting against another that was dismissive and sceptical about its severity. The comment section of Reddit provided a platform for expression of ideas and support for each of these groups. In spite of the pervasiveness of misinformation, the community remained prudent with the news they received, as reflected from the upvotes and downvotes.

This study has demonstrated that Reddit can be employed to investigate levels of public awareness and sentiments which follow real-time events of the pandemic. The government and authorities can use these to understand the common misinformation, talking points and misunderstandings – and explore various channels to proactively advise and educate the public to fight against “fake news”. Likewise, they can also use such a forum to understand people's expectations and anxieties, and respond to these in a prompter manner. Timely intervention in public discourse may also help limit the politicisation of the virus

to prevent mixed messaging and quell the public's concerns with authoritative information and rigorous scientific studies.

5.2 Limitations and future work

In this work we study the social media discourse regarding an entire year, in contrast to most prior works which were restricted to shorter time window. The nature of the platform (with no length limit for most practical purposes, and pseudo-anonymity) also provided us access to in-depth and candid discussions. Nevertheless, there are several limitations to this study, that should be addressed in future by the broader research community, including ourselves.

Firstly, this study only focused on submissions and comments of r/Coronavirus. While it is the largest COVID-19-related community on Reddit, there are other subreddits that contained COVID-19 news and may provide additional venues of data. The methods used in this study can be extended to geographical-based subreddits to accurately mine sentiments of the corresponding demographic since government restrictions varied in time across geographic areas.

This leads to the second point, which is that research data may be expanded to cover sources of other languages to gain broader insights, particularly in a global pandemic unbounded by borders.

Furthermore, even though multiple vaccines have emerged, they vary in their efficacy – and this concern is particularly aggravated by the advent of mutated versions of the virus itself. Moreover, there is a huge extent of heterogeneity and latency in rolling out the vaccine across different countries. As such, there have been major new waves in many countries beyond the time frame studied in this work. While our methodology would apply, the concerns and thus the themes could be different during the second year of the pandemic. We provide the data curated in this work (https://doi.org/10.21979/N9/0LGZYN), so that future studies can pick-up where this study terminated, and can readily use our data for extension or comparison.

Lastly, although the natural language processing tools we have used, e.g., LDA model, have shown to be effective in topic extraction, one may want to apply other tools to gain further or more precise insight. E.g., the scientific quality

and consistency of the themes may be further improved to reduce human bias when labelling topic themes. Future studies may also perform analysis on the secondary themes in the documents for a finer granularity of details.

6 REFERENCES

Bhat M, Qadri M, Beg N-A, et al (2020) Sentiment analysis of social media response on the Covid19 outbreak. Brain Behav Immun 87:136–137. https://doi.org/10.1016/j.bbi.2020.05.006

Bonta V, Kumaresh N, .N J (2019) A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. 1–6

Boon-Itt S, Skunkan Y (2020) Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. JMIR Public Health Surveill 6:e21978. https://doi.org/10.2196/21978

Cucinotta D, Vanelli M (2020) WHO Declares COVID-19 a Pandemic. Acta Biomed 91:157–160. https://doi.org/10.23750/abm.v91i1.9397

Dagan I, Lee L, Pereira F (1997) Similarity-Based Methods For Word Sense Disambiguation. arXiv:cmp-lg/9708010

FDA (2020) FDA Takes Key Action in Fight Against COVID-19 By Issuing Emergency Use Authorization for First COVID-19 Vaccine. In: FDA. https://www.fda.gov/news-events/press-announcements/fda-takes-key-action-fight-against-covid-19-issuing-emergency-use-authorization-first-covid-19. Accessed 21 Mar 2021

Hasell J, Mathieu E, Beltekian D, et al (2020) A cross-country database of COVID-19 testing. Sci Data 7:345. https://doi.org/10.1038/s41597-020-00688-8

Hutto CJ, Gilbert E (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: Eighth International AAAI Conference on Weblogs and Social Media

Kanji JN, Zelyas N, MacDonald C, et al (2021) False negative rate of COVID-19 PCR testing: a discordant testing analysis. Virology Journal 18:13. https://doi.org/10.1186/s12985-021-01489-0

Kastrenakes J (2020) Reddit reveals daily active user count for the first time: 52 million. In: The Verge. https://www.theverge.com/2020/12/1/21754984/reddit-dau-daily-users-revealed. Accessed 14 May 2021

Low DM, Rumker L, Talkar T, et al (2020) Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: an observational study. PsyArXiv

McKelvey T (2020) Coronavirus: Why are Americans so angry about masks? BBC News

Newman N (2020) Reuters Institute Digital News Report 2020. Reuters Institute 112

Perez S (2018) Twitter's doubling of character count from 140 to 280 had little impact on length of tweets. In: TechCrunch. https://social.techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/. Accessed 20 Mar 2021

Röder M, Both A, Hinneburg A (2015) Exploring the Space of Topic Coherence Measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, Shanghai China, pp 399–408

Taylor DB (2021) A Timeline of the Coronavirus Pandemic. The New York Times

Xue J, Chen J, Chen C, et al (2020) Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PLOS ONE 15:e0239441. https://doi.org/10.1371/journal.pone.0239441

Yin H, Yang S, Li J (2020) Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media. arXiv:200702304 [cs]

Yourish K, Lai KR, Ivory D, Smith M (2020) One-third of all US coronavirus deaths are nursing home residents or workers. New York Times 5:

Appendix 1 Submission Topic Distribution across the Year

Fig. 26 Submission topics distribution across the year

Appendix 2 Comment Examples and Themes

Theme	Comment
Food and supplies	Our local Walmart is out of just about everything but if you go to any other store there are aisles filled with stuff. Have you checked everywhere? Even local restaurants are selling stuff like flour sugar and eggs.
Angry and hateful comments	Then blame shitty media not WHO when they're literally doing everything right and still getting blame. Would you prefer they just lie to everyone not say anything about this and then get called out for that further down the line and still get a load of people blaming them for everything that happened? They actually can't win right now can they.
Comparison of COVID-19 with other strains of flu/viruses	There are currently no vaccines against coronaviruses. You must be thinking of the flu shot. It's not for a coronavirus it's for an influenza virus It's a different virus. For the common cold there is no vaccine and it's not actually caused by one specific virus. Most of them are caused by rhinoviruses but also some coronaviruses cause it as well. There are different viruses that can cause common cold. So even if you get a cold it's not a certainty you will get any kind of immunity against COVID because it's possible your illness was not even caused by a coronavirus but by something completely different.
Personal preventive measures	I heard that surgical masks are still more effective than cloth masks in protecting the one wearing the mask and that putting tape around the mask to create a seal provide more protection. Wearing a face shield over the mask can also help protect you better.

Handling of COVID-19 by US and China	Don't trust what you see about China as truth. Everything is censored by the Chinese government before any of us sees anything. It's basically a facade to give the impression the China is doing well. That's why the world is in this big mess the Chinese tried to cover up their virus and pretended it didn't exist.
Negative outlook towards the situation (fear, pessimism, etc)	I tell people that I know personally that are not taking this as seriously that it's going to get real very quickly when you know someone who knows someone with the virus. Then the actual reality will set in when you know someone with it. Now I'm not all doom and gloom about the situation but let s stop pretending this isn't bad. It's not the apocalypse but things are getting pretty shook up.
Government response	Florida is suppressing COVID deaths on a level that would make China jealous. Here are some of the steps Florida is taking. Anyone who dies from coronavirus and is not a Florida resident does not get counted. Firing the person in charge of the Florida Health Department coronavirus tracker because they won't manipulate the data.
Scientific studies and sources	What bothers me is the lack of scientific studies to support this claim. I clicked in hoping to hear about a sound study that supported this claim but there was nothing.
Infection and death rates	Stat I haven t seen anyone mention. This past week was the first to see an increase in deaths as a percentage of the total since the beginning of March. After the initial increase it's been slowly decreasing every week Cases were the same until about three weeks ago when they started increasing again. Now this week deaths increased as well. Cases have

	increased as a percent of the total each week since then so we can expect deaths will see increasing percentages each week. This week saw deaths next week will probably be following week. And that's only if the percentages increase slowly.
Miscellaneous topics	It can be done. But they are getting hit a lot harder than we are so far, That's the only reason it makes sense for them,
Putting lives at risk	If you get infected, it most likely won't impact you too harshly. But if you get infected and infect someone else who's not so well off as you are, you're potentially putting others' lives at risk. I wouldn't want that on my conscious. Is your work really worth people potentially dying for?
Staying at home	If schools are closed in your area be smart this isn't the time to go out and about having fun. This is about staying home going out only when needed. This is to keep your family safe. Think about the elderly and the immunocompromised. Be safe.
Virus transmission from animals	Actually no. A lot of human diseases evolved alongside human beings and humans are the reservoir hosts. Diseases that came from animals recently are called zoonotic. Covid is a zoonotic disease.
Accuracy of tests	Tests can definitely give false positives and they do that very often. A test that is likely to give many false positives is one that is considered sensitive. A test that is likely to give false negatives is a test that is considered specific. Even if a test were specific however there's always operator error.
Travel restrictions and lockdown	I think population density and spread matters as well. In Canada there's only a handful of big metro areas

	with lots of empty space in between. Imported cases generally would go through Toronto or Vancouver airport as most flights from overseas land at one of those two.
Shortage of medical resources	And even if there are enough beds there aren't enough doctors. And even if there are enough doctors there's not enough equipment.
Long-term effects	There are long term effects to having COVID that are still unknown. You may not die but some long-term effects, we've already seen include brain damage heart defects and lung damage.
Measure of time	I bet two weeks ago they thought it was going away in two weeks. Today they think it’s going away in days. In days they’ll be sure it’s going away in six months.
Economy, jobs, and income	I'm one of the people with cut hours, I'm losing half my pay as my company struggles to survive payroll tax cut would help people who lose their job completely are guaranteed some sort of assistance, do I not deserve one?
Effectiveness and availability of vaccines	If any concerns about the safety and efficacy of the vaccine arise, it won't be approved. Anything approved will be safe.

Table 16 Comment examples and themes

Appendix 3 Comment Topic Distribution across the Year

Fig. 27 Comment topics distribution across the year