Health Information Science and Systems

RESEARCH

Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning

Yang Liu[1], Christopher Whitfield[1], Tianyang Zhang[1,2], Amanda Hauser[3], Taeyonn Reynolds[4] and Mohd Anwar[1*]

Abstract

Purpose: It has been over a year since the first known case of coronavirus disease (COVID-19) emerged, yet the pandemic is far from over. To date, the coronavirus pandemic has infected over eighty million people and has killed more than 1.78 million worldwide. This study aims to explore "*how useful is Reddit social media platform to surveil COVID-19 pandemic?*" and "*how do people's concerns/behaviors change over the course of COVID-19 pandemic in North Carolina?*". The purpose of this study was to compare people's thoughts, behavior changes, discussion topics, and the number of confirmed cases and deaths by applying natural language processing (NLP) to COVID-19 related data.

Methods: In this study, we collected COVID-19 related data from 18 subreddits of North Carolina from March to August 2020. Next, we applied methods from natural language processing and machine learning to analyze collected Reddit posts using feature engineering, topic modeling, custom named-entity recognition (NER), and BERT-based (Bidirectional Encoder Representations from Transformers) sentence clustering. Using these methods, we were able to glean people's responses and their concerns about COVID-19 pandemic in North Carolina.

Results: We observed a positive change in attitudes towards masks for residents in North Carolina. The high-frequency words in all subreddit corpora for each of the COVID-19 mitigation strategy categories are: Distancing (DIST)—"*social distance/distancing*", "*lockdown*", and "*work from home*"; Disinfection (DIT)—"*(hand) sanitizer/soap*", "*hygiene*", and "*wipe*"; Personal Protective Equipment (PPE)—"*mask/facemask(s)/face shield*", "*n95(s)/kn95*", and "*cloth/gown*"; Symptoms (SYM)—"*death*", "*flu/influenza*", and "*cough/coughed*"; Testing (TEST)—"*cases*", "*(antibody) test*", *and* "*test results (positive/negative)*".

Conclusion: The findings in our study show that the use of Reddit data to monitor COVID-19 pandemic in North Carolina (NC) was effective. The study shows the utility of NLP methods (e.g. cosine similarity, Latent Dirichlet Allocation (LDA) topic modeling, custom NER and BERT-based sentence clustering) in discovering the change of the public's concerns/behaviors over the course of COVID-19 pandemic in NC using Reddit data. Moreover, the results show that social media data can be utilized to surveil the epidemic situation in a specific community.

Keywords: COVID-19, Social media, Natural language processing, Named-entity recognition, Topic modeling, Sentence clustering

Introduction

According to official reports from the Centers for Disease Control and Prevention, the COVID-19 pandemic has caused 19,232,843 confirmed cases and 334,029 deaths in the United States as of December 30th, 2020.[1] As the novel coronavirus pandemic continues to affect people's lives, their concerns and discussions on the epidemic

*Correspondence: manwar@ncat.edu
[1] Human-Centered AI (HC-AI) Lab, North Carolina A&T State University, Greensboro, NC 27411, USA
Full list of author information is available at the end of the article

[1] https://covid.cdc.gov/covid-data-tracker/.

continue on social media. People take to social media to express their concerns about many issues including public health, politics, society, environment, etc.

As of November 2020, Reddit ranks as the No. 7 most visited website in North America and No. 18 in global internet engagement, according to Alexa Internet.[2] Each subreddit is a community on the Reddit social media platform created and organized by users. Participants discuss topics of common interest or concern in the subreddit. Using both Reddit Application Programming Interface (API) and the Python Reddit API Wrapper (PRAW), text can be collected from subreddits. In this study, we use PRAW to scrape data from subreddits such as the title, comments, and the body of a specific post. Then we use Natural Language Processing (NLP) [1], a set of methods for automatic manipulation of natural language, to analyze the data collected from Reddit.

Machine learning has been successfully applied to a wide range of information retrieval, data mining, and social media text analysis tasks. Using the unsupervised machine learning technique of topic modeling, we found the topics of discussions that the people of North Carolina were most interested in regarding COVID-19 pandemic. We also compared the topics of discussion and the change in topics over time across subreddits for multiple cities.

The remainder of the paper is organized as follows: After the Problem Statement, Motivation and Contributions, the Related Work section surveys literature related to this study. The Methodology introduces techniques of data collection, data preprocessing, word embedding, cosine similarity, named-entity recognition, topic modeling, and BERT-based sentence clustering. The Results section presents the results of this study followed by a discussion. The final section provides the limitations and conclusions of this study.

Problem statement, motivation and contributions

This study aims to explore “how useful is Reddit social media platform to surveil COVID-19 pandemic?” and “how do people’s concerns/behaviors change over the course of COVID-19 pandemic in North Carolina?”. To achieve the research aims, we applied methods from natural language processing and machine learning to analyze collected Reddit posts using feature engineering, cosine similarity measures, LDA topic modeling, custom named-entity recognition, and BERT-based sentence clustering. Using these methods, we were able to gather people’s concerns about and their responses to this pandemic in North Carolina. The main contributions of the paper are as follows:

- We built a cleaned corpus of COVID-19 pandemic-related posts from North Carolina subreddit communities using various NLP techniques.
- We developed a custom NER system to assess the uptake of mitigation measures against the spread of COVID-19 disease.
- We extracted how people’s concerns/behaviors changed about the pandemic using an LDA-based topic model and BERT-based sentence clustering.
- We verified the effectiveness of applying Reddit data to monitor the COVID-19 pandemic in North Carolina.

Related work

Social media is widely used as a platform for people to post and share their personal opinions and feelings. For decades, researchers have used social media data for public opinion research and election results prediction [2–6], public health surveillance [7–10], marketing research [11, 12], etc. Reddit is a platform that shares content in text, pictures, or videos [13]. In this study, we use Reddit as a social media platform to collect data.

Using Natural language processing techniques to analyze social media data is becoming increasingly widespread [14]. NLP methods are very useful to extract information from multitudinous social media data. Farzindar and Inkpen showed how innovative NLP approaches can integrate appropriate linguistic information into social media monitoring [15]. In our research, we utilized an open-source Python library called the Natural Language Toolkit (NLTK) for data preprocessing.

Information extraction is one of the main tasks of natural language processing, which is the process of identifying the most important content within documents or topics. Debasmita et al. [16] presented an information retrieval system on a corpus of scientific articles related to COVID-19 using graph-based clustering on a network of articles in their corpus, and performed an extractive summarization using BERT and PageRank methods. Miller [17] reported a Python based RESTful service that utilizes the BERT model and K-Means clustering for extractive summarization on lectures. Milad et al. [18] demonstrated that contextualized representations extracted from the pre-trained deep language model BERT can be effectively used to measure the similarity between sentences and to quantify the informative content.

COVID-19 is currently affecting every country in the world and has led to lockdown measures across the countries to slow the spread of the pandemic. In terms

[2] https://www.alexa.com/siteinfo/reddit.com.

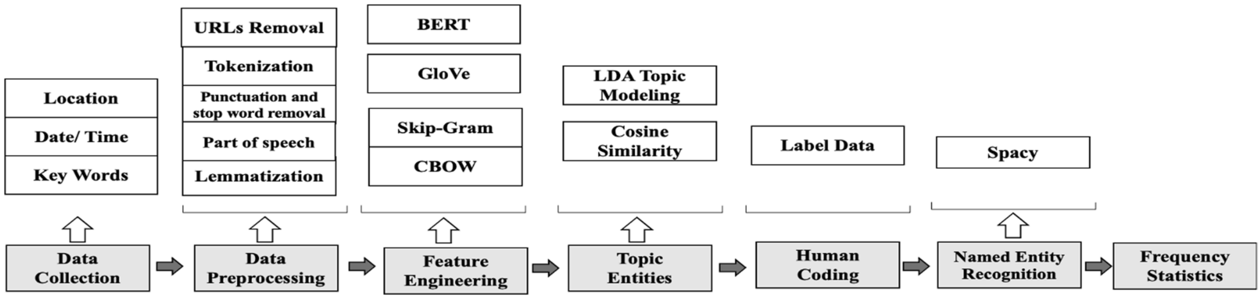


Fig. 1 Methodological workflow

of the number of cases, the United States is one of the most affected countries. By the end of December 2020, more than 524,200 confirmed cases have been reported in North Carolina, and more than 3600 deaths was attributed to COVID-19.[3] Reddit data have recently been used to track health-related discussions for public health applications [19–22], to observe mental health discourse and health anxiety during COVID-19 [23–25], and to track citizens' concerns during the COVID-19 public health crisis [26, 27]. In this research, we utilized several NLP approaches including cosine similarity, LDA topic modeling, custom named-entity recognition (NER), and BERT-based sentence clustering to discover the public's concerns over the course of the COVID-19 pandemic in North Carolina.

Methodology

The block diagram in Fig. 1 shows the following methodological workflow of our research: text collection (Reddit), text preprocessing (removal of URLs, lowercasing, tokenization, stop word removal, part-of-speech tagging, and lemmatization), feature engineering (CBOW, Skip-Gram, Glove, and BERT), topic entities discovery (Cosine Similarity and LDA topic modeling), custom NER, and frequency statistics.

Data collection

We collected data from 18 location specific subreddits for 12 cities, 3 regions, and 3 for the entire state of North Carolina: Asheville (r/asheville), Chapel Hill (r/chapelhill), Charlotte (r/Charoltte), Cary (r/Cary), CoronaNC (r/CoronaNC), Durham (r/bullcity), Elizabeth City (r/elizabethcity), Eastern NC (r/ENC), Fayetteville (r/fayettenam), Greenville (r/greenvilleNCarolina), Greensboro (r/gso), Raleigh (r/raleigh), Wilmington (r/Wilmington), Winston-Salem (r/winstonsalem), North Carolina (r/NorthCarolina), NorthCarolinaCOVID (r/NorthCarolinaCOVID), Triangle Area (r/triangle), and Western NC (r/WNC). These posts from March 3rd, 2020 (North Carolina Identifies First Case of COVID-19) to August 31st, 2020 with titles including one of the following keywords: *coronavirus, corona virus, COVID-19,* or *SARS-CoV-2.*

We used Pushshift.io Reddit API[4] to search for and record the data that met our data collection requirements as shown in Fig. 1. Then we extracted the unique post IDs from the subreddits. Using the post ID and the Python Reddit API Wrapper (PRAW), we extracted the post title, body, and comments. The extracted data of each post from all subreddits were then saved into one text file.

Relatedly, we performed additional data collection for our NER experiment using the aforementioned techniques. To train our NER model, we needed a large amount of data similar to the previously collected data. Thus, we scraped data from the subreddits of the three major COVID-19 hotspots as of August 1st, 2020. The subreddits include Arizona, Florida, Texas, CoronavirusAZ, coronavirusflorida, and CoronaVirusTX. The rationale was to select heavily populated areas to ensure we had enough data to adequately annotate and train our model.

Data preprocessing

The data preprocessing step is important as it will eliminate some of the noise and inconsistencies in the data [28]. The preprocessing steps were done on each line from the text file to extract and clean each title, body, and comment separately.

- **Removal of URLs** URL does not provide any important information and deleting URLs does not significantly affect the text information.

[3] https://covid19.ncdhhs.gov/dashboard.

[4] https://github.com/pushshift/api.

- **Tokenization** This simply breaks the text down into individual words. We completed this step using the word_tokenize function from the NLTK library.
- **Punctuation and stop word removal** Punctuation and stop words do not provide any meaning to the text and deleting punctuation does not meaningfully affect the text.
- **Part of speech (POS) tagging** POS tagging gives some contextual information about the word. To complete this step, we used the pos_tag function in NLTK. This function returns a list of tuples with the first entry being the word and the second entry being the POS tag.
- **Lemmatization** Lemmatization is the process of removing the affixes from a word by finding the word and its corresponding POS in a dictionary. Root words have different affixes but essentially the same meaning. To complete this step, we used the WordNetLemmatizer from the NLTK library.

Word embedding and cosine similarity

Word embedding is a type of text representation in which words with the same meaning have similar numerical values. In other words, word embedding is a technique for mapping the words from the dictionary to vectors of real numbers. For word embedding, we used Word2Vec [29] and Global Vectors for Word Representation (GloVe) model [30]. Word2Vec is a two-layer neural network that processes text by "vectorizing" words [31]. Its input is a text corpus (our preprocessed text in this case), and its output is a set of vectors (feature vectors that represent words in the original corpus). Word2Vec can embed data by using either of the two architecture methods: Continuous Bag of Words (CBOW) and Skip-Gram. CBOW is considered a faster method, however Skip-Gram does a better job with less frequent words. CBOW takes the words surrounding a context word and tries to predict the correct context word by probability. In the Skip-Gram model, the target words are inputted into the network and the model outputs probability distributions. For each target position, we get probability distributions for each word in the corpus. In the model, each word is encoded using one-hot encoding. One hot encoding is when the integer encoded variable is removed and a new binary variable is added for each unique integer value [32]. The output is equipped with a softmax regression classifier which is a generalization of logistic regression that is used for multi-class classification. It is different from logistic regression (LR) as LR uses binary numbers for their target variable, whereas softmax regression allows handling of many available cases.

The GloVe model captures the global corpus statistics (word-word and co-occurrence matrix), at the beginning of word embedding. Once completed, the co-occurrence probabilities can then be examined to formulate the cost function. The cost function measures the performance of a machine learning model for a given dataset. It calculates the error between expected values and the values that were produced. There are many different parameter options available during implementation including vector dimension and window size. The similarity between words during word embedding are computed using cosine similarity. Cosine similarity [33] measures the similarity between two vectors of an inner product space. It is estimated by the cosine of the angle between two vectors and decides if two vectors are pointing generally in a similar way. It is frequently used to gauge document similarity in text analysis.

Named entity recognition

Named entity recognition (NER) is the process of identifying and classifying certain words or names in a text into predefined categories [34]. To perform custom NER on our dataset, we chose to build a custom model with 5 categories. The 5 categories are distancing (DIST), disinfection (DIT), personal protective equipment (PPE), symptoms (SYM), and testing (TEST). We decided to construct our own labelled dataset using a portion of the raw text corpus from all of the NER related subreddits. The initial corpus contained 705,525 sentences. Using a keyword search method, we extracted 13,829 sentences containing relevant terms that were covered under our predefined categories. Through a combination of automation and manual configuration, we structured the data to prepare it for labelling. The tokens representing each word from the sentences were placed vertically in a column which yielded us 309,772 words to label. At random, 70% of the constructed corpus was kept as training data and the remainder was reserved for evaluation (30%). The tokens were annotated using the BILOU (Beginning, Inside, Last, Outside and Unit) [35] format. The BILOU format labels a token B-label if it is the first token in a multi-word named entity, I-label if the token is in a named entity but is not the first or last token, L-label if it is the last token in a multi-word named entity, O if it is not in a named entity, or U-label if it is a single word named entity [36]. The custom NER model that we trained was based on spaCy's multi-task, OntoNotes-trained Convolutional Neural Network which uses GloVe vectors that were trained using Common Crawl [37] corpus.

Table 1 Number of posts distribution of 18 subreddits in six months for the three NC landform distributions

Subreddits	Members	March	April	May	June	July	August	Total # of posts	Landform distributions			
									Mountain (Western)	Pied-mont (Central)	Coast (Eastern)	Other
r/asheville	26,620	1671	1047	628	588	671	330	4935	X			
r/bullcity	16,422	590	84	361	380	405	280	2100		X		
r/cary	2603	15	2	0	0	0	3	20		X		
r/chapelhill	6629	40	1	0	3	0	3	47		X		
r/Charlotte	58,773	1441	686	1430	428	441	368	4794		X		
r/CoronaNC	2593	252	168	139	108	82	85	834				X
r/elizabethcity	117	16	5	0	0	0	0	21			X	
r/ENC	411	5	0	0	0	0	0	5			X	
r/fayettenam	2180	48	27	2	2	8	28	115		X		
r/greenvilleNCarolina	870	4	0	0	6	0	0	10			X	
r/gso	9547	308	222	67	38	28	7	670		X		
r/NorthCarolina	90,677	896	1307	956	1049	1167	1124	6499				X
r/NorthCarolinaCOVID	1278	126	64	59	33	43	27	352				X
r/raleigh	64,580	1825	812	802	636	584	777	5436		X		
r/triangle	30,541	570	131	72	39	78	110	1000		X		
r/Wilmington	9114	403	209	143	202	21	29	1007			X	
r/winstonsalem	7330	255	104	105	57	37	8	566		X		
r/WNC	2524	39	12	23	13	4	0	91	X			
Total	332,809	8504	4881	4787	3582	3569	3179	28,502				

Topic modeling

Topic modeling is a method of unsupervised learning which aims to group documents into different topics, which is similar to clustering methods for numeric data [38]. There are multiple different topic modeling algorithms, however, for this study we chose to use Latent Dirichlet Allocation (LDA) [39]. The two main assumptions that guide LDA are that each document is a mixture of topics, and each topic is a mixture of words, thus, the two main parts in LDA are the words contained in each document and the words contained in each topic [38]. LDA randomly assigns each word to a topic then computes two probabilities to update the words in each topic over multiple iterations. From there, the documents are grouped into different topics in which the topics are comprised of high probability keywords.

BERT-based information extraction

In our approach, we attempt to find the people's concerns and key points from their Reddit posts which are related to COVID-19. We use the Bidirectional Encoder Representations from Transformers (BERT) [40] language model to capture the context in which sentences appear within Reddit posts. BERT was pre-trained on large text corpora (Wikipedia and BookCorpus) and fine-tuned on our Reddit dataset. Then we do the average pooling on BERT sequence of hidden states at the output of the last layer to obtain sentence level embeddings. We also try to capture the people's concerns during two three-month periods. So, we group our Reddit data by period and perform K-means clustering on each group data.

Results

Data collection and data preprocessing

Once data collection was complete, we combined the titles, bodies, and comments for each subreddit which represents a post. Table 1 depicts the total number of members and posts for each of the three North Carolina landform distributions, where we classified the 18 subreddits as Mountain (Western), Piedmont (Central), Coast (Eastern) and other.

In Fig. 2, there are 332,809 members in 18 subreddits. An average of 12 members contributed 1 post about COVID-19. The top three contribution rate of posts of subreddit: r/CoronaNC (3 ppl/post), r/NorthCarolina-COVID (4 ppl/post), and r/asheville (5 ppl/post); The last three contribution rate of posts of subreddit: r/Chapehill (141 ppl/post), r/Cary (130 ppl/post), and r/greenvilleN-Carolina (130 ppl/post).

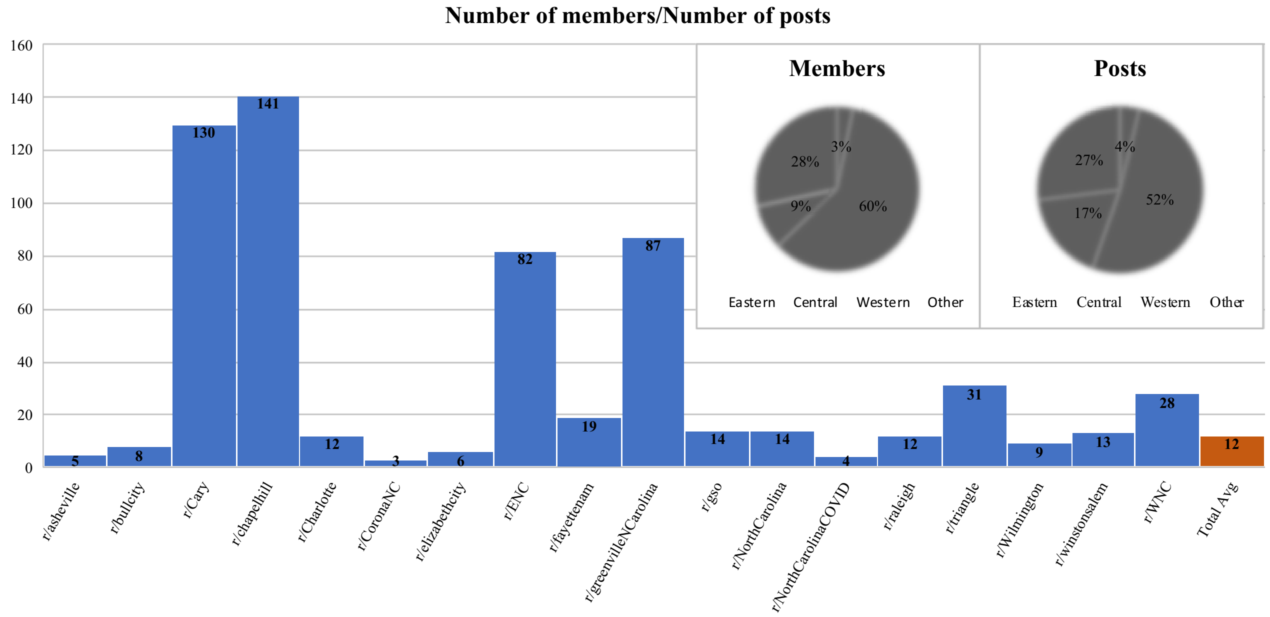


Fig. 2 One post per number of members in each subreddit. The number of people provide 1 post in each subreddit. There are a total of 332,809 members in 18 subreddits. The left Pie chart is the percentage of members based on geography classification; the right Pie chart is percentage of posts based on geography classification (Example: In subreddit of *r/asheville*, one post per 5 members.)

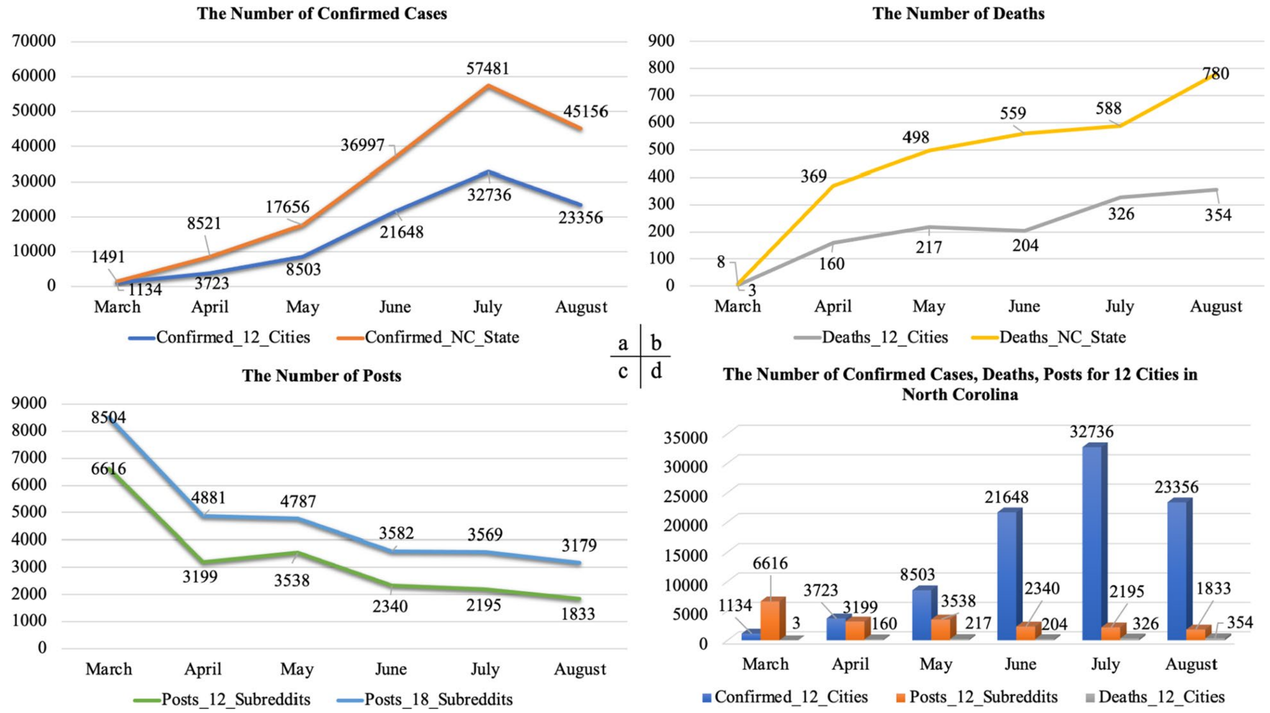


Fig. 3 Distribution of the number of confirmed cases (**a**), deaths (**b**) and posts (**c**) from March to August

Table 2 The five most similar words to Gloves, Soap, Fever, Test, and Lockdown across the three different algorithms (CBOW, Skip-Gram, and GloVe)

Gloves			Soap			Fever			Test			Lockdown		
CBOW	Skip-Gram	GloVe	CBOW	Skip-Gram	GloVe	CBOW	Skip-Gram	GloVe	CBOW	Skip-Gram	GloVe	CBOW	Skip-Gram	GloVe
Save	Practice	Wear	Alcohol-based	Water	Alcohol-based	Infected	Negative	Cough	Kit	Currently	r/coronavirussc	Admit	Reasonable	California
Clean	Sanitize	Useless	Refrain	Sanitizer	Sleeve	Cough	Cough	Negative	Positive	Lab	r/coronavirusalabama	Eviction	Possibly	Similar
Completely	Wear	Sanitize	Squirt	Alcohol-based	Water	Thousand	Breath	Shortness	Case	Result	Positive	Strike	Relatively	Monger
Apart	Hygiene	Mask	Wipe	Bottle	Gallon	Symptom	Shortness	Ache	Confirm	Kit	Kit	Course	Stand	Compare
Homemade	Shake	Touch	Towel	Often	Hearsay	Yesterday	100.4	100.4	cdc	cdc	Roadblock	Vulnerable	ppl	Martial

In Fig. 3, the trend of confirmed cases in North Carolina, as provided by North Carolina Department of Health and Human Services (NCDHHS),[5] was consistent with the trend of confirmed cases observed in our 12 location-specific subreddits. Regarding NC COVID-19 death trends, the data (see footnote 5) is consistent with the trend of deaths in our 12 location-specific subreddit data. The trend of the number of posts of 12 subreddits from March to August was consistent with the trend of the number of posts of 18 subreddits from March to August.

Word embedding and cosine similarity

All but three parameters were assigned their default values for each model. We considered several values for vector dimension, window size, and word count. Regarding the CBOW and Skip-Gram models, the optimal parameters for this study were 400 for vector dimension, 5 for window size, and 5 for minimum word count. For the GloVe model, we used a vector dimension of 400, window size of 15, and minimum word count of 5. For each of our three word embedding models, the five most similar words to *Gloves*, *Soap*, *Fever*, *Test*, and *Lockdown* were computed using cosine similarity, as shown in Table 2.

Named entity recognition

We represented the five categories as Distancing (DIST), Disinfection (DIT), Personal Protective Equipment (PPE), Symptoms (SYM), and Testing (TEST). We removed the irrelevant words and combined the similar words, then we chose the top 3 words for each category. The results are shown in Table 3. The high-frequency words in all subreddit corpora for each category are as follows: Distancing (DIST)—"*social distance/distancing*", "*lockdown*", and "*work from home*"; Disinfection (DIT)—"*(hand) sanitizer/soap*", "*hygiene*", and "*wipe*"; Personal Protective Equipment (PPE)—"*mask/facemask(s)/face shield*", "*n95(s)/kn95*", and "*cloth/gown*"; Symptoms (SYM)—"*death*", "*flu/influenza*", and "*cough/coughed*"; Testing (TEST)—"*cases*", "*(antibody) test*", *and* "*test results (positive/negative)*". Given the total number of test results combined for the 6 subreddits in Table 3, the average number of positive results during the first three-month period is 71.3% and 28.7% for negative, the average number of positive results during the second three-month period are 74.4% and 25.6% for negative.

Topic modeling

The NCDHHS recommends people practice 3Ws (Wear mask, wait 6 feet apart, and wash hands) if they leave home (see footnote 5). Therefore, we separate the dataset into two groups to compare people's adherence to the recommendations during two time periods. The first group contains the data from March, April, and May; and the other group contains the data from June, July, and August. For the remainder of this section, the period from March to May will be referred as the first trimester, and the period from June to August will be referred as the second trimester. After using LDA topic modeling, we obtained 5 topics for each group where each topic contains the top 9 keywords. The size of the word is determined by the word's importance in that topic. The sizes of the words between word clouds do not signify their importance relative to one another. The word clouds representing each topic of 6 subreddits (Asheville (asheville), Charlotte (Charlotte), Greensboro (gso), Raleigh (raleigh), Wilmington (Wilmington), North Carolina (NorthCarolina), and an aggregation (NC_All) that includes all 18 subreddits are shown in Fig. 4.

During Asheville's first trimester, there is no *Wash*, however, one topic mentions *Wear* (mask) and another topic mentions *Wait* (stay home). As indicated by the emphasis on the keywords *work*, *pay*, *business*, and *home*, people are more concerned about working, business, and their homes. Concerning the second trimester for Ashville, three of the five topics emphasize the word *mask*, which is part of the 3Ws *Wear*. During the first trimester in Charlotte (r/Charlotte), people talked about *Wait* (work home) and *Wear* (wear masks), however, no topics related to *Wash* was mentioned. In Charlotte's second trimester, 4 topics contained *Wear* (masks). In the first trimester in Greensboro (r/gso), people are less concerned with precautionary measures of COVID-19. Moreover, people did not talk about any 3Ws during the second trimester. During the first trimester in Raleigh (r/raleigh), people talked about the effects of COVID-19 (such as cases and deaths) *Wash* (hand), and *Wait* (work, home). During the second trimester, no *Wash* is mentioned, and people start to talk about *Wear* (masks). Regarding Wilmington, there is very little mentioning of "*social*" and *"wear"* during the first trimester. However, there are three topics containing *Wear* (masks) during the second trimester. During the first trimester in the subreddit representing the entire state of North Carolina (r/NorthCarolina), the people discuss very little about how they can prevent COVID-19 transmission. There is only one topic that mentions "*stay home*" and "*mask*", however, two topics contain *Wear* (masks) during the second trimester. Regarding the combined 18 subreddits (NC_All), there is no topic that contains *Wear* or *Wash*, yet one topic contains *Wait* (work from home) during the first trimester. During the second trimester, there is one topic that heavily emphasize *Wear* (mask).

[5] https://covid19.ncdhhs.gov.

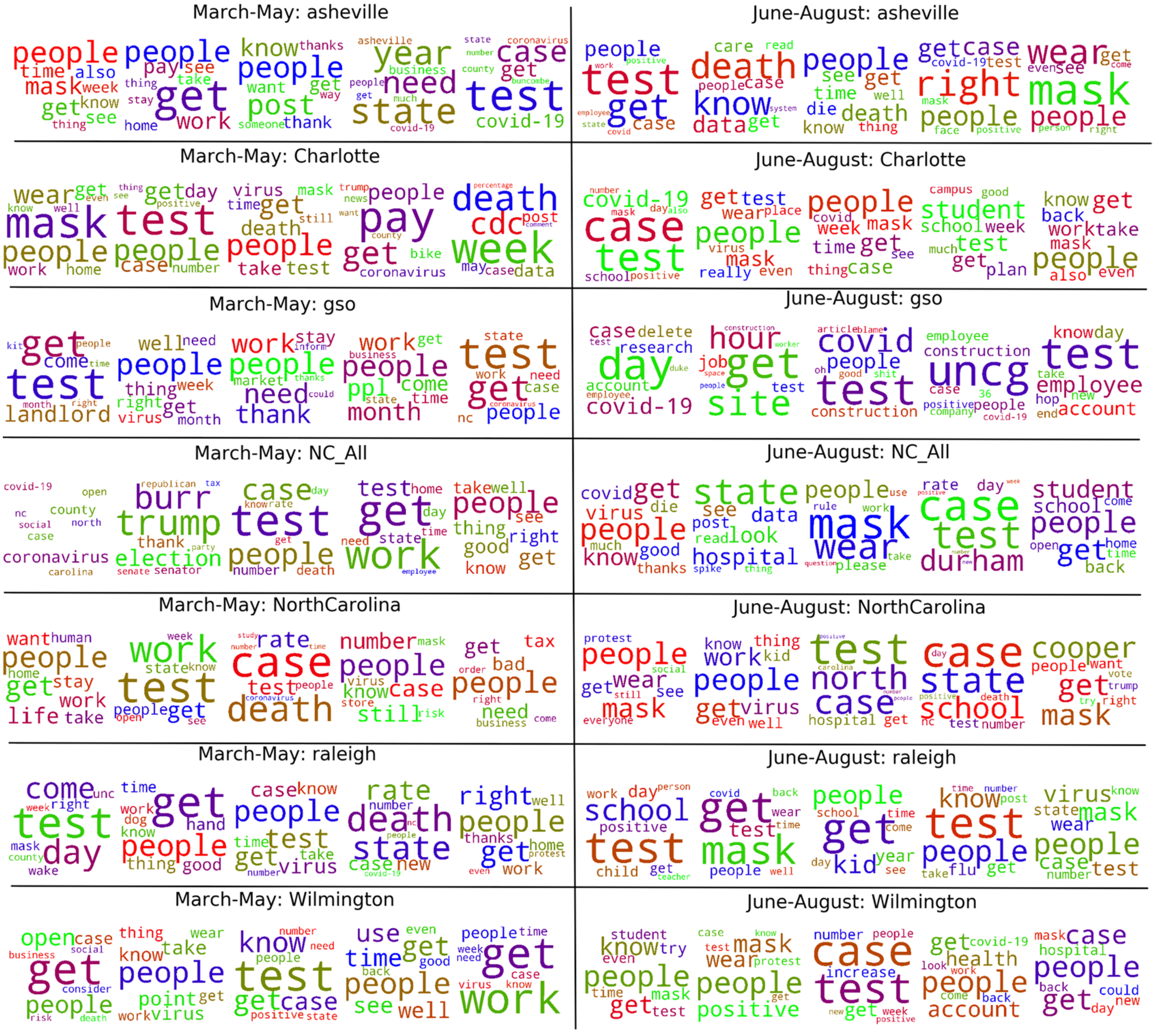
Table 3 Identification of entities for 3 mitigation types (distancing, disinfection, and PPE), and 2 detection types (symptoms and testing)

Categories	Asheville				Categories	Charlotte				Categories	Greensboro			
	March to May		June to August			March to May		June to August			March to May		June to August	
	Entity name	# of entities	Entity name	# of entities		Entity name	# of entities	Entity name	# of entities		Entity name	# of entities	Entity name	# of entities
DIST	**Social distanceing**	87	**Lockdown**	135	DIST	**Social distanceing**	140	**Social distanceing**	220	DIST	**Social distanceing**	8	**Social distanceing**	28
	Lockdown	53	Social distanceing	105		Lockdown	110	Lockdown	146		Work from home	2	Lockdown	13
	Work (from) home	6	Work (from)/ stay home	10		Work (from) home	18	Work (from) home	69		Work time	1	Work from home	5
DIT	(Hand) sanitizer(s)/ soap	9	**(Hand) sanitizer/soap**	50	DIT	(Hand) sanitizer/ soap	10	**(Hand) sanitizer/ soap**	116	DIT	Impact	1	**(Hand) sanitizer/soap**	21
	Hygiene	10	Wipe	12		**hygiene**	11	Wipe	33		Wipe	1	Wipe	11
	Wipe	3	Bleach	11		wipe	3	Lysol	25		Profit	1	Hygiene	7
PPE	**Mask/ facemask(s)/ face shield**	973	**Mask/ facemask(s)/ face shield**	574	ppe	**Mask/ facemask(s)/ face shield**	952	**Mask/ facemask (s)/ face shield**	707	ppe	**Mask/ facemask(s)/ face shield**	117	**Mask/ facemask(s)/ face shield**	125
	n95(s)/kn95	46	Glove	37		n95(s)	24	n95(s)	51		Cloth	3	Glove	9
	Glove	17	n95(s)/kn95	32		Cloth/gown	9	Cloth/ gown	36		Glove	3	n95(s)	9
SYM	**Death**	229	**Death**	225	Sym	**Death**	320	**Death**	511	SYM	**Flu/influenza**	7	**Flu/influenza**	43
	Flu/influenza	99	Flu/influenza	211		Flu/influenza	81	Flu/influenza	316		Death	4	Death	29
	Coughed	30	Coughed	62		Cough	25	Cough	108		Coughed	2	Coughed	15
TEST	**Cases**	320	Cases	436	Test	**Cases**	529	Cases	992	Test	Cases	12	Cases	122
	(Antibody) test	277	**(Antibody) test**	648		(Antibody) test	324	**(Antibody) test**	1170		**(Antibody) test**	21	**(Antibody) test**	188
	Test result	260	Test result	208		Test result	182	Test result	421		Test result	16	Test result	80

Table 3 (continued)

Categories	North Carolina				Categories	Raleigh				Categories	Wilmington			
	March to May		June to August			March to May		June to August			March to May		June to August	
	Entity name	# of entities	Entity name	# of entities		Entity name	# of entities	Entity name	# of entities		Entity name	# of entities	Entity name	# of entities
DIST	**Social distanceing**	185	Social distanceing	201	DIST	**Social distanceing**	174	**Social distanceing**	286	DIST	**Social distanceing**	15	Social distanceing	19
	Lockdown	141	**Lockdown**	240		Lockdown	122	Lockdown	159		Lockdown	7	**Lockdown**	23
	Work (from) home	6	Work (from) home	50		Work (from) home	30	Work (from) home	39		Social worker	2	Work (from) home	4
DIT	Hygiene	13	**(Hand) sanitizer(s)/ soap**	63	DIT	**(Hand) sanitizer(s)/ soap**	33	**(Hand) sanitizer/ soap**	130	DIT	Disaster	1	**(Hand) sanitizer(s)/ soap**	7
	Wipe	10	Hygiene	16		Wipe	15	Wipe	50		Hand sanitizer	1	Hygiene	2
	(Hand) sanitizer(s)/ soap	22	Wipe	15		Bleach	6	Bleach	28		–	–	Bleach	1
PPE	**Mask/ facemask(s)/ face shield**	2012	**Mask/ facemask(s)/ face shield**	551	PPE	**Mask/ facemask(s)/ face shield**	1984	**Mask/facemask (s)/ face shield**	1046	PPE	**Mask/ facemask(s)/ face shield**	82	**Mask/ facemask(s)/ face shield**	66
	n95(s)/kn95	51	n95(s)/kn95	27		n95(s)/kn95	59	n95(s)/kn95	64		n95(s)/kn95	1	n95	4
	Cloth/gown	42	Glove	55		Cloth	21	Cloth/ gown	47		Clown	1	Glove	15
SYM	**Death**	441	**Death**	639	SYM	**Death**	583	**Death**	461	SYM	Death	19	**Death**	45
	Flu/influenza	182	**Flu/influenza**	451		**Flu/influenza**	263	Flu/influenza	384		Coughed	3	Flu/influenza	44
	Coughed	47	Coughed	98		Coughed	70	Coughed	147		Flu/influenza	3	Coughed	13
TEST	**Cases**	948	Cases	1037	Test	**Cases**	1118	Cases	849	Test	**Cases**	53	Cases	82
	(Antibody) test	692	**(Antibody) test**	1311		(Antibody) test	1054	**(Antibody) test**	1187		Test	31	**(Antibody) test**	174
	Test result	583	Test result	545		Test result	1043	Test result	651		Test result	41	Test result	57

Most distinct and frequently mentioned entities are in bold


March-May: asheville

June-August: asheville

March-May: Charlotte

June-August: Charlotte

March-May: gso

June-August: gso

March-May: NC_All

June-August: NC_All

March-May: NorthCarolina

June-August: NorthCarolina

March-May: raleigh

June-August: raleigh

March-May: Wilmington

June-August: Wilmington

Fig. 4 Word clouds representing each topic found using LDA Topic modeling. The larger the word is the more significant it is within that topic

As opposed to the first trimester, the users from all subreddits (except Greensboro) pay more attention to *Wear* (mask) during the second trimester. Although not part of the 3Ws, it was uniquely observed that all subreddits during the entire six-month period contains at least one topic that emphasizes the keyword "*Test*".

BERT-based information extraction

We use BERT-base-uncased as our initial weight and we fine-tuned it with a total of 14,500 steps using our Reddit dataset. BERT-base model contains 12 layers, 768 hidden units in each layer, 12 attention heads per unit, and a total number of 110 million parameters. After fine-tuning the BERT model, we use its encoder to do the feature extraction. After the feature extraction step, each token is represented as a contextualized embedding with a size of 768. Next, a contextualized representation is computed for each sentence by averaging over all the representations of the tokens that belong to each sentence. Based on each contextualized embedding, we apply K-means cluster algorithms to cluster our data into 3 categories. As for the results in Table 5 (Appendix), during the first trimester in Asheville, people were more concerned about the spread of virus and its impact on people's lives. During the second trimester, people turn their focus on *COVID-19* testing. During the first trimester in Charlotte

Table 4 Comparison of state-of-the-art methods

Objective	References	Data source	Method
To measure and monitor citizens' concern levels using public sentiments in Twitter data	Chun et al. [10]	Twitter	NLP and case fatality rate (CFR)
To retrieval articles related to COVID-19	Das et al. [16]	A corpus of scientific articles	Graph community detection and Bio-BERT embeddings
To utilize NLP for the analysis of public health applications	Conway et al. [21]	Reddit, Microblog, Instagram, etc.	Literature review
To characterize the media coverage and collective internet response to the COVID-19 in four countries	Gozzi et al. [22]	Reddit and Wikipedia	Linear regression model, nonnegative matrix factorization
To characterize people's responses to COVID-19 on two Reddit communities	Zhang et al. [23]	2 subreddits on Reddit	Classification, Fightin' words model,
To leverage NLP to characterize changes in mental and non-mental health support groups during the initial stage of the pandemic	Low et al. [24]	Reddit	NLP, unsupervised clustering, topic modeling, Similarity
To predict the general sentiment polarity of the COVID-19 related news on Reddit before a news article is published	Dheeraj [25]	Reddit	Sentiment analysis
To understand the patient mental health through the stages of COVID-19 illness	Murray et al. [26]	Reddit	Topic modeling, sentiment analysis, clustering
To understand the public's concerns around coronavirus and identify future opportunities for medical experts to leverage the Reddit in communicating with the general public	Lai et al. [27]	1 subreddit on Reddit	Retrospective content analysis
To track public priorities and concerns regarding COVID-19	Stokes et. al [28]	Reddit	Topic modeling
To explore "how useful is Reddit social media platform to surveil COVID-19 pandemic?" and "how do people's concerns/behaviors change over the course of COVID-19 pandemic in North Carolina?	Our paper	18 subreddits of North Carolina on Reddit	NLP, word embedding, similarity, topic modeling, custom NER, BERT-based clustering, K-means clustering

and Greensboro, lockdown and spread of virus are two hot topics that people heavily discussed. In Charlotte's second trimester, people talk more about reopening, however people in Greensboro tend to talk about the impact of *COVID-19* on their lives. After analyzing all of our acquired North Carolina Reddit posts, we found that reopen and the spread of the virus are the most discussed topics during the entire 6 months.

Discussion

In our dataset, the Piedmont (Central) region accounts for 9 of the 18 subreddits and provides 53% of the overall posts. The Coastal Plains (Eastern) makes up 4 of the 18 subreddits and provides 4% of the total number of posts. The Mountain (Western) region accounts for 2 of the 18 subreddits and consists of 17% of the total number of posts. Given that the Piedmont region represents 53% of the overall posts and Raleigh represents 19% (highest), the fact that Asheville (second highest) represents 17% of the overall posts is highly significant. Thus, Asheville is nearly as good a representation for the western area of NC as Raleigh is for the Piedmont. Conversely, the total number of posts for the eastern region of NC is limited, however, Wilmington provides the most for the region with roughly 3.5% of the overall posts.

In Table 2, we selected *Glove*, *Soap*, *Fever*, *Test*, and *Lockdown* according to the five COVID-19 mitigation strategy categories: Personal Protective Equipment (PPE), Disinfection (DIT), Symptoms (SYM), Testing (TEST), and Distancing (DIST). The purpose of using three word embedding methods (CBOW, Skip-Gram, and GloVe) was to determine which method most effectively encodes COVID-19 related words to vectors whereby the cosine similarity scores were determined and the performance for each approach was assessed. For Table 2, the Skip-Gram and CBOW models appear to give good results throughout the entire table. However, the GloVe model appears to be inconsistent based on the results it produced. For the word fever, it appears to give good results and produces results that are similar to those found in the CBOW and Skip-Gram models. However, for the word test, the first two results produced by the GloVe model are subreddit names and do not provide any useful information to analyze. Thus, it seems that the Skip-Gram and CBOW models would be preferable for finding similar words.

We compared with other Reddit-based COVID-19 related research in Table 4. In our research, we collected posts from 18 location-specific subreddits for 12 cities, 3 regions, and 3 for the entire state of North Carolina, micro-communities within the Reddit platform, as a data source to monitor the COVID-19 pandemic in North Carolina. To reiterate, we compared people's thoughts, behavior changes, discussion topics, and the number of confirmed cases and deaths, we applied methods from natural language processing and machine learning to analyze collected Reddit posts using feature engineering, topic modeling, custom named-entity recognition, and BERT-based (Bidirectional Encoder Representations from Transformers) sentence clustering. Moreover, we verified the effectiveness of applying our obtained Reddit data to monitor the COVID-19 pandemic in North Carolina.

Limitations

There were a few limitations noted in this research. First, the period of our dataset is from March 3, 2020, through August 31, 2020. We did not collect the posts after August 2020 in this research. Second, we collected data from 12 location-specific subreddits and 6 independent communities comprised of multiple North Carolina cities. Although we selected as many representative North Carolina communities as possible, not every region in North Carolina has a subreddit community. Additionally, we cannot guarantee everyone who posted in the subreddit community still lived in these areas at the time of posting. Finally, our Reddit corpus only contains the posts written in English, therefore the results are limited to users who post in English.

Conclusion

In this study, we used six months of Reddit data to survey the COVID-19 pandemic in North Carolina by employing NLP, cosine similarity, LDA topic modeling, custom NER, and BERT-based sentence clustering. Our study monitored changes in public behavior during the COVID-19 pandemic in North Carolina. During the first trimester, the public was most concerned with reducing the spread of COVID-19 by adhering to social distance guidelines and washing hands. Over the course of the second trimester, we further observed a positive change in attitudes towards masks for residents in North Carolina.

The findings in our study show that the use of Reddit data to monitor COVID-19 pandemic in North Carolina is effective. The study further shows the effectiveness of NLP, cosine similarity, LDA topic modeling, custom NER and BERT-based sentence clustering in discovering how the public's concerns/behavioral changed over the course of the COVID-19 pandemic in North Carolina using Reddit data. The results show that the representative social media data can be utilized to surveil the epidemic situation in a specific community.

Appendix

See Table 5.

Table 5 Sample of BERT sentences clustering on different topics of subreddits

Subreddit	Time period	Topics	Sentence sample
Asheville	March–May	Concerns	*What's your overwhelming desire* that is worth infecting and killing people? I'm DYING to know?
		Spread of virus	***Try *to keep 6 ft distance* between you and other people.**The WHO has said that the virus most easily spreads through the air when you're closer than *6 ft* to an infected person. Try to *keep your distance*, especially on public transit
		Impacts	That means right now the average North Carolinian on *unemployment* gets just under $ 2300, spread out over two months. Looking at the chart above, you can tell that's not going to be enough
	June–August	Concerns	I don't understand *why people won't take it seriously*. like okay, maybe it won't kill you but what about your neighbors?! if you don't care about them what about your at risk family and friends?!
		Impacts	In June of 2020 what are we afraid of? Hospitals have had time to prepare. Cases will *naturally go up with more people working*. It should not be a surprise. Cases are up but NC death rates are at the lowest levels
		Testing	Literally the least shocking thing ever. And yet Buncombe County has suspended all *community testing* until an undisclosed time in August. Why aren't they mandating testing for all children returning to school? No way will I be sending my kids to the slaughter for some shoddy political agenda
Charlotte	March–May	Lockdown	Doubt it, remember when Florida *opened* their beaches? It's been 2 weeks and no spike
		Impacts	So, in a few weeks, US *deaths* per capita will likely look even worse relative to other wealthy countries than they do now. One reason is our confirmed infection rate is high given the lower level of testing. That means that actual infection rate is even higher, and that is a good predictor of future deaths 2–3 weeks from now
		Spread of virus	Ummm…we have a lower infection rate because of the *stay at home*
	June–August	Spread of virus	In my view, the only good news from these charts is that growth appears linear rather than exponential. However, we shouldn't be complacent in thinking that this is an inherent quality of the virus. If we relax controls enough and a seasonal affect is removed, we will likely see *exponential breakout*
		Reopen	I think it would be nearly *impossible to re-close* anything that's open right now … I mean, look at how much anger is already happening about asking to just wear a mask that practically has no impact on your lifestyle at all
		Lockdown	We all stayed *locked down* for weeks on end. The massive reduction in vehicular traffic back then is incontrovertible evidence of compliance with the initial lockdown
Greensboro	March–May	Testing	There's no third option. It's reclosures or much more *testing*
		Lockdown	Go do some real reporting and find out why we *closed the country down* for a virus that can't even come within a million cases of what the "scientists" estimated as deaths in America alone
		Spread of virus	One scenario without lockdown: 1k get it in March, 2k in April, and the remaining 7k in May. Given this scenario, we peak in May with 7k cases all at once
	June–August	Politics	I'm not trying to change anyone's mind on masks. There's no changing anyone's mind about it if we have conflicting guidance between our doctors/scientists and our *politicians*
		Lockdown	So, this is not regular time. Everything isn't hunky dory. We don't need another person to get infected and then go out and infect other people. *Stay home*. Do something else. Because even though the virus might not kill you, it will permanently damage you with all sort of horrible things that will shorten your life. And you will also become another case that will keep this country from being able to interact with the rest of the world
		Impacts	The $600 Pandemic Unemployment Assistance (PUA) from the federal government runs out on 7/31. Your regular state benefits will continue past that date
Raleigh	March–May	Impacts	We have to *live with coronavirus* as a reality in our world, since neither of the steps above will get rid of it. What does that world look like?
		Reopen	They *don't plan on closing now*. Saying that the CDC doesn't recommend it at this time….and 2 weeks isn't long enough to be effective anyway
		Lockdown	My point was they were smart. They are now prepared for even bigger *lockdowns*
	June–August	Spread of virus	What you see now, however, is a clearly disturbing trend: *percent positives are going up* *even as number of tests* is going up. This is the opposite of what one would expect if things were really plateaued / stable / declinin
		Impacts	Still, *N95 are hard to find* and anything is better than nothing
		Reopen	Bull and Bear is operating way outside the guidelines. Gyms are the last place to be *open* if the virus spreads like we think it does

Table 5 (continued)

Subreddit	Time period	Topics	Sentence sample
Wilmington	March–May	Reopen	We know who is at risk, yet we *mandated a lockdown* for all. Not the best idea to approach the situation
		Testing	Widespread *testing* + *antibody testing* would be ideal. The reality is is that in two months we've tested 75,000/10,000,000
		Lockdown	No. *No lockdown*
	June–August	Spread of virus	New Hanover has *192 active cases* currently, Brunswick 164
		Reopen	Other countries were told to stay home and they did. Now they're *getting back to normal much quicker*
		Testing	When you look at the percentage of positives against the *amount of tests* administered, the infection rate in North Carolina is going down…very slowly. I guess technically, we did have 1 day that was 6.8% positive, and we are now at 6.9% of tests given are positive…but a 0.1% fluctuation isn't where I'd starting talking about the infection rate going up
North Carolina	March–May	Reopen	I want NC to "*reopen*"
		testing	And it frankly pisses me off that there isn't way more *testing*. You want to get people back out sooner and resuscitate the economy? Then test people! Find the ones who are infected and quarantine them, let everyone else out. And test everyone again and again until we're sufficiently past this. But running more tests would reveal the full extent of how much our governments dropped the ball on this, so of course they won't do it
		Spread of virus	These people are underestimating how easily this *virus is spread*. No doubt there are infected people in that crowd and tonight there will be more
	June–August	Spread of virus	* NC is currently *#9 in the nation for total coronavirus cases* > NC is also ranked #10 by total population, so I don't see that as a red flag. If we consider the number of cases per 100k, we're # 25th
		Reopen	You do realize that people have *to leave their homes for groceries, work and medical resources*. There is no possible way to keep everyone home ever
		Impacts	As far as the costs. What are the economic costs of 231,000,000 *infections* to our economy? What are the costs of the long term health effects of 231 million infections with a disease known to have high rates of long term neurologic, cardiac, and pulmonary damage?

Funding

The funding was provided by National Science Foundation.

Author details

[1]Human-Centered AI (HC-AI) Lab, North Carolina A&T State University, Greensboro, NC 27411, USA. [2]University of Massachusetts Amherst, Amherst, MA 01003, USA. [3]North Carolina State University, Raleigh, NC 27695, USA. [4]Elizabeth City State University, Elizabeth City, NC 27909, USA.

Received: 9 January 2021 Accepted: 15 June 2021

Published online: 25 June 2021

References

1. Calvo RA, Milne DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng. 2017;23(5):649–85.
2. Metaxas PT, Mustafaraj E, Gayo-Avello D. How (not) to predict elections. In: Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. IEEE; 2011, p. 165–171.
3. Shi L, Agarwal N, Agrawal A, Garg R, Spoelstra J. Predicting US primary elections with Twitter; 2012. http://snap.stanford.edu/social2012/papers/shi.pdf.
4. Ramteke J, Shah S, Godhia D, Shaikh A. Election result prediction using Twitter sentiment analysis. In: Proceedings of the 2016 international conference on inventive computation technologies (ICICT), vol 1. IEEE; 2016, p. 1–5.
5. Bermingham A, Smeaton A. On using Twitter to monitor political sentiment and predict election results. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011); 2011, p. 2–10.
6. Anstead N, O'Loughlin B. Social media analysis and public opinion: the 2010 UK general election. J Comput Mediat Commun. 2015;20(2):204–20.
7. Yang M, Li Y, Kiang MY. Uncovering social media data for public health surveillance. In: PACIS; 2011, p. 218.
8. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q. 2014;92(1):7–33.
9. Paul MJ, Sarker A, Brownstein JS, Nikfarjam A, Scotch M, Smith KL, Gonzalez G. Social media mining for public health monitoring and surveillance. In: Biocomputing 2016: Proceedings of the Pacific symposium; 2016, p. 468–479.
10. Chun SA, Li ACY, Toliyat A, Geller J. Tracking citizen's concerns during COVID-19 pandemic. In: proceedings of the 21st Annual International Conference on Digital Government Research; 2020, p. 322–323.
11. Alalwan AA, Rana NP, Dwivedi YK, Algharabat R. Social media in marketing: A review and analysis of the existing literature. Telemat Inf. 2017;34(7):1177–90.
12. Hays S, Page SJ, Buhalis D. Social media as a destination marketing tool: its use by national tourism organisations. Curr Issues Tourism. 2013;16(3):211–39.
13. Liu Y. A comparative study of vector space language models for sentiment analysis using reddit data (Doctoral dissertation, North Carolina Agricultural and Technical State University); 2020
14. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.; 2009.
15. Farzindar A, Inkpen D. Natural language processing for social media. Synth Lect Hum Lang Technol. 2015;8(2):1–166.
16. Das D, Katyal Y, Verma J, Dubey S, Singh A, Agarwal K, et al. Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-bert embeddings. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020; 2020.
17. Miller D. Leveraging BERT for extractive text summarization on lectures. arXiv preprint arXiv:1906.04165; 2019.

18. Moradi M, Samwald M. Clustering of deep contextualized representations for summarization of biomedical texts. arXiv preprint arXiv:1908.02286; 2019.
19. Hemalatha I, Varma GS, Govardhan A. Preprocessing the informal text for efficient sentiment analysis. Int J Emerg Trends Technol Comput Sci (IJETTCS). 2012;1(2):58–61.
20. Park A, Conway M. Tracking health related discussions on Reddit for public health applications. In AMIA Annual Symposium Proceedings, vol 2017. American Medical Informatics Association; 2017, p. 1362.
21. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearbook Med Inf. 2019;28(1):208.
22. Gozzi N, Tizzani M, Starnini M, Ciulla F, Paolotti D, Panisson A, Perra N. Collective response to media coverage of the COVID-19 pandemic on reddit and Wikipedia: mixed-methods analysis. J Med Internet Res. 2020;22(10):e21597.
23. Zhang JS, Keegan BC, Lv Q, Tan C. A tale of two communities: characterizing reddit response to covid-19 through/r/china flu and/r/coronavirus. arXiv preprint arXiv:2006.04816; 2020.
24. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. J Med Internet Res. 2020;22(10):e22635.
25. Dheeraj K. Analysing COVID-19 news impact on social media aggregation. Int J Adv Trends Comput Sci Eng. 2020;9(3):2848–55.
26. Murray C, Mitchell L, Tuke J, Mackay M. Symptom extraction from the narratives of personal experiences with COVID-19 on Reddit. arXiv preprint arXiv:2005.10454; 2020.
27. Lai D, Wang D, Calvano J, Raja AS, He S. Addressing immediate public coronavirus (COVID-19) concerns through social media: utilizing Reddit's AMA as a framework for Public Engagement with Science. PLoS ONE. 2020;15(10):e040326.
28. Stokes DC, Andy A, Guntuku SC, Ungar LH, Merchant RM. Public priorities and concerns regarding COVID-19 in an online discussion forum: longitudinal topic modeling. J Gen Internal Med. 2020;35(7):2244–7.
29. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst. 2013;26:3111–9.
30. Pennington J, Socher R, Manning CD. Glove: global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014, p. 1532–1543.
31. Nicholson C. A beginner's guide to neural networks and deep learning; 2019.
32. Brownlee J. Why one-hot encode data in machine learning. Mach Learn Mastery; 2017
33. Jones WP, Furnas GW. Pictures of relevance: a geometric analysis of similarity measures. J Am Soc Inf Sci. 1987;38(6):420–42.
34. Utkarsh K. Named Entity Recognition using Bidirectional LSTM-CRF. Retrieved July 05, 2020, from https://medium.com/@utkarsh.kumar2407/named-entity-recognition-using-bidirectional-lstm-crf-9f4942746b3c; 2020
35. Kapadia S. Topic modeling in python: latent dirichlet allocation (LDA). Retrieved January 02, 2021, from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0; 2020
36. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360; 2016
37. Panchenko A, Ruppert E, Faralli S, Ponzetto SP, Biemann C. Building a web-scale dependency-parsed corpus from CommonCrawl. arXiv preprint arXiv:1710.01779; 2017.
38. Kulshrestha R. A beginner's guide to latent dirichlet allocation (LDA). Medium. https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2; 2020
39. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
40. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1710.01779; 2018