A Path to an AI Strategy for Pandemic and Disease Surveillance

Anasse Bari, Deepti Saravanan, Jahnavi Swetha Pothineni et al.

Initial Draft - Work in Progress

New York University

We propose in this study an international AI strategy for disease surveillance that has the potential to detect, predict disease outbreaks, and prevent pandemics. We design an AI framework that combines passive and active disease surveillance. Active disease surveillance presented here encapsulates AI tools that leverages alternative data sources to epidemiology such search engines search queries, human mobility data, and microblogs among others. Further, active surveillance will also consist of a data ingestion tool that will collect data and detect patterns from workers that interact frequently with large number of citizens such as teachers, pharmacists, supervisors, human resources, and others. Existing passive surveillance tools will join forces with an AI based active disease surveillance for the purpose of creating early alters signals for governments, international organization, and healthcare decision makers to act swiftly and control an outbreak before it turns into a pandemic. We present in this work the first experimental AI tools, a data architecture, and data strategy for international global pandemic surveillance, as well as call for international players and organizations to play role in the proposed framework. The analytics components adopted in this framework consists of data collection, data cleaning, data fusion and data structuring. The AI capabilities proposed are natural language processing, causality models, supervised and unsupervised machine learning models.

Section 1: Introduction & Definitions

Disease surveillance is the systematic collection, analysis and interpretation of real-time data to understand what's happening currently during a pandemic to draft policies and measures accordingly. It could also be viewed from a different aspect as studying the pattern of the past pandemic to be better prepared for the future. This is immensely important to track the spread of diseases and curb outbreaks.

Disease surveillance could be achieved in two different ways (or a combination of these two): Active surveillance and Passive surveillance, as described by Bill Gates in his book 'How to Prevent the Next Pandemic'. Passive Surveillance involves collecting data from physicians and medical practitioners on reported cases. Though this process is less resource-intensive, we might miss out large amounts of data as not all cases are necessarily reported. This could be overcome by incorporating Active surveillance.

Active surveillance is the process of proactively tracking the spread of diseases via regular surveys. This process is resource-heavy but it helps in identifying and incorporating unreported cases.

Disease surveillance holds a huge significance in detecting and responding to emerging threats to public health. It also helps in evaluating the effectiveness of public health interventions and vaccination drives. It is a key component of the public health sector.

Section 2: Lessons Learned from the COVID-19 Pandemic

COVID-19 has been a global crisis affecting multiple aspects of life. The most important lesson learnt is the importance of preparedness for public health emergencies. Robust surveillance systems, adequate resources and accessibility come a long way in helping to curb the spread of disease.

Scientific research plays a critical role in addressing public health challenges and developing vaccines and diagnostic tools. It is also important for global cooperation to work together on battling the pandemic.

The other important lesson learned is the interconnection between human, animal and environmental health. It is imperative to recognize the interdependence of these systems and promote well-being across all these sectors, especially because a significant percentage of pandemic is caused by zoonotic virus, which infects both animals and humans.

Our study takes inspiration from these lessons and proposes an approach of incorporating both active and passive surveillance into account, while also studying the possible causal relationship between zoonotic infections in animals and humans.

Section 3: A Brief Survey of Existing Surveillance Efforts:

The use of google trends and social media data for disease surveillance has become increasingly popular in recent years. Researchers have used these alternate data to track the spread of infectious diseases such as influenza, Ebola, and COVID-19. In the study, Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada by Yousefinaghani S et al.,[1] they have evaluated digital data streams like Twitter and Google Trends to identify early warning signals of COVID-19 outbreaks in Canada and the US. The study finds that symptoms-related tweets and Google searches on symptoms were effective in predicting initial waves of outbreaks up to a week earlier in Canada and 2-6 days earlier in the US. An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time by Nicole E [2], evaluated digital data streams as early indicators of state-level COVID-19 activity and proposed a means of harmonizing these data streams to identify future outbreaks. The findings suggest that combining health and behavioral data can help identify disease activity changes weeks before observation using traditional epidemiological monitoring. Tim Mackey et al.[3] proposed an unsupervised machine learning to detect and characterize user-generated conversations that could be associated with COVID-19-related symptoms, experiences with access to testing, and mentions of disease recovery.

In the paper, A decision support framework for prediction of avian influenza [4], they have proposed surveillance tools to predict avian influenza outbreaks and the development of a decision support framework for aiding decision-makers in assessing future risks of disease events. The framework proposed uses a pre-built knowledge base to detect events and answer queries, with an average sensitivity and specificity of 69.70% and 85.50%, respectively, and has the potential to assist health care authorities in early control of emergency situations. The study by Israel Edem Agbehadji et al.,[5] highlights the implications of using nature inspired computing in the accurate detection of pandemic cases and optimized contact tracing.

Section 4: An AI Disease Surveillance Framework and Tools

4.0 System Architecture

We propose a big data analytics solution serving as a one-stop solution for understanding everyday trends related to various diseases such as COVID-19 and MPox. Each disease consists of multiple components that show different ongoing trends. This together would help in understanding any possible correlation between different unusual activities that might point to possible spread of a new disease or viral variant.

4.1 Components

The dashboard for different diseases would comprise components focusing on different aspects of the disease, including search for symptoms, preventive measures, treatments, transmission, economy and policies (such as unemployment fund). In addition, the dashboard would also include trends in the increase of related viral infections in animals, where abnormal spikes might serve as an early alert signal for potential transmission of the infection to humans.

4.2 Data Sources

For the active surveillance part, we plan to incorporate alternative data sources such as Google Trends Search, Reddit and other media resources.

We also leverage reliable data sources such as WHO and CDC for the number of covid cases and deaths, and also for zoonotic infection trends in the animals.

4.3 Players

In addition to the alternative data sources for the active surveillance part, our study also injects everyday data from players who interact with the general public such as teachers, pharmacists, supervisors, human resources, epidemiologists, local healthcare workers, public health officials and others.

Section 5: Proof-of-Concept:

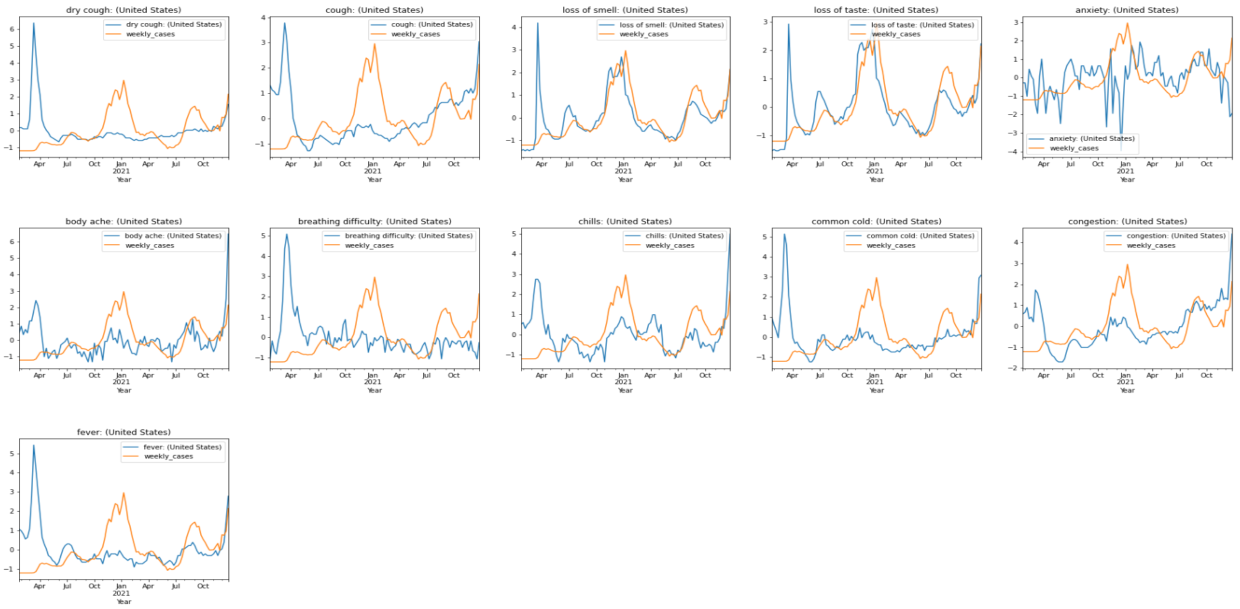
1. Google Trends

Using the Google trends, interest of keywords over time, which can be classified into five broad categories are extracted.

Symptoms

Keywords used are – **'fever', 'cough', ‘dry cough’, 'breathing difficulty', 'body ache', 'loss of taste', 'loss of smell', 'congestion', 'common cold', 'anxiety', ‘chills’**

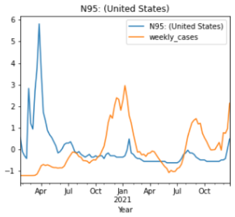
Google trend analysis of covid symptoms and cases

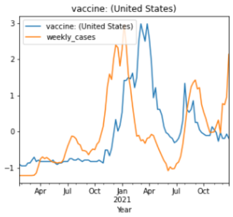


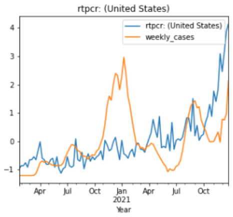
Prevention

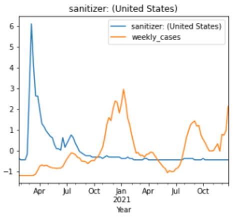
Keywords used are – **'vaccine', 'sanitizer', 'RTPCR', 'N95', 'Cloth face mask'**

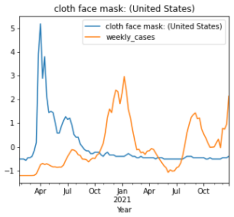
Google trend analysis of covid prevention terms and cases





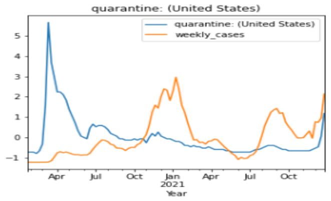


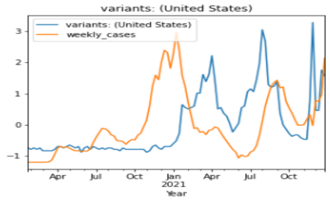




Transmission

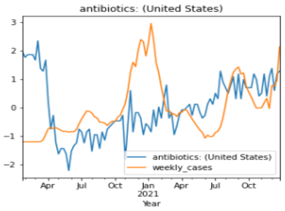
Keywords used are – **'variants', 'quarantine'**

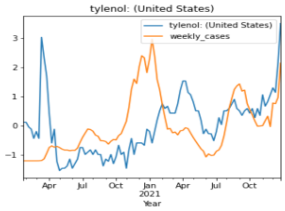


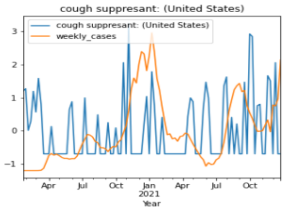


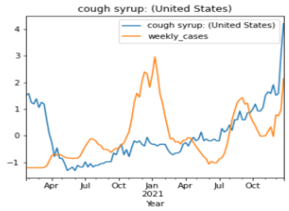
Treatment

Keywords used are – **'tylenol', 'cough suppressant', 'antibiotics', 'cough syrup'**



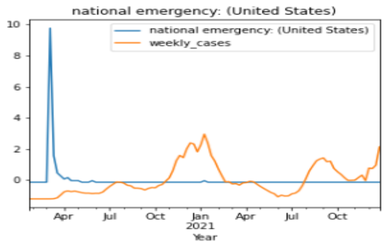


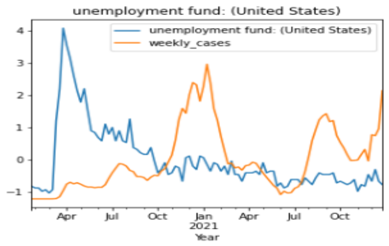




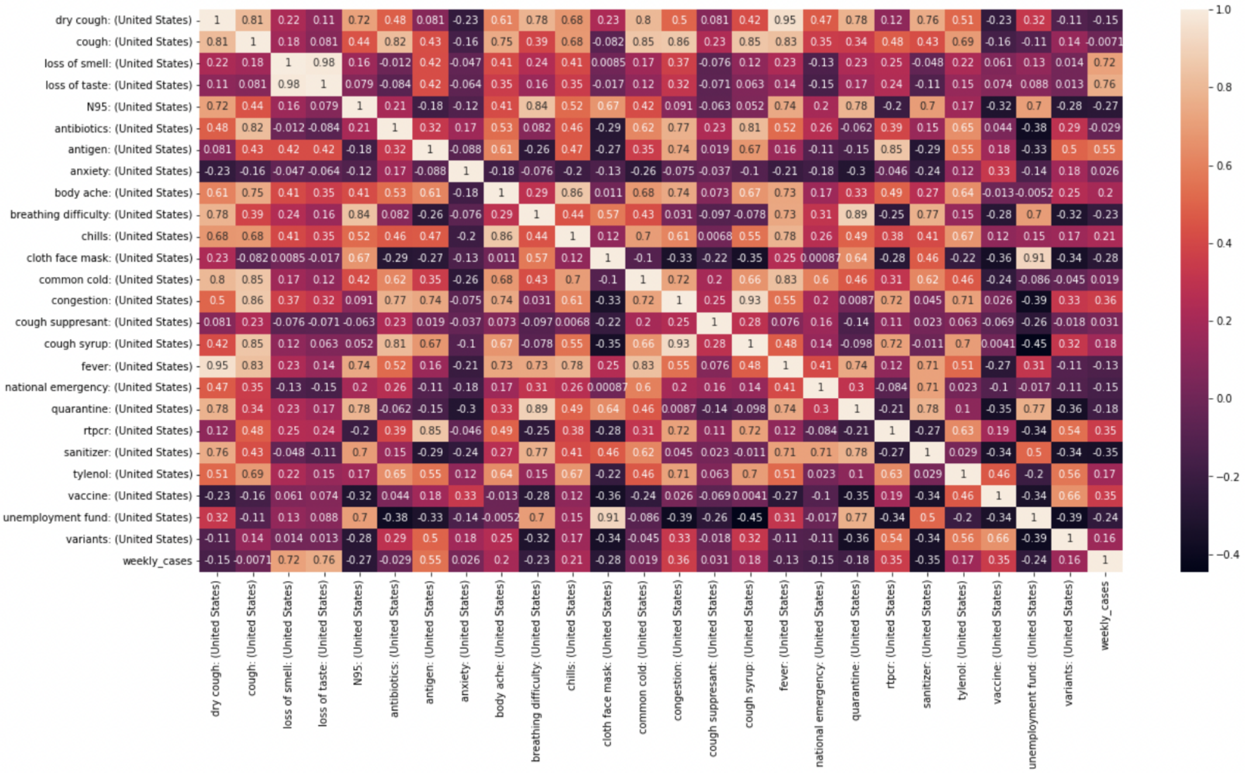
Economy

Keywords used are – **'unemployment fund', 'national emergency'**



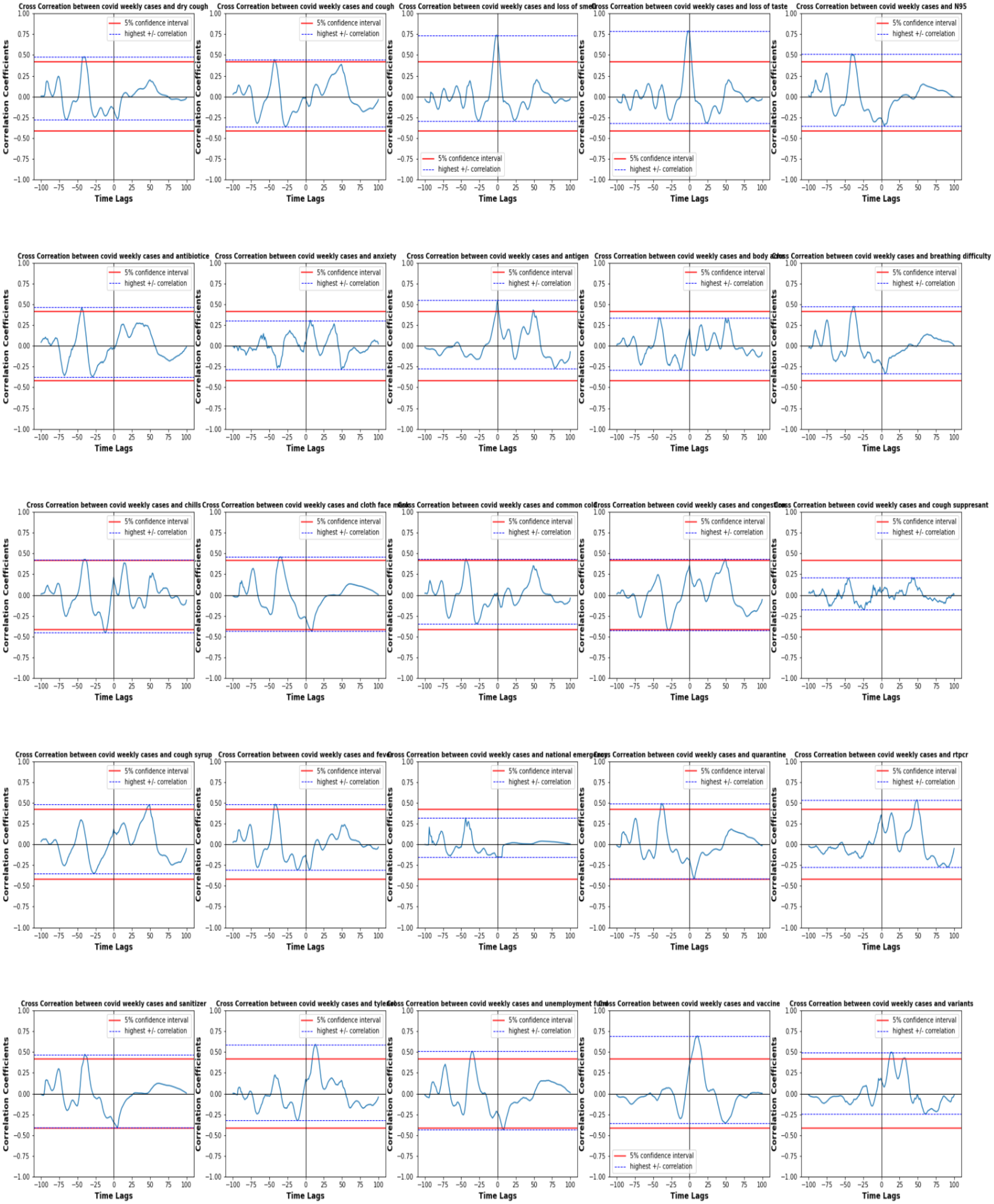


Correlation of the keywords



From the above heatmap, we can see that loss of smell, loss of taste and antigen have shown highest correlation with weekly covid cases.

Cross-Correlation of key words:



If the positive correlation occurs in the negative lag time of x, this means that there is a weak positive correlation between the google trends interest of that feature today and the covid cases x weeks later. Similarly, If the positive correlation occurs in the positive lag time of x, this means that there is a weak positive correlation between the google trends interest of that feature today and the covid cases x weeks before.

From the above graphs, we can see that loss of smell and loss of taste feature have a positive correlation with the covid cases with a lag of one week.

Stationary Test using Augmented Dickey-Fuller Test

Feature	p-value	Status
dry cough	0.027	Stationary
cough	0.409	Non-stationary
loss of smell	0.457	Non-stationary
loss of taste	0.051	Non-stationary
anxiety	0.007	Stationary
body ache	0.99	Non-stationary
breathing difficulty	0.001	Stationary
chills	0.904	Non-stationary
common cold	0.99	Non-stationary
congestion	0.99	Non-stationary
fever	0.89	Non-stationary
N95	0.00	Stationary
vaccine	0.153	Non-stationary
rtpcr	1.0	Non-stationary
sanitizer	0.177	Non-stationary
cloth face mask	0.00	Stationary
antibiotics	0.87	Non-stationary
tylenol	0.974	Non-stationary

cough suppresant	0.00	Stationary
cough syrup	0.99	Non-stationary
national emergency	0.00	Stationary
unemployment fund	0.081	Non-stationary
quarantine	0.673	Non-stationary
variants	0.460	Non-stationary

The above features which have p-value greater 0.05 are non-stationary. The non-stationary features are converted to stationary by differencing. After converting the above features to stationary, we have applied Granger's causality test to determine whether the features can help in the forecast of covid cases.

Granger Causality Test

Feature	F-Score	lag	p-value
dry cough	0.0355	1	4.5828
cough	15.3316	1	0.0002
loss of smell	11.3542	1	0.0012
loss of taste	34.1443	1	0.0000
anxiety	8.2802	2	0.0006
breathing difficulty	0.2810	1	0.5976
common cold	11.0765	1	0.0013
N95	4.8793	3	0.0038
vaccine	4.3417	1	0.0405
rtpcr	4.1702	1	0.0446
sanitizer	0.3393	1	0.5619

cloth face mask	0.7649	2	0.4690
antibiotics	0.4034	1	0.5273
tylenol	6.7084	1	0.0115
cough suppresant	1.2056	2	0.3053
national emergency	13.4678	1	0.00
unemployment fund	6.5339	2	0.0024
variants	3.3566	2	0.0402

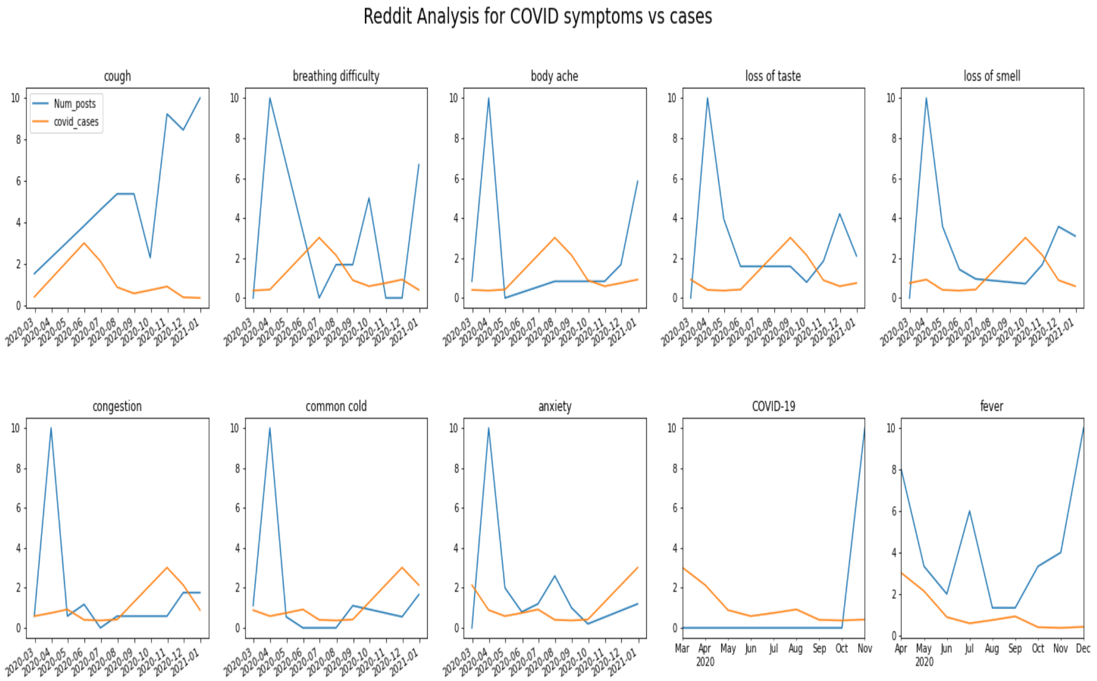
Cough, loss of smell, loss of taste, anxiety, common cold, N95, vaccine, rtpcr, tylenol, national emergency, unemployment fund and variants have p-value less than 0.05. This indicates that they could be useful in forecasting the number of covid cases.

2. Reddit:

Using the Reddit API, posts from the subreddit ‘CoronavirusUS’ were extracted based on keywords that can be put into five broad categories:

Symptoms

Keywords used are – **'COVID-19', 'fever', 'cough', 'breathing difficulty', 'body ache', 'loss of taste', 'loss of smell', 'congestion', 'common cold', 'anxiety'**



Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Reddit Posts	-3.97	0.0015
Number of Covid Cases	-1719896699504085.00	0.00

Since the p-value is less than 0.05 for both the features, **they are both stationary time series**.

Cointegration Test

The output is as follows:

t-statistic: -3.848
p-value: 0.01167

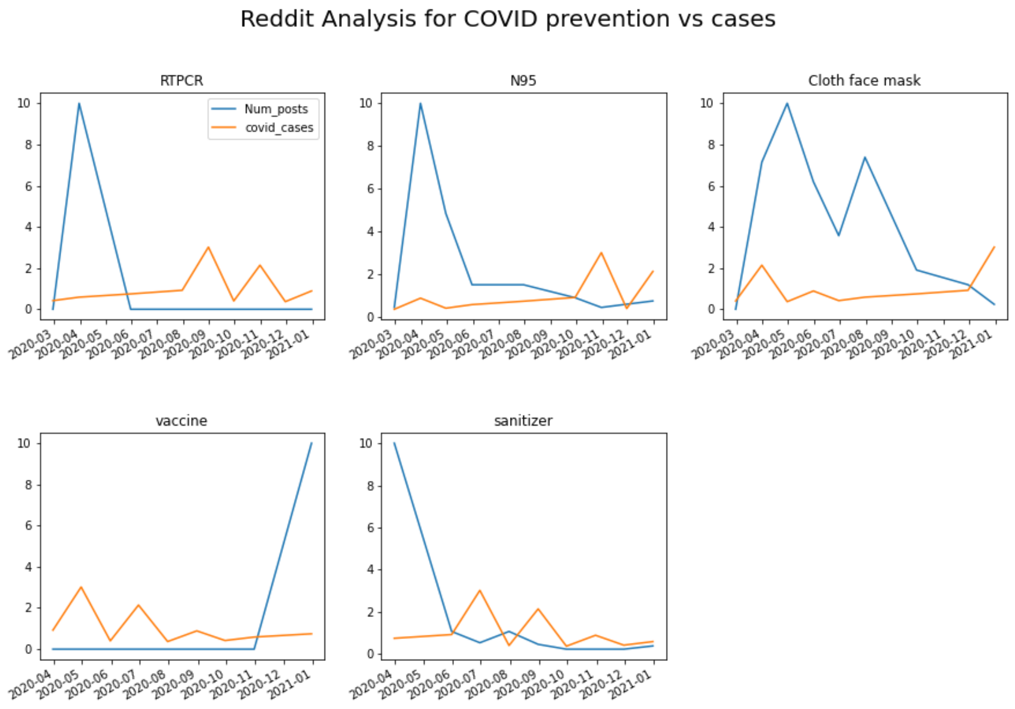
Since the p-value is less than 0.05, the timeseries of number of covid cases reported in humans and the number of reddit posts having the symptom keywords **are cointegrated**.

Granger Causality Test

The result **observed no causality relationship** between the number of reddit posts having the symptom keywords and the number of covid cases reported in humans. **The reverse causality is also not true.**

Prevention

Keywords used are – **'vaccine', 'sanitizer', 'RTPCR', 'N95', 'Cloth face mask'**



Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Reddit Posts	-4.601	0.00013
Number of Covid Cases	-2695843039549202.5	0.00

Since the p-value is less than 0.05 for both the features, **they are both stationary time series**.

Cointegration Test

The output is as follows:

t-statistic: -4.605

p-value: 0.00082

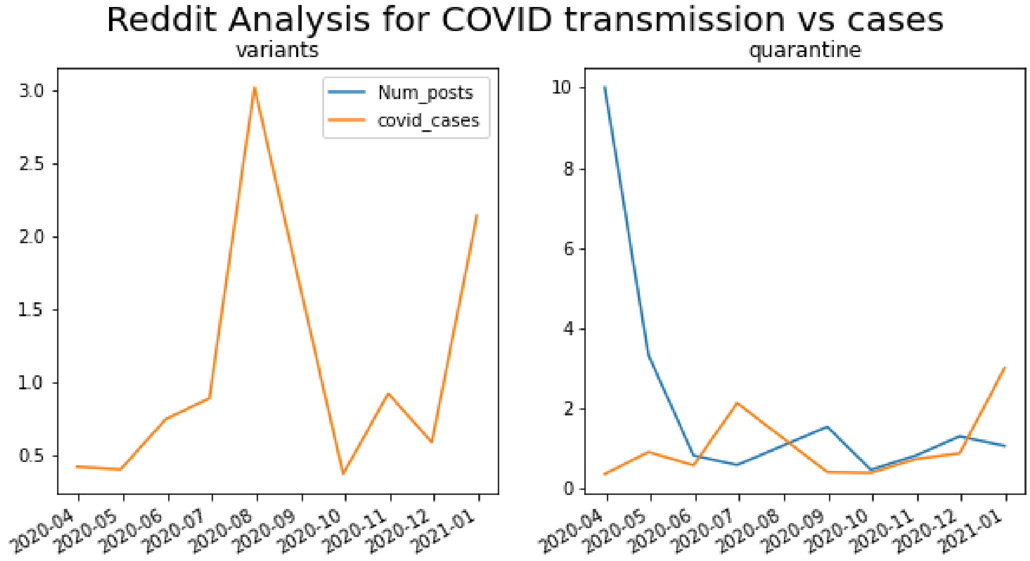
Since the p-value is less than 0.05, the timeseries of number of covid cases reported in humans and the number of reddit posts having the prevention keywords **are cointegrated**.

Granger Causality Test

The result **observed no causality relationship** between the number of reddit posts having the prevention keywords and the number of covid cases reported in humans. **The reverse causality is also not true.**

Transmission

Keywords used are – **'variants', 'quarantine'**



Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Reddit Posts	-4.72	0.000076
Number of Covid Cases	-1.876	0.343

Since the p-value is greater than 0.05 for the feature **number of covid cases feature,** it is a **non-stationary time series in this context.**

Cointegration Test

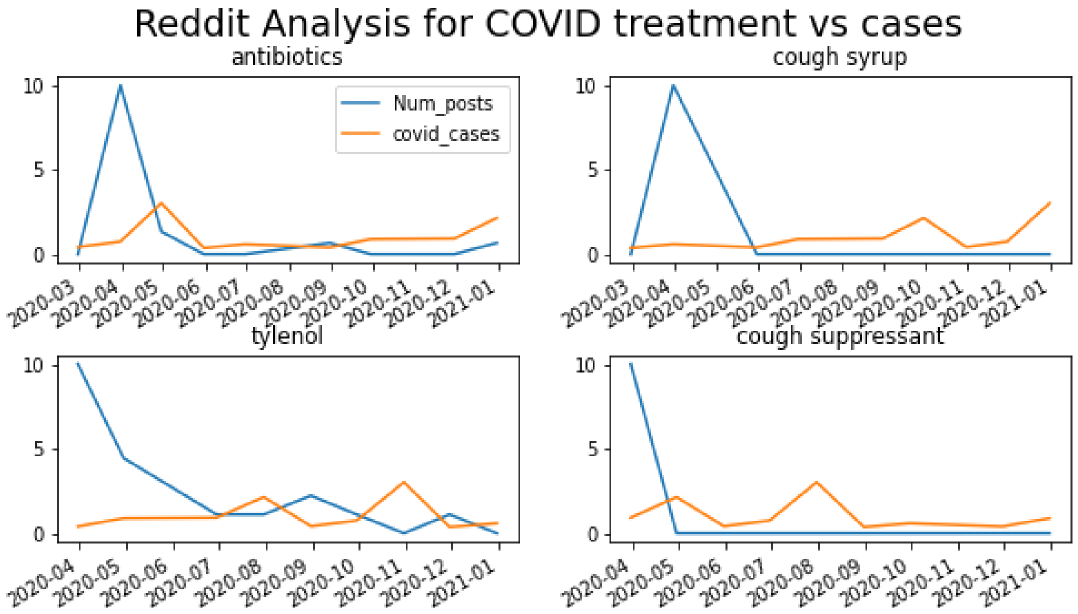
The output is as follows:

t-statistic: -4.581
p-value: 0.0009

Since the p-value is less than 0.05, the timeseries of number of covid cases reported in humans and the number of reddit posts having the transmission keywords **are cointegrated**.

Treatment

Keywords used are – **'tylenol', 'cough suppressant', 'antibiotics', 'cough syrup'**



Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Reddit Posts	-2.18	0.2145
Number of Covid Cases	-1175544860456421.0	0.00

Since the p-value is greater than 0.05 for the feature **number of reddit posts with treatment keywords,** it is a **non-stationary time series in this context.**

Cointegration Test

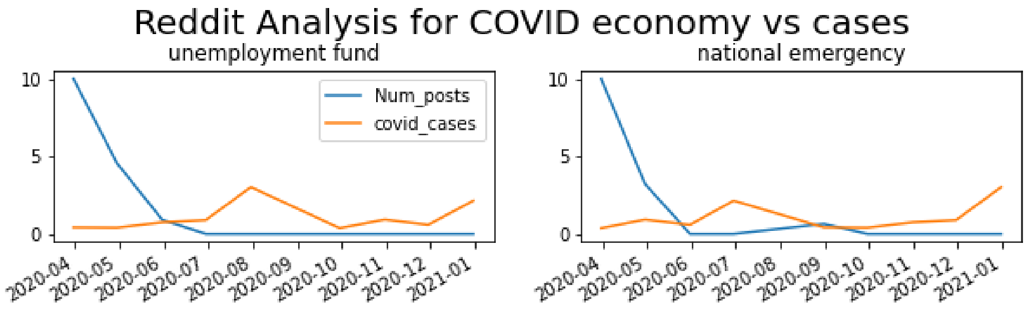
The output is as follows:

t-statistic: -3.848
p-value: 0.01167

Since the p-value is less than 0.05, the timeseries of number of covid cases reported in humans and the number of reddit posts having the treatment keywords **are cointegrated**.

Economy

Keywords used are – **'unemployment fund', 'national emergency'**



Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Reddit Posts	-2.584	0.096
Number of Covid Cases	-19390016410522.35	0.00

Since the p-value is greater than 0.05 for the feature **number of reddit posts with economy keywords,** it is a **non-stationary time series in this context.**

Cointegration Test

The output is as follows:

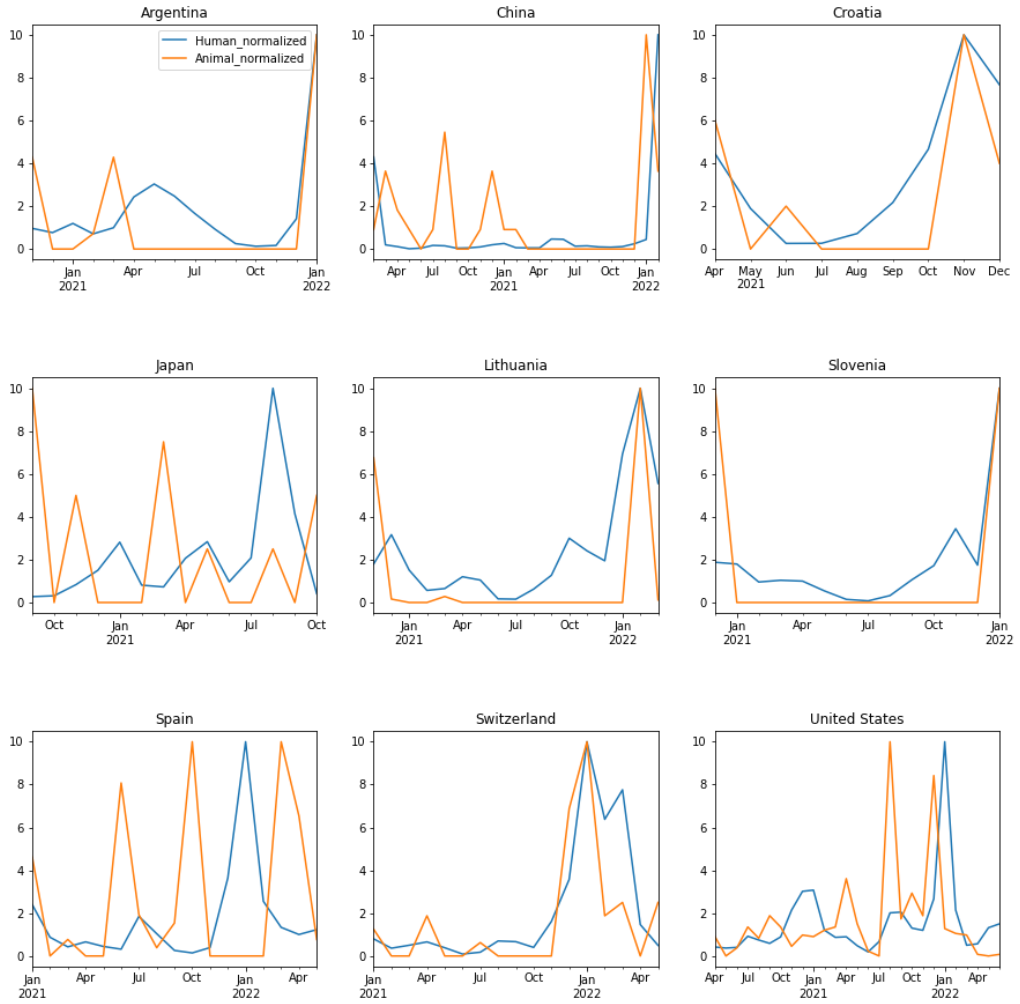
t-statistic: 0.0
p-value: 0.986

Since the p-value is greater than 0.05, the timeseries of number of covid cases reported in humans and the number of reddit posts having the economy keywords **are not cointegrated**.

3. Zoonotic Infections

The initial analysis focused on the number of reported cases of SARS virus in animals and humans in a sample few countries across the world to study if there is any observable correlation.

Number of SARS cases in humans and animals



As can be seen from the above plots for a few countries, the peaks of covid cases in humans mostly follow peaks in SARS cases in animals.

Stationary Test using Augmented Dickey-Fuller Test

Features	**ADF Statistic**	**p-value**
Number of Human Cases	-7.56	0.00
Number of Animal Cases	-7.56	0.00

Since the p-value is less than 0.05 for both the features, **they are both stationary time series**.

Cointegration Test

The output is as follows:

t-statistic: -8.696
p-value: 5.12e-13

Since the p-value is less than 0.05, the timeseries of number of covid cases reported in humans and the number of SARS cases reported in animals **are cointegrated**.

Granger Causality Test

The result, as can be seen from the table, **observed causality** between the number of SARS cases reported in animals and the number of covid cases reported in humans with a **maxlag of 4 months** (best case scenario of high chi2 value with the highest significance). That is, *when there is a peak in number of cases in animals, we can expect a similar peak in the number of cases in humans roughly after 4 months*. **The reverse causality is not true.**

Maxlag (in months)	**Chi2 value**	**p-value**
1	6.283	0.012
2	9.892	0.007
3	15.049	0.0018
4	**17.760**	**0.0014**
5	18.129	0.0028

Section 6: Experiments and Results
Section 6: Conclusion and Future Work
Section 7: References