Nowcasting COVID-19 hospitalizations using Google Trends and LSTM

Guillaume Derval , Vincent François-Lavet , Pierre Schaus
ICTEAM, UCLouvain, Belgium
{first, second}@uclouvain.be

Abstract

The Google Trends data of some keywords have strong correlations with COVID-19 hospitalizations. We attempt to use these correlations and show an experimental procedure using a simple LSTM model to nowcast hospitalization peaks using Google Trends data. Experiments are done on French regions and on Belgium. This is a preliminary work, that would need to be tested during a (hopefully non-existing) second peak.

1 Introduction

Multiple models have been proposed to estimate various metrics of the ongoing COVID-19 crisis, such as the number of infected individuals, the R_0 indicating the exponential growth factor of these infections. Beyond estimation of these metrics, some models have also attempted to make predictions on these metrics.

The number of hospitalizations is one of these metrics. In Belgium and in France, the countries on which we experiment in this paper, hospitalization is one of the most consistent widely available metrics of the state of the epidemic. The number of individuals tested positive for the virus is dependent on the way testing is done (in Belgium and France, this changed multiple times, depending on the countries' capacities), and the number of deaths imputed to the virus is not reliable (depending on the country, only tested positive individuals count in this metric) nor comparable between countries.

Forecasting the daily hospitalizations in a given region is thus of a vital importance:

- Detecting peaks would allow evaluating early governmental measures, such as confinement, and to adjust them
- Moreover, it would give some days to hospitals to prepare for the incoming flow of patients.
- After a peak, a model estimating future hospitalizations can serve as a warning method for a second peak.

We propose to use Google Trends to predict hospitalizations one week in advance. Google Trends is a tool linked to Google Search, a well-known web search engine. Terms typed into the search engine are logged, and aggregated result of the frequency of searches of these terms are given as time series.

Search engine logs have been used previously in related context, including in the now-defunct Google Flu[Ginsberg *et al.*, 2009], but also for other purposes such as migration forecasts [Böhme *et al.*, 2020].

The question we aim to answer is

Given the history of some keywords on Google Search, what will be the hospitalizations in seven days?

In order to do this, we first discuss about data acquisition and preprocessing, then show a short analysis of this data. We then experiment with simple LSTM[Hochreiter and Schmidhuber, 1997] models and demonstrate that this approach gives good results.

Research using the same principles is being conducted independently by various teams [Ortiz-Martínez *et al.*, 2020; Ayyoubzadeh *et al.*, 2020; Mavragani and Gkillas, 2020].

2 Data acquisition and preprocessing

2.1 Google Trends

Google Trends differentiate between *keywords*, which are literal words typed as-is in the search engine (for example, ”truck” and ”Truck” are different keywords), and *subjects*, which are actually group of terms. As an example, the ”truck” subject regroup both the keywords ”truck” and ”Truck”, but also ”trucks” and ”camion” (translation of truck in French). Using subject thus allows having a multilingual approach, which is useful given the countries we use in our experiments. Belgium has three national languages. Subjects are referenced in Google Trends via an *id*, which seems to always begin by ”/m/”.

We selected some subjects related to symptoms, illnesses and other medical aspects. They are listed in Table 1.

The data is extracted from Google Trends via their web interface. A query to Google Trends consists of mainly two parameters: a list of keywords/subjects, and a time window. The data is returned in a normalized form: the biggest frequency over all the horizon and keywords/subjects is defined to be 100, and every data point is rounded to the nearest integer. Moreover, the time resolution depends on the time window size. To ensure that we always have one point per day, we apply the following algorithm:

Name
Symptoms
Cough
Sore Throat
Quarantine
Respiration
Anosmia
Olfaction
Ageusia
Influenza-like illness
Hospital
Tissue
Diarrhea
Migraine
Lung
Uncomfortable
Fever
Taste
Temperature
Pharmacy

Table 1: Selected Google Trends Subjects

- Download each month separately. This gives daily data points, but each month is normalized to 100, which makes them uncomparable.
- Download the same data for a year. This aggregates the data points weekly, and again is normalized to 100.
- Renormalize the data of each month such that their highest point for a week (mean of the 7 data points for a week) has the right scale.

We thus obtain a time series with a time resolution of one day, normalized to 100 (a small error, due to rounding, occurs here).

As explained above, it is possible to do a query with multiple keywords/subjects at a time, allowing comparing the frequency of multiple keywords/subjects between each other. This is not needed for the purpose of the data, and comes with a price in terms of precision. We thus query each subject individually, and each one is normalized to 100 w.r.t itself.

In order to avoid the effect of the weekend and other repeating cycles impacting the data, we use a rolling, centered, average on a 7-days period.

The data is downloaded and processed for the 22 French regions and Belgium (as a whole). Note that as some French department merged in 2016, and as Google did not adapt its Trend tool, we had to aggregate some data point to reform the new regions (we use the mean of the data points over the old, merged regions, which is not accurate as they have different populations but is sufficient in practice).

2.2 Hospitalizations

We use data from governmental sources [Sciensano, 2020; Santé publique France, 2020]. As explained in the introduction, we do not use the total number of hospitalized persons but rather the number of new entrance each day. As week-ends and other weekly cycle have an influence on the data, we apply a centered rolling average of 7 days on the data.

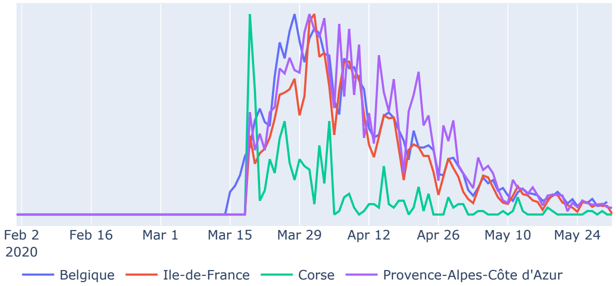

Figure 1: Normalized hospitalization curves for some regions (without rolling average)

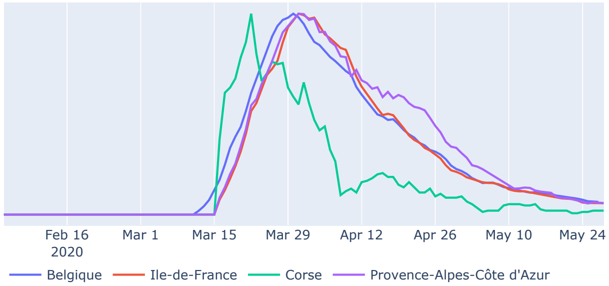

Figure 2: Normalized hospitalization curves for some regions (with a rolling average of 7 days)

We, moreover, normalize the data for each department between 0 and 1, by dividing by the maximum number of hospitalizations in the department. This allows avoiding predicting absolute numbers, which are dependent on the population, its repartition, and other factors that will not be given in inputs of the models presented hereafter. Instead, the models are designed to detect peaks and trends in the slope of the hospitalizations.

This is done in a preprocessing step, for the whole time series of given regions. Note that this does not forbid models to predict peaks greater than 1, which means they have a bigger incidence than the previously seen peaks.

As the daily report of the COVID-19 hospitalizations started with some lag in France and Belgium. Therefore we prepend the official data with zeroes (i.e. no hospitalizations) until the first of February. This should allow the models to fit a longer period without the presence of the virus. This is again an approximation, particularly on the days immediately before the beginning of the reports, but its influence should be minor.

3 Overview of the data

At the moment of writing this paper, we currently lie at the end of the first (and hopefully last) wave of the COVID-19 pandemic in France and Belgium. See Figure 1 and 2 for plots of the hospitalizations in different regions.

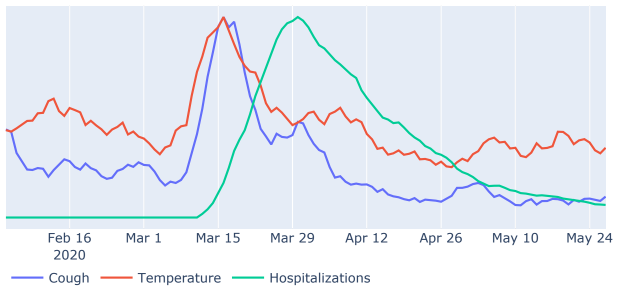

Figure 3: Trends for the subject *Cough* and *Temperature*, compared to the hospitalization curve (in Belgium)

Subject	Opt. delay	Opt. correlation
Diarrhea	13	0.92
Ageusia	4	0.85
Pharmacy	10	0.84
Respiration	7	0.84
Fever	14	0.83
Anosmia	7	0.77
Symptoms	16	0.77
Cough	13	0.76
Olfaction	5	0.74
Lung	9	0.74
Temperature	14	0.73
Quarantine	15	0.71
Migraine	44	0.64
Sore Throat	15	0.55
Taste	0	0.47
Hospital	18	0.46
Influenza-like illness	18	0.31
Uncomfortable	11	0.29
Tissue	44	-0.05

Table 2: Optimal delays that maximize subject correlation with the hospitalization curve

Note that the peaks happen approximately at the same time.

A first interesting analysis is to detect if, alone, a Google Trends subject can be correlated with the time series of hospitalizations with a given delay. Let us take, as an example, the subject *Cough* and *Temperature* in Belgium, and compare them to the hospitalization curve in Figure 3.

There is a delay of approximately 14 days between the peaks of the curves. We measured this delay by selecting the one that corresponds to the maximum correlation between the hospitalization curve and the delayed curve. For all the chosen subjects from Table 1, we obtain the delays and correlations in Table 2.

This shows that the Google Trends subjects have indeed correlations with the hospitalization, at various delays. A sudden increase in these subjects can be used to raise a first alarm. As a warning, we note that the peaks occur in a third of the subjects with a delay around 14 days, which makes the date of the peaks in the searches around the 13th of March, which is the data of the beginning of the confinement in Belgium. These analyses will need to be replicated when a second peak occurs to check for possible hidden correlations. This is a known and important limitation of this exercise.

4 Method and models

In order to forecast changes in the hospitalization curve and its value at seven days, we experiment with LSTM-based models in this section.

We use a simple many-to-many LSTM model in the following form:

1. A first LSTM layer with a LSTM nodes
2. A second LSTM layer with b LSTM nodes
3. n_{dense} dense layers containing between $n_{\text{dense nodes}}$ nodes each
4. A final dense layer returning one output, the (normalized) estimation of the hospitalization at D+7.

There are multiple additional hyperparameters such as the regularization type and weight, the optimizer to be used, the number of epochs, and the type of activation in the dense layers.

The dataset was split randomly into three parts: the training set (seven regions), the validation set (three regions) and the testing set (four regions). The training set is obviously used to train the models, and the validation set is used by a random grid search algorithm to find the best hyperparameters.

The dataset is given time point per time point to the model, each time as a vector containing the Google Trends subjects values on this particular day, the day of the week (all these features are renormalized between -1 and 1), and the hospitalizations seen this day. The model outputs a single number: the estimated hospitalizations seven days later.

We use the Mean Squared Error (MSE) as error metric and loss for the optimization.

5 Results

During our experiments, the model that performed best with respect to the validation MSE is the following:

- $a = 10$, $b = 0$ (no second layer), $n_{\text{dense}} = 0$ (no additional layers)
- Regularization L2 with weight $1e-3$.
- The Adam[Kingma and Ba, 2014] optimizer
- 500 epochs
- Default activation on the LSTM nodes.

This model was used to compute the metrics (Mean Square Error (MSE) and Mean Absolute Error (MAE)) shown for each region in Table 3. Figure 4 shows the results obtained for each region in the validation and test set.

For most regions but *Corse*, the model is able to reproduce the peaks. *Corse* is indeed a particular region: disconnected from France by the sea, its epidemic dynamics have been different, with the peak occurring very soon after the reporting started. In this case, the model is not able to learn efficiently, and mostly reproduces the given curve with a 7 days delay.

This either shows that the model is not able to learn in this case, or that the trends in Corse are different.

Set	Region	MSE	MAE
Train	Centre-Val de Loire	5.84e-3	0.048
	Bourgogne-Franche-Comté	1.70e-3	0.024
	Auvergne-Rhône-Alpes	1.00e-3	0.017
	Alsace-Champagne-Ardenne-Lorraine	1.46e-3	0.021
	Pays de la Loire	2.10e-3	0.029
	Belgique	0.55e-3	0.014
	Ile-de-France	1.26e-3	0.019
	Overall	1.99e-3	0.025
Val.	Corse	33.62e-3	0.106
	Nord-Pas-de-Calais-Picardie	5.97e-3	0.048
	Bretagne	1.79e-3	0.027
	Overall	13.79e-3	0.060
Test	Languedoc-Roussillon-Midi-Pyrénées	1.82e-3	0.026
	Aquitaine-Limousin-Poitou-Charentes	3.57e-3	0.038
	Provence-Alpes-Côte d'Azur	2.28e-3	0.031
	Normandie	2.36e-3	0.030
	Overall	2.51e-3	0.031

Table 3: Error metrics

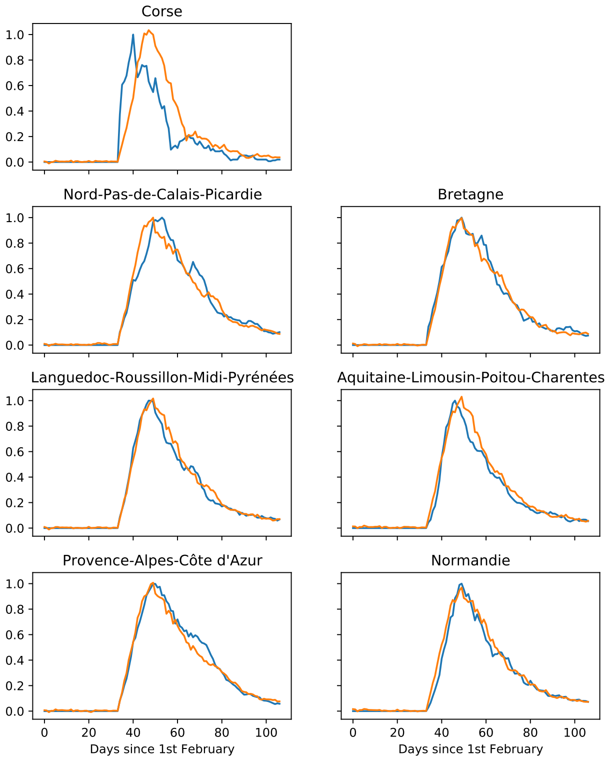

Figure 4: Result for the validation set (first three plots) and test set (bottom four plots). In blue the ground truth, and in orange the prediction at 7 days.

6 Conclusion, limitations and future work

The results presented above are encouraging but non-conclusive, and should be taken with utter caution. The limitations are mainly caused by the proximity of all the hospitalization time series used in this work and the peaks occurring mostly at the same time in the different regions taken into account. Expanding the dataset to other countries and regions could solve this problem, but is not straightforward: different cultures use words differently in the same context. This paper shows preliminary work, and an extension of the dataset is the next step do to in future work. Moreover, using more complex forms of validation, and more complex models should also be attempted, along with an analysis of the usage of the features by the trained models.

Source code

The source code and the data used are available on Zenodo: https://zenodo.org/record/3880044 .

References

[Ayyoubzadeh *et al.*, 2020] Seyed Mohammad Ayyoubzadeh, Seyed Mehdi Ayyoubzadeh, Hoda Zahedi, Mahnaz Ahmadi, and Sharareh R Niakan Kalhori. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR public health and surveillance*, 6(2):e18828, 2020.

[Böhme *et al.*, 2020] Marcus H. Böhme, André Gröger, and Tobias Stöhr. Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142:102347, January 2020.

[Ginsberg *et al.*, 2009] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, February 2009.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. Publisher: MIT Press.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[Mavragani and Gkillas, 2020] Amaryllis Mavragani and Konstantinos Gkillas. On the predictability of covid-19 in usa: A google trends analysis. 05 2020.

[Ortiz-Martínez *et al.*, 2020] Yeimer Ortiz-Martínez, Juan Esteban Garcia-Robledo, Danna L. Vásquez-Castañeda, D. Katterine Bonilla-Aldana, and Alfonso J. Rodriguez-Morales. Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia. *Travel Medicine and Infectious Disease*, April 2020.

[Santé publique France, 2020] Santé publique France. Données hospitalières relatives à l'épidémie de COVID-19 - data.gouv.fr, 2020. Library Catalog: www.data.gouv.fr.

[Sciensano, 2020] Sciensano. Epistat – COVID-19 Monitoring, 2020.