# Extracting Physical Events from Digital Chatter for Covid-19

Vikram Nagapudi
*ArchBishop Mitty High School*
*San Jose, USA*
vikramnagapudi2004@gmail.com

Ameeta Agrawal
*Department of Computer Science*
*Portland State University*
*Portland, USA*
ameeta@cs.pdx.edu

Nirupama Bulusu
*Department of Computer Science*
*Portland State University*
*Portland, USA*
nbulusu@pdx.edu

2021 IEEE International Conference on Smart Computing (SMARTCOMP) | 978-1-6654-1252-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMARTCOMP52413.2021.00082

***Abstract*—By June 3, 2021, the US experienced over 33 million total cases of Covid-19, surpassing 592,000 deaths. In response, the Centers for Disease Control and Prevention (CDC) advised masking, social distancing and avoiding mass gatherings. In this work, we seek to automatically identify physical mass gathering events including dates and locations from digital chatter, i.e., social media data. We also study spread and sentiment associated with such large gathering events, finding a moderate negative correlation between large public gatherings, overall sentiment, and reported Covid-19 case numbers post event.**

***Index Terms*—event identification, social media, Covid-19, natural language processing**

## I. Introduction

Natural Language Processing (NLP) can play a vital role in combating the Covid-19 global pandemic [1], enabling automated extraction of critical information from large unstructured text volumes, like electronic health and clinical trial records, news and adverse event reports, scientific pre-prints, and social media [2]–[4]. In this paper, we focus on automating the study of *physical large gathering events*, their perceived sentiment, and their impact on the spread of Covid-19 using NLP on widely available but noisy social media data.

CDC guidelines for Covid-19 advise avoiding mass gatherings that exponentially increase cases, or "super-spreader events" [5], [6]. Understanding such events is an ongoing research effort, with outbreaks studied at specific sites like choirs, food processing plants, sport arenas and university campuses [7]–[9]. Because smartphone data availability as in [9] is restricted, previous studies on profiling large events have, employed time and labor-intensive alternate methods [5], [10].

Economists, journalists, and sociologists seeking to study how large events impact Covid-19 spread face multiple challenges. How to select events to profile without human bias? How to study known (e.g., conventions) and spontaneous events (e.g., protests)? How to account for the phase lag between event occurrence and county-level case detection, caused by delays in Covid-19 symptom onset and testing? In each case, experts must manually select events, ascertain location, county case data, and then analyze their impact, limiting study scope to specific populations, locations or events (e.g. Trump rallies [10], Sturgis motorcycle rally [5]).

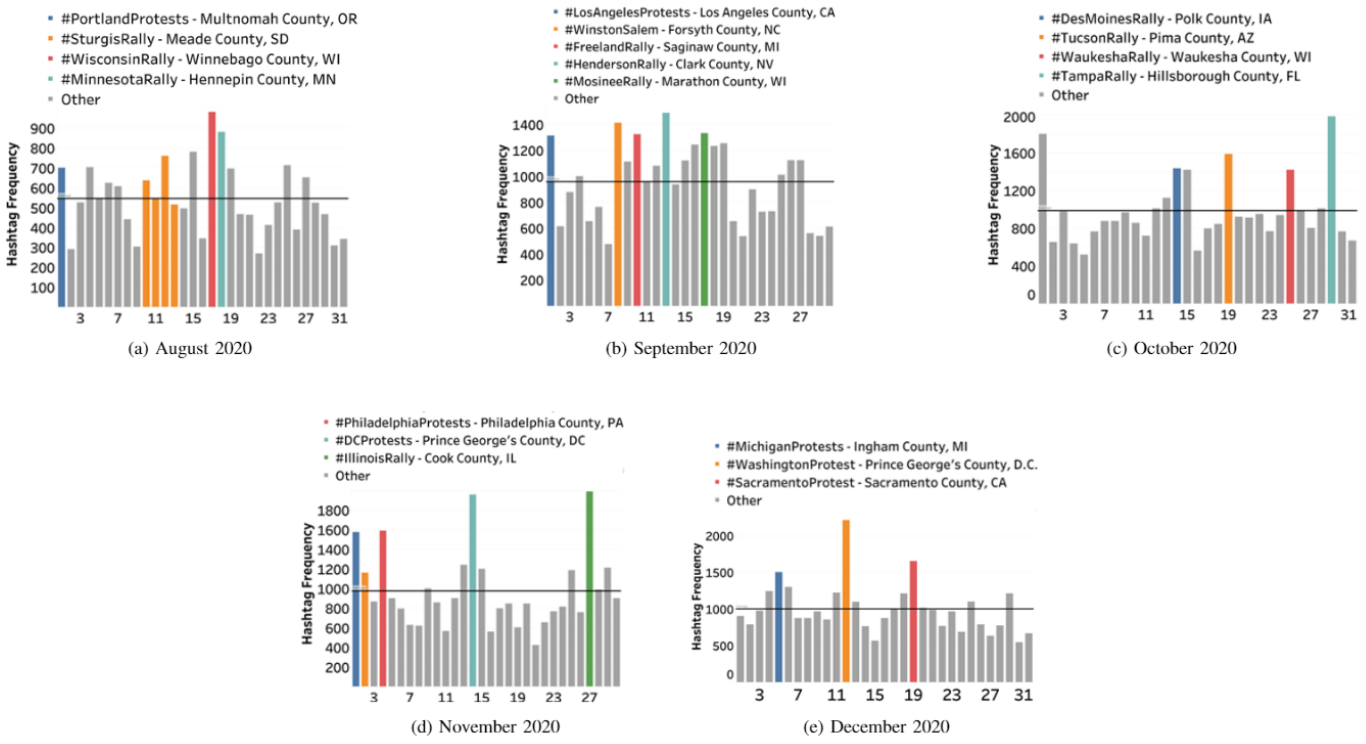
To analyze sentiment of tweets on Covid-19, [11] examined over 13 million tweets in 2020 and found through topic modeling, ten pandemic themes most important to Twitter users. They found that negative tweets covered racism, spread of cases, and symptoms, while positive tweets covered prevention, governments' response, treatment, and recovery. They did not address issues related to large gatherings or key events.

Prior work explores either the impact of mass gatherings on Covid-19 spread, or analyzes sentiments regarding Covid-19. Bridging this gap, our work uses social media tweets to automatically identify and analyze the sentiment for physical large gathering events (i.e., crowds). This is non-trivial due to two challenges. First, we must identify pertinent events from a small set of initial keywords in social media. In contrast to regular event identification, that includes even non-physical or virtual key events [12], the novelty of our work lies in focusing specifically on events associated with gatherings in the physical space. Second, we must combine disparate data sets, while determining the offset between events and infections. We compute the correlation between sentiment indicators and daily new cases of Covid-19 over a two week period in the event-hosting counties. Our purpose is to enable modeling Covid-19 case spread solely by harnessing social media data to gather a better understanding of mass gatherings.

Our automated approach, precluding manual event identification and human bias, is scalable, agile, rich and extensible. Identifying analysis-worthy events nationwide, it provides fine-grained information on event time, location, and social media sentiment, not achievable with other methods. To our best knowledge, this is the first work to propose an automated NLP system for simultaneously extracting physical mass gathering events *and* analyzing sentiment. Our contributions are:

- We develop a methodology for automatically identifying large gathering events in the physical world including dates and location, from digital chatter, i.e., social media (Twitter).
- We verify that the output of our automatically extracted events actually constitutes large gathering events in the physical world based on ground-truth data.
- We extract county-level Covid-19 statistics for identified event locations and compute the sentiment score for tweets referring to the event. We also model the relation between the sentiment and Covid-19 case counts for a two week period following the event and find a moderate negative correlation of -0.527, and note that highly negative sentiment

978-1-6654-1252-0/21/$31.00 ©2021 IEEE
DOI 10.1109/SMARTCOMP52413.2021.00082




(a) August 2020 (b) September 2020 (c) October 2020

(d) November 2020 (e) December 2020

Fig. 1. Large Gathering Event Extraction Aug.—Dec. 2020

values were generally associated with increased case counts.

## II. Methodology

### A. Data Collection From Heterogeneous Sources

We use tweet IDs from the Covid-19 Twitter dataset [13] to extract actual text and date for tweets (USA origin, English language) spanning August to December 2020, resulting in a total of 2.5 million tweets. Covid-19 statistics were obtained from Johns Hopkins Coronavirus Resource Center [14].

### B. Identifying Large Gathering Events

Intuiting that large gathering events in the physical world might be correlated with increased chatter in the digital world, we propose the large gathering event identification algorithm:

- We define a set of generic hashtags and keywords that represent the concept of large gatherings, e.g., '*rally*', '*gathering*', '*crowd*', and some others. Using these seed words, we filter out about 210,000 tweets.
- For all tweets thus filtered, we perform hashtag expansion by extracting all the hashtags present in the set of tweets, enabling us to not only associate large gathering events with generic hashtags such as '*#rally*' but also incorporate event-specific hashtags including '*#SturgisRally*', '*#pdxprotests*', '*#DNC*'. The resulting top 50 most frequent hashtags constitute our set of large gatherings hashtags (LGHashtags).
- Revisiting our original set of 2.5 million tweets, we retrieve tweets containing any hashtag from our LGHashtags list, thus, collecting 47,000 more tweets, leading to a total of 257,000 large gatherings tweets (LGTweets). An example extracted tweet: "*There are A LOT of people here in Portland dancing & chanting in front of the Federal Courthouse #BlackLivesMatter #PDXprotests #PortlandProtest*".
- For each day, we compute the total frequencies of all large gathering hashtags. From such a histogram (see Fig. 1), a large gathering event is considered to occur on date(s) when hashtag mentions exceed the monthly average frequency.
- Having identified large gathering events and their dates, the final step is to automatically extract the physical event *location*. Therefore, from all the tweets for each potential large gathering event date, we parse all the hashtags to retrieve only those with a mention of a city, county or state by consulting the US Cities Database[1], thus, finally extracting large gathering events, dates, and locations (LGEvents). Note that generic hashtags such as '*#DNC*' get filtered out at this step as they do not contain any location name.

### C. Modeling the Spread and Sentiment

We map each location identified in the previous step to county-level to obtain Covid-19 statistics from the JHU CSSE repository[2], for a two week time period following the event. The spread is considered to be the maximum value of the 7-day rolling average of the new case counts within this period. As an additional component, we also analyze the sentiment score of event-specific tweets. First we fine-tune a pretrained BERT_base model [15] on a dataset of Covid-19 tweets with sentiment labels (positive/negative)[3] and use it to

[1] {https://simplemaps.com/data/us-cities}
[2] https://github.com/CSSEGISandData/COVID-19
[3] https://www.openicpsr.org/openicpsr/project/120321/version/V5/view

predict the sentiment score of all tweets containing particular event hashtag (e.g. '*#PortlandProtests*').

## III. Results

### A. Extracting Large Gathering Events

The frequency histograms (see Figure 1) highlight some automatically extracted large gathering events with dates and locations. Verification with ground truth data of news articles reporting upon key events, shows that our extracted events correspond to actual physical large gathering events on those specific dates (e.g., Sturgis rally from Aug. 7-17, Tampa rally on Oct. 29, etc.), validating effectiveness of our approach in automatically identifying large gatherings.

### B. Modeling the Spread and Sentiment of Covid-19 after Large Gathering Events

Table I summarizes spread and sentiment results for large gathering events over a subsequent two week period. 'Sent.' and 'Spread' denote sentiment score and maximum 7-day rolling average of case counts within a two week period post event respectively. 'Sentiment of Random Subset' denotes a control value, i.e., the average sentiment of 5000 randomly sampled *non*-large gatherings tweets. Key highlights are indicated in boldface. We observe a moderate negative Pearson's correlation coefficient $\rho$ = -0.527 between the sentiment and the spread. We also compute the average sentiment score of 5000 randomly sampled *non*- large gathering events tweets to serve as a control value and find it to be quite neutral (-0.02). In general, highly negative sentiment values for event-specific tweets were associated with increased case counts. Though most events were associated with negative sentiment, we note an outlier in the case of 'Michigan Protests' with a positive sentiment and a negative case count. We need more research to understand the confounding factors of this phenomenon and why certain events cause more impact than others.

TABLE I
Sentiment and spread correlation of large gathering events.

| Date | Event | Sent. | Spread |
|---|---|---|---|
| Aug 01 | Portland Protests | -0.07 | 5.2 |
| Aug 09 | Sturgis Rally | **-0.35** | **648.5** |
| Aug 17 | Wisconsin Rally | -0.11 | 68.8 |
| Aug 18 | Minnesota Rally | -0.11 | 11.7 |
| Sep 01 | Los Angeles Protests | -0.09 | 0.0 |
| Sep 08 | Winston-Salem Rally | -0.10 | 0.0 |
| Sep 10 | Freeland Rally | -0.10 | 15.1 |
| Sep 13 | Henderson Rally | -0.17 | 72.4 |
| Sep 17 | Mosinee Rally | -0.31 | **165.7** |
| Oct 14 | Des Moines Rally | -0.12 | 70.8 |
| Oct 19 | Tucson Rally | -0.22 | 138.1 |
| Oct 25 | Waukesha Rally | -0.14 | 83.2 |
| Oct 29 | Tampa Rally | -0.34 | 43.1 |
| Nov 01 | Portland Riots | **-0.36** | 98.0 |
| Nov 02 | Scranton Rally | -0.09 | 30.8 |
| Nov 04 | Philadelphia Protests | -0.21 | 116.0 |
| Nov 14 | Washington Protests | -0.15 | 37.4 |
| Nov 27 | Illinois Rally | -0.32 | 0.72 |
| Dec 05 | Michigan Protests | **0.06** | **-0.6** |
| Dec 12 | Washington Protests | -0.09 | 2.8 |
| Dec 19 | Sacramento Protests | -0.15 | 16.7 |
| Sentiment of Control Random Subset | | -0.02 | - |
| Pearson's Correlation Coefficient | | -0.527 | |

## IV. Conclusions

We presented a novel methodology for automatically identifying physical mass gatherings, dates and locations from social media data. A majority of large events, particularly with negative sentiment, are followed by Covid-19 case increases in their geographic vicinity, though extent of increase is variable. In future work, we hope to improve our model by (i) incorporating more data sources, e.g., news articles to cover hybrid mass gatherings, (ii) automatic grouping of similar events, given impact variance across single events, and (iii) adding a predictive model and dashboard for Covid-19 spread.

## References


[1] J. H. CRC, "Covid-19 map. johns hopkins coronavirus resource center," Reuters, 2021, accessed: 2021-01-26. [Online]. Available: https://coronavirus.jhu.edu/map.html

[2] B. Jiménez Gutiérrez, J. Zeng, D. Zhang, P. Zhang, and Y. Su, "Document classification for COVID-19 literature," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Jul. 2020.

[3] D. Das, Y. Katyal, J. Verma, S. Dubey, A. Singh, K. Agarwal, S. Bhaduri, and R. Ranjan, "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*, Jul. 2020.

[4] T. Möller, A. Reina, R. Jayakumar, and M. Pietsch, "COVID-QA: A question answering dataset for COVID-19," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, Jul. 2020.

[5] D. Dave, D. McNichols, and J. J. Sabia, "The contagion externality of a superspreading event: The sturgis motorcycle rally and covid-19," *Southern economic journal*, 2020.

[6] K. K. Martin Enserink and N. Desai, "The science of superspreading," Science Magazine, 2020, accessed: 2021-01-26.

[7] C. I. e. a. Hamner L, Dubbel P, "High sars-cov-2 attack rate following exposure at a choir practice — skagit county, washington," *Morb Mortal Wkly Rep 2020*, vol. 69, no. 19, pp. 606–610, 2020.

[8] K. L. e. a. Leclerc QJ, Fuller NM, "What settings have been linked to sars-cov-2 transmission clusters? [version 2; peer review: 2 approved]." *Wellcome Open Res 2020*, p. 83, 2020.

[9] J. E. Harris, "Geospatial analysis of the september 2020 coronavirus outbreak at the university of wisconsin–madison: Did a cluster of local bars play a critical role?" National Bureau of Economic Research, Tech. Rep., 2020.

[10] B. D. Bernheim, N. Buchmann, Z. Freitas-Groff, and S. Otero, *The effects of large group meetings on the spread of COVID-19: The case of Trump rallies*, 2020.

[11] 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/

[12] S. Zong, A. Baheti, W. Xu, and A. Ritter, "Extracting covid-19 events from twitter," *arXiv preprint arXiv:2006.02567*, 2020.

[13] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalin, and G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration," 2020.

[14] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.