Using Event-Based Web-Scraping Methods and Bidirectional Transformers to Characterize COVID-19 Outbreaks in Food Production and Retail Settings

Joseph Miano[1,2], Charity Hilton[1](✉), Vasu Gangrade[3], Mary Pomeroy[3], Jacqueline Siven[3], Michael Flynn[3], and Frances Tilashalski[3]

[1] Georgia Tech Research Institute, Atlanta, GA 30318, USA
Charity.Hilton@gtri.gatech.edu
[2] Georgia Institute of Technology, Atlanta, GA 30332, USA
[3] Centers for Disease Control and Prevention, Atlanta, GA 30333, USA

Abstract. Current surveillance methods may not capture the full extent of COVID-19 spread in high-risk settings like food establishments. Thus, we propose a new method for surveillance that identifies COVID-19 cases among food establishment workers from news reports via web-scraping and natural language processing (NLP). First, we used web-scraping to identify a broader set of articles (n = 67,078) related to COVID-19 based on keyword mentions. In this dataset, we used an open-source NLP platform (ClarityNLP) to extract location, industry, case, and death counts automatically. These articles were vetted and validated by CDC subject matter experts (SMEs) to identify those containing COVID-19 outbreaks in food establishments. CDC and Georgia Tech Research Institute SMEs provided a human-labeled test dataset containing 388 articles to validate our algorithms. Then, to improve quality, we fine-tuned a pretrained RoBERTa instance, a bidirectional transformer language model, to classify articles containing $\geq$1 positive COVID-19 cases in food establishments. The application of RoBERTa decreased the number of articles from 67,078 to 1,112 and classified ($\geq$1 positive COVID-19 cases in food establishments) articles with 88% accuracy in the human-labeled test dataset. Therefore, by automating the pipeline of web-scraping and COVID-19 case prediction using RoBERTa, we enable an efficient human in-the-loop process by which COVID-19 data could be manually collected from articles flagged by our model, thus reducing the human labor requirements. Furthermore, our approach could be used to predict and monitor locations of COVID-19 development by geography and could also be extended to other industries and news article datasets of interest.

Keywords: COVID-19 · Public health · Web-scraping · Natural language processing

A. Tucker et al. (Eds.): AIME 2021, LNAI 12721, pp. 187–198, 2021.
https://doi.org/10.1007/978-3-030-77211-6_21

1 Introduction

Though several studies have captured outbreaks of COVID-19 among employees in congregate settings [18,22,31], to our knowledge few reports [14,28,32] comprehensively characterize outbreaks among essential workers in food production and retail settings. Local news reports can be an important source in identifying outbreaks. This project applies machine learning-based web-scraping and a RoBERTa (A Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach) [21] language model to locate, quantify, and characterize outbreaks among food system workers via local news reports. Our system highlights the utility of media reports in characterizing COVID-19 outbreaks and can be extended to other industries and congregate settings.

2 Background

2.1 COVID-19 and Food Establishments

Understanding the scope of outbreaks of COVID-19 identified in non-healthcare settings, such as food settings, is crucial to both informing public health decision-making and prevention messaging tailored to reducing transmission among worker populations and identifying potential health equity issues in workplaces. Though outbreaks in food settings have been reported through various state surveillance mechanisms [1,2,7], in web-based portals[1] and in scientific reports [3,14,20,28,32], which specifically highlight incidence in meat processing facilities, little to no systematic and ongoing data collection exists for all food system sectors, including restaurants or grocery settings. Additionally, public, and occupational health data collection systems often do not collect information on social determinants of health (such as race, ethnicity, nativity, language spoken, etc.) in workplaces [26]. Workers in food settings are predominantly critical infrastructure workers, who contribute to the security and well-being of our food supply. Collecting these data may aid in the early identification of workplace transmission of SARS-CoV-2, the virus that causes COVID-19. This creates potential opportunities for successful mitigation efforts and improving overall worker safety and health.

2.2 Bidirectional Transformers

Transformer language models have seen widespread development and use [13,21,24,30] over the past several years and achieve state-of-the-art results across various natural language processing tasks, like text generation and classification [12,29,34]. Introduced in 2017 by Vaswani et al. [30], the Transformer is a purely attention-based model (i.e., with no convolutional layers or recurrence) that is faster to train and achieves better results than previous techniques on a

[1] https://public.tableau.com/profile/leah.douglas#!/vizhome/CumulativeCovid-19casesbysector/Dashboard1.

language translation task [30], and its architectural details can be found in the original 2017 paper [30].

Several variations on the original Transformer architecture and training process have been developed since 2017 [12,13,21,30], including Bidirectional Transformer models like BERT (Bidirectional Encoder Representations from Transformers) [13] and RoBERTa [21]. BERT extended previous transformer-based models [24,30] by incorporating bidirectional pre-training; i.e., BERT is pre-trained using the masked language model (MLM) objective, by which it learns to predict a masked word given the context words that come before and after [13]. Devlin et al. [13] also leveraged next sentence prediction (NSP) to pre-train BERT, which is an objective by which the model attempts to predict the next sentence given a current sentence input [13]. Together, these pre-training objectives build a robust language model that can be fine-tuned for a target task like text classification [29].

RoBERTa [21] shares the same model architecture as BERT but modifies the BERT pre-training process to omit the NSP task and leverages additional data during pre-training (e.g., CC-News, which consists of 63 million English news articles) not used by BERT. The result is a model that outperforms BERT across several tasks [21] and that may be especially suited to the task of news article classification, due to the news article data used during RoBERTa pre-training.

3 Methods

Our proposed method (Fig. 1) leverages NLP-Based Web-Scraping and a RobERTa model [21] to classify news articles as containing mention of ≥1 cases of COVID-19 in food establishments. Once a model is trained and evaluated, a daily list of relevant news articles could be output for human review.

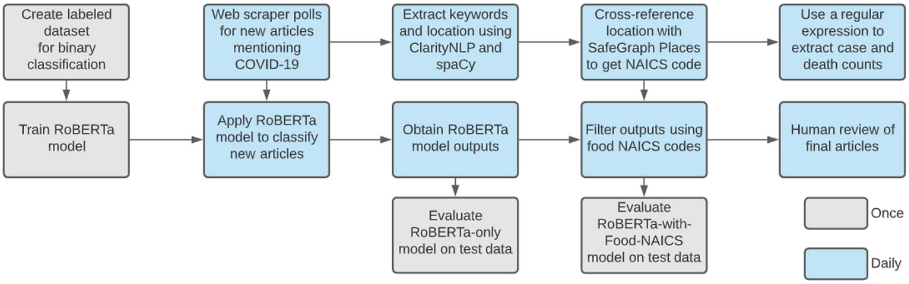
Create labeled dataset for binary classification
Train RoBERTa model
Web scraper polls for new articles mentioning COVID-19
Apply RoBERTa model to classify new articles
Extract keywords and location using ClarityNLP and spaCy
Obtain RoBERTa model outputs
Evaluate RoBERTa-only model on test data
Cross-reference location with SafeGraph Places to get NAICS code
Filter outputs using food NAICS codes
Evaluate RoBERTa-with-Food-NAICS model on test data
Use a regular expression to extract case and death counts
Human review of final articles
Once
Daily

Fig. 1. A flow diagram of what our method would look like if it were deployed. In our experiments, we ran each of these steps on our overall dataset once, but we did not deploy our approach to produce real-time daily outputs.

3.1 NLP-Based Web-Scraping

Since April 2020, we tracked keyword mentions of COVID-19 in the news media from articles reported from March 15, 2020, to September 30, 2020 to develop a COVID-19 news dataset. From aggregating lists of news media sources from web sites, such as Wikipedia, we created a list of over 1,700 news sources [8–10]. Each news source was visited daily by an automated Python script, using a web-scraping library to discover publicly available news articles mentioning COVID-19 [25]. Next, we ran each article through two NLP pipelines, ClarityNLP and spaCy [15–17].

NLP Pipelines. ClarityNLP is an open-source project written in a modern, Python-based stack, and it leverages a user-friendly query language, NLPQL [16,17]. ClarityNLP allows for building custom NLP pipelines with a focus on biomedical text. To build our COVID-19 news dataset, we utilized ClarityNLP's NLPQL query language to find non-negated mentions of keywords relevant to our dataset. These keywords were deemed relevant to potential COVID-19 outbreaks in congregate settings, such as food processing facilities, long-term care facilities, and correctional facilities. Each keyword was reviewed by SMEs from the CDC. Additionally, we developed a regular expression algorithm to extract case counts and death counts from each news article.

We utilized the advanced NLP library, spaCy to identify the location of the news article [15]. SpaCy provides a named-entity recognition feature to extract categories of entities from sections of text. We used the location (LOC), organization (ORG), geographic (GPE), and facility (FAC) categories to identify the location of the potential COVID-19 outbreak.

Location Identification and NAICS Codes. We cross-referenced the location entities found by spaCy against the news source's location, and the SafeGraph Places dataset [4]. SafeGraph is a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places via the Placekey Community. In addition to name and address, the SafeGraph Places dataset provides the North American Industry Classification System (NAICS) Code for each location. The NAICS system is a hierarchy used to classify industries for analysis and publication [5]. At the time of building our COVID-19 news dataset, there were 5.4 million places in the SafeGraph Places dataset across the United States.

Persistence. Finally, we saved the metadata, case/death count, NAICS code, and location variables for each news article in a MySQL database. Currently, there are over 93,000 keyword mentions in our news dataset. These articles were the basis for training the RoBERTa model[2].

[2] At the time we trained the RoBERTa model, we evaluated just over 67,000 articles.

3.2 Data Validation

We validated the extracted variables and removed duplicate facilities mentioned in multiple articles by manual review. We specified "potential" outbreaks if there were $\geq$2 positive cases identified as employees of a single food setting. However, we could not always confirm that the infection was attributable to workplace exposure as the sources for the data were only news articles. We also could not verify the reporting mechanism of the case; whether a case was identified as a case by confirmatory laboratory testing or whether it was self-reported. News articles may not always include how the establishment was made aware of employee cases–whether through laboratory confirmation of a diagnosis or through self-reporting. Self-reporting without lab confirmation might have led to introduction of biases which could lead to misclassification (e.g., an employee reports COVID-19-like symptoms to workplace but has not confirmed infection [11]). While validating, we also extracted more qualitative information; for example, if these establishments opted to close their doors for 1–2 weeks after an employee or employees were identified as having COVID-19.

3.3 Applying RoBERTa to Article Classification

Implementation. We implemented our RoBERTa model training and evaluation using Python with the PyTorch [23] and HuggingFace [34] libraries between October 2020 and January 2021.

Training. We trained a RoBERTa model to automatically assign a positive tag to an article, based on its contents, if it contains mention of $\geq$1 cases of COVID-19 in food establishments, which we also combined with a NAICS code filter for food-related industries. To train our RoBERTa model, we first randomly divided our overall human-labeled dataset (n = 2061) into a training dataset (n = 1189), validation dataset (n = 471), and test dataset (n = 388). The training dataset is what the RoBERTa model sees at every training iteration in order to update its weights. In contrast, the validation dataset determines which model weights should be kept after completing all the training epochs (i.e., we keep the model weights with the lowest loss on the validation dataset). The test dataset is used for model evaluation, which we discuss in the next section. We used cross-entropy loss, with class weights set to be balanced using scikit-learn's compute_class_weight function as our loss function and trained for 300 epochs at a batch size of 32 using the Adam optimizer [19]; here, an epoch is defined as a single complete pass through the training dataset (i.e., the model seeing every training example once constitutes one epoch) and the batch size refers to the number of training examples used to update the model's weights at each iteration of training.

Evaluation. We evaluated our models by computing various metrics on our human-labeled test dataset, including accuracy, precision, recall, and F1-score (terms are defined below). We also counted the number of positive examples flagged by our models in the overall 67,078 articles; for this evaluation, an article is considered a ground truth positive example if it contains mention of $\geq$1 positive cases of COVID-19 in food establishment employees and a ground truth negative example otherwise. Accuracy was computed as the number of correctly classified examples divided by the total number of examples classified (n = 388 in the test dataset). Precision is computed as: $\left(\frac{\text{true positives}}{\text{true positives+false positives}}\right)$, where a true positive refers to a ground truth positive example, as defined above, that the model correctly classifies as positive, a true negative refers to a ground truth negative example that the model correctly classifies as negative, a false positive refers to a ground truth negative example that the model falsely classifies as positive, and a false negative refers to a ground truth positive example that the model falsely classifies as negative. Next, recall is computed as: $\left(\frac{\text{true positives}}{\text{true positives+false negatives}}\right)$. Finally, the F1-score is computed as: $\left(\frac{2}{\frac{1}{\text{precision}}+\frac{1}{\text{recall}}}\right)$.

4 Results

4.1 Web-Scraping, Manual Data Validation, and Visualization

This paper summarizes results from a web-scraping analysis of news articles and media reports, where one or more employees in specific food settings were identified as having COVID-19. Through an automated web-scraping process followed by manual cleaning and data validation, we identified 276 facilities with 30,734 employees who reportedly tested positive for COVID-19, according to the news articles (see Table 1)[3].

Ninety percent of these cases were identified among the animal slaughtering and processing facility workers, with some plants reporting as many as 1,000 positive cases at a single location. Although most articles were about restaurants among the consumer-facing establishments, news articles reported significantly more employees per facility with COVID-19 in grocery store settings. News reports mentioned only five food distribution centers, but they had 151 employees with COVID-19. Of note, articles reported that most restaurants opted to close their doors for 1–2 weeks after an employee or employees were identified as having COVID-19. We also categorized the outbreaks by geographic region

[3] The US map, which can be accessed at: https://www.google.com/maps/d/u/0/viewer?mid=1ymY4bzI70AOCeFzRYvfe4HPWVvgPBoJh&ll=45.80359787060013%2C-114.35715944999998&z=4 shows the food-setting locations found using manual validation of news articles. The legend on the left within the map shows the different types of food settings based on NAICS codes. The user may access more details about each facility including the title, a link to the article, and descriptors including case and death counts by clicking on a chosen location on the map.

Table 1. Below, we show a summary of the reported facility types organized by NAICS code descriptors. This includes facility type, number of facilities (with ≥1 case reported), number of potential outbreaks (≥2 cases) of COVID-19, and number of cases linked to the type of facility.

NAICS code	Industry classification/type of facility	No. of facilities	No. of potential outbreaks	No. of cases
3116	Animal slaughtering and processing	132	130	27704
7225	Restaurants and other eating places	65	13	91
4451	Grocery stores	30	15	254
7224	Drinking places (alcoholic beverages)	13	7	42
3114	Fruit and vegetable preserving	9	9	561
3117	Seafood product preparation and packaging	6	6	283
4244	Grocery and related product merchant wholesalers	5	5	151
3119	Other food manufacturing	4	4	312
1113	Fruit and tree nut farming	2	2	77
1119	Other crop farming	2	2	131
7223	Speciality food services (food trucks)	2	1	3
3112	Grain and oilseed milling	1	1	13
1122	Hog and pig farming	1	1	60
1151	Support activities for crop production	1	1	6
1112	Vegetable and melon farming	1	1	14
Unk	Unknown	2	2	1032
	Total	**276**	**200**	**30734**

(interactive map) and industry sector (e.g., food processing, restaurant, grocery stores), and number of cases/deaths[4].

4.2 RoBERTa Model Evaluation

Using our RoBERTa model enabled article classification at 88.4% accuracy with a 62.8% F1 score (Fig. 2). The combined RoBERTa-with-Food-NAICS model achieved the best performance (Fig. 2). Furthermore, RoBERTa-with-Food-NAICS flagged 1,112 articles as positive of the 67,078 total articles, which is less than the 4681 flagged by Food-NAICS-only (Fig. 2). However, the recall was lower for RoBERTa-with-Food-NAICS than for Food-NAICS only (Fig. 2, 79.2% vs. 91.7%).

Next, we extended our initial dataset of 67,078 news articles to include news articles through to the end of 2020, thus adding 17,170 new news articles. We used this expanded dataset of 84,248 articles to generate Fig. 3, which

[4] We developed a public dashboard of news article keywords of COVID-19 in food processing facilities from March 15-September 30, 2020. It is available at https://public.tableau.com/profile/charity.hilton\#!/vizhome/COVID-19NewsReportsaboutFoodSettings/COVID-19NewsMap.

shows the daily number of articles flagged as positive (i.e., containing mention of ≥1 COVID-19 cases in food establishments) by considering all articles, Food-NAICS-only, and combined RoBERTa-with-Food-NAICS. To generate Fig. 3, we applied our RoBERTa model trained on the original 67,078 articles to the expanded dataset, which spans from March 15, 2020 to December 31, 2020. When considering all articles, the maximum number of daily articles was 2315, which occurred on May 28, 2020, and the mean number of daily articles is 277 (Fig. 3). Food-NAICS-only has a maximum number of daily articles of 143 (on May 27) and a mean of 25 articles, while RoBERTa-with-Food-NAICS model has a maximum number of daily articles of 40 (May 29) and a mean of 4 (Fig. 3).

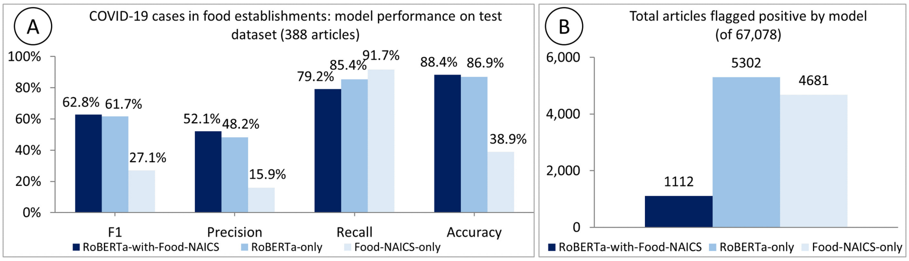

Fig. 2. In (A), we show the precision, recall, F1 score, and accuracy for 3 variations of our model on a held-out test dataset of 388 articles (48 positive examples and 340 negative). In (B), we apply our models to our overall dataset and display how many articles are flagged as containing mention of ≥1 COVID-19 cases in a food establishment.

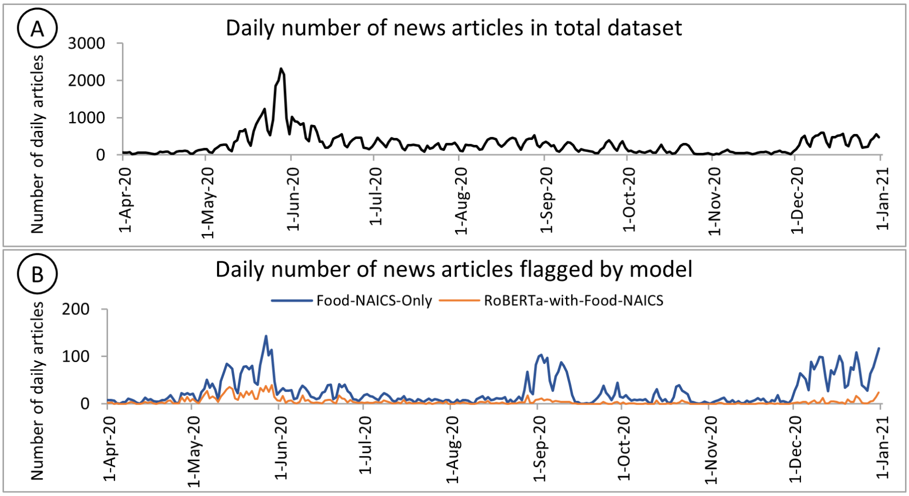

Fig. 3. In (A), we show the number of daily news articles without any NAICS or model-based filtering (i.e., the total number of articles). In (B), we show the number of daily news articles using a food NAICS filter only (blue) and a combined RoBERTa model output and food NAICS filter (orange). (Color figure online)

5 Discussion

5.1 RoBERTa Model and NAICS Codes

Because RoBERTa-with-Food-NAICS flagged a total of 1,112 articles and Food-NAICS-only flagged 4,681 articles, RoBERTa-with-Food-NAICS reduces the amount of human manual review needed for validation. The recall being lower for RoBERTa-with-Food-NAICS than for NAICS-only means that RoBERTa with NAICS misses more positive examples than NAICS-only; however, this is counterbalanced by the greater precision of RoBERTa with NAICS.

One limitation of the combined approach is that the NAICS code is not always available, in which case RoBERTa-only (i.e., without the food NAICS filter) could be used. Another limitation is the inability to differentiate between true transmission among workers in a food setting versus exposure within the community, particularly in geographic locations with sustained COVID-19 transmission. A third limitation is the likely differences in case definitions reported by various news sources. There were no efforts made to account for inaccuracies in reporting and all "news" sources were treated equally. Future efforts are needed to compare results from the proposed model to proven methods of public health surveillance relying on case investigation and contact tracing.

5.2 COVID-19 Surveillance

Some of the earliest reported cases of COVID-19 in food settings were identified in meat processing facilities in early April 2020 [14]. During that time, this web-scraping platform was in the early stages of development. In fact, a number of outbreaks had occurred within food settings by the time the tool was fully developed (see Fig. 3). Though we could not detect outbreaks using this systematic web-scraping approach during the early months of the COVID-19 pandemic, we were able to capture and quantify these outbreaks retroactively. This methodology can prove invaluable in addressing any time lapse between outbreak and response; during infectious disease outbreaks it takes significant time and effort to ramp up the public health response, yet news agencies often have the capacity to quickly collect and disseminate reports on cases and localized outbreaks. Because early detection of outbreaks is crucial for effective mitigation and successful public health coordination [33], a proactive approach in using this methodology and addressing that time lapse may be a useful tool to aiding in earlier outbreak detection, particularly for novel pathogens for which no transitional surveillance methods or reporting requirements exist.

This methodology could also be used to identify health equity issues within these worker populations during infectious disease outbreaks. Using an Occupational Health Equity framework (OHE), i.e., that social, economic, environmental, and structural factors impact work-related disparities in illness and disease in avoidable ways [6], this methodology could be used to identify the frequency of OHE-relevant mentions in news articles. By leveraging the by-the-minute reports of news agencies to compare outbreak data and OHE variables, we can aid in early identification of potential compounding social factors affecting disease transmission. Such early identification is vital; as we have seen during the COVID-19 pandemic, rates of infection and disease can be heavily impacted by the interactions between type of work and social variables (like race, ethnicity, and migration status). To properly address how social and structural vulnerabilities impact work-related illness and disease we must enhance our data collection through both innovative targeted efforts and expansion of current surveillance efforts. By developing new ways to incorporate race, ethnicity, work arrangement, and other variables, into OSH surveillance we can better understand how these variables interact to affect different worker populations [27].

Our findings improve our understanding of outbreaks in food settings and help identify health equity issues and health disparities among essential workers. Ideally this methodology may produce more focused and industry-specific public health interventions and provide meaningful feedback to partners, including federal agencies and policymakers.

6 Conclusion

By combining NLP-based web-scraping and a RoBERTa language model, our approach enables an efficient human-in-the-loop process to identify COVID-19 outbreaks in food establishments based on news article contents, which addresses the limitation of current COVID-19 surveillance largely underreporting occupational data. We found that combining the RoBERTa outputs with a NAICS code filter for food settings yielded better performance on our dataset than using only the RoBERTa outputs or only the NAICS code filter. Our approach could be applied to help predict geographic areas of concern for COVID-19 outbreaks based on early event-based reporting in news articles. Other future directions include extending the RoBERTa method to improve its performance without requiring the NAICS code (e.g., by gathering more human-labeled data and/or applying self-supervised learning) and applying our approach to other industries or news article datasets. Future studies are warranted for true validation of the surveillance model results against existing public health surveillance tools.

7 Disclaimer

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC).

References

1. Coronavirus - outbreak reporting. www.michigan.gov/coronavirus/0,9753,7--406-98163_98173_102057---,00.html
2. COVID-19 outbreaks: department of health: state of Louisiana. https://ldh.la.gov/index.cfm/page/3997

3. Investigating and responding to COVID-19 cases in non-healthcare work settings. https://www.cdc.gov/coronavirus/2019-ncov/php/open-america/non-healthcare-work-settings.html
4. Places schema. https://docs.safegraph.com/docs/places-schema
5. What is a NAICS code and why do i need one? September 2018. https://www.naics.com/what-is-a-naics-code-why-do-i-need-one/
6. CDC - NIOSH program portfolio: occupational health equity: program description, December 2019. https://www.cdc.gov/niosh/programs/ohe/default.html
7. https://www.doh.wa.gov/Portals/1/Documents/1600/coronavirus/data-tables/StatewideCOVID-19OutbreakReport.pdf, December 2020
8. List of newspapers in the united states (2020). https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_States
9. List of radio stations in the united states (2020). https://en.wikipedia.org/wiki/Lists_of_radio_stations_in_the_United_States
10. List of television stations in the united states (2020). https://en.wikipedia.org/wiki/Lists_of_television_stations_in_the_United_States
11. Althubaiti, A.: Information bias in health research: definition, pitfalls, and adjustment methods. J. Multi. Healthc. **9**, 211 (2016)
12. Brown, T.B., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Dyal, J.W.: COVID-19 among workers in meat and poultry processing facilities–19 states, April 2020. MMWR Morb. Mortal. Wkly Rep. **69** (2020)
15. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: industrial-strength natural language processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
16. Georgia Tech Research Institute: Clarity NLP (2018). https://github.com/ClarityNLP
17. Georgia Tech Research Institute: Clarity NLP documentation (2018). https://clarity-nlp.readthedocs.io/en/latest/
18. Kakimoto, K., Kamiya, H., Yamagishi, T., Matsui, T., Suzuki, M., Wakita, T.: Initial investigation of transmission of COVID-19 among crew members during quarantine of a cruise ship–Yokohama, Japan, February 2020
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Krebs, C.: Guidance on the essential critical infrastructure workforce: ensuring community and national resilience in COVID-19 response. Cybersecurity and Infrastructure Security Agency (CISA) **5** (2020)
21. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
22. McMichael, T.M.: COVID-19 in a long-term care facility—king county, Washington, February 27–March 9, 2020. MMWR Morb. Mortal. Wkly Rep. **69**, 339 (2020)
23. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8026–8037 (2019)
24. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
25. Richardson, L.: Beautiful soup documentation, April 2007

26. Rodriguez-Lainz, A., et al.: Collection of data on race, ethnicity, language, and nativity by us public health surveillance and monitoring systems: gaps and opportunities. Public Health Rep. **133**(1), 45–54 (2018)
27. National Academies of Sciences, Engineering, and Medicine and others: A smarter national surveillance system for occupational safety and health in the 21st century. National Academies Press (2018)
28. Steinberg, J., et al.: COVID-19 outbreak among employees at a meat processing facility–South Dakota, March-April 2020. Morb. Mortal. Wkly Rep. **69**(31), 1015 (2020)
29. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
31. Wallace, M.: Public health response to COVID-19 cases in correctional and detention facilities—Louisiana, March–April 2020. MMWR Morb. Mortal. Wkly Rep. **69** (2020)
32. Waltenburg, M.A., et al.: Update: COVID-19 among workers in meat and poultry processing facilities–united states, April-May 2020. Morb. Mortal. Wkly Rep. **69**(27), 887 (2020)
33. Wilson, K., Brownstein, J.S.: Early detection of disease outbreaks using the internet. Cmaj **180**(8), 829–831 (2009)
34. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)