Using Natural Language Processing to Identify COVID-19 Spread Factors From the Literature to Assist With Mitigation of Future Outbreaks

Krishna Patel, Tuan-Dung Le, Thanh Duong, and Dr. Thanh Thieu

Introduction

The novel coronavirus pandemic has put the world under new normality. In an effort to join hands fighting the pandemic, we set a goal to investigate geospatial, societal, and infrastructure factors that affect the spread of SARS-CoV-2. We use Natural Language Processing (NLP) to extract entities and relations associated with the spreading of the virus from the COVID-19 Open Research Dataset (CORD-19). We not only extract information about SARS-CoV-2 but also extract information on other coronavirus strains in CORD-19. Next, we aggregate the extracted information into a knowledge graph. This graph embodies factors that affect pandemic spread throughout the history of humankind. We hypothesize that the knowledge graph will provide interpretable insights into the rapid spread of COVID-19. The insights can then be used for policy making, increased preparedness, and prevention of any future pandemic outbreak.

Methodology

Stage 1
- **Data preprocessing: retrieve dataset, parse dataset, clean noise**

Stage 2
- **Extract COVID-19 related entities**

Stage 3
- **Extract relations between entities**

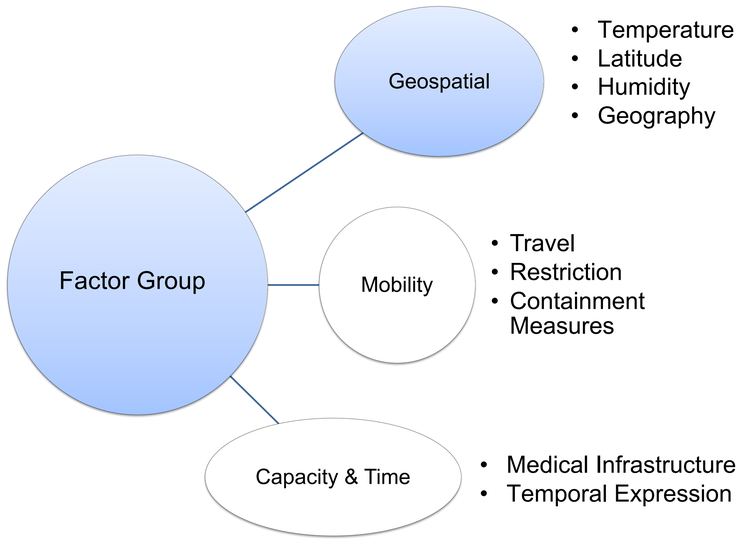
Stage 4
- **Create a knowledge graph from the extract entities and relations**

Data Collection

Statistics (Version 06-21-2021)

Total number of papers: 659,502, of which 307,698 since 2019 → Dropped 96,589 articles with missing abstract → Dropped 40,265 articles with duplicate abstract → Final dataset: 170,844 paper abstracts

Problem Formulation



Implementation

Human Annotations

(temperature)
The hot weather might makes some people, especially those with comorbidities or older ages,
(spread_outcome)
develop aggressive inflammation that ends up with complications and mortality.

MAIN BODY: A very rapid spread and high mortality rates have characterized the COVID-19
(latitude) (temperature)
pandemic in countries north of the equator where air temperatures have been seasonally low.

(spread_cause)
(temperature)
The results show that average temperature, minimum temperature, and air quality were
(Geo)
significantly associated with the spread of COVID-19 in LAC .

Supervised Machine Learning

spaCy

NLTK

Results

Trained 20 documents → 2 Annotators → Evaluated 5 documents

	Double Annotation A1-A2			A1 – Curation			A2 - Curation		
	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
Temperature	0.4	0.3	0.35	0.58	0.82	0.68	0.6	0.65	0.60
Humidity	0.56	0.17	0.26	0.26	0.85	0.4	0.75	0.75	0.75
Latitude	0.49	1.0	0.66	1.0	0.88	0.94	0.56	1.0	0.71
Geography	0.69	0.89	0.78	0.9	0.82	0.86	0.65	0.76	0.7
Spread_cause	0.09	0.21	0.13	0.97	0.67	0.8	0.05	0.07	0.06
Spread_event	0.0	0.0	0.0	0.86	0.61	0.71	0.05	0.58	0.1
Spread_outcome	0.05	0.13	0.07	0.48	0.92	0.63	0.32	0.32	0.39

	Precision	Recall	F-1
Total	0.3684	0.1892	0.2500

Conclusion

- Many direct relationships are seen between geospatial entities and SARS-CoV-2 related entities.
- The annotation stage shows numerous instances of geography, temperature, latitude, and humidity as a direct factor in the spread of SARS-CoV-2.
- Further research would involve extracting the relation between entities, along with creating a knowledge graph to exhibit the relationship.
- In addition to researching the effects of societal and infrastructure factors that affect the spread of SARS-CoV2.

References

1. Lucy Lu et al. (2020). {CORD-19}: The {COVID-19} Open Research Dataset. *Proceedings of the 1st Workshop on {NLP} for {COVID-19} at {ACL} 2020*. Retrieved June 01, 2021, from https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

2. AI, A. I. F. (2021, July 13). *COVID-19 Open Research Dataset Challenge (CORD-19)*. Kaggle. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

3. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."