A Deep Learning framework to model Influenza/Flu predictions using aggregated Google Search query data
Final Report
Team : Kingpins

Sagar Satyanarayana
Mohammad Turab Ali
Garima Silewar
Nikhil Kumar Mutyala

December 5 2019

Contents

List of Tables

List of Figures

1 Introduction

Influenza outbreaks cause up to 500,000 deaths a year worldwide, and an estimated 3,000–50,000 deaths a year in the United States [1]. The ability to effectively prepare for and respond to outbreaks heavily relies on the availability of accurate real-time estimates and the existing methods remains limited [2, 3]. Traditional flu surveillance systems, such as Center for Disease Control and Prevention's (CDC) influenza reports lag behind real-time by one to two weeks, whereas information contained in internet users' search activity is available in near real-time [4].

This project proposes and implements a framework for predicting Influenza Like Illness (ILI) activity using trends in Google Search activity.

1.1 Problem Description

Building a real time framework using online activity data (like Google Search Trends) using Deep Learning model in-place of traditional Auto-Regressive (AR) forecasting methods in the United States.

2 Literature review

2.1 Need for internet user's search activity trends in epidemiological predictions

Traditional methods of data collection in epidemiological studies need heavy resources in terms of logistics, time, as well as human and material resources, so leading the way to searching alternative strategies for collecting data [5]. Since internet has increasingly become a meaningful health resource for both laypeople and health professionals, internet-derived information has been recognized as a surrogate tool for estimating epidemiology and gathering data about patterns of disease and population behavior [6]. Internet query platforms, which allows to interact with internet-based data, have been considered a source of potentially useful and accessible resources, especially aimed to identify outbreaks and implement intervention strategies [7]. The US Institute of Medicine (IOM) has also recently acknowledged that the use of internet data in health care research holds promise, and may also "complement and extend the data foundations that presently exist"[8]. Numerous studies have also suggested great potential of these big data sets to detect/manage epidemic outbreaks [9, 10, 11, 12].

2.2 Internet query platforms

Big data sets are constantly generated nowadays as the activities of millions of users are collected from Internet-based services. In recent years, methods that harness Internet-based information have also been proposed, such as Google, Yahoo, and Baidu Internet searches, Twitter posts, Wikipedia article views, clinicians' queries, and crowdsourced self-reporting mobile apps such as Influenzanet (Europe), Flutracking (Australia), and Flu Near You (United States). Among them, GFT `https://ai.googleblog.com/2014/10/google-flu-trends-gets-brand-new-engine.html` has received the most attention and has inspired subsequent digital disease detection systems.

2.3 Accurate estimation of influenza epidemics (ARGO)

Yang et al., in their paper 'Accurate estimation of influenza epidemics using Google search data via ARGO' [3] have developed an Autoregression model with Google search data (ARGO). ARGO outperforms all previously available Google-search based models including the 2014 Google Flu Trends. ARGO captures the changes in people's online search behaviour over time as well as incorporates seasonality in influenza epidemics. ARGO is self-correcting, scalable, flexible and robust making it a potentially powerful tool which can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

2.4 A Deep Learning Framework in Epidemiological studies

Yuexin et al., in their paper 'Deep Learning for Epidemiology Predictions' [13] derived a deep learning framework to predict epidemiology profiles in the time series perspective. The deep learning framework consisted

of two types of neural networks: Recurrent Neural Networks (RNNs) to capture long term correlation in the data and Convolutional Neural Network (CNNs) to fuse information from different data sources. This model was tested against standard Autoregressive (AR) methods and Gaussian Process Regression (GPR) and performed consistently better than the latter two. However, these models were trained and tested using historical data and not real-time search trends.

2.5 Difference/Novelty

Traditionally epidemiological predictions are modeled from time-series perspective using Autoregressive (AR) methods and its variants like Gaussian Process Regression (GPR) methods. These methods make use of historical data to make usually a linear (or a pre-defined non-linear kernel) predictions to capture spatio-temporal patterns. This method is popular due to the reason that epidemiological predictions are usually weekly sampled statistics which provides limited training instances. But, as the data availability grows in size and diversifies its sources, obtaining training instances is not an issue. This project is novel in a way that we propose use of Neural Network to model multi-variate time-series data using real time google search queries instead of traditional AR methods.

Yuexin et al., in their paper 'Deep Learning for Epidemiology Predictions' [13] have used a **univariate** Deep learning framework to model influenza trends just from historical trends data. We would model a **multivariate** Deep learning framework with Google Search Query's as independent variables with instances being their frequencies at time t. This allows us to predict ILI activity using correlated search queries (similar to supervised linear regression) rather than forecasting ILI itself depending on its previous values (time-series forecasting).

3 Methods

3.1 Data Collection

Weekly Influenza Like Activity data is obtained from CDC `https://www.cdc.gov/flu/weekly/`. This data contains weighted and unweighted cases of Influenza Like Infections (ILI) from 2010 to 2019 for every state in the U.S. CDC complies actual number of cases of Influenza from clinics and labs and release a weekly report which usually runs two weeks behind real-time.

The weighted ILI activity data is used to find the correlated google search queries in the U.S from Google Correlate. Since, Google Correlate has stopped providing support, we have used the search queries derived by Yang et al., [3] for 2004 to 2014 ILI activity. This data will be used to train our model.

3.2 Model

In this project, we are inspired by S. Yang et al.[11] and World Health Organization. Based on their outstanding work, we will provide our new idea of implementing a deep learning framework for predicting ILI activity. In order to train our model, first we obtain the data from Center for Disease Control and Prevention's (CDC) ILI activity reports. Now using this time stamp, we search for the google correlate queries on influenza like illness(ILI) regarding influenza/ flu. We validate the timestamps of both CDC and google correlate. This gives X_d, where $d = \{1, 2, 3, ...$ number of correlated google queries$\}$ is the number of correlated search queries.

Using ILI data from 2004 to 2014, we obtained the correlated queries shown in the table below 1.

influenza type a	get over the flu	type a influenza	flu care
symptoms of flu	treating flu	i have the flu	how long contagious
flu duration	flu vs. cold	taking temperature	fight the flu
flu contagious	having the flu	flu versus cold	reduce a fever
flu fever	treatment for flu	bronchitis	cure the flu
treat the flu	human temperature	how long flu	medicine for flu
how to treat the flu	dangerous fever	flu germs	flu length
signs of the flu	the flu	cold vs. flu	cure flu
over the counter flu	remedies for flu	flu and cold	exposed to flu
how long is the flu	influenza a and b	thermoscan	low body
symptoms of the flu	contagious flu	flu complications	early flu symptoms
flu recovery	how long does the flu last	high fever	remedies for the flu
cold or flu	fever flu	flu children	flu report
flu medicine	oscillococcinum	the flu virus	incubation period for flu
flu or cold	flu remedies	how to treat flu	break a fever
normal body	how long is flu contagious	pneumonia	flu contagious period
is flu contagious	flu treatments	flu headache	influenza incubation period
treat flu	influenza symptoms	flu cough	cold versus flu
body temperature	cold vs flu	ear thermometer	flu in children
is the flu contagious	braun thermoscan	how to get rid of the flu	what to do if you have the flu
reduce fever	fever cough	flu how long	medicine for the flu
flu treatment	signs of flu	symptoms of bronchitis	flu and fever
flu vs cold	how long does flu last	cold and flu	flu lasts
how long is the flu contagious	normal body temperature	over the counter flu medicine	incubation period for the flu
fever reducer	get rid of the flu	treating the flu	do i have the flu

Table 1: ILI activity correlated Google Search Queries for 2004 -2014

Now the obtained data is trained using the data from 2004 to 2014 on ILI activity. The correlated data and cdc data are merged into one dataframe based on the columns. The required features are extracted from the merged file. The data is split into training and validation sets. We imported Keras for building the model. We are using long short term memory(LSTM). This LSTM has loss as 'mean square error(mse)', optimizer as 'rmsprop' and metrics is set to 'mae'. We got obtained parameters as 241,001. The parameters are tuned to get the best model. Now we fit the network using 'model.fit' with epochs set to 500 and batch size of 52.Now we build the history of successive mean K-fold validation scores for no of epochs and the average MAE history is obtained. Now we plot the validation scores for MAE by epoch. Now the loss is plotted for training and validation. We observed that at 350 epochs, the model performs best. Now the data is tested on the log transformed values of CDC data. The data is tested on 'year', 'week' and ' WEIGHTED ILI'. For this model, the tested mae score is 1.10, test mse score is 2.28. A plot of Predicted Values v/s Actual Observations is plotted to obtain the accuracy of the model.

3.3 Model Architecture

The weighted influenza cases will be our target values (labels) y_t. The vector of log transformed of Google Search Queries at time t will be the independent variables X_t. This is under the assumption that X_t depends only on the ILI activity at the same time, this follows the intuition that flu occurrence causes people to search flu-related information online. With these, we can model our time-series data similar to a supervised linear regression model. We prefer using a Deep Learning model here instead of traditional AR model because from our literature review we have reasons to believe Deep Learning models can out perform AR models when dealing with multiple inputs (multivariate models).

Our model is selected with an objective to handle two important objectives;

- It should be able to handle and learn from multiple input variables as well as learn temporal dependencies in the time-series: We propose a LSTM layer and a Dense layer to achieve this as they can be seamlessly modeled to multivariate problems [14].

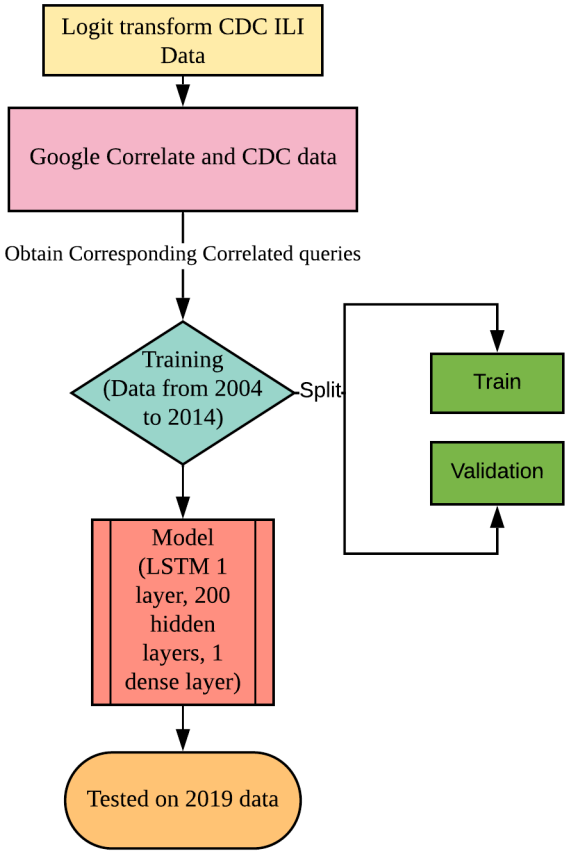

Figure 1: Architecture of the model

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 200)               240800
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 201
=================================================================
Total params: 241,001
Trainable params: 241,001
Non-trainable params: 0
_________________________________________________________________
None
```

Figure 2: Summary of the model

3.4 Model Parameters

Data from 2004-2014 is divided into training and validation sets, and the model is fit with initial conditions of 50 hidden layers in LSTM and the number of hidden layers is increased to see it it leads to a decrease in validation loss. We see the model performance improve after 100. We chose 200 as our number of hidden layers, this value is not increased further because we have limited data instances of about 573 from 2004-2014 and increasing the hidden layer may cause over-fitting. Figure 3 shows the Mean Absolute Validation error and figure 4 shows the variation of the training and validation losses across epochs. We seem an improvement in validation loss (loss function = MSE) at around 350 epochs and choose this to be our epochs value. The batch size of 52 is chosen. This equals to the number of weeks in a year. The internal state of the LSTM in Keras is reset at the end of each batch [14], Since our data is weekly time-series we hope to capture any seasonal trends throughout the year.

The model is compiled using MSE loss function, rmsprop optimizer and MAE as the metric.

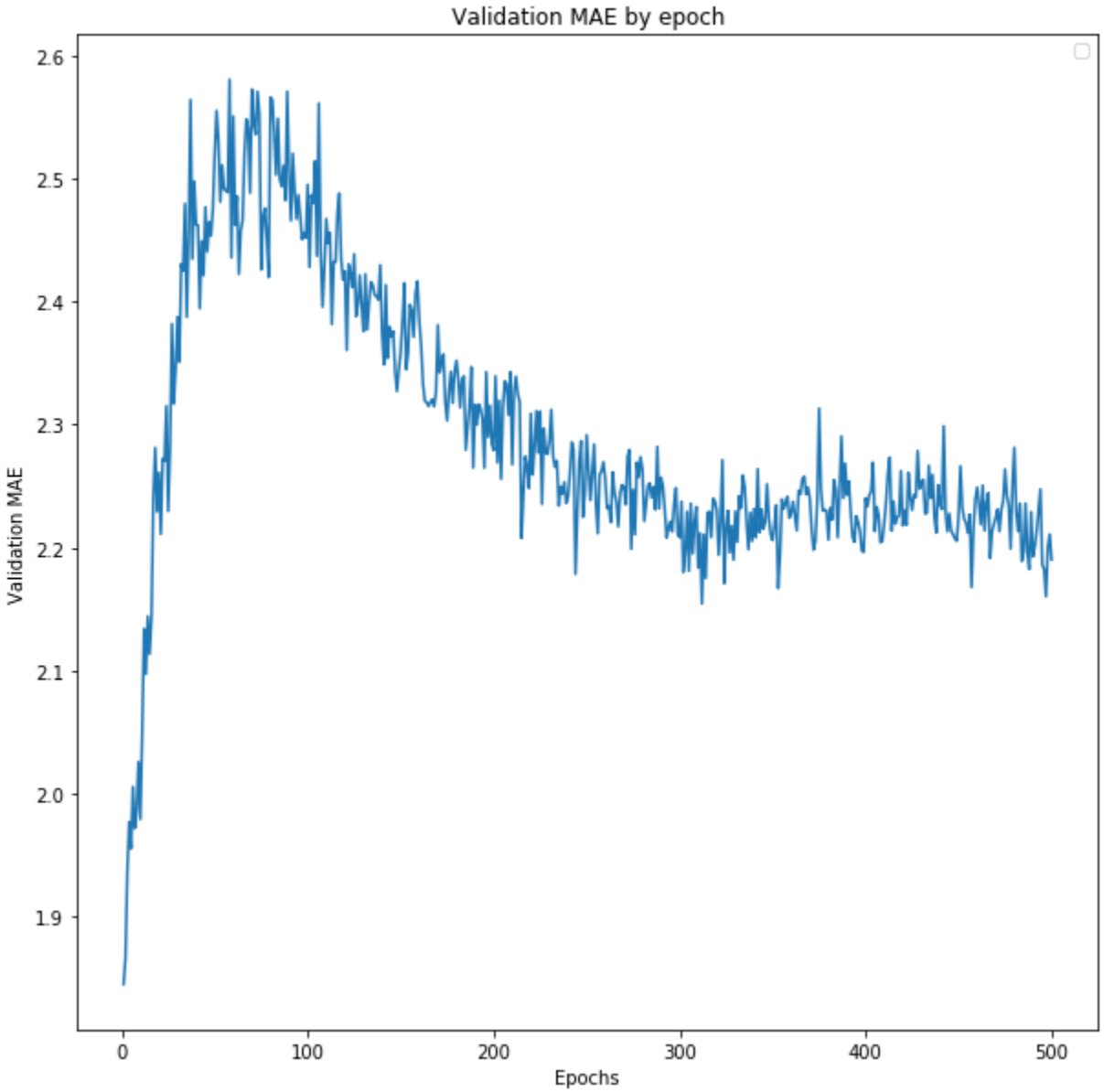

Figure 3: Mean Absolute Validation Error

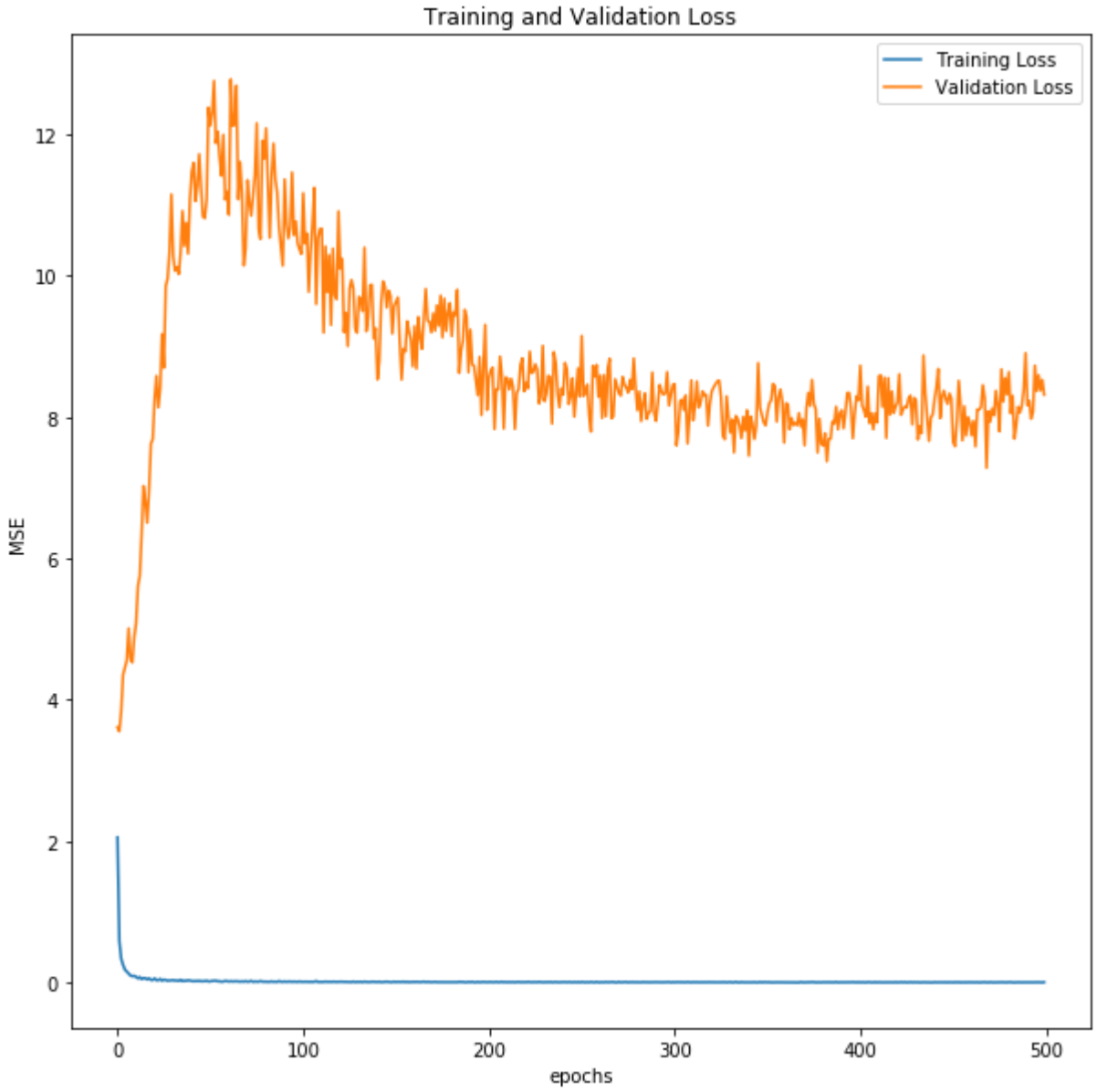

Figure 4: Training and Validation MSE

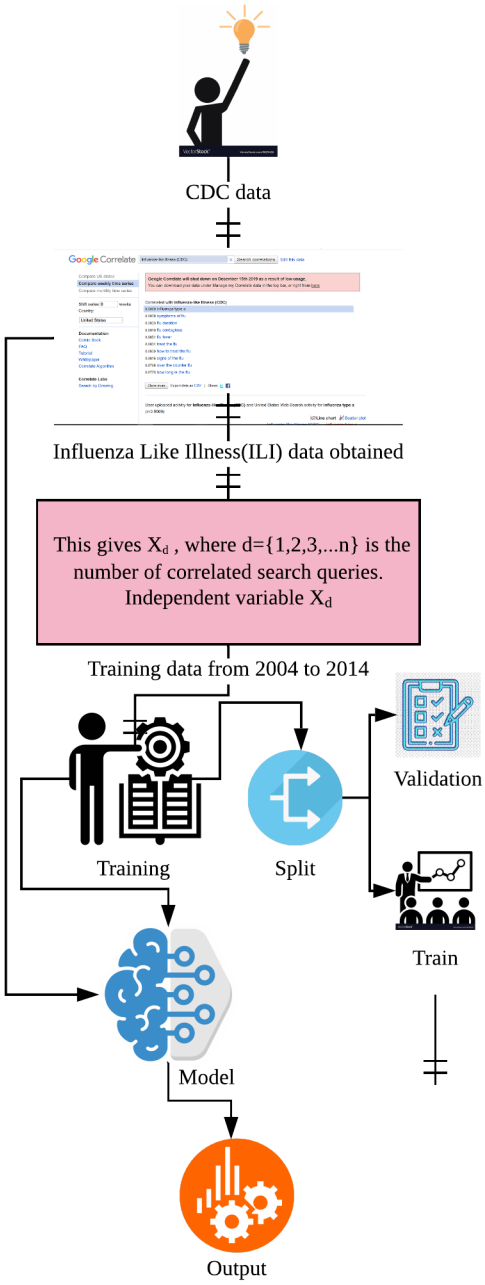

Figure 5: Project Pipeline

4 Results and Conclusion

4.1 Evaluating the model for 2019 ILI Predictions

We test the model on 2019 ILI activity data from CDC to see how well the model performs today (since the correlation was derived using 2004-2014 data).

The model performed with a Test MAE of 1.1 and Test MSE of 2.3.

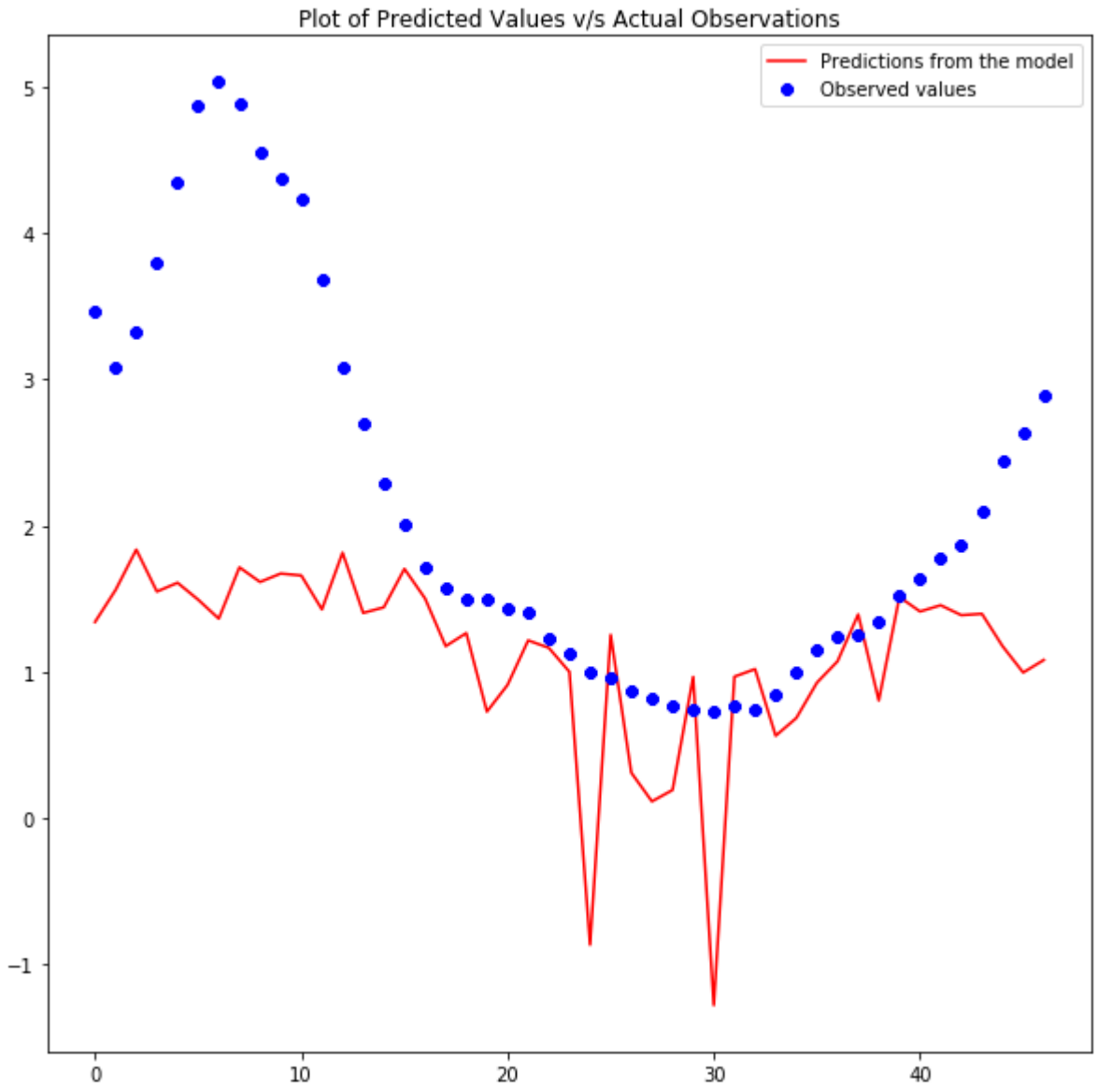

Figure 6: Plot of Predicted Values v/s Actual Observations 2019 for 2019 ILI activity

4.2 Conclusion

By being able to predict the influenza outbreak much before the CDC comes out with actual results , we could actually foresee the influenza outbreak and people can be made aware of this in real time which could save many lives.The test MAE and MSE values that we got are very promising given the fact that we did not have data from 2015 to 2019 ,revealing that our model could actually perform pretty good.

5 Future Work

- Evaluate different model parameters and architecture of the model to improve predictions.
- Run the model for state-wise and country-wise influenza predictions and map the trends.
- Evaluate the model performance for different countries.
- Building the web-scraping program interface to automatically pull weekly search queries.

- Develop a pipeline for real-time FlLu-tracking.

6 Response to Feedback

We have worked according to the feedback received from Dr. Lee for the Proposal in the Mid-Progress Report, and we got positive feedback for the Mid-Progress Report. Please see 'Response to the feedback' section for the detailed response.

7 Work-log

Task	Sagar Satyanarayana	Garima Silewar	Nikhil Kumar Mutyala	Mohammad Turab Ali
Literature Review	✓			
Data Collection	✓	✓		
Project Proposal	✓	✓	✓	✓
Data Preprocessing	✓	✓		
Model Selection and Parameter Tuning	✓			
Model Testing	✓	✓		
Feature Selection and Extraction from Google Correlate		✓		
Mid-Progress Report	✓	✓	✓	✓
Model Training	✓			
Model Evaluation and Parameter Tuning	✓			
Final Report	✓	✓	✓	
Poster	✓	✓	✓	

Figure 7: Work logs

8 Response to the feedback

- How is our model different from ARGO?
  We are proposing the use of Neural Networks to model multi-variate time-series data using real time google search queries instead of traditional AR methods which is used in ARGO.
- Which deep learning model are we going to use?
  Our proposed model will be using the following Deep Learning Models: LSTM (recurrent networks) Gated recurrent unit(recurrent networks) CNN.
- Specific Plan for our deep learning model?
  For our Deep learning Model we are planning to follow the below plan:
  - It will handle and learn from multiple input variables by using LSTM model.
  - It will learn temporal dependencies in the time-series: by using a Gated Recurrent Unit.
  - We will also try to add another layer of Convoluted Neural Networks to capture local dependencies.

References

[1] World Health Organization, "Influenza (seasonal) (world health org, geneva), fact sheet 211." https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal), October 1, 2019.

[2] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. 20425–20430, 2012.

[3] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

[4] S. Yang, M. Santillana, J. S. Brownstein, J. Gray, S. Richardson, and S. C. Kou, "Using electronic health records and internet search information for accurate influenza forecasting,"

[5] A. Ekman and J.-E. Litton, "New times, new needs; e-epidemiology," *European Journal of Epidemiology*, vol. 22, pp. 285–292, May 2007.

[6] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection — harnessing the web for public health surveillance," *New England Journal of Medicine*, vol. 360, no. 21, pp. 2153–2157, 2009. PMID: 19423867.

[7] M. Salathé, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani, "Digital epidemiology," *PLOS Computational Biology*, vol. 8, pp. 1–3, 07 2012.

[8] G. Cervellin, I. Comelli, and G. Lippi, "Is google trends a reliable tool for digital epidemiology? insights from different clinical settings," *Journal of Epidemiology and Global Health*, vol. 7, no. 3, pp. 185 – 189, 2017.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data,"

[10] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using Internet Searches for Influenza Surveillance," *Clinical Infectious Diseases*, vol. 47, pp. 1443–1448, 12 2008.

[11] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in china with search query from baidu," *PLOS ONE*, vol. 8, pp. 1–7, 05 2013.

[12] M. J. Paul, M. Dredze, and D. a. Broniatowski, "Twitter improves influenza forecasting.," *PubMed*, vol. 6, 2014.

[13] Y. Wu, Y. Yang, H. Nishiura, and M. Saitoh, "Deep learning for epidemiological predictions," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, (New York, NY, USA), pp. 1085–1088, ACM, 2018.

[14] J. Brownlee, *Deep Learning for Time Series Forecasting*.