Forecasting influenza-like illness rates using online search data : A Literature Survey

Samuel Bouilloud

Abstract: Prediction of influenza rates and outbreaks have become crucial in our world, as the need for public entities to better allocate their resources increases. Governments are used to track influenza-like illness (ILI) rates, through reports from health professional in the countries. However, these methods are too slow and can't even tell us what the current state is. In our modern world, internet is an obvious input to nowcast these rates, but also to predict the future epidemics. Indeed, several studies attempted to gather internet-based data, like search queries or tweets, to combine them with the existing ILI data. Thus, they aimed to forecast influenza activity, including peak intensities and timings. The goal of this literature survey is to review the different methods attempted to nowcast and forecast influenza, as studies have shown that the use of internet data through several machine learning models is crucial in nowadays predictions. We will evaluate and compare the different methods, showing that even if they are capable of accurate forecasts, further research is needed to precisely capture influenza outbreaks.

1. Introduction

As influenza is a worldwide matter of public health, it is critical and needs all the attention possible by public entities. For these to better allocate their resources, our main goal here is to be able to accurately forecast the rates of influenza-like illness, and more particularly in the event of an epidemic. At the moment, the only accurate rates of influenza we have are the ones provided by entities like the Centers for Disease Control and Prevention (CDC) in the United States. Their Influenza-like Illness Surveillance Network (ILINet) releases each week the US rates. However, the main problem is that these rates are released with a 1-2 week lag, and can be later updated. For that reason, researchers started to think about the first task that comes to mind, the nowcast of ILI rates. In order to accurately do that, they needed more real-time information, and used online data, which was proven to be very useful. Indeed, Google search queries, Twitter tweets or even Wikipedia access rates, were very largely used on that task.

The main present goal is, of course, to predict the future rates of influenza, in any place in the world, with past and current data from that region. We will see that, even if some real-time electronic records are starting to emerge in some places, we need the online data to get accurate forecasts of ILI rates. The main, and more difficult task, is to predict the peak times and heights during influenza outbreaks and for that, several methods can be used. For more information, the forecasting task is summarized in figure 1.

We will here review the state of the art for online ILI forecast, as a big amount of papers exist, with very different approaches. First, we will see the basis that was brought to this research field and that opened the path for many more studies, with Google Flu Trends and the nowcasting task. Next, we will focus on how many more different type of internet-based data can be used and improve the forecasting of influenza, and then evaluate the main machine learning methods that were used and tested. Finally, we will take a look at improvements that are being researched, particularly in the pre-selection of the data, but also about a more spatial approach of the forecasting task.

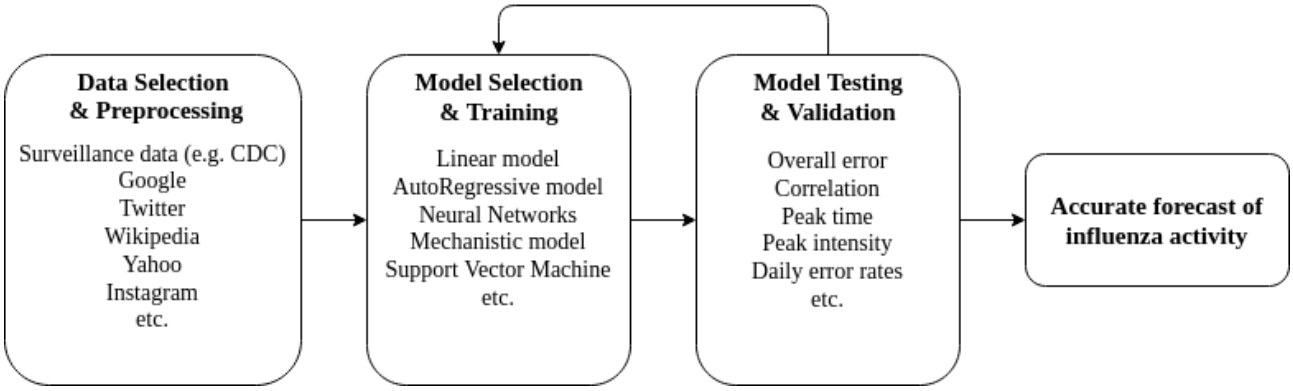

Figure 1: Summary of the forecasting task with online data

2. Nowcasting & Using Google search queries

In this first section, we will see what were the first and most basic iterations of influenza nowcasting, as they were imagined by Google in 2009, when they launched Google Flu Trends, a public platform to get indications about the current state of the flu. Indeed, knowing that people usually search online when they are ill, Google used its own data to come up with game-changing methods, opening a research field where many more professional would commit.

2.1. The baseline : Google Flu Trends

In the first paper introducing Google Flu Trends (GFT), Ginsberg et al. [1] used the previous assumptions to train a machine learning linear model with their search engine query data, in 9 regions of the US. Indeed, they first gathered the rates per queries for each available week (2003-2008) and used a linear model to check which queries related the most to the true ILI rates :

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

where P is the predicted ILI percentage, Q the selected query rate and ε an error term. Thus, the values of β_0 and β_1 are the values we need. With that, they tested each query and could choose the best 45 queries, by checking correlations with the true ILI rates. Then, they used the aggregate of these queries rates per week to train the linear model (i.e. compute the unknown coefficients) and perform accurate nowcasts.
Even this paper being at the time a clear progress in the field, the results are far from perfect. Indeed, GFT completely mispredicted the 2012-13 flu epidemic season, as the media coverage was huge and related google activity as well. To explain that, there are multiple methodical causes. First, the model was validated on one year only (2007-2008), which proved not to be enough, and it was mainly assessed using correlation. For that reason, even if the curves had similarities in shape, the model undervalued a lot the ILI rates, a clear weakness in the event of an epidemic.

2.2. Improving the nowcasting task

Taking into account the failure of GFT, Lampos et al. [2] attempted to improve the nowcasting of ILI rates using Google search queries. Indeed, they used individual queries rates per week (as a vector) on a linear regularized model, called Elastic Net. The regularization adds some penalty on the features, reducing the weights of some of them to 0 (i.e. the irrelevant queries). Furthermore, they also explored with a nonlinear model, by dividing the queries (selected from the Elastic Net model) into clusters, in order to reduce the impact of singular outliers. They used Gaussian Processes (GPs), a model that seeks to learn the function that links a prior distribution (computed from the queries vector) to an output value, our ILI rate. Finally, they augmented the two approaches by inputting them into an

autoregressive model (explained in section 3), that takes into account previous available CDC data. Like that, the paper creates a relationship between two sources of data, from present and past, in order to nowcast the influenza rates.
The results from this paper showed that these models alone are a lot more performing than the baseline GFT, but that they still struggle to capture influenza peaks, particularly in the event of epidemic (2009-10, 2012-13). Furthermore, the autoregressive addition clearly improved the models, especially the GPs, allowing it to get the 2009-10 peak. We see from this paper that past data obviously matters to "predict" the current rates, alongside present internet data, but also that influenza rates evolution is not necessarily a linear relation. Finally, the feature selection and queries clustering was here more conclusive, with the right queries having the most impact, making the data pre-selecting a crucial step in predicting influenza.

3. Forecasting & The use of other internet-based data

In this section, we will introduce the forecasting task, along with the use of other types of internet-based data, like tweets and Wikipedia access rates. We will see that this new type of data enhances the possibilities for our task.
First, Paul et al [3] thought about a way to improve influenza nowcasting and forecasting, by using Twitter data, alongside GFT data. Here, they present two versions of their linear autoregressive model: the first predicts the ILI rates of a certain week $w + k$ (when $k = 0$, it is nowcasting) by using only previous weeks CDC's data; and the second one that adds the influenza tweets rates. Although their results show that the use of this type of data improves both nowcasting and forecasting, this paper presents some approximations. Indeed, not mentioning the fact that they only trained the model on three seasons, the cross-validation does not respect time periods (e.g. trained on 2013-14 data, tested on 2012-13), and doesn't take into account the fact that the way people use Twitter changed a lot from 2006 to 2014. Moreover, it bases its results on GFT data, that we have seen mispredicted influenza a lot, leaving us to relativize on the results presented, that could have been superior.
In addition, McIver et al. [4] used Wikipedia access rates to enhance the nowcasting possibilities of influenza. Indeed, they selected a group of English pages related to ILI and collected their number of views per week. Then, they used the data in a generalized linear model (GLM) assuming a Poisson distribution and into a regularized LASSO model. Both methods got similar results compared to GFT. Although they estimated an overall less accurate ILI value, the peaks were better forecasted. However, it is important to note that the Wikipedia pages included were only accessed 41% of the time from the US. When predicting ILI rates based on United States CDC, this is a clear weakness.
Some other type of internet-based data were also tested, like Yahoo search queries by Polgreen et al. [5] or even, more recently, Instagram pictures by Gencoglu et al. [6].

4. Experiments with other Machine Learning methods

Next, we will review several other main types of approaches that were used to perform ILI rates prediction and compare them where appropriate.

4.1. AutoRegressive models

We have previously seen that Lampos et al. [2] mentioned autoregression, using a modified AutoRegressive Moving Average (ARMA) model. For our case, the AR component would be past ILI rates, while the moving average component would be some mean zero Gaussian noise. In the paper, their modification adds the internet based data (exogenous input, thus ARMAX), giving the equation :

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-1} + \sum_{i=1}^{D} w_i h_{t,i} + \varepsilon_t$$

where, for a week t, y_t is the ILI rate, ε_t is the Gaussian noise and h_t the output from an independent model (Elastic Net and GPs here). Thus, ϕ_i, θ_i and w_i are the coefficients we want to learn to perform accurate predictions. We saw that this autoregression addition clearly improves the independent models, showing that the near-term previous data matters when nowcasting a given week. Furthermore, Yang et al. [7] developed a generalized framework, ARGO (AutoRegression with GOogle search data), that directly takes log-transformed search frequencies with ILI activity level. This model, that also includes regularization, outperforms all the other models assessed, in correctness and robustness. Indeed, by adding regularization terms, the model auto-adjusts the different queries used in time, as opposed to the previous paper that used the prediction of another independent model as input. Finally, methods can also consider the seasonal information. Zhang et al [8] used a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model to predict ILI rates. Even if the use of cross-hemisphere data and the results don't allow us to formulate conclusions, we can see that taking the repeating cycles into account is encouraging and further studies are needed.

4.2. Mechanistic models

Kandula et al. [9] tried to forecast influenza with another type of method, model-inference models. Indeed, it uses the assumption that the population is divided into several states : susceptible-exposed-infectious-recovered-susceptible (SEIRS). By choosing initial parameters and modeling the relations between these states, they could predict the proportions of each one, for a whole season. To do that, they simulated using Kalman Filtering (EAKF), a data assimilation method that takes observations as a prior distribution and computes the posterior distribution, repetitively for each week. The results show that this type of approach is good at near-term forecasts but fails at predicting intensities of peak weeks. It is understandable, as by modeling the evolution of the states, it is easier to predict tomorrow (e.g. if there is no epidemic), but harder to get how big and fast-growing an outbreak is.

4.3. Bayesian Model Averaging

Xu et al. [10] introduce various new methods in their paper, where they perform forecasts in Hong Kong. Indeed, they use linear and autoregressive models (GLM, LASSO, ARIMA), and Deep Learning; but their strength is the use of all models at once with Bayesian Model Averaging (BMA). It uses the predictions of each individual model to compute their weights in the combined prediction of ILI rates for a certain week. The method allows a better comprehension of the dynamic relationships between models and outperforms all the individual methods in forecasting by correcting their errors, especially for peak weeks. It is also worth noting that this paper uses meteorological data, clearly closely related to influenza. Moreover, the previously seen research by Kandula et al. [9] also used this type of super-ensemble, on six dynamical models and two statistical ones, including for example the K-nearest neighbors (KNN) method. Their results confirm our learning, with better overall scores than individual models and better forecast of the peak week intensity. However, it wasn't the best at predicting the timing of peak weeks, and is not conclusively superior to individual approaches.

4.4. The use of Neural Networks

Here, we will see that the great breakout of neural networks in the field of machine learning also had an impact on time series forecasting, and on our influenza forecasting.
First of all, Ahmed et al. [11] used neural nets to perform forecasts on business-type time series. Indeed, the first one was the most simple form of neural networks, a multilayer perceptron (MLP), where nodes are organized in layers and nodes on each layer receive input signals from the nodes on the previous layer. By inputting our data on the first layer, we get an output value, and the objective is to find the perfect input weights on each node. They also used several other modified types of neural nets, like bayesian neural networks, radial basis function neural networks or generalized regression neural networks. The results from this paper, that also uses other methods like GPs and KNN, suggest that the simple MLP has the best performance. However, we can note that they only used a basic model

for each of the methods considered, used very few data points to train them (63 to 108) and did not do an in-depth comparison of these (only one measure to assess the models performance). On the other side, it is worth to mention that they tested a few types of preprocessing on the data, backing up our point that every model's performance relies heavily on how and what data is used.
More focused on our task, Volkova et al. [12] developed a way to nowcast and forecast (up to 4 weeks) ILI rates for military populations with a new type of neural network, based on Long Short Term Memory (LSTM) units. Indeed, they used known ILI rates and Twitter data to imagine a two-branch neural network, one for each data source. These two branches are merged before the output layer, allowing the choice of using only one of the data sources. The other particularity, the LSTMs, are nodes that basically have a memory, using it to modify their state. The results show that LSTMs not only outperformed the other models used (Support Vector Machines and AdaBoost), but that they also capture very well the influenza peaks (timing and magnitude) when nowcasting. Unfortunately, we can't know from this paper how LSTMs really evaluate outbreaks when forecasting. Moreover, this study was conducted on military populations across different locations, and further studies are needed to see how it performs with a country's population.

5. Research - An optimal feature selection

We have seen that the models used to forecast influenza rates around the world heavily rely on what data is used and, more particularly, how data is used. Although it really depends on the model used, some research of the field focused on this aspect, in order to perform predictions in an optimal way. In their paper, Lampos et al. [13] developed a way to select search queries using neural embeddings. Indeed, they first created the embedding space using tweets, inserting the search queries afterwards, in order to fill previously created concepts with similar search queries. Subsets of these concepts were then evaluated using Elastic Net or GPs, to get the optimal textual features. Moreover, this technique was then combined with the more classical way of checking the correlation with the true ILI rates, resulting in an hybrid method outperforming other methods, particularly when GP regression is used. This study clearly shows that feature selection is crucial and opens the path for further research, in order to forecast ILI rates with optimal input data.

6. Research - A spatial approach

In this section, we will see that some studies attempted to use spatial information to better predict influenza rates, as epidemics can quickly move in time and space. Indeed, Davidson et al. [14] tried to use networks with GFT data. First, they used CDC confirmed ILI rates to measure the weights between each node (region), by performing cross-correlations, and used each region's GFT measure as a sum over each incoming weight. By using this type of preprocessed data, the model performs better in epidemic phases, as relations between regions are now accounted. However, the use of a basic linear model and GFT data are downsides and results aren't as good on a whole year.
To continue with this type of research, Lu et al. [15] used ILI rates, Google search queries and electronic health records to combine network approaches and an autoregressive model (ARGO). Their network uses near-past ILI data from each other state to predict the rates of a given week. Although they unfortunately didn't use weightings to model relation between states, we can see that, by combining this type of data with the online data (ARGO model), the results are encouraging. Indeed, the combined ARGO-Net model improves prediction accuracies, even if further studies are needed to show the real value of the networks approach.

7. Conclusion

We reviewed the basis of influenza nowcasting and forecasting, along with the usage of different types of internet-based data. We showed that this type of data is nowadays crucial for accurate results, and

evaluated the strengths and weaknesses of multiple machine learning methods.
Let's first note that, in this task, data needs to be considered very carefully, as we don't always know how reliable public data it is. Indeed, CDC's ILI rates could sometimes be biased if practitioners fail to provide accurate numbers; or Google frequency rates falsified by intense media coverage, as it happened before (e.g. 2009 epidemic). Moreover using time scattered data is always a difficult task, like for example the fact that the usage of internet services really changed throughout the years. Here, we tried to give appropriate comparisons between methods, but for these reasons, and because data collecting and preprocessing was never the same, we couldn't concretely contrast models based on their results.
Concerning these methods, we first saw that linear models were at one point the basis of influenza forecasting, but that they are too limited for accurate results. In addition, mechanistic approaches are good for near-term predictions but fail to capture epidemics, while multiple models averaging can combine the strengths of several methods but still require further research. Moreover, autoregression performs very well, but sub-models like seasonal autoregression (SARIMA) always need more testing. We also saw that the more encouraging results are the ones from neural nets, that can capture relations that other models can't, and are likely to be continuously tested on this task. Finally, the network approaches we reviewed really seem to be the future of this research field, as they could model relations between countries/regions and have localized predictions, while accurately tracking the evolution of an epidemic.
We can say that overall, more research is needed, especially for data selection and preprocessing methods, that are key in this task. Additionally, neural networks methods seem to have the best results when forecasting influenza epidemics, but the small set of papers that used this approach shows us that studies are still going on. The fact that no public entities are yet largely using these techniques to manage their actions and resources really shows us that machine learning models still can't predict influenza peaks accurately enough.
The future work on this research field will most likely include autoregressive models, with all their different derivations, and neural nets. Now that influenza forecasting is accurate enough in "normal" times and to capture outbreak timings, the main goal is to precisely compute epidemic intensities, in any region of the world.

References

[1] *J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2008) 1012–4. doi:10.1038/nature07634.*

[2] *V. Lampos, A. Miller, S. Crossan, C. Stefansen, Advances in nowcasting influenza-like illness rates using search query logs, Scientific Reports 5. doi:10.1038/srep12760.*

[3] *M. Paul, M. Dredze, D. Broniatowski, Twitter improves influenza forecasting, PLoS currents 6. doi:10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.*

[4] *D. Mciver, J. Brownstein, Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time, PLoS computational biology 10 (2014) e1003581. doi:10.1371/journal.pcbi.1003581.*

[5] *P. Polgreen, Y. Chen, D. Pennock, F. Nelson, Using internet searches for influenza surveillance, Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 47 (2008) 1443–8. doi:10.1086/593098.*

[6] *O. Gencoglu, M. Ermes, Predicting the flu from instagram.*

[7] *S. Yang, M. Santillana, S. C. Kou, Accurate estimation of influenza epidemics using google search data via argo, Proceedings of the National Academy of Sciences 112 (47) (2015) 14473–14478. arXiv:https://www.pnas.org/content/112/47/14473.full.pdf, doi:10.1073/pnas.1515373112.*

[8] *Y. Zhang, L. Yakob, M. B. Bonsall, W. Hu, Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data, Scientific Reports 9 (2019) 2045–2322. doi:10.1038/s41598-019-39871-2.*

[9] *S. Kandula, T. Yamana, S. Pei, W. Yang, H. Morita, J. Shaman, Evaluation of mechanistic and statistical methods in forecasting influenza-like illness, Journal of The Royal Society Interface 15 (2018) 20180174. doi:10.1098/rsif.2018.0174.*

[10] *Q. Xu, Y. Gel, L. Ramirez, K. Nezafati, Q. Zhang, K.-L. Tsui, Forecasting influenza in hong kong with google search queries and statistical model fusion, PLoS ONE 12. doi:10.1371/journal.pone.0176690.*

[11] *N. Ahmed, A. Atiya, N. Gayar, H. El-Shishiny, An empirical comparison of machine learning models for time series forecasting, Econometric Reviews 29 (2010) 594–621. doi:10.1080/07474938.2010.481556.*

[12] *S. Volkova, E. Ayton, K. Porterfield, C. Corley, Forecasting influenza-like illness dynamics for military populations using neural networks and social media, PLOS ONE 12 (2017) e0188941. doi:10.1371/journal.pone.0188941.*

[13] *V. Lampos, B. Zou, I. Cox, Enhancing feature selection using word embeddings: The case of flu surveillance, 2017. doi:10.1145/3038912.3052622.*

[14] *M. Davidson, D. Haim, J. Radin, Using networks to combine big data and traditional surveillance to improve influenza predictions, Scientific reports 5 (2015) 8154. doi:10.1038/srep08154.*

[15] *F. Lu, M. Hattab, C. Clemente, M. Biggerstaff, M. Santillana, Improved state-level influenza nowcasting in the united states leveraging internet-based data and network approaches, Nature Communications 10. doi:10.1038/s41467-018-08082-0.*