

CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: B. Aditya Prakash
Scribe: Aishwarya Vijaykumar Sheelvant

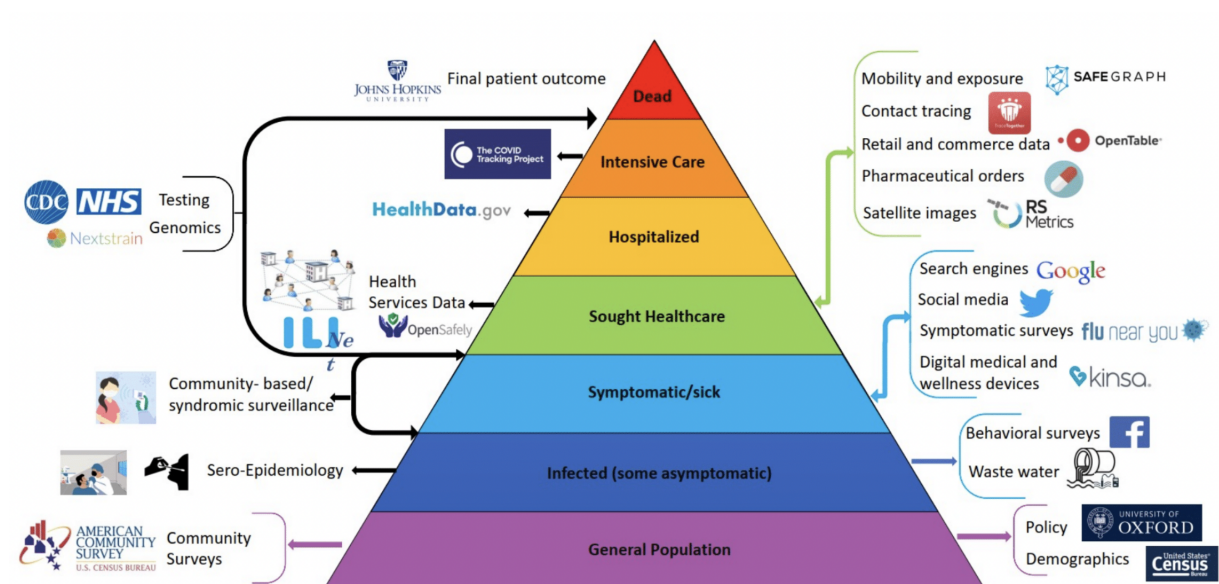
October 13, 2022
Lecture 14 : TITLE Surveillance-II

1 Introduction

In this lecture, the main topics of discussion are the Surveillance techniques for the Symptomatic/sick layer of the surveillance pyramid using surrogate data. Machine learning and Statistical models can be used for epidemiological surveillance and predicting the present data based on the history: Nowcasting. In this lecture we discuss the pitfalls and multiple case studies of Surveillance.

2 Surveillance

Surveillance pyramid: Surveillance can be performed on different parts of the pyramid.



We need different techniques to handle data. We use ML techniques for surveillance of search engines, social media, symptomatic surveys, digital medical and wellness devices. There are positive, negative, and ugly aspects to these surrogate data sources. The distributional shifts have an impact on the ML techniques.

Distributional changes in public health can have a significant impact on search query results. Many of these sources are extremely sensitive to the distribution. A lot of these sources are very sensitive to the distribution.

Proposals for flu surveillance:

Search queries : "Miley Cyrus cancels concert over Flu"

The search queries can be really affected by distributional shifts in public health.
 OTC medication sales: The sales could be affected by factors such as hoarding, discounts, can be affected by inflation.
 Wikipedia: Lack of patient data.
 Self reinforcing and self defeating prophecies The required reading provided talks extensively about pitfalls of using such digital data and one of the core challenges is that the data that receives attention is not the output of instruments that are reliable and valid.

No solution has yet been identified. Without human interaction, there is no dataset that can be used that will produce the desired outcomes. Caution has be taken while implementing this approach.

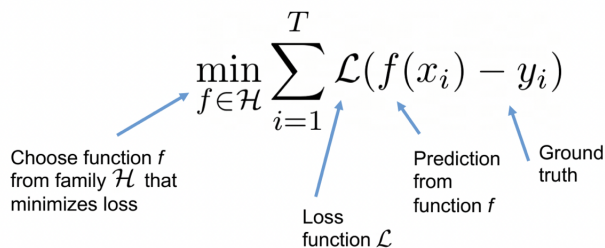
3 Nowcasting

The idea of Nowcasting is to predict the present rather than the future. This is because many socio-economic public health indicators do not maintain present data. For example, to get data of how many active COVID cases are present at the moment would be impossible. There's no way to answer such questions. Governments periodically release data which are lacking. This is true not just for health related data but even for social media data. We employ surrogate data sources to approximate the present for this reason.

The idea behind nowcasting is to use statistical and machine learning models to select from a family of functions the optimal function to approximate the forecast target given the input data. Due to the fact that we are anticipating the current behavior based on past experience, nowcasting is not all that different from forecasting. Using the current data, forecasting involves predicting the subsequent k values. In terms of ML models it's not that different from forecasting but in practice a few techniques are more popular in nowcasting than in forecasting (Regression) and nowcasting is also much more reliant on data sources and indicators than forecasting.

Nowcasting with Statistical & ML Models

- Intuition:
 - Find the best function from a family of functions that approximate forecast target given input data.
 - Best approximate is found using past training data.



4 Google Flu Trends

It was a pretty influential system when it was published in 2009. It was a nowcasting system for monitoring health-seeking behavior through google queries. 50 million candidate queries were narrowed down to set of 45 queries that most accurately fit CDC ILI data in the US. Google gives information on a keyword's most popular search patterns. Therefore, GFT looked for the queries that had a good correlation with the CDC ILI data. Queries correlated with the flu season were hand pruned. It was a combination of both statistical and manual effort where they pruned large queries into robust sets of queries. Relative query volume were used as independent variables. Initially they started out with 45 queries and then they started to categorize into classes of queries. The model was a relatively straightforward linear model. Not revealing the features or queries was a key design element. One of the reasons for this is to avoid introducing bias to the general population, as well as the fact that some of the questions were embarrassing. They had a correlation-based automated selection system. It was an effective model. Google created a dashboard so the trends were visible to everyone.

Model in Google Flu Trends

- Simple linear model for nowcasting ILI
- Use search logits of query fractions as features
$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon$$
 - P = ILI (physician visits)
 - Q = Fraction of search queries that are ILI-related
- ILI (Flu) -related queries
 - Automated selection based on proprietary set of keywords

Figure 1: GFT model

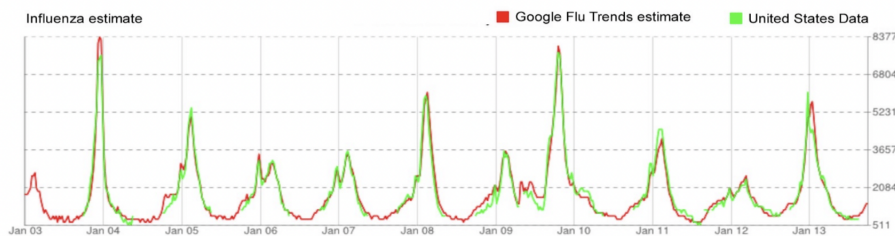


Figure 2: Output of GFT in comparison to CDC

The model was effective upto 2009 H1N1 pandemic. However during H1N1 pandemic, GFT could not capture the changing trends. GFT completely missed the first wave of the pan-

demical and portrayed misleading correlations. GFT evaluated at 3 geographic scales and densely populated areas and it could not extrapolate to other regions. In 2012-2013, GFT overestimated the intensity of H3N2.

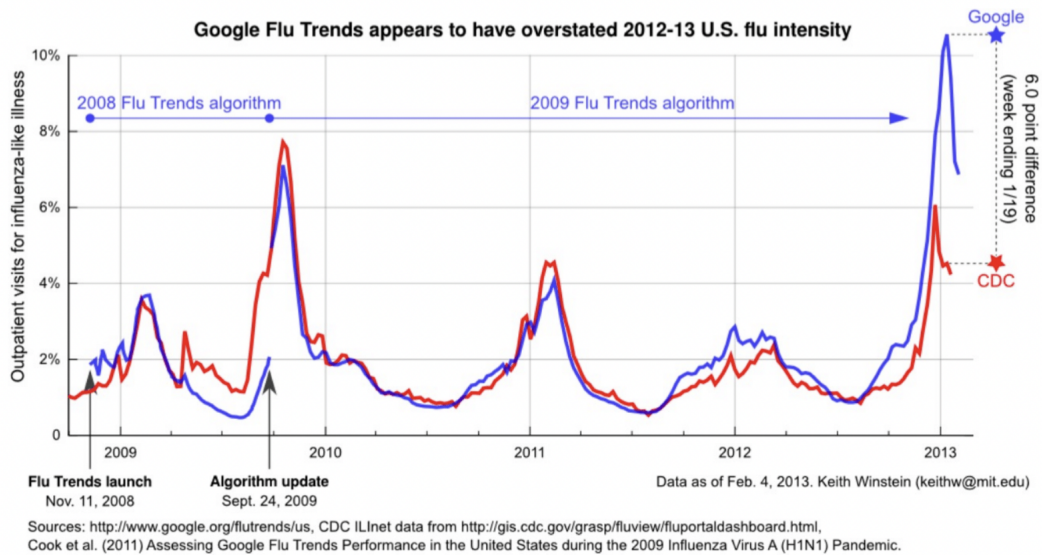


Figure 3: Output of GFT in comparison to CDC during H3N2

GFT had problems with news articles. The search algorithm was not static and as a result health seeking behavior was also changing. The other problem was search terms lacked transparency so they could not find what was wrong in the search queries. By 2013 estimates were off by 30% and then GFT was shut down as people realized that it was no longer reliable. Google asked independent research groups to work on it.

The main improvements to GFT were : Ignoring inorganic queries which heightened media coverage-This was found using spike detectors. Handling drift by retraining the model every year and using regularized parameters.

The dictionary used worked well in English speaking countries. There were no resources for other languages.

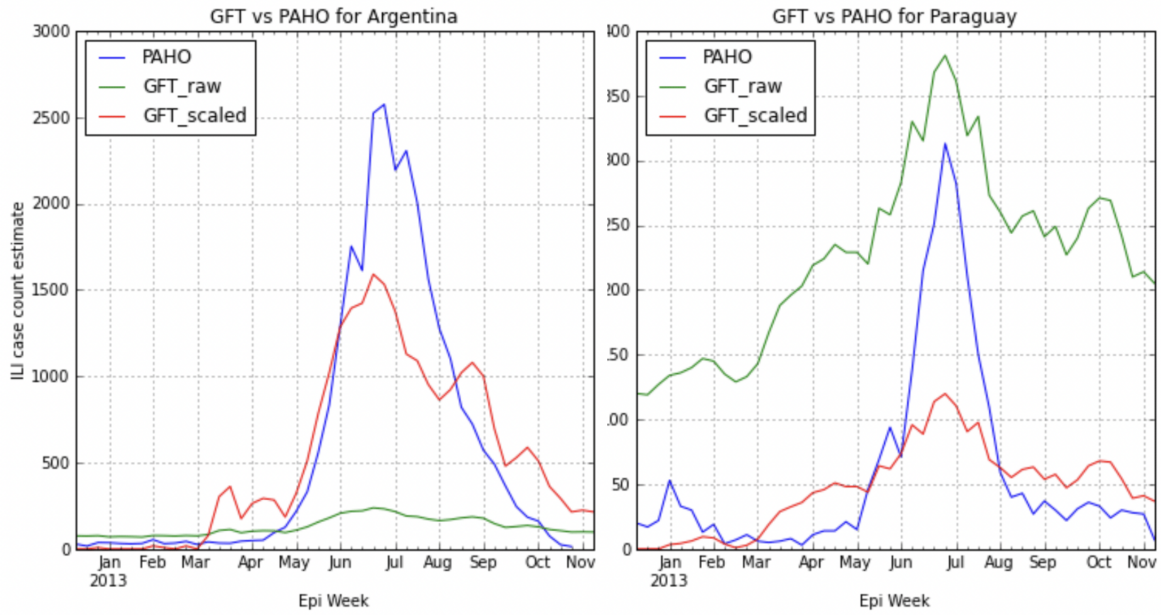
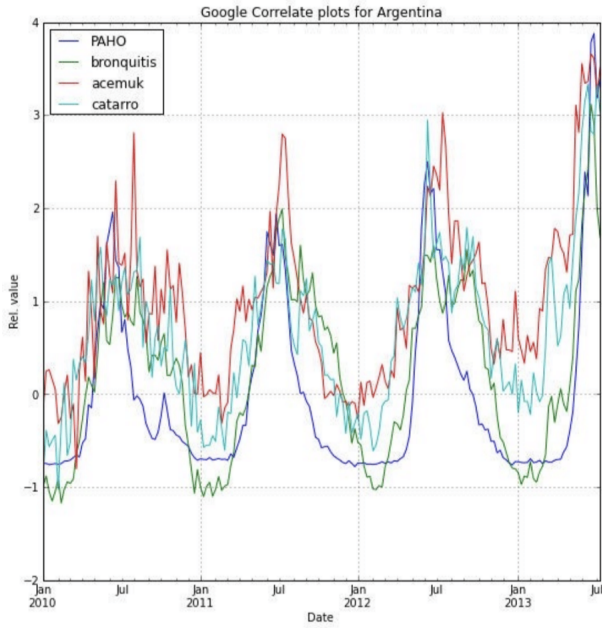


Figure 4: GFT for Latin American countries

Therefore a new dictionary had to be designed. The following approaches were used to build a dictionary:

Query expansion method : to define own queries which starts with queries collected by health ministry. Datasets related to flu were collected.

Google correlate : used to correlate keyword search query volume with PAHO time series. After these improvisations were done GFT worked well with different languages and with query expansion.



Symptomatic words:
 “bronquitis”, “catarro”, “tos seca”
 (whooping cough)

Medicinal words: “acemuk”,
 “claritromicina” (clarithromycin)

Interesting words: ginger
 (“jengibre”), leave letter (“letra de
 deja”)

Figure 5: GFT after new dictionary had been introduced

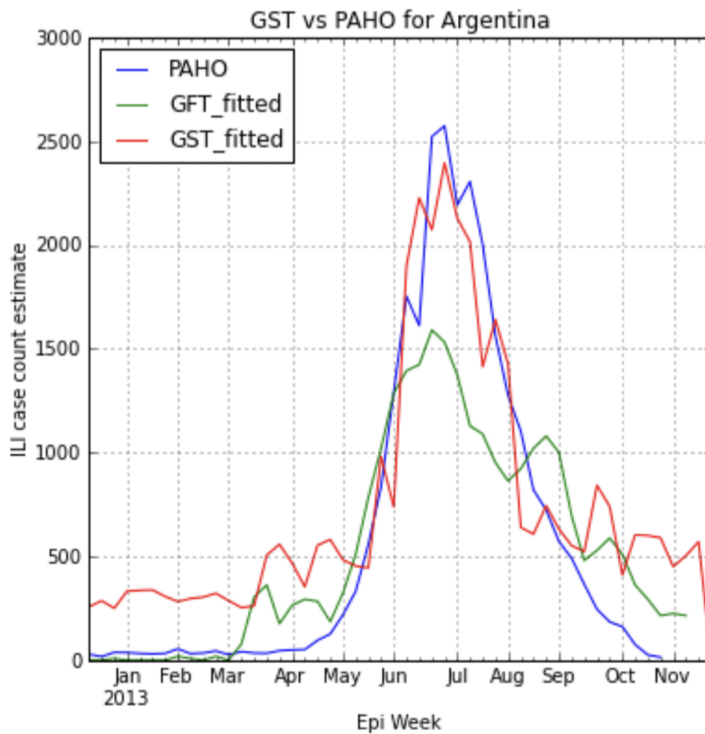


Figure 6: GFT after new dictionary had been introduced

5 Food borne illness detection

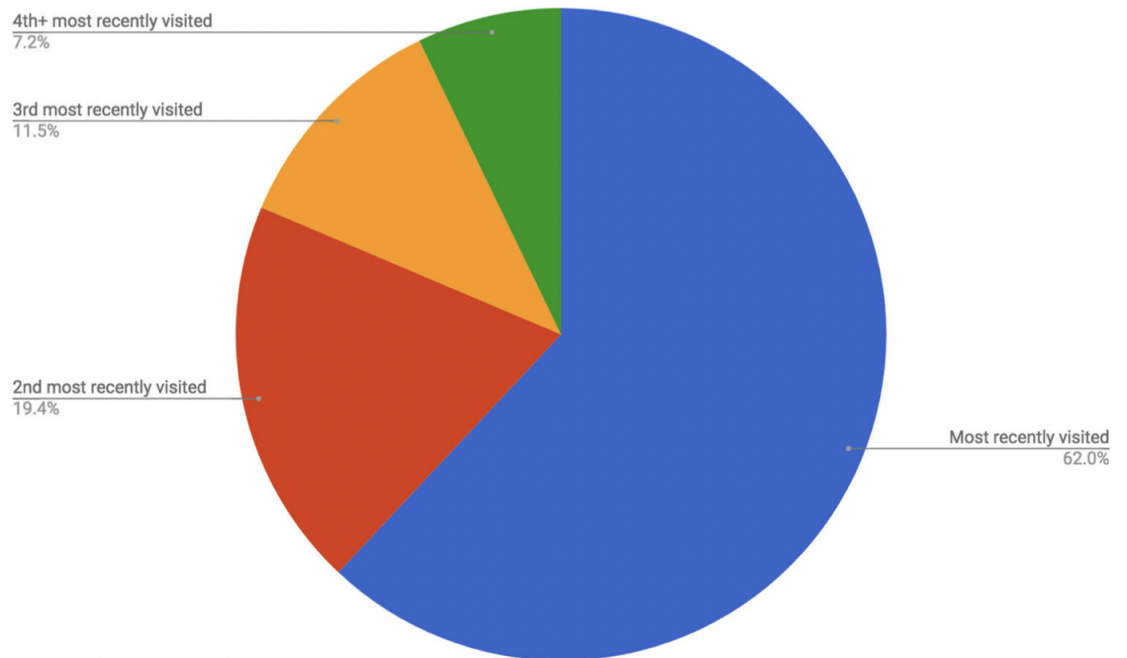
Another area where the application of search queries can be seen is Food borne illness: The idea is to apply Machine learning models to Google search and location logs to find out which restaurants have poor sanitation and are causes of food borne illness. Few of the approaches which can be followed are checking the previous location data, mapping the history of check ins of an infected person.

The approach followed was:

- Identification of queries indicative of food borne illnesses.
- Find the restaurants visited in aggregate by the users who issued the queries.
- Calculate for each restaurant the proportion of people who visited the restaurant and showed symptoms of food borne illness.

One of the challenges in this approach was that the user may have visited multiple restaurants. An algorithm to narrow down to one restaurant based on voting system or something similar to a message passing system had to be built. For each restaurant the calculated the proportion of users who visited it and then showed symptoms of illness. The important part of this study is it used passive data (location) collected from users. Collecting passive data comes with greater responsibility. They developed a privacy preserving machine learning classifier approach which leverages a collection of signals beyond the query itself, such as results shown in response to the search query. The location history was also privacy reserved. This study was implemented in Las Vegas and Chicago. This method has a more targeted approach to testing the restaurants which caused a significant reduction in the amount of work needed. The results of this study is as follows:

Results



Frequency with which illness can be attributed to recently visited restaurants, among FINDER restaurants. $N = 132$

6 Nowcasting with twitter

In the light of the success of the previous approach many people began to use social media data for nowcasting. This study focuses on adopting GFT like ideas to forecast ILI case counts using Twitter. This study uses geolocation to narrow down to regions of interest and document filtering to identify tweets related ILI .Regression was then performed and it was observed that multiple keyword independent variables performs better than simple linear regression which was used in GFT.Lasso based linear regression model was used to predict the number of cases.

Twitter data worked well during the H1N1 pandemic.Geolocation tweets were collected based on US home locations containing specific flu related keywords.Filtering was done on queries by removal of stopwords and processing with stemming.Support vector regression was performed to map dictionaries to CDC ILI rates and a new dictionary was created. Since the decline of GFT due to data grips and distributions,people started to focus more on content analysis and therefore coding rules were used to categorize the tweets.An example of the categories can be found below.

Table 1. Descriptions and Examples of Content Categories.

Content	Description	Example Tweets
Resource	Tweet contains H1N1 news, updates, or information. May be the title or summary of the linked article. Contents may or may not be factual.	"China Reports First Case of Swine Flu (New York Times): A 30-year-old man who flew from St. Louis to Chengdu is.. http://tinyurl.com/rdbhcg " "Ways To Prevent Flu http://tinyurl.com/r4l4cx #swineflu #h1n1"
Personal Experience	Twitter user mentions a direct (personal) or indirect (e.g., friend, family, co-worker) experience with the H1N1 virus or the social/economic effects of H1N1.	"Swine flu panic almost stopped me from going to US, but now back from my trip and so happy i went :-)" "Oh we got a swine flu leaflet. clearly the highlight of my day" "My sister has swine flu!"
Personal Opinion and Interest	Twitter user posts their opinion of the H1N1 virus/situation/news or expresses a need for or discovery of information. General H1N1 chatter or commentary.	"More people have died from Normal Flu than Swine flu, its just a media hoax, to take people's mind off the recession" "Currently looking up some info on H1N1" "Swine flu is scary!"
Jokes/Parody	Tweet contains a H1N1 joke told via video, text, or photo; or a humorous opinion of H1N1 that does not refer to a personal experience.	"If you're an expert on the swine flu, does that make you Fluent?"
Marketing	Tweet contains an advertisement for an H1N1-related product or service.	"Buy liquid vitamin C as featured in my video http://is.gd/y87r #health #h1n1"
Spam	Tweet is unrelated to H1N1	"musicmonday MM lamarodom Yom Kippur Polanski Jay-Z H1N1 Watch FREE online LATEST MOVIES at http://a.gd/b1586f "

The hypothesis was that only a subset of tweets were actually useful and the rest of it was just spam. It was found that 52.6% of the tweets were about news and information and 4.5% were misinformation. The tweets were passed through multiple filters to find out if they were useful or just unrelated information. After applying these filters regression was performed.

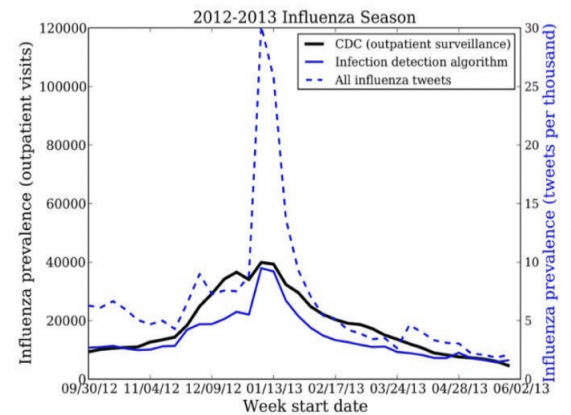
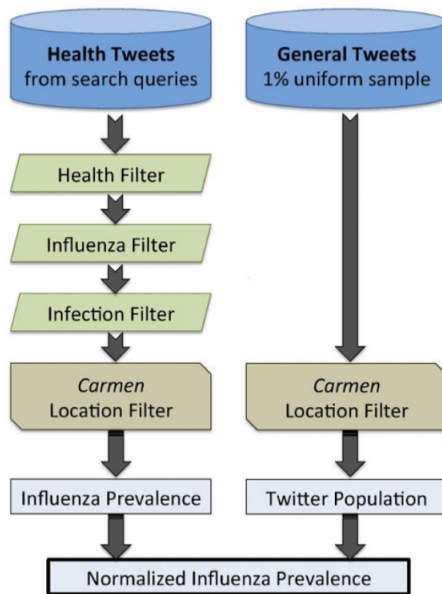


Figure 7: Filters used to find the related tweets

From the graph it can be seen that the Infection detection algorithm follows a similar trend to the CDC data. Many other approaches have followed this method of filtering out unrelated data. For example a paper by Lamb in 2003, tries to develop further distinctions

by trying to distinguish between infection vs concerned awareness by building meaningful classifiers and building parts of speech templates from world class features.

Class Name	Words in Class
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Possession	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
Past Tense	was, did, had, got, were, or verb with the suffix "ed"
Present Tense	is, am, are, have, has, or verb with the suffix "ing"
Self	I, I've, I'd, I'm, im, my
Others	your, everyone, you, it, its, u, her, he, she, he's, she's, she, they, you're, she'll, he'll, husband, wife, brother, sister, your, people, kid, kids, children, son, daughter

Figure 8: Parts of speech templates

The paper titled "Flu Gone Viral" by the professor, proposes a temporal topic model for inferring the biological state of the user and an EM algorithm for modelling the hidden epidemiological state of the user(S,E,I,R). The Hidden Flu State from Tweets(HSFTM) model generates the state form the tweet and then the topic from the word. Then EM algorithm is used to infer the topic distributions and state transition probabilities. This method may suffer from large noisy vocabulary and can be improvised by introducing an already curated list of keywords from an expert.

- There are different states in an infection cycle.
- SEIR model:
 1. Susceptible
 2. Exposed
 3. Infected
 4. Recovered

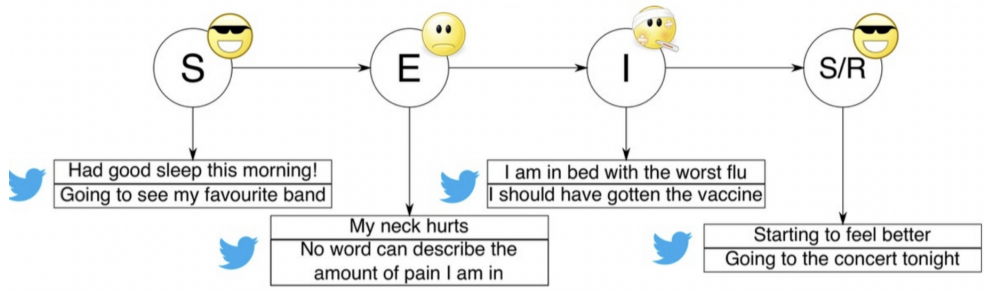


Figure 9: SEIR model

• Generating tweets

Generate the state for a tweet
Generate the topic for a word

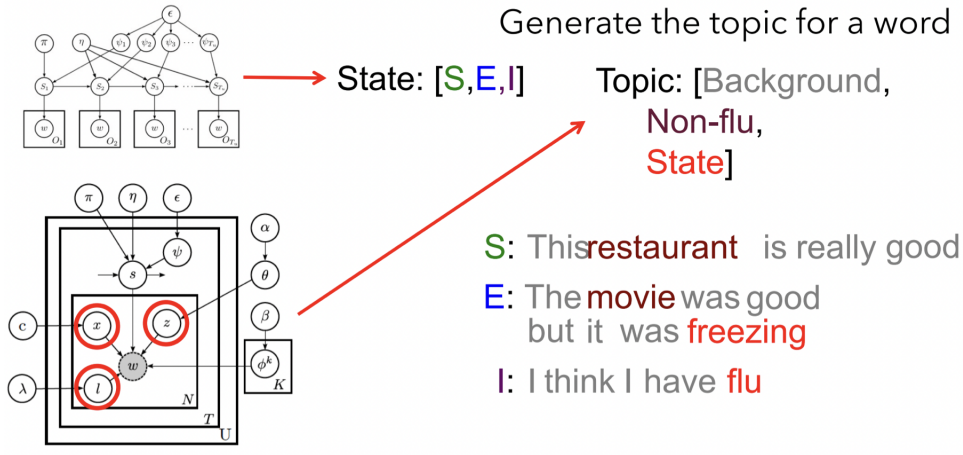


Figure 10: HFSTM

Observations: The most probable words in each state:

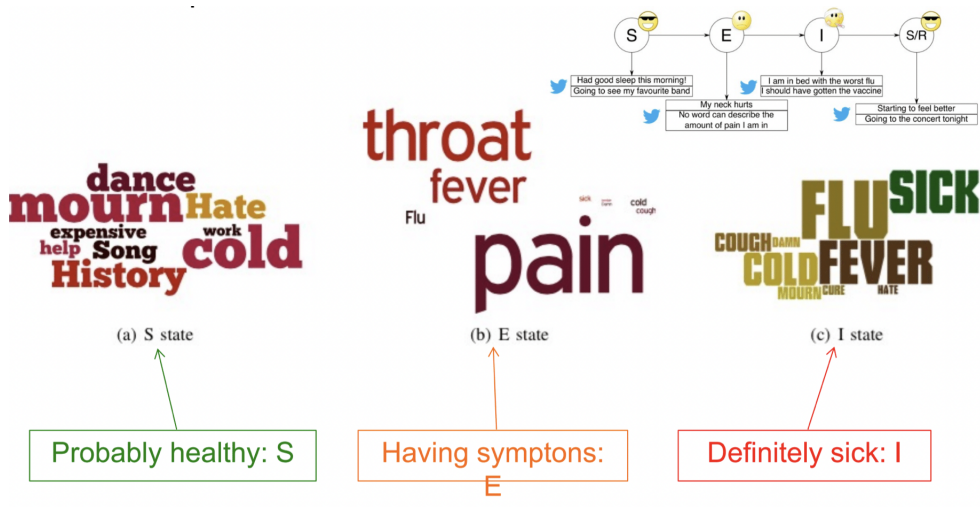


Figure 11: Learned word distributions

Data obtained from PAHO was considered as ground truth. The number of keywords were counted as features and the ground truth curve was regressed. Google flu trends data was used to regress the PAHO curve. Using the HSFTM, they distinguished the states of the keywords, and only the keywords in I state were identified the again used to regress to PAHO. The model can learn transition probabilities (S,E,I,R). This model is observed to learn transitions well.

7 Tracking COVID-19 with Online Search

This study uses time series of online search query frequencies to gain insights about the prevalence of COVID 19 in multiple countries. They first built unsupervised modeling techniques based on associated symptom categories identified by United Kingdom’s National Health Service and Public Health England and then they created an online search time series. One of the challenges was to clean the data. They tried to minimize an expected bias in these signals caused by public interest. Symptom categories were weighted based on their reported ratio occurrence in cases of COVID-19. They reduced the effect of news via autoregression. Their study confirms the unsupervised approach’s insights and demonstrates how early warnings may have been gathered from areas that had already felt the effects of COVID-19. Then they conducted a correlation and regression analysis to uncover potentially useful online search queries that refer to underlying behavioural or symptomatic patterns in relation to confirmed COVID-19 cases. The output of this model provides useful insights including early warnings for potential disease spread, and showcases the effect of physical distancing measures.

- Reduce effect of news via autoregression. For the weighted score of symptom-related online searches g :

$$g = g_p + g_c$$

Where infected (g_p) and concerned (g_c) users. Then, there exists a constant $\gamma \in [0, 1]$ such that

$$g_p = \gamma g \quad \text{and} \quad g_c = (1 - \gamma)g.$$

On any given day the proportion of news articles about the COVID-19 pandemic is $m \in [0, 1]$, then:

$$\arg \min_{w, v, b_2} \frac{1}{N} \sum_{t=1}^N \text{AR}(g, m): \text{autoregressive function on } g \text{ and } m$$

$$(g_t - w_1 g_{t-1} - w_2 g_{t-2} - v_1 m_t - v_2 m_{t-1} - v_3 m_{t-2} - b_2)^2,$$

Figure 12: Model formulation

- Online searches precede the reported confirmed cases by 16.7 (10.2–23.2) and deaths by 22.1 (17.4–26.9) days.

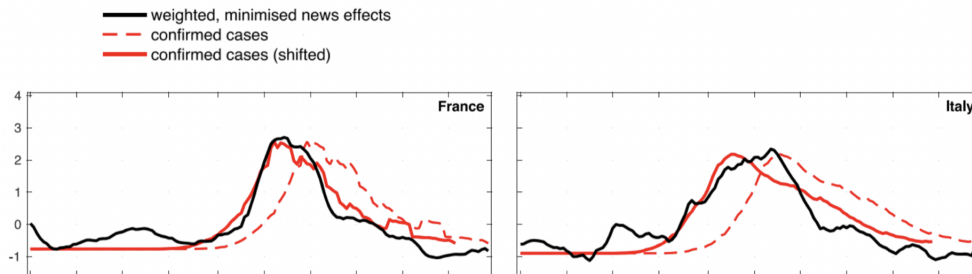


Figure 13: Early warning signal by the model

- Correction to weighted score only when news media signal helps

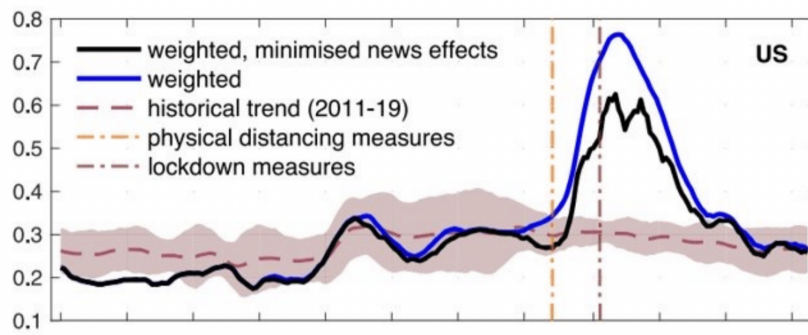


Figure 14: Results of the model