Pragmatic Design and Application of a Named Entity Recognition Pipeline to Assist Contact Tracers during the COVID-19 Pandemic

[1]Majid Afshar, MD, MSCR; [1]Iain McConnell, PhD; [1]John Caskey, PhD; [1]Madeline Oguss, MS; [1]Meysam Ghaffari, PhD; [2]Dmitriy Dligach, PhD; [3]Rachel Kulikoff; [3]Brittany Grogan; [3]Crystal Gibson, PhD; [4]Elizabeth Wimmer; [4]Traci E DeSalvo, MS; [1]Matthew Churpek MD, MPH, PhD
[1]University of Wisconsin-Madison, WI; [2]Loyola University Chicago, IL; [3]Public Health Madison Dane County, Madison, WI; [4]Wisconsin Department Health Services, Madison, WI

Introduction: As of February 2021, the state of Wisconsin (WI) tested over three million individuals for SARS-CoV-2 and nearly 564,000 tested cases confirmed positive. At the county level, health departments have found that free text fields from the COVID-19 Initial Case Interview (contact tracing) forms contain the most crucial information to identify potential organizations and locations where transmission of the virus occurred when individuals were infectious. Public health workers have encountered a high case-load and are overwhelmed with an abundance of free-text information in the interview forms. Current methods to mine the free text fields are manual, without rapid and systematic methods for finding transmission clusters. We employed methods in natural language processing with pre-trained neural language models for named entity recognition (NER) from the Wisconsin Electronic Disease Surveillance System (WEDSS) for potential organizations and locations to assist in contact tracing efforts and reduce the burden on the health departments.

Methods: WEDSS is a secure, web-based system designed to facilitate reporting, investigation, and surveillance of communicable diseases including data on testing for SARS-CoV-2 cases since the outbreak began in February 2020. A total of 281 fields were extracted from WEDSS for our analyses, including 26 character string fields from the county-level data related to information on potential contacts.

For the pipeline development, all 26 character string fields were concatenated into one document with model runs at the case-level. The WordPiece tokenizer was used to build groups of 512 tokens from each document that were fed into the Bidirectional Encoder Representations from Transformers (BERT) neural language model[1]. The BERT-base-cased model was fine-tuned on the English version of the standard CoNLL-2003 Named Entity Recognition (NER) shared task[2]. The pipeline was downloaded from HuggingFace[3] and reports an F1 score of 91.3 on the CoNLL-2003 test dataset. Post-processing of the NERs included the removal of frequently occurring NERs (i.e., “Wisconsin”, “GMT”) identified from 12 months of Case Interview forms and removal of duplicate NERs. For all case IDs that had the same NER reported, the average predicted probability was provided as the score for the likelihood of identifying it as a person, organization, location, or miscellaneous.

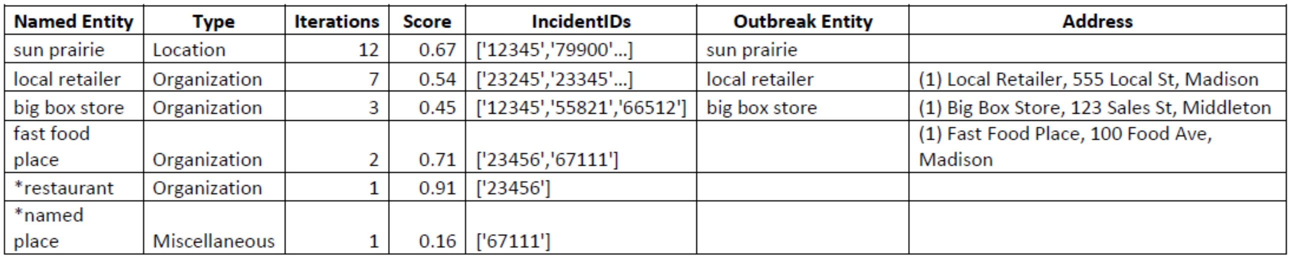
Weekly reports were built using the date nearest to contracting the virus or testing positive. The goal of the report was to identify NERs associated with a cluster outbreak. A cluster was defined as two or more cases associated with the same NER in a seven-day period. Each cluster in the report also included the associated case IDs to guide contact tracers. In some cases, an outbreak investigation was already performed so the NER pipeline was also applied to the OutbreakID fields, and locations. Matching outbreak NERs were also reported to prevent duplicate effort by the contact tracers.

Many of the NERs contained common business names that have multiple locations within a county (i.e., “Culvers”, “Walmart”). Therefore, we performed unsupervised machine learning on the case IDs associated within each NER cluster. The latitude/longitude coordinates for each case ID in the cluster were extracted from WEDSS and a k-means approach was used to identify the centroid coordinates for the cluster of case IDs for a particular organization or miscellaneous NER. Data from the US Census Bureau for 2018 showed that the commute distances for over two-thirds of the businesses in three major metropolitan areas in Wisconsin were between 0 and 24 miles (https://onthemap.ces.census.gov). Therefore, we set a search radius of 14 miles from the centroid coordinates, and our pipeline ran the Google Places API to perform a search for businesses matching the NER terms within this radius that matched our internal database of organizations. Ultimately, the algorithm provided the most likely business address for the NER. If an exact match could not be made, then the top three results were filtered using greedy regular expressions. The address results were merged into the report to further aid contact tracers.

All confirmed and probable cases from Dane County were initially examined. Validation of results was compared against a sampling of known outbreaks with OutbreakID organization names provided by Public Health Madison Dane County staff for November 2020. Addresses identified from the centroid coordinates were also validated against semi-structured fields from WEDSS containing work locations, school locations, and businesses visited. The analyses were all performed using Python Version 3.8.5 (Python Software Foundation). The Institutional Review Board of the University of Wisconsin approved this study and a Data Use Agreement was established between the Wisconsin Department of Health Services and the University of Wisconsin.

Results: The validation sample of Dane County data was composed of 48 probable or confirmed cases of SARS-CoV-2 with 7881 total BERT tokens and 1020 unique BERT tokens. The longest field was InvestigationNotes with a median token count of 126.5 (Interquartile Range 67.0-232.5). The NER pipeline captured 100% (n=15) of NERs associated with an ongoing outbreak investigation. An additional 29 NERs that qualified as cluster outbreaks were identified as potential undetected outbreaks to assist contact tracers. In the end, a sample report with 44 NERs linked to potential cluster outbreaks was provided to Dane County for November 2020. A de-identified sample report is shown in the **Table** demonstrating a match across named entities from the NER pipeline with NERs from known outbreaks (outbreak entity) as well as a new fast food restaurant that may be an unrecognized cluster.

Table. Weekly deidentified report for contact tracers by county

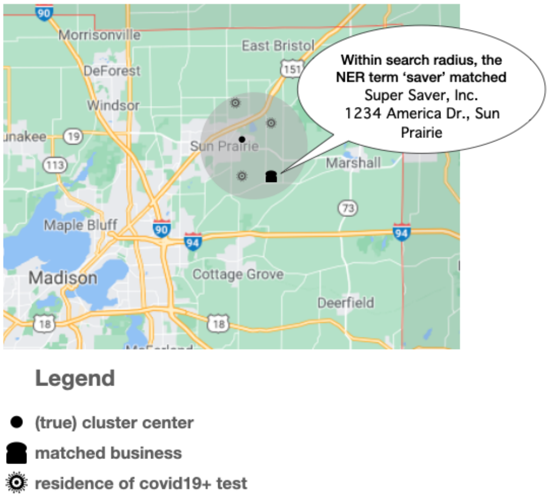


Named Entity	Type	Iterations	Score	IncidentIDs	Outbreak Entity	Address
sun prairie	Location	12	0.67	['12345','79900'...]	sun prairie	
local retailer	Organization	7	0.54	['23245','23345'...]	local retailer	(1) Local Retailer, 555 Local St, Madison
big box store	Organization	3	0.45	['12345','55821','66512']	big box store	(1) Big Box Store, 123 Sales St, Middleton
fast food place	Organization	2	0.71	['23456','67111']		(1) Fast Food Place, 100 Food Ave, Madison
*restaurant	Organization	1	0.91	['23456']		
*named place	Miscellaneous	1	0.16	['67111']		

Named Entity = Name Entity Recognition; Iterations = unique occurrences of cases; IncidentID = unique case IDs for lookup by contact tracer; Outbreak entity = known outbreak exposures; Address = matched NER using longitude/latitude for address from k-means clustering from Google Places API; score = average predicted probability from the classifier for the type of NER. Iterations = unique mentions of NER across available case IDs from the reporting period. *NERs that did not qualify as cluster outbreaks (needing $\geq$2 case IDs) with less than 2 iterations.

Of the known outbreaks with business names as the NER (n=7), a match was made between the business name from the NER pipeline to the business NERs identified in the search radius of the centroid coordinates, and an exact address was provided for 100% of the matched cases. Among the remaining NERs, an additional 84% (n=37) had an address identified for potential outbreak locations to assist contract tracing efforts. The **Figure** demonstrates how the search algorithm performs to find the centroid coordinates among a cluster of case IDs and match to the business name within the search radius. Of the addresses provided in our report, 75% (n=28) were also found in the semi-structured WEDSS fields to further validate our pipeline.

Figure. Sample of cluster outbreak with centroid coordinates for extracting matched organizations within search radius



Example of single NER with 3 case IDs. The coordinates of each case ID are identified, and the cluster center derived is from k-means clustering across the case IDs. All organization names are identified within 14 mile radius of centroid coordinate and run through NER pipeline. Any matched NER from the map to the case interview data are reported with an address.

<u>Discussion/Conclusion</u>: The beta version of our NER pipeline was able to extract large amounts of surveillance data and summarize a report to highlight potential outbreaks and their associated addresses. The report was designed in weekly intervals by county so a systematic approach can be shared for any county in the state of Wisconsin. Our initial retrospective validation work with colleagues in Dane County demonstrated our pipeline did not miss any ongoing outbreak investigations and may have identified additional noteworthy cluster outbreaks. We are currently piloting our pipeline with Dane County for prospective validation and to gather qualitative data if any meaningful improvements are made in improving accuracy and saving time during contact tracing efforts.

Our work derives from a statewide database so it may be scaled across counties to assist other health departments in Wisconsin. Our pipeline is also publicly available at [GitHub repo to be shared if accepted] for other states to adapt to their workflow. Further, our pipeline may also be applied for other communicable disease and surveillance efforts.

<u>References</u>

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* CoRR http://arxiv.org/abs/1810.04805
2. Tjong Kim Sang, Erik F., and De Meulder, Fien (2003) *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition* Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL https://www.aclweb.org/anthology/W03-0419 p.142-147
3. https://huggingface.co/dslim/bert-base-NER