Surveillance of COVID-19 Pandemic using Social Media: A Reddit Study in North Carolina

Christopher Whitfield  
Computer Science  
North Carolina A&T University  
Greensboro NC USA  
cewhitfield@aggies.ncat.edu

Yang Liu  
Computer Science  
North Carolina A&T University  
Greensboro NC USA  
yliu2@aggies.ncat.edu

Mohd Anwar  
Computer Science  
North Carolina A&T University  
Greensboro NC USA  
manwar@ncat.edu

ABSTRACT
Coronavirus disease (COVID-19) pandemic has changed various aspects of people’s lives and behaviors. At this stage, there are no other ways to control the natural progression of the disease than adopting mitigation strategies such as wearing masks, watching distance, and washing hands. Moreover, at this time of social distancing, social media plays a key role in connecting people and providing a platform for expressing their feelings. In this study, we tap into social media to surveil the uptake of mitigation and detection strategies, and capture issues and concerns about the pandemic. In particular, we explore the research question, “how much can be learned regarding the public uptake of mitigation strategies and concerns about COVID-19 pandemic by using natural language processing on Reddit posts?” After extracting COVID-related posts from the four largest subreddit communities of North Carolina over six months, we performed NLP-based preprocessing to clean the noisy data. We employed a custom Named-entity Recognition (NER) system and a Latent Dirichlet Allocation (LDA) method for topic modeling on a Reddit corpus. We observed that mask, flu, and testing are the most prevalent named-entities for “Personal Protective Equipment”, “symptoms”, and “testing” categories, respectively. We also observed that the most discussed topics are related to testing, masks, and employment. The mitigation measures are the most prevalent theme of discussion across all subreddits.

CCS CONCEPTS
• Web Mining • Content Analysis

KEYWORDS
COVID-19 surveillance, Social Media, Reddit, Natural Language Processing, Named-Entity Recognition, Topic Modeling, LDA

1 Introduction
The COVID-19 disease has led to a global pandemic and the latest global public health crisis. At the time of writing this paper, global cases have surpassed 172 million (over 33 million US cases), and there are more than 3 million deaths worldwide (over 597 thousand US deaths)1. As a result, governments worldwide have developed new policies and mandates to help mitigate the spread of the disease. Specifically, Centers for Disease Control and Prevention (CDC) recommended wearing a mask, washing hands, and watching social distance. These recommendations are also implemented by various state and city ordinances.

Although research shows that the mitigation strategies are effective to flatten the transmission curve, public uptake of these strategies makes a huge difference. However, there is a gap in understanding the public uptake of mitigation measures or public concerns due to the pandemic. At this time of social isolation, social media has become the platform of choice for many people to express themselves which can be a valuable resource for study. Social media platforms have a strong worldwide presence of approximately 3.6 billion users as of January 2020 which accounts for nearly 49% of the global population per statistica.com.

During times of public health crises, social media platforms such as Facebook, Twitter, and Reddit are inundated with people’s perceptions, opinions, and concerns. Consequently, large data sources are created that allow researchers opportunities to mine meaningful information using various methodologies. Increasingly, researchers are using Natural Language Processing (NLP) methods such as topic modeling and Named-Entity Recognition to mine public health-related data. Similarly, we leveraged the publicly available posts from Reddit, a popular social media platform, to surveil public uptake and concerns relating to the mitigation measures and burden of the disease.

In this study, we collected posts from location-specific subreddits, micro-communities within Reddit platform, as a data source to monitor the COVID-19 pandemic in North Carolina. We employed NLP techniques, particularly LDA topic modeling and custom NER, to capture issues and concerns and examine the prevalent mitigation strategies about the pandemic. More
specifically, we wanted to answer the research question “how much can be learned regarding the public uptake of mitigation strategies and concerns about COVID-19 pandemic by using natural language processing on Reddit posts?”. The main contributions of this paper are as follows:

- We built a comprehensive cleaned corpus of COVID-19 pandemic-related posts from North Carolina subreddit communities using various NLP techniques.
- We trained and evaluated an extended, custom NER model to assess the uptake of mitigation measures against the spread of COVID-19 disease.
- We extracted people’s concerns about the pandemic using an LDA-based topic model, revealing interesting side topics related to the pandemic.

The rest of this paper is organized as follows. Section 2 outlines work related to this study. In Section 3, we present methods for constructing datasets and NLP-based experiments. Section 4 presents the results and limitations of this study, and in Section 5, we conclude and outline future work.

2 Related Work

NLP-based approaches to mine social media text include discourse analysis [1], sentiment classification [2], and topic modeling. Social media text mining was utilized to study public health crises such as Ebola and Zika. For instance, studies examined Twitter posts to identify the public’s concerns about the Zika [3] and Ebola [4] epidemics in the United States. The authors in [5] examined the public discourse and emergent themes regarding Malaria by leveraging text mining algorithms.

Concerning the COVID-19 public health crisis, recent studies used twitter data to track public opinions [10], reactions [6], perceptions [7], emotions [8], and concerns [9-13]; to detect topic and sentiment dynamics [14]; to track discourse, mental health status and symptoms [15, 16]; and to detect self-reported symptoms, access to tests, and recovery efforts of users [17]. Twitter has proven to be an alternative source for the surveillance of public health data, however, tweets contain limited contextual data due to its character length restriction. As such, other social media platforms have garnered the attention of researchers to exploit context rich data such as Sina Weibo and Reddit. For instance, [18] investigated Chinese citizens’ attention to COVID-19 related events by analyzing Sina microblog data. On the other hand, Reddit data have recently been used to observe mental health discourse and COVID-19 related health anxieties [19, 20], to track health related discussions for public health applications [21, 22], to track public COVID-19 concerns [23], and to analyze opioid related discussions [24].

Topic modeling, specifically Latent Dirichlet Allocation (LDA), and Named-Entity Recognition (NER) are increasingly used for NLP-based social media text analysis [25-28]. Regarding public health, Collier et al. [29] employed topic modeling and NER to detect and track the distribution of infectious disease outbreaks, demonstrating the suitability of these NLP techniques. Similarly, we analyzed Reddit data from four North Carolina subreddits using LDA topic modeling and a custom NER system to capture public concerns and mitigation strategies for the pandemic. To our knowledge, this is the first study to use both LDA topic modeling and a custom NER system on the Reddit posts to monitor the COVID-19 pandemic in North Carolina.

3 Methodology

We used application programming interfaces (APIs) and a set of predefined search terms (“corona virus”, “Coronavirus”, “COVID-19” and “SARS-CoV-2”) to extract 122,249 comments from 2,319 Reddit posts from four location-specific North Carolina subreddit communities from March 1, 2020 through August 31, 2020. To understand the public uptake of mitigation measures, detection strategies, and public concerns regarding the pandemic in North Carolina, we developed a custom NER model and an LDA-based topic model. Our custom NER model identifies key mitigation and detection measures in unstructured Reddit posts using three mitigation named-entity categories: distancing (DIST), disinfectant (DIT), and personal protective equipment (PPE); and two detection named-entity categories: symptoms (SYM) and testing (TEST). Our datasets and code are publically available at https://github.com/NCAT-NLP/ACM-BCB

3.1 Dataset Construction

The dataset construction involves data collection from Reddit platform, data preprocessing, building a corpus for topic modeling, and custom NER dataset construction.

3.1.1 Application Programming Interfaces. To programmatically access Reddit posts, we used Python Reddit API Wrapper 2 and Python Pushshift.io API Wrapper 3. In this study, the primary attributes concerning a Reddit/subreddit instance are ‘id’ and ‘name’. Regarding submission objects, the primary attributes are ‘id’, ‘created_utc’, ‘title’, and ‘num_comments’. More over, ‘submission (parent post)’, ‘body’, and ‘replies’ are the primary attributes for comment objects. To capture public concerns and mitigation strategies for the pandemic. To our knowledge, this is the first study to use both LDA topic modeling and a custom NER system on the Reddit posts to monitor the COVID-19 pandemic in North Carolina.

3.1.2 NC Dataset. To scrape and store the Reddit data, we generated Python scripts. Using our Reddit instance, we queried posts and comments containing the aforementioned COVID-19 keywords from March 1st through August 31st 2020 from the subreddits: r/Charlotte, r/raleigh, r/gso (Greensboro), and r/NorthCarolina. The first three subreddits represent communities of the top three populated cities in North Carolina, and r/NorthCarolina represents the subreddit for the entire state. Next, we looped through the query results (omitting duplicates) and saved the comment objects to a list. Using the list of comment objects, we then retrieved each comment’s parent submission (Reddit post) id, submission date, submission title, and the entire comment thread (including number of comments) associated with the parent post. Finally, we converted the acquired data to a table arranged by date, title, comment body, and number of comments. The final composition of our NC dataset is described in Table 1.

---

2 https://praw.readthedocs.io/en/latest/
3 https://github.com/dmarx/psaw
datasets were labeled using the BILUO annotation scheme. The composition of the resulting dataset is provided in Table 2 below.

### Table 2: Number of labeled entity tags

| Entity Label | # of Tags | Entity Label | # of Tags |
|--------------|-----------|--------------|-----------|
| DIST         | 1,354     | DIST         | 882       |
| TEST         | 5,240     | TEST         | 1,612     |
| SYM          | 3,519     | SYM          | 1,574     |
| DIT          | 1,562     | DIT          | 701       |
| PPE          | 3,873     | PPE          | 2,010     |
| O            | 202,853   | O            | 99,236    |

### 3.2 Named-Entity Recognition (NER)

NER is a sub-task of Information Extraction (IE) which aims to classify certain named-entities found within an unstructured body of a text corpus [30]. Entities found using the NER task are typically classified into pre-defined categories such as person, location, organization, medical codes, disease names, etc. NER models are required to be evaluated in order to check the validity of their performance by comparing the outputs typically against human-annotated tags. The comparisons are generally quantified simultaneously by attempting to correctly recognize a detected instance’s boundary and its entity type. Consequently, accurate predictions using the performance metrics below are assessed:

\[
P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = 2 \times \frac{P \times R}{P + R}
\]

The custom NER models trained for this study were based on spaCy’s multi-task convolutional neural network (CNN) which was trained using the OntoNotes corpus, and contain GloVe vectors [31] that were trained on Common Crawl.

#### 3.2.1 Training and Evaluation.

To train our custom NER models, we first created labels from our predetermined NER categories, provided a model name, and created an NER pipeline. We followed spaCy’s recommendations for fine-tuning the following hyperparameters: training iterations, batch size, and dropout rate. A batch size begins at a user defined minimum and each batch increases until it reaches a user defined maximum threshold. Dropout is a stochastic regularization technique that aims to reduce overfitting in neural networks by temporarily removing neurons during training [32]. Custom NER model 1 was trained with 30 iterations, a compounding batch size from 4 to 32, and a dropout rate of 0.5. Custom NER model 2 was trained with 50 iterations, a batch size from 1 to 16, and a dropout rate of 0.35. The adjusted hyperparameters used to train Custom NER model 3 were number of iterations and dropout rate which were 100 and an incremental decay from 0.6 to 0.35, respectively.

#### 3.2.2 Entity Detection.

To detect our named-entities, we performed custom NER task on our cleaned NC_dataset. We then extracted and stored the named-entities and their corresponding

---

4 https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html
5 https://www.emeditor.com/
6 https://spacy.io/api/annotation#biluo
7 https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf
8 https://spacy.io/api/annotation#biluo
3.3 Topic Modeling

A topic model is a statistical model that clusters Documents into topics by discovering hidden semantic structures in a text corpus. Although there are other topic modeling approaches such as Author-Topic Model (ATM) [33], the most common type of topic model is latent Dirichlet allocation (LDA), which requires the selection of a $k$ value that indicates the number of underlying topics in a text corpus [34]. Initially, LDA assigns each word in a corpus to a random $k$ topic, which is then iteratively updated based on the prevalence of each word across the $k$ topics. Once optimized, LDA employs the Term Frequency-Inverse Document Frequency (tf-idf) metric which assigns probabilities based on the number of occurrences of a word in a Document, and then offsets the probabilities according to the total number of Documents. The topic assignments are continuously updated until a user defined threshold is reached or until the iterations no longer have much impact on the probability assignments to the words in the corpus.

In this study, we designed and implemented our topic models using Python’s Scikit-learn library, assigning $k = 5$ topics to surveil the underlying topics of discussion of North Carolina residents during COVID-19 pandemic from their Reddit posts in four location-specific NC subreddits. Additionally, we obtained the word clouds of the top 15 frequent keywords for each topic and recorded the topic assignment frequency for each Document to aid in topic and theme interpretations.

3.3.1 Design and Implementation. The feature extraction algorithm that we used for this study was Python’s CountVectorizer from the Scikit-learn library. The parameters used were max_df = 0.90 and min_df = 3, by which terms that had a Document frequency strictly higher than 90% were ignored when building the vocabulary, and terms that had a Document frequency strictly lower than 3 were ignored when building the vocabulary, respectively. We then implemented our LDA model with Scikit-learn’s LatentDirichletAllocation class using the recommended optimization parameters. The number of $k$ topics were 5 (n_components) and the number of high frequency keywords were 15. By first conducting three preliminary runs using $k = 5$, 7, and 10, we determined that $k = 5$ yielded the most consistent number of coherent topics across all subreddits. Once the analysis completed, we exported the unique vocabulary size for the combined Documents and the high probability keywords for each of the $k$ topics to a text file. Additionally, the word clouds representing the high probability keywords for each of the $k$ topics were captured and saved as image files. Furthermore, the original datasets were appended with two additional columns that 1) recorded the topic number that each Document belonged to and 2) the probability that each Document was about the assigned LDA topic. The resulting text files, high frequency wordcloud images, and extended datasets were then used to interpret each of the $k = 5$ topics and their general themes.
Table 4: Identification of entities for 3 mitigation types (distancing, disinfectant, and PPE), and 2 detection types (symptoms and testing). Most distinct and frequently mentioned entities, and their contributing percentages are presented (Heading represents the four NC subreddits).

| Entity Type | Entity Name | # of Entities | Entity Type | Entity Name | # of Entities | Entity Type | Entity Name | # of Entities | Entity Type | Entity Name | # of Entities |
|-------------|-------------|---------------|-------------|-------------|---------------|-------------|-------------|---------------|-------------|-------------|---------------|
| **DIST**    | social distancing | 377 (52%)     | DIST        | social distancing | 500 (58%)     | DIST        | social distancing | 38 (67%)     | DIST        | social distancing | 446 (51%)     |
|             | lockdown     | 262 (36%)     |             | lockdown     | 293 (34%)     |             | lockdown     | 13 (23%)      |             | social distancing | 360 (41%)     |
| **DIT**     | hand sanitizer | 71 (27%)      | DIT         | hand sanitizer | 107 (27%)     | DIT         | hand sanitizer | 11 (19%)      | DIT         | hand sanitizer | 46 (23%)      |
|             | wipe         | 44 (17%)      |             | wipe         | 70 (17%)      |             | wipe         | 16 (27%)      |             | wipe         | 25 (12%)      |
| **PPE**     | mask         | 1944 (85%)    | PPE         | mask         | 3900 (86%)    | PPE         | mask         | 290 (82%)     | PPE         | mask         | 3188 (89%)    |
|             | glove        | 115 (5%)      |             | glove        | 202 (4%)      |             | glove        | 17 (5%)       |             | glove        | 80 (2%)       |
| **SYM**     | flu          | 349 (48%)     | SYM         | flu          | 543 (47%)     | SYM         | flu          | 49 (45%)      | SYM         | flu          | 505 (56%)     |
|             | cough        | 131 (18%)     |             | cough        | 221 (19%)     |             | cough        | 18 (17%)      |             | cough        | 142 (17%)     |
| **TEST**    | testing      | 1382 (61%)    | TEST        | testing      | 2051 (51%)    | TEST        | testing      | 190 (59%)     | TEST        | testing      | 1842 (57%)    |
|             | antibody test | 67 (3%)       |             | antibody test | 146 (4%)      |             | antibody test | 8 (2%)        |             | antibody test | 76 (3%)       |

Figure 2: Number of detected entities per entity type per subreddit

4.2.1 Detected Entity Analysis. Table 4 above lists the top two detected named-entities for each subreddit. Across all subreddits, social distancing, hand sanitizer, and mask were the most prevalent entities for DIST, DIT, and PPE, respectively. Moreover, flu and testing were the most prevalent COVID-19 detection related entities for the SYM and TEST categories, respectively. Regarding COVID-19 detection based entities, testing was the named-entity that encompassed the majority of mentions across all North Carolina subreddits with an average of 57.5%. Furthermore, we found that antibody test was the most prevalent detected named-entity concerning specific types of COVID-19 detection tests. This observation is significant as antibody tests can aid with tracking the spread of COVID-19, which according to [35], can lead to more accurate representations of the COVID-19 pandemic. Concerning the COVID-19 mitigation entities, mask denotes an overwhelming average of 85.5% detected across all subreddits. Thus, we focused on mask to conduct further analysis to examine the context in which it was being discussed.

After first extracting all mask related sentences (3,917) from the NorthCarolina subreddit, we conducted sentiment analysis using NLTK’s sentiment package to gauge sentiment. Once duplicate and incomplete sentences were removed, the results from the analysis yielded 1,340 neutral, 1,282 positive, and 1,202 negatively labeled sentences. In solely looking at labels, the comments would lean slightly neutral (positive when disregarding neutral), however, with no indication of strength. Thus, we took the average compound scores of all comments which resulted in a score of 0.007. Since a score of 1 would indicate highly positive and -1 highly negative, it is thus inferred that the comments from the NorthCarolina subreddit based on masks are essentially neutral. Due to the overwhelming neutrality of the comments, we focused on the comments labeled negative to further explore the context in which the users/comments were discussing masks. As we manually inspected the sentences, we begin to notice reoccurring themes and labeled them accordingly as referenced in Table 5.

We found that sentiment alone does not necessarily accurately depict the overall discourse or tone. For example, many of the comments that were in fact accurately labeled as negative were about adherence to wearing masks and mask related policies. In these instances, the users’ delivery was negative (e.g. “I always wear my mask, these freaking morons should do the same”), yet the discourse was in favor of masks. Moreover, other comments were about businesses and/or law enforcement enforcing the NC mask mandate and/or mask wearing in general. Other seemingly contrary themes for comments labeled negative include exemptions and mask efficacy. Conversely, themes that could be deemed neutral or associated with negativity regarding masks include noncompliance, politicizing masks, ineffectiveness, lack of enforcement, sarcastic remarks, selfishness, mask shaming, shortages, skepticism, anti-maskers, wearing masks improperly, hypocrisy, and discomfort.
Table 5: Reoccurring themes from mask related comments

| Themes       | Frequency |
|--------------|-----------|
| noncompliance| 186       |
| compliance   | 146       |
| politicized  | 117       |
| ineffective  | 115       |
| effective    | 66        |
| nonenforcement| 63     |
| enforcement  | 62        |
| sarcasm      | 62        |
| selfishness  | 59        |
| mandate      | 50        |
| shaming      | 43        |
| shortage     | 40        |
| skepticism   | 37        |
| anti-mask    | 32        |
| improper use | 27        |
| hypocrisy    | 25        |
| exemptions   | 24        |
| uncomfortable| 20        |

The final NER analysis that we conducted was to monitor the evolution of the most prevalent detected entities from the NorthCarolina subreddit dataset from March through August. As shown in Figure 3, hand sanitizer is the entity that has the least substantial progression during our observable time span. In fact, distancing (DIT) related entities are the least mentioned entities throughout all datasets. Flu and social distancing follows similar patterns in terms of month-to-month trends; with social distancing slightly peaking above flu in June. Testing is by far the entity that was discussed the most at the onset of the COVID-19 pandemic in the NorthCarolina subreddit and remained fairly steady until May. There was a modest decline from May to June, remained steady until July, then had a comparable decline from July to August as it did from May through June.

4.3 Topic Modeling Results

Prior to performing LDA topic modeling, we extracted features from the Documents in our NC_dataset. As a result, the number of unique vocabulary words for each of the subreddits are as follows: r/Charlotte – 10,744, r/raleigh – 14,031, r/gso – 3,143, and r/NorthCarolina – 13,611. Table 6 depicts the interpreted topics and themes for each subreddit and Figure 4 shows the word cloud representing the 15 high frequency words used to shape each of the $k = 5$ topics.

Figure 5(a)-(d): Number of assigned topics and their overall discussion percentages. Refer to Figure 4 and Table 6 for indexing.

We found that in nearly all cases, the top 15 frequent keywords sufficiently shaped clearly defined topics. In terms of topic frequency (Figure 5(a)-(d)), COVID-19 testing/cases/measures (57%) was the most discussed topic in r/Charlotte followed by Jobs/Work from Home (40%) and Wearing Masks (2%). Social justice (1%) was the least discussed. Regarding r/raleigh, we found that Wearing Masks (48%) was the most discussed topic followed by COVID-19 testing/cases/measures (24%), Shopping (18%), and Education (10%). Pets was the least discussed topic, representing less than 1% of the overall discussions. The most discussed topic in r/gso was COVID-19 testing/cases/measures (46%) followed by Wearing Masks (39%), housing (9%), and dining (4%). The least discussed topic in r/gso was Education (2%). Regarding the number of assigned prevalent topics.
Discussed in r/NorthCarolina, Masks / COVID-19 testing/cases/measures (61%) was the highest followed by Unemployment (28%) and Politics (7%). The lowest measurable topic discussed in r/NorthCarolina was Education (4%).

Table 6: Interpreted LDA topics with categories

| Subreddit Name | Topic Number | Topic Name | Category                  |
|----------------|--------------|------------|---------------------------|
| r/Charlotte    | 1            | Jobs / Work From Home | Employment               |
|                | 2            | Wearing Masks (mandate) | COVID-19 Mitigation Measures |
|                | 3            | Wearing Masks | COVID-19 Mitigation Measures |
|                | 4            | Social Justice/Law Enforcement | Law Enforcement |
|                | 5            | COVID-19 testing/cases/measures | Spread of COVID-19 Disease |
| r/raleigh      | 1            | Pets | Pets                       |
|                | 2            | Education | Education                  |
|                | 3            | Shopping | Business                   |
|                | 4            | Wearing Masks | COVID-19 Mitigation Measures |
|                | 5            | COVID-19 testing/cases/measures | Spread of COVID-19 Disease |
| r/gso (Greensboro) | 1         | Dining | Business                   |
|                | 2            | Housing | Housing                    |
|                | 3            | COVID-19 testing/cases/measures | Spread of COVID-19 Disease |
|                | 4            | Wearing Masks | COVID-19 Mitigation Measures |
|                | 5            | Education | Education                  |
| r/NorthCarolina | 1           | Masks / COVID-19 testing/cases/measures | COVID-19 Mitigation Measures / Spread of COVID-19 Disease |
|                | 2            | Thales Academy | Education                  |
|                | 3            | Unemployment | Employment                  |
|                | 4            | Education | Education                   |
|                | 5            | Politics/2020 Election | Politics                   |

Table 7: Dominant NC subreddit monthly side topics

| Month  | Topic Name                                |
|--------|-------------------------------------------|
| March  | Coronavirus Insider Trading Scandal       |
| April  | Unemployment                              |
| May    | Unemployment                              |
| June   | Law Enforcement/Social Justice            |
| July   | Education                                |
| August | 2020 Election/Early Voting                |

Figure 6: Top 15 frequency words associated with Table 7

To finalize our LDA topic modeling analysis, we used our NorthCarolina subreddit dataset as representation for the state of NC to determine the most prevalent side topics discussed on a per month basis. Our impression was that by dividing the dataset into months and performing LDA topic modeling with \( k = 2 \) topics on each, then one topic would encompass COVID-19 related keywords and the other would produce the most prevalent side topic. This proved true in all but two cases where instead of acquiring a COVID-19 related topic for the month of May, we received a topic that was mostly incoherent. The other instance was July where two COVID-19 related topics emerged, however, one of them had a distinguishable co-occurring topic (Education). Disregarding the incoherent and COVID-19 related topics, the results for the prevalent side topics on a month-to-month basis are shown in Table 7 and Figure 6.

4.4 Discussions

Our multi-pronged NLP study proved highly effective in surveilling the public uptake of mitigation/detection strategies and public concerns regarding the COVID-19 pandemic. In this study, our custom NER model effectively detected prevalent COVID-19 mitigation and detection-based entities from four location-specific North Carolina based subreddits. Moreover, the LDA topic modeling portion of this study also proved effective in determining the primary concerns of North Carolinians during the COVID-19 public health crisis. It was shown that the top discussion percentages in terms of topic frequency was related to COVID-19 mitigation and detection strategies. Furthermore, our LDA method detected interesting side topics related to unexpected consequences of the COVID-19 pandemic. For example, the citizens of Charlotte faced a heavier burden on jobs than in Raleigh or Greensboro, and the pandemic significantly impacted shopping and institutions of higher learning in Raleigh. The most significant indirect impacts of the COVID-19 pandemic in the subreddits of Greensboro and North Carolina were related to housing and unemployment, respectively. It was additionally shown that our LDA method was effective in revealing the most prevalent side topics in the North Carolina subreddit on a month-to-month basis, heavily corresponding with real-time events that occurred in North Carolina during those months.
5 Conclusion

In this study, we used six months of Reddit data to survey the COVID-19 pandemic in North Carolina by employing LDA topic modeling and a custom NER system with much success. Based on our sample population, we found that the North Carolina public is most concerned with mitigating the spread of COVID-19 by wearing masks, adhering to social distancing guidelines, and using hand sanitizer. We further observed that the most prevalent topics of discussion among the North Carolina public were wearing masks and the spread of COVID-19. Our experiments have demonstrated 1) the effectiveness of using a custom NER system on Reddit posts to monitor prevalent COVID-19 detection and the uptake of mitigation strategies in North Carolina, and 2) the efficacy of employing LDA topic modeling to discover the underlying concerns of North Carolinians during COVID-19 pandemic using Reddit posts.

REFERENCES

[1] Lipizzi, C., Dessavre, D. G., Iandoli, L., & Marquez, J. E. R. (2016). Towards computational discourse analysis: A methodology for mining twitter backchanneling conversations. Computers in Human Behavior, 64, 782-792.

[2] Jelodar, H., Wang, Y., Orji, R. and Huang, H. 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. arXiv preprint arXiv:2004.11695.

[3] Glowacki, E.M., Lazard, A.J., Wilcox, G.B., Mackert, M. and Bernhardt, J.M. 2016. Identifying the public's concerns and the Centers for Disease Control and Prevention’s reactions during a health crisis: An analysis of a Zika live Twitter chat. American journal of infection control 44, 1709-1711.

[4] Lazard, A.J., Scheinfeld, E., Bernhardt, J.M., Wilcox, G.B. and Suran, M. 2015. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. American journal of infection control 43, 1199-1111.

[5] Boit, J. and El-Gayar, O. 2020. Topical mining of malaria using social media. A text mining approach. In Proceedings of the 53rd Hawaii International Conference on System Sciences.

[6] Long, Z., Alharthi, R. and El Saddik, A. 2020. NeedFull –a Tweet Analysis Lipizzi, C., Dessavre, D. G., Iandoli, L., & Marquez, J. E. R. (2016). Towards computational discourse analysis: A methodology for mining twitter backchanneling conversations. Computers in Human Behavior, 64, 782-792.

[2] Jelodar, H., Wang, Y., Orji, R. and Huang, H. 2020. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. arXiv preprint arXiv:2004.11695.

[3] Glowacki, E.M., Lazard, A.J., Wilcox, G.B., Mackert, M. and Bernhardt, J.M. 2016. Identifying the public's concerns and the Centers for Disease Control and Prevention’s reactions during a health crisis: An analysis of a Zika live Twitter chat. American journal of infection control 44, 1709-1711.

[4] Lazard, A.J., Scheinfeld, E., Bernhardt, J.M., Wilcox, G.B. and Suran, M. 2015. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. American journal of infection control 43, 1199-1111.

[5] Boit, J. and El-Gayar, O. 2020. Topical mining of malaria using social media. A text mining approach. In Proceedings of the 53rd Hawaii International Conference on System Sciences.

[6] Long, Z., Alharthi, R. and El Saddik, A. 2020. NeedFull –a Tweet Analysis Lipizzi, C., Dessavre, D. G., Iandoli, L., & Marquez, J. E. R. (2016). Towards computational discourse analysis: A methodology for mining twitter backchanneling conversations. Computers in Human Behavior, 64, 782-792.