Leveraging Natural Language Processing to Mine Issues on Twitter During the COVID-19 Pandemic

Ankita Agarwal[§]
Department of Computer Science
Wright State University
Dayton, Ohio 45435 USA
agarwal.15@wright.edu

Preetham Salehundam[§]
Department of Computer Science
Wright State University
Dayton, Ohio 45435 USA
salehundam.2@wright.edu

Swati Padhee
Department of Computer Science
Wright State University
Dayton, Ohio 45435 USA
padhee.2@wright.edu

William L. Romine
Department of Biological Sciences
Wright State University
Dayton, Ohio 45435 USA
william.romine@wright.edu

Tanvi Banerjee
Department of Computer Science
Wright State University
Dayton, Ohio 45435 USA
tanvi.banerjee@wright.edu

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378028

***Abstract*—The recent global outbreak of the coronavirus disease (COVID-19) has spread to all corners of the globe, introducing numerous social challenges. Twitter platforms have been used to identify public opinion about events at the local and global scale. In this study, we constructed a system to identify the relevant tweets related to the COVID-19 pandemic throughout January 1st, 2020 to April 30th, 2020 and explored topic modeling to identify the most discussed topics and themes during this period. Additionally, we analyzed the temporal changes in the topics with respect to the events that occurred. We found eight topics were sufficient to identify the themes in our corpus. The dominant topics were found to vary over time and align with the events related to the COVID-19 pandemic.**

***Index Terms*—Coronavirus disease 2019 (COVID-19), machine learning, natural language processing, public health, social media, topic modeling**

I. Introduction

The coronavirus "SARS-CoV-2", a deadly acute atypical respiratory disease, took over the world from December 2019. The disease caused by this virus was called Coronavirus disease 2019 (COVID-19), and a pandemic was declared by the World Health Organization (WHO). The disease was found to be highly contagious, with a reproduction rate of 2.2 (R_0) [1].

Social media allow access to timely information about disease symptoms and its prevention. Recent studies [2], [3] have shown that social media play an essential role in understanding public attitudes and behaviors during crises. In the case of strong emotional reactions, media coverage of the pandemic may influence public sentiments [4].

Efficient identification of thoughts, attitudes, feelings, and concerns about the COVID-19 pandemic can help policymakers, health care professionals, and the public identify concerns and address them [5]. Recent studies have identified the topics of discussions, concerns, and controversies surrounding COVID-19 in social media [6], [7].

We investigate how these conversations are a part of the discourse for different events unfolding during COVID-19. We explore: (1) The feasibility of differentiating between relevant and non-relevant tweets with respect to the COVID-19 pandemic; (2) trending topics in public discussions on Twitter related to the COVID-19 pandemic; and (3) the relationship between COVID-19 events and the trending topics on Twitter.

II. Related Works

A. Relevant Tweet Extraction

Previous works have focused on analyzing online behavior and reactions to the spread of COVID-19 [6], or investigating conspiracy theories and social activism [8], [9]. Often studies ignore the need to preprocess the tweets to filter irrelevant tweets [10]. Relevant tweets provide both situational and actionable information to aid decision-making and immediate assistance to the affected people [11], [12]. While [13] have identified tweets relevant to influenza, [14] have worked on filtering out noisy tweets by identifying relevant tweets for the Zika virus using text-based features. Wahbeh et al. [15] have utilized a qualitative approach to identify relevant tweets during COVID-19. We experiment with machine learning and deep learning language models to filter out the relevant tweets for COVID-19.

B. Topic Modeling and Analysis

Topic modeling has been applied in areas like health informatics [16] and social media networks [17] in order to organize and summarize large textual information and to uncover hidden thematic structures in a collection of documents [18]. Recent works with disease include topic modeling on documents related to the Ebola [19], Zika virus [20], dengue [21], and seasonal influenza [22] epidemics. Others have explored the feasibility of topic modeling on data from Facebook [23], Weibo [24] and Google trends [25]. Medford et al. [26] studied the change in topics throughout January

[§] co-first authors

14th to 28th, 2020 using Twitter data related to COVID-19. Ordun et al. [27] explored a time series based analysis, and Yin et al. [28] explored Dynamic Topic Modeling to study the trending topics in tweets for the pandemic. Chandrasekaran et al. [29] outlined the topics of discussion on Twitter related to the COVID-19 pandemic and the sentiment change in each topic over weeks. Building on prior work, we attempt to understand how public discussions on social media about COVID-19 resonate with the related events.

III. Data

A. Data Collection

We utilized the COVID-19 pandemic-specific keywords like “coronavirus”, “covid-19”, “sars-ncov” to extract tweets posted from January 1st 2020 to April 30th 2020. An open-source crawler library, getOldTweets[1], and Twitter Streaming API were used to extract the tweets. The getOldTweets library was used to overcome the daily quota limitations of the Twitter API. This study was limited to English language tweets alone, and a total of 957,923 tweets were collected and preprocessed as described in Section III-C. Consequently, 866,527 unique tweets were obtained[2].

B. Data Annotation

When tweets were collected, there were still tweets not related to COVID-19, which contain the specified keywords. We randomly sampled 1,500 unique tweets out of 866,527 total tweets, and three annotators labeled them as “Relevant” or “Irrelevant”. A tweet was labeled as “Relevant” if it contained information about the spread, cause, effect, opinion, sentiment, emotion with regards to COVID-19; otherwise, it was labeled as “Irrelevant”. Out of 1,500 tweets, 1,154 (77%) tweets were labeled as “Relevant” and 346 (23%) as “Irrelevant”, which served as our dataset for building a relevancy classifier.

C. Data Preprocessing

To understand the context and semantics of the content, we first converted the tweets to lowercase and then removed the emoticons, stop words, punctuations, and URLs, including symbols like “#” ,“@” as well as special characters. We also removed the keywords used to collect the tweets to avoid any bias. Ten percent of the total collected data were duplicate and filtered out to obtain 866,527 unique tweets. We further tokenized the tweets using the Penn Treebank tokenizer, which uses regular expressions to tokenize texts similar to the tokenization used in the Penn Treebank[3]. These tokens were then passed through a Porter stemmer[4] to reduce the word library size.

IV. Methodology

Fig. 1 presents an overview of the sequence of steps involved in our analysis. To design the relevancy classifier, two baseline traditional machine learning algorithms using Term Frequency - Inverse Document Frequency (TF-IDF) embeddings and contextual language model embeddings were evaluated and further discussed in Sections IV-A and IV-B. We then performed topic modeling on the predicted relevant tweets to estimate the probability of a tweet belonging to a unique topic and understand the hidden semantic similarities between the tweets.

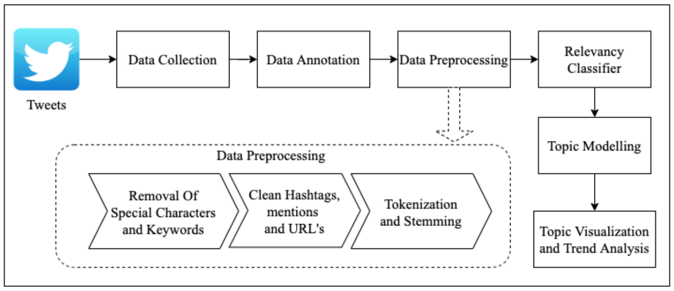
Tweets
Data Collection
Data Annotation
Data Preprocessing
Relevancy Classifier
Topic Modelling
Topic Visualization and Trend Analysis
Data Preprocessing
Removal Of Special Characters and Keywords
Clean Hashtags, mentions and URL's
Tokenization and Stemming

Figure 1. Overview of the Framework to explore most discussed topics about the COVID-19 pandemic on Twitter.

A. Feature Representation

We utilized two types of features in this study:

- *Context-Free Embedding:* The vector representation of TF-IDF attempts to evaluate the importance of a word by reducing the weight of words that occur more frequently in the data set and increasing the weight of less frequent words. However, TF-IDF still fails to capture the meaning (context) of the same word at different places (polysemy).
- *Language Models (Contextual Embedding):* Recent developments in language modeling enabled capturing the context of a word by looking at an input sequence and deciding at each step which other parts of the sequence are essential. Bidirectional Encoder Representations from Transformers (BERT) introduced a novel technique of masked language models (MLM) and next sentence prediction [30] to learn the distributional contextual representation of words.

B. Relevancy Classification

We used the 1,500 annotated tweets to build a relevancy classifier. Due to class imbalance (Relevant:77%, Irrelevant:23%), we used the synthetic minority over-sampling technique (SMOTE) [31] to generate synthetic samples of minority class and train the relevancy classifier on 1,154 relevant and 346 irrelevant tweets. After oversampling, we split the dataset into train, test, and validation subsets with a distribution of 75%, 15%, and 10%, respectively. We used a 5-fold cross-validation technique to confirm the generalizability of the model. Both the feature representations discussed in Section IV-A were used to train two traditional classification

[1] https://github.com/Mottl/GetOldTweets3

[2] https://github.com/preetham-salehundam/COVID-TOPIC-ANALYSIS

[3] https://web.archive.org/web/19970614160127/http://www.cis.upenn.edu/ treebank/

[4] https://tartarus.org/martin/PorterStemmer/

algorithms: Support Vector Classification (Linear SVC) and Logistic Regression (LR) to filter the relevant tweets for COVID-19 from the collected tweets. TF-IDF vector representations were generated using Scikit-Learn[5] pipeline and BERT based sentence embeddings were generated using Sentence-Transformers[6]. A pre-trained *'bert-large-uncased'* [32] model with an embedding dimension of 768 was used to generate the embeddings from the tweets. Additionally, we performed a Principal Component Analysis (PCA) on the embeddings and reduced the dimensionality of the feature space to 300 to reduce the chances of overfitting.

C. Topic Modeling

Latent Dirichlet Allocation (LDA) [18] is an unsupervised probabilistic model to automatically identify the hidden themes (topics) that are the best representation of the data set. After preprocessing the relevant tweets, we used a coherence score to find the optimal number of topics [33]. The Coherence Model[7] was used for this purpose based on which an 8-topic LDA model was selected.

V. RESULTS

A. Relevancy Classifier (Research Question 1)

The performance of the SVC classifier with BERT embedding was better than the other settings (Table I). The best performing SVC classifier on BERT embeddings showed precision and recall of 93% and classification accuracy of 93% in the test set and precision and recall of 95% with a classification accuracy of 95% in the validation set. This model was selected to generate predictions on the entire data set of 865,027 remaining tweets (excluding the 1,500 annotated tweets). A total of 688,825 (79.6%) tweets were classified as relevant tweets. We used this dataset for topic modeling, visualization, and trend analysis.

Table I
RELEVANCY CLASSIFIER RESULTS.
(ACC = ACCURACY, F1 = F1 SCORE, P = PRECISION, R= RECALL)

	Test Set				**Validation Set**			
	TFIDF Vectors							
Model	Acc(%)	F1(%)	P(%)	R(%)	Acc(%)	F1(%)	P(%)	R(%)
LR + TFIDF	90	90	91	90	93	93	92	93
SVC + TFIDF	87	87	88	87	89	89	89	88
	Contextual BERT Embeddings							
LR + BERT	88	87	87	88	89	89	89	88
SVC + BERT	**93**	**92**	**93**	**92**	**95**	**95**	**95**	**95**

B. Topic Analysis (Research Question 2 and 3)

The coherence score increased until eight topics and then gradually decreased; hence we chose eight to be the optimal number of topics for our data set. The percentage of tweets discussing each topic in the corpus is shown in Table II.

Table II
TOPIC DISTRIBUTION IN THE DATA SET

Topic Number	**Theme**	**Tweets(%)**
1	Pandemic impact and reopenings	11.7
2	Government response	12.8
3	Health workers and authorities	14.5
4	Federal help and quarantine	6.0
5	Origin of novel coronavirus	12.5
6	People's thoughts during COVID-19	21.1
7	Case statistics	14.5
8	Lockdown and stay at home	6.9

The results of running an LDA model on all the relevant tweets for eight topics and manually curated themes summarizing the top ten most relevant terms along with some representative tweets for each topic are discussed in Table III. We discuss the themes for each topic as curated based on the top 10 relevant terms below.

- **Pandemic impact and reopenings:** As shown in Table III, the keywords and representative tweets for Topic 1 represent the significant impact of COVID-19 on the economy, businesses, education, employment, and other aspects of public life. It also contained tweets that discussed the possibilities and consequences of reopening certain closed businesses due to the COVID-19 pandemic.
- **Government response:** Topic 2 was identified as the government response during the COVID-19 pandemic, which might indicate the social media discussions about the response of all government agencies over COVID-19 from all over the world during the press briefings.
- **Health workers and authorities:** During the COVID-19 pandemic, the role of health officials, organizations, and workers have been prominent, and the keywords in Topic 3 imply the same. Moreover, it also covered the pharmaceutical interventions associated with COVID-19.
- **Federal help and quarantine:** In order to facilitate the multiple crisis scenarios caused by this global pandemic, various federal organizations came up with relief plans for proper pandemic disaster management to help their citizens. At the same time, people were quarantined to control the spread of the virus. The keywords in Topic 4 suggest the theme 'Federal help and quarantine' to contain such tweets.
- **Origin of novel coronavirus:** The keywords in Topic 5 mostly pointed towards the tweets mentioning the origin of novel coronavirus, which was believed to be the Hubei province of China, from where it spread all across the world.
- **People's thoughts during COVID-19:** Topic 6 represented a significant portion of the tweets discussing

[5] https://scikit-learn.org/stable/modules/classes.html
[6] https://github.com/UKPLab/sentence-transformers
[7] https://radimrehurek.com/gensim/models/coherencemodel.html

Table III
THEMES RELATED TO EACH TOPIC, WORDS AND REPRESENTATIVE TWEETS FOR EACH TOPIC

Topic Number	Theme	Words	Representative tweets
1	Pandemic impact and reopenings	president, lockdown, business, american, reopen, economy, close, school, child, leader	• WHO advice for international travel and trade in relation to the outbreak of pneumonia caused by a new #coronavirus in #Wuhan, #China. • China businesses, markets reopen, but many people stay home as coronavirus cases on rise again
2	Government response	trump, pandemic, people, governor, force, white, video, happen, america, donald	• Unfortunately under this trump administration, I don't believe anything they say¡No reason for Americans to panic': White House seeks to calm fears over coronavirus. • Xi Jinping has turned invisible during China's coronavirus epidemic, likely to cover his back in case things go badly wrong
3	Health workers and authorities	state, pandemic, health, worker, response, american, youtube, public, outbreak, briefing	• As public health officials work to respond to the Coronavirus in Washington, the number of contacts the patient came in close contact with rises to 50. • World Health Organization discusses the novel coronavirus USA TODAY
4	Federal help and quarantine	crisis, claim, give, federal, quarantine, community, high, help, life, provide	• UK is badly prepared for recession • PANIC!!!THEIR GREATEST FEAR. PUBLIC AWAKENING Bill Gates said movements towards nationalism worsened government response to the coronavirus crisis - Business Insider
5	Origin of novel coronavirus	vaccine, house, virus, pandemic, china, april, never, social, medium, chinese	• Appears the cause of the #Wuhanpneumonia is as suspected, a #novel coronavirus. Curious since it appears to have connection to seafood market if it is zoonotic in etiology. Perfect example of why #OneHealth and #healthsecurity are important. • Chinese government scientists have identified a new coronavirus as the cause of a mystery pneumonia outbreak in Wuhan
6	People's thoughts during COVID-19	people, go, think, thing, would, everyone, really, get, right, still	• 3700 people were presumed dead from COVID-19 just like that. Something is fishy, are we just inflating figures for political reasons... • We have more than enough data. We know that basically EVERY projection was wrong about the coronavirus. So....now that we have MORE THAN ENOUGH INFO...IS THE CORONA VIRUS WORSE THAN THE FLU?
7	Case statistics	death, case, test, report, number, patient, hospital, total, update, positive	• At least 835 coronavirus cases diagnosed in China, 25 dead • #BREAKING UK virus death toll rises by 888 to 15,464: health ministry
8	Lockdown and stay at home	realdonaldtrump, order, china, record, increase, company, stayhome, south, fight, lockdown	• Coronavirus: China puts millions in lockdown amid rising deaths via @YouTube why Pakistan Govt does not buy screening virus equipments from China and why let people die in large. • We may all have second and third waves. These can only be controlled by lockdown measures. Add the fact that Covid19 may be endemic. So if we can't find a vaccine, then we have to develop a new way of living. A new world awaits and it's not going to be good.

various aspects of people's thoughts during the COVID-19 pandemic as it progressed.

- **Case statistics:** We observed that Topic 7 could be identified as case statistics during COVID-19 as it predominantly depicted the set of tweets discussing the number of people affected by COVID-19, number of deaths and number of recovered patients due to this pandemic.
- **Lockdown and stay at home:** Lockdown and stay at home advisories had been issued during the COVID-19 pandemic to reduce the exponential spread of the virus. Our topic model identified tweets that talked about these advisories as observed by the generated keywords in Topic 8.

Based on the keywords in each topic, individual tweets were labeled with the prominent topic using our LDA model, and the percentage of tweets belonging to each topic for every week was calculated.

In order to understand the trends in the topics, we utilized the events[8] during COVID-19, as listed by The American Journal of Managed Care (AJMC)[9] and discuss the details in Section VI-A3. Fig. 2 shows the dominant topics every week, possibly due to the influence of these events.

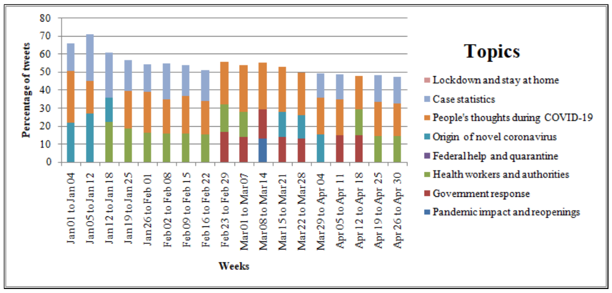

Figure 2. Prominent topics trending with time

VI. DISCUSSION

A. Principal Findings

1) Data Collection and Annotation Analysis: In the early weeks of January, we observed that the tweets with keywords like coronavirus and COVID-19 were very low compared to later weeks. The volume of tweets increased from the 4th week of January and continued to remain considerably high throughout February and decreased during April. It was also observed that the ratio of relevant to irrelevant tweets in the entire dataset (3.9) was similar to the manually labeled dataset (3.4), which shows that the distribution of the labeled data is similar to the data annotated by the classifier.

[8] https://git.io/JTpug
[9] https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

2) Relevancy Classifier Analysis: It was observed that some of the tweets collected using COVID-19 specific keywords were irrelevant to the COVID-19 pandemic. A Support Vector Classifier was used to design a relevancy classifier which could filter out the irrelevant tweets, and tweets like “olcokcevat malnla cmertsrrnla cimri olmal veren azizsr veren zelil olur”, “nojo demais dessa pessoa” which were tweeted using English alphabets.

3) Topic Analysis: Overall, among all the eight topics, People’s thoughts during COVID-19, Case statistics, and Health workers and authorities were predominantly discussed. This may be due to the fact that as COVID-19 has been declared a pandemic by World Health Organization[10], rising cases and the role of health agencies were the main concern over social media. Several conspiracy theories were given by people [9], and their reactions on economic, societal, and political areas came to fore throughout the pandemic [34]. We summarize our key observations of the topics with respect to the events[11] during COVID-19 below:

- **Jan 01 to Jan 11:** During the first two weeks of January, the World Health Organization had announced the news about the clusters of pneumonia like cases of unknown origin in Wuhan, China and travel restrictions were a concern. As a result, among all the topics discovered on our data set, “People’s thoughts during COVID-19” was the most discussed one. Tweets on Twitter during this week like “Only very limited information is available, but this looks like an outbreak of what could be a new respiratory (corona?)virus” mainly talked about some symptoms and general perceptions about this virus by the people. Topics like “Origin of novel coronavirus” and “Case statistics” were also prominent as a majority of the tweets talked about the origin of this virus and the number of people infected with it by that time.
- **Feb 23 to Feb 29:** According to CDC, as COVID-19 was heading towards a pandemic stage, “Government response” topic emerged as one of the prominent topics. Tweets like “You thought the coronavirus was bad? Just look forward to the next 8 months of it being a political weapon used by the US presidential candidates” depicted the political impact of COVID-19 at this time.
- **Mar 08 to Mar 14:** On March 11, WHO declared this novel coronavirus as a global pandemic. As a result, the topic, “Pandemic impact and reopenings” emerged as one of the highly discussed topics of this time. People started to discuss the effect of this pandemic as evident by the tweet “Stock market news live: Oil crashes, stock futures crater on coronavirus, crude war fears #stocks #markets”. Press briefings by the government officials and their response on COVID-19 became a prominent topic on social media.
- **April 12 to April 18:** During this time, US President Donald Trump started planning for the reopening of the economy according to the “gating criteria” . Besides, he announced the halt of funding to the World Health Organization (WHO) as he held the organization responsible for not giving proper information about the pandemic earlier[12]. As a result, “Health workers and authorities” topic emerged once again as a dominant topic along with “Government response”.
- **April 19 to April 30:** During the last weeks of April, people started deferring their treatments due to cost concerns. “Health workers and authorities” was still one of the predominantly discussed topics as NIH trials had shown the drug, remdesivir, to be effective for the treatment of COVID-19. Promoting telehealth as mentioned in the tweet “NEW: Bipartisan lawmakers back efforts to expand telehealth services for seniors to help combat the coronavirus” might have also led to the increased discussion related to the “Health workers and authorities” topic during this time.

B. Limitations

As the COVID-19 pandemic is a global pandemic, the English language constraint restricted our data collection to users who tweeted in the English language alone, and as a result, the findings of this study may not be generalizable worldwide. We cannot make causal claims about the correlation between certain topics and events mentioned in this paper. However, these findings help describe the relationship between COVID-19-related timeline events and discussions on Twitter, and exploration of the usefulness of Twitter data for actual detection of events related to COVID-19 is an essential next step.

VII. Conclusion

While urgent actions are needed to mitigate the potentially devastating effects of COVID-19, they can be supported by understanding the behavioral and social impact on the people. Because the pandemic imposes significant psychological burdens on individuals, insights from analyzing public conversations’ trends can be used to help align public emotions with the recommendations of epidemiologists and public health experts. To understand the possible effect of media coverage on public emotions, we present an analysis linking the trending topics of conversations on Twitter to events during the pandemic. Such analysis can provide insights so that a positive frame could be designed in the media coverage of the pandemic to educate the public and relieve negative emotions while increasing compliance with public health recommendations. The officials monitoring the spread of the pandemic can judge the influence of some false news or rumors related to COVID-19 as they could monitor the topics trending at a given time and thus provide appropriate information Empirical and qualitative evaluations of our analysis indicated that our analysis is trustworthy and we hope our data will contribute to precise and definitive social data mining to assist humanitarian organizations during global pandemics.

[10] https://covid19.who.int/

[11] https://git.io/JTpug

[12] https://www.bbc.com/news/world-us-canada-52289056

References

[1] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Trop. Med. Int. Health*, vol. 25, no. 3, pp. 278–280, Mar. 2020.

[2] Z. Shah and A. G. Dunn, "Event detection on twitter by mapping unexpected changes in streaming data into a spatiotemporal lattice," pp. 1–1, 2019.

[3] M. S. Steffens, A. G. Dunn, K. E. Wiley, and J. Leask, "How organisations promoting vaccination respond to misinformation on social media: a qualitative investigation," *BMC Public Health*, vol. 19, no. 1, p. 1348, Oct. 2019.

[4] J. J. Van Bavel, K. Baicker, P. S. Boggio, V. Capraro, A. Cichocka, M. Cikara, M. J. Crockett, A. J. Crum, K. M. Douglas, J. N. Druckman *et al.*, "Using social and behavioural science to support covid-19 pandemic response," *Nature Human Behaviour*, pp. 1–12, 2020.

[5] O. Belkahla Driss, S. Mellouli, and Z. Trabelsi, "From citizens to government policy-makers: Social media data analysis," *Government Information Quarterly*, vol. 36, no. 3, pp. 560 – 570, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0740624X18302983

[6] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study (preprint)."

[7] L. Chen, H. Lyu, T. Yang, Y. Wang, and J. Luo, "In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for COVID-19," Apr. 2020.

[8] E. Ferrara, "What types of COVID-19 conspiracies are populated by twitter bots?" Apr. 2020.

[9] S. Shahsavari, P. Holur, T. R. Tangherlini, and V. Roychowdhury, "Conspiracy in the time of corona: Automatic detection of covid-19 conspiracy theories in social media and the news," Apr. 2020.

[10] F. R. Lamy, R. Daniulaityte, R. W. Nahhas, M. J. Barratt, A. G. Smith, A. Sheth, S. S. Martins, E. W. Boyer, and R. G. Carlson, "Increases in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis," *International Journal of Drug Policy*, vol. 44, pp. 121–129, 2017.

[11] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events," 2010.

[12] S. Padhee, T. K. Saha, J. Tetreault, and A. Jaimes, "Clustering of social media messages for humanitarian aid response during crisis," *arXiv preprint arXiv:2007.11756*, 2020.

[13] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter," in *Proceedings of the 2011 Conference on empirical methods in natural language processing*, 2011, pp. 1568–1576.

[14] R. Muppalla, M. Miller, T. Banerjee, and W. Romine, "Discovering explanatory models to identify relevant tweets on zika," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2017, pp. 1194–1197, Jul. 2017.

[15] A. Wahbeh, T. Nasralah, M. Al-Ramahi, and O. El-Gayar, "Mining physicians' opinions on social media to obtain insights into COVID-19: Mixed methods analysis (preprint)."

[16] A. Zalewski, W. Long, A. E. W. Johnson, R. G. Mark, and L.-W. H. Lehman, "Estimating patient's health state using latent structure inferred from clinical time series and text," *IEEE EMBS Int Conf Biomed Health Inform*, vol. 2017, pp. 449–452, Feb. 2017.

[17] D. A. Ostrowski, "Using latent dirichlet allocation for topic modelling in twitter," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, Feb. 2015, pp. 493–497.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, "Detecting themes of public concern: a text mining analysis of the centers for disease control and prevention's ebola live twitter chat," *Am. J. Infect. Control*, vol. 43, no. 10, pp. 1109–1111, Oct. 2015.

[20] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What are people tweeting about zika? an exploratory study concerning its symptoms, treatment, transmission, and prevention," *JMIR Public Health Surveill*, vol. 3, no. 2, p. e38, Jun. 2017.

[21] P. Missier, A. Romanovsky, T. Miu, A. Pal, M. Daniilakis, A. Garcia, D. Cedrim, and L. da Silva Sousa, "Tracking dengue epidemics using twitter content classification and topic modelling," in *Current Trends in Web Engineering*. Springer International Publishing, 2016, pp. 80–92.

[22] I. Kagashe, Z. Yan, and I. Suheryani, "Enhancing seasonal influenza surveillance: Topic analysis of widely used medicinal drugs using twitter data," *J. Med. Internet Res.*, vol. 19, no. 9, p. e315, Sep. 2017.

[23] A. S. Raamkumar, S. G. Tan, and H. L. Wee, "Measuring the outreach efforts of public health authorities and the public response on facebook during the COVID-19 pandemic in early 2020: Cross-Country comparison (preprint)."

[24] X. Han, J. Wang, M. Zhang, and X. Wang, "Using social media to mine and analyze public opinion related to COVID-19 in china," *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, Apr. 2020.

[25] S. Bhattacharya and S. Singh, "Visible insights of the invisible pandemic: A scientometric, altmetric and topic trend analysis," Apr. 2020.

[26] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An "infodemic": Leveraging High-Volume twitter data to understand public sentiment for the COVID-19 outbreak."

[27] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, UMAP, and DiGraphs," May 2020.

[28] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," Jul. 2020.

[29] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, "Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study," *J Med Internet Res*, vol. 22, no. 10, p. e22624, Oct 2020. [Online]. Available: http://www.jmir.org/2020/10/e22624/

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[32] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[33] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," *conference on Web search and data . . .*, 2015.

[34] G. Bonaccorsi, F. Pierri, M. Cinelli, A. Flori, A. Galeazzi, F. Porcelli, A. L. Schmidt, C. M. Valensise, A. Scala, W. Quattrociocchi, and F. Pammolli, "Economic and social consequences of human mobility restrictions under covid-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 530–15 535, 2020. [Online]. Available: https://www.pnas.org/content/117/27/15530