Predicting New COVID-19 Cases with Machine Learning Algorithms and finding mitigation strategies

Oded Salton & Ben Nadler | Associate Professor Roi Reichart | Technion – Israel Institute of Technology

Introduction

Since December 2019 we are facing a global pandemic. The rate of infection and death caused by the virus worried government and citizens globally. We know today that those countries who took this threat seriously, evaluated the situation, and responded quickly were the ones succeeded the most in slowing down the infection rate of the virus and death rate of the population.

In this paper we will attempt to predict the daily status of new cases in terms of new cases per million people as a classification problem, using structured and unstructured machine and deep learning models and algorithms.

Data

We chose to construct our dataset based on multiple data tables we gathered from ourworldindata.org/coronavirus. We took data from March to July as the training data and from then on we treated the data of August as test data.
We chose 11 countries which we thought can represent the virus spread behaviour and actions taken against it globally and generalize our model.

Dataset

Time based COVID-19 data from March to August – 1694 samples

Features

- Daily and cumulative deaths, deaths per million, Population in Millions
- Parks, Retail, Transit, Workplace, Groceries and Pharmacies visitor change %
- International travels, workplace, schools, transportation closure policies
- Public Gathering, Campaigns, Testing and Lockdown policies

Labels

0 = No new cases today, **1** = 1-10 new cases per million, **2** = 11-100 new cases per million, **3** = 101-250 new cases per million, **4** = 250+ new cases per million

Objective and Metrics

Learning Objective

Predicting new COVID-19 cases per million people each day

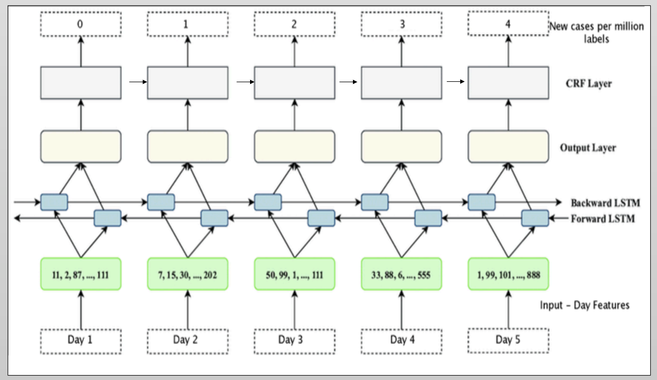
Evaluation Metrics

- Accuracy - Percentage of correct labelling
- F1 - Weighted average of the precision and recall metrics, which contribute equally
- Precision - Ratio between true classified days with label i and number of days which were classified as I
- Recall - Ratio between the correct classified days with label i and number of days which are true labeled as i

Models and Algorithms

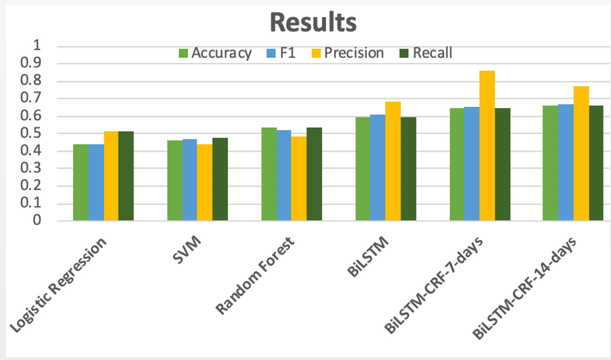
Unstructured Approach	
Model	Algorithm
All Countries	SVM
All Countries	Logistic Regression
All Countries	Random Forest

Stuctured Approach	
Model	Algorithm
All Countries	BiLSTM
Advanced	
Linear Chain CRF 7 Days	BiLSTM-CRF + Viterbi
Linear Chain CRF 14 Days	BiLSTM-CRF + Viterbi



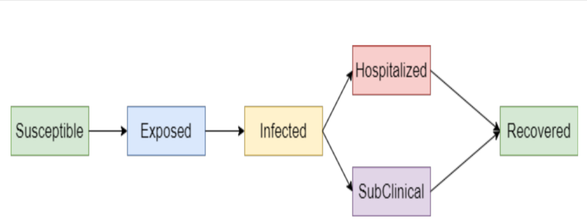
Experiments and Results

Classifier	Accuracy	F1	Precision	Recall
Logistic Regression	0.443	0.441	0.512	0.513
SVM	0.463	0.468	0.439	0.474
Random Forest	0.538	0.517	0.486	0.536
BiLSTM	0.598	0.612	0.682	0.598
BiLSTM-CRF-7-days	0.643	0.651	0.859	0.643
BiLSTM-CRF-14-Days	0.664	0.669	0.769	0.664



Creative Part

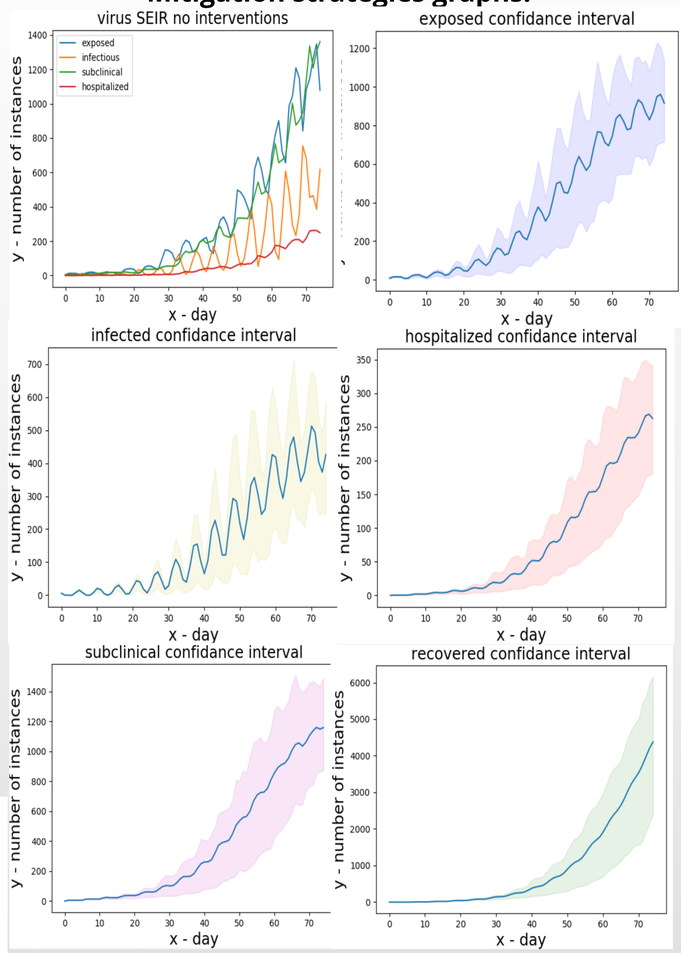
We Used the S,E,I,H,Sc,R model on generated community:



We simulate 88 different combinations of infected individuals in t=0 and ran the simulation for 75 days on our community graph with 4 different mitigation strategies.
Quarantine Elders : remove all connections from the elders
Quarantine Connected: remove all connections from the most connected individuals
Quarantine lowest CC: remove all connections from nodes with lowest Clustering coefficient.
Remove lowest NO relations: remove the relations with lowest neighborhood overlap.

Results – Creative Part

Mitigation strategies graphs:



Conclusions

- Difficult learning objective, with lack of data due to the fact that the pandemic began only 10 months ago. More reliable data is needed.
- LSTM can greatly improve performance for sequence labelling tasks.
- 14 days provided the best results on most metrics followed by 7 days, with a large margin from the unstructured models.
- We can reduce up to 90% of the number of cases by removing 11% of the graph relations with a winning strategy of quarantine for most connected individuals.