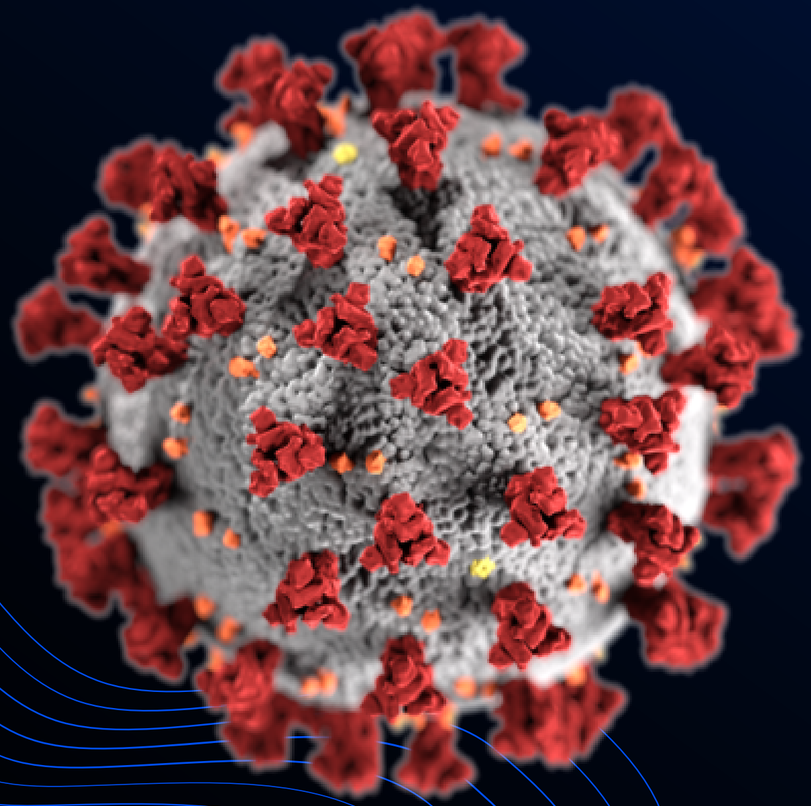
COVID-19 GLOBAL OUTLOOK: MODELLING, PREDICTION AND SENTIMENT ANALYSIS

MARCO DIBUONO
BARBARA TARANTINO

BRIEF INTRODUCTION



WHAT IS COVID-19?

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome. The first disease was identified in Dicember 2019 in Wuhan.
On March 11 2020, after 118,000 people being infected in 114 Countries, and causing the death of 4,291 people, COVID-19 has been recognized as a pandemic.
Today, April 13,2020, the pandemic infected 1,854,464 people in 185 countries, causing the death of 114,331 people.

PURPOSE OF OUR RESEARCH

Starting from a global exploratory analysis, then we focus on virus' modelling and prediction for the countries with the largest number of confirmed cases.
For modelling, we implemented SIR Model, with some extensions and, for prediction, Logistic and Gompertz model.
At the end, we choose the best model based on R2 score, we check the forecasts by country and show some insights from Sentiment Analysis.

DATASET OVERVIEW

COVID-19 GLOBAL FORECASTING

In the context of the global COVID-19 pandemic, Kaggle has launched several challenges in order to provide useful insights about the virus. This is the case of the COVID19 Global Forecasting, in which participants are encouraged to fit worldwide data in order to predict the pandemic evolution and determine which factors impact the transmission behavior of COVID-19.

CBC NEWS CORONAVIRUS/COVID-19 ARTICLES (NLP)

Has the news media been overreacting or under-reacting during the development of COVID-19? What are the media's main focuses? How is the news correlated to public reactions or policy changes? We might find many insights with more than 3,500 CBC news articles' dataset that spans from 2020-01-08 to 2020-03-27.

OTHER DATA SOURCES

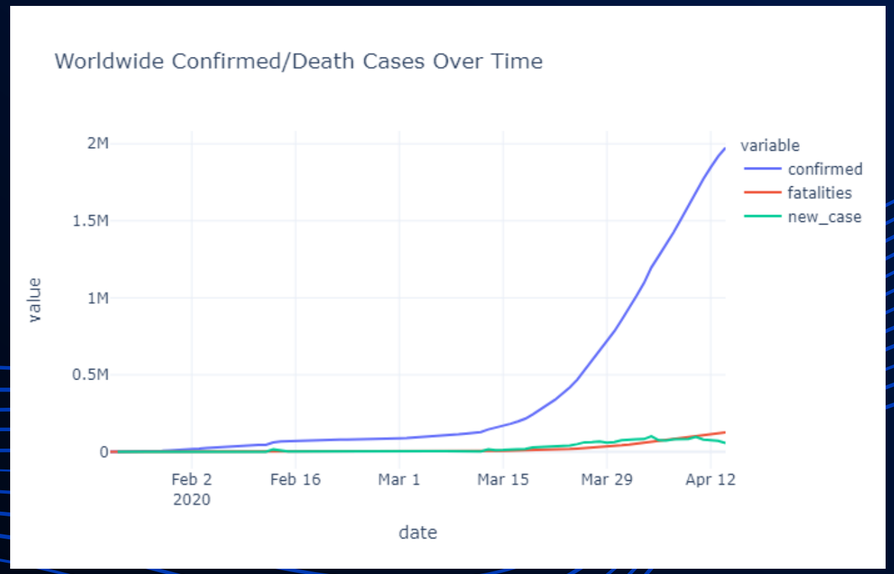
US State Code and Population by Country-2020 have been used in order to provide Country-wise information.

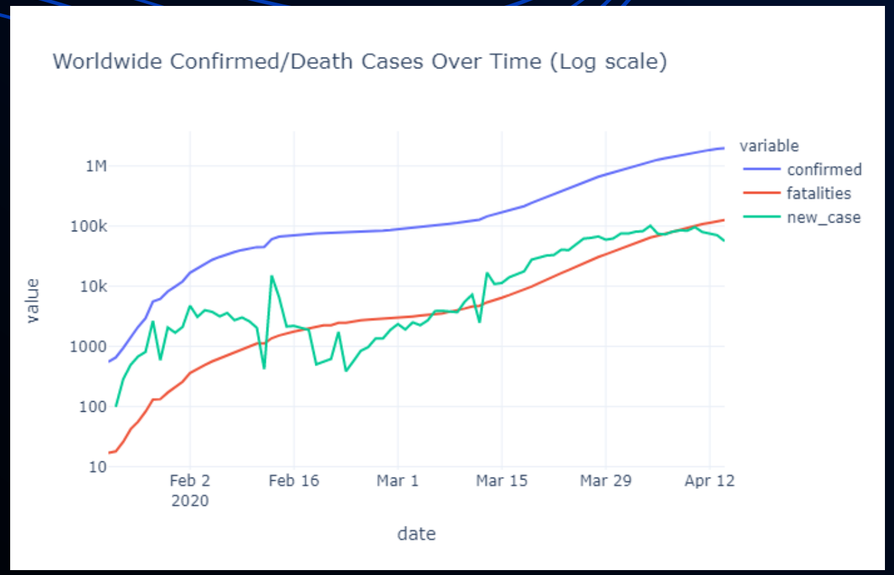
*All the datasets have been provided by Kaggle.

EXPLORATORY DATA ANALYSIS

WORLDWIDE TREND

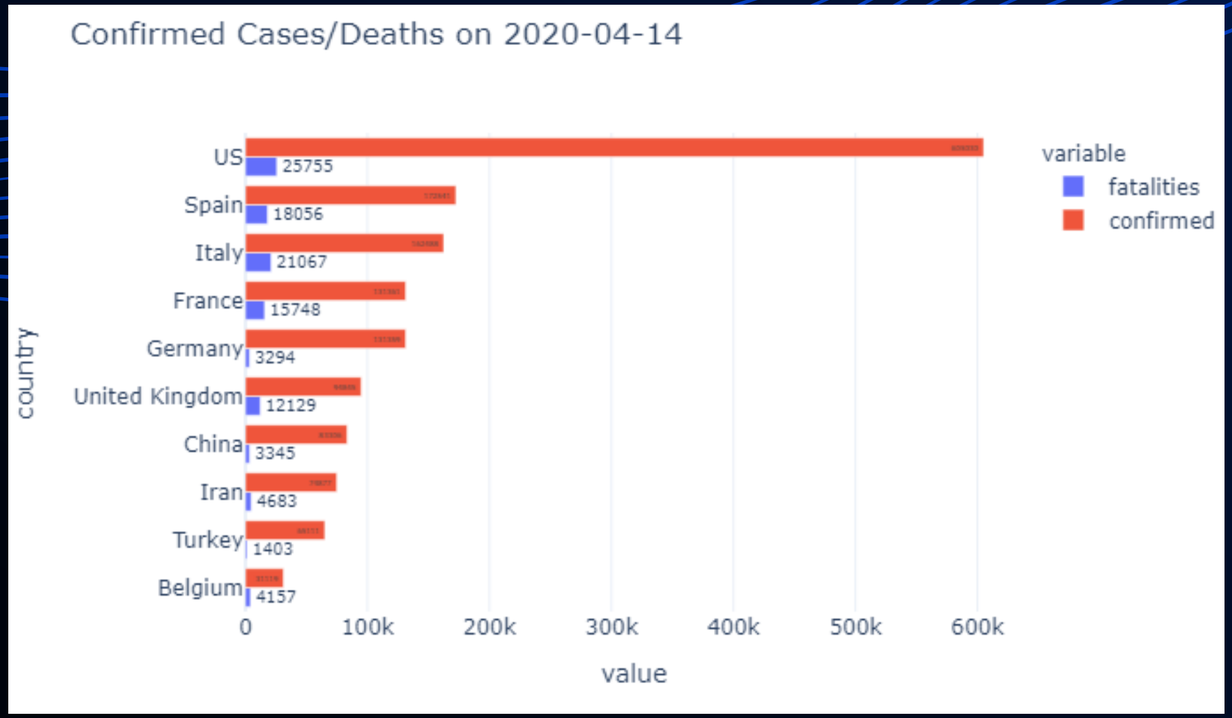
The confirmed cases' curve is exponentially growing till April 11th without signal of deceleration. This is probably due to the influence of US, since New York state has become the epicentre of the outbreak in the US, recording more than 180,000 of the country's nearly 530,000 cases.
It can be noticed that, despite the Lockdown policy in Europe or US, the speed of confirmed cases growth rate slightly increases when compared with the beginning and end of March



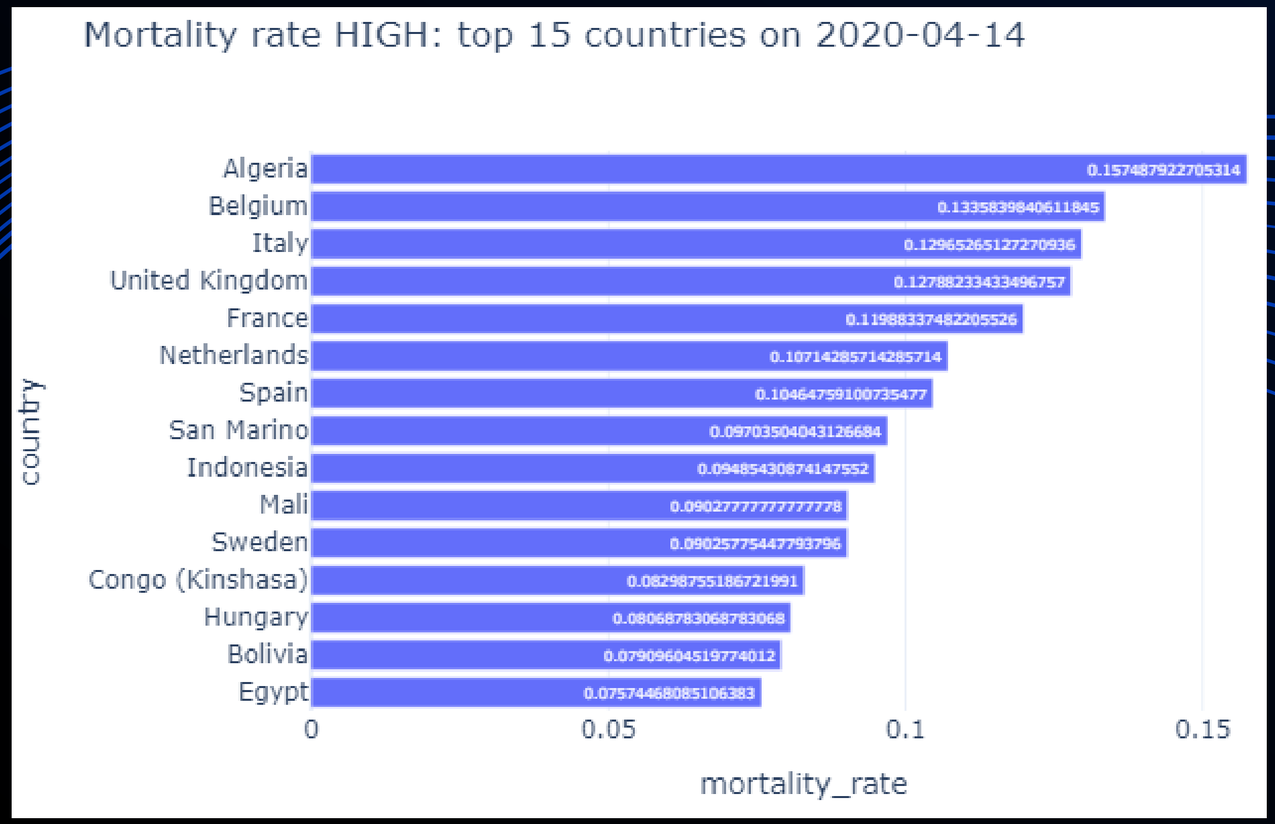


COUNTRY-WISE GROWTH

For a detailed view, cases by country are shown.
COVID-19 has spread to 185 countries around the world, with the most affected ones being the United States, Spain, Italy, France, Germany, United Kingdom, China, Iran, Turkey and Belgium. In terms of confirmed cases, US and many Europe countries are in the top, overtaking the number of confirmed cases in China.



COUNTRY-WISE GROWTH



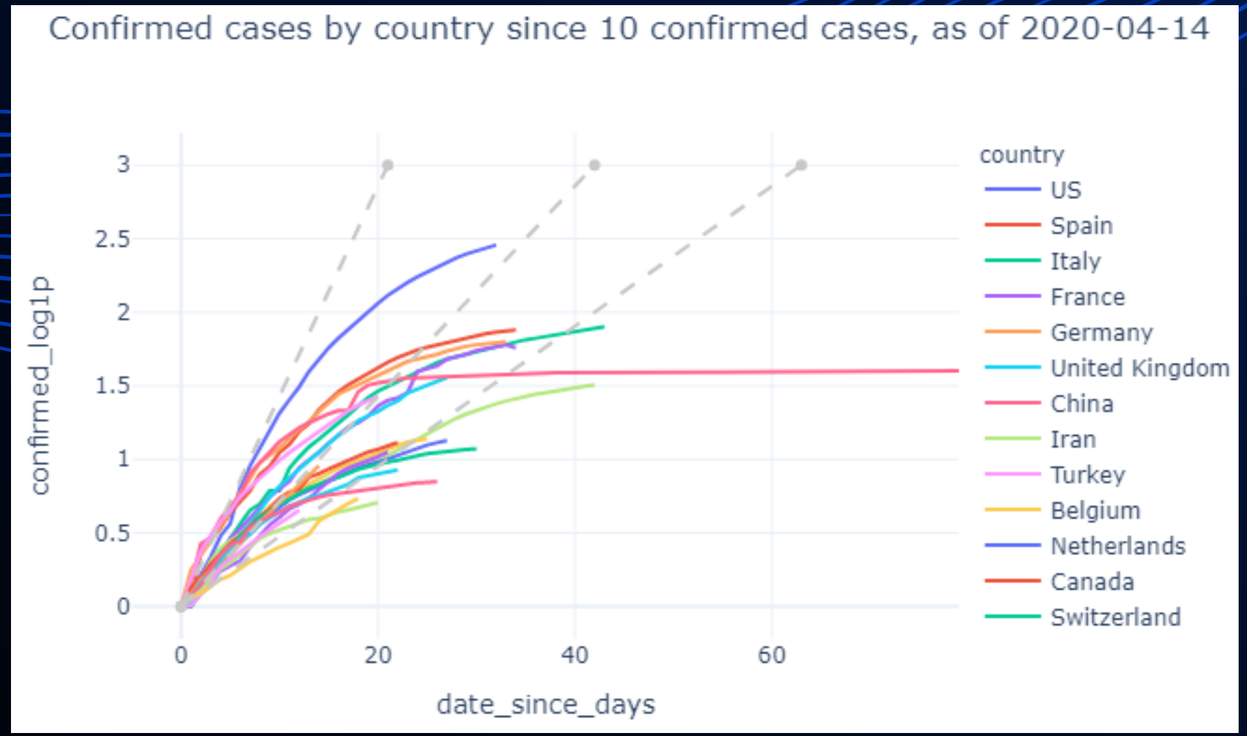
In terms of mortality rate by country, even if Algeria is on the top, Italy is in the most serious situation, since its mortality rate is over 10% as of 2020/3/28, causing the death on average of 600/700 people per day.

Instead, many Middle Eastern countries have statistically lowest % of population ages 65 and above, compared to the overall population, resulting in low mortality rate.

COUNTRY-WISE GROWTH

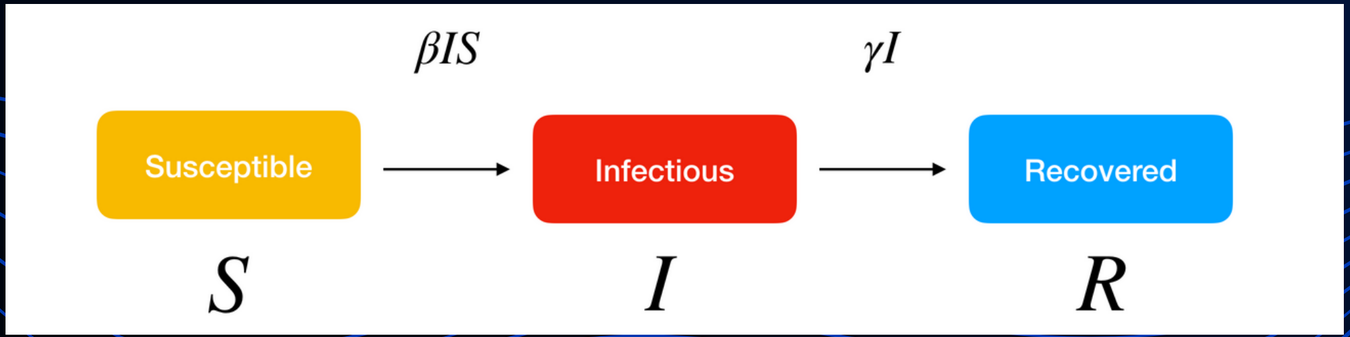
Are confirmed cases increasing at different rates in different countries?
The starting point for each country is the day that particular country had reached 10 total confirmed confirmed from COVID-19. On the x-axis we see the days since the 10th confirmed case, and on the y-axis, the total number of confirmed cases in log10 scale.

It can be noticed that US' trajectory is the steepest compared to the other countries and China's and South Korea's confirmed cases seem to have reached a steady state, thanks to the National responses including lockdowns, control over population and frequent swab tests.



MODELLING

SIR MODEL



Traditional infectious disease prediction models mainly include differential equation prediction models that establish a differential equation in order to reflect the dynamic characteristics of infectious diseases according to population growth, the occurrence of diseases and the laws of transmission within the population. The currently widely studied and applied models include the SIR, where individuals are divided into different categories, and each category is in a state, respectively: S (Susceptible), I (Infectous) and R (Recovered).

SIR model is a framework describing how the number of people in each group can change over time, according to β (Contagion Rate) and γ (Recovery Rate).

SIR MODEL

The SIR model differential equation is defined, adding the cumulative deaths X(t) to the model, that is equal to the number of cumulative deaths on day t-1 plus the number of newly infected 13 days prior multiplied with the case fatality rate α. Since the number of confirmed cases is far from the real number (not the whole population is getting tested), the number of deaths have been used in order to find the parameters for days infectious, R_0 and CFR.

SIR- MODEL WITH LOCKDOWN

Many countries implement National intervention policies, like lockdowns in order to reduce the contagion. The lockdown greatly reduces the basic reproduction number R_0, transforming this parameter into R0,2 that comes into effect on day L (i.e., lockdown).

- max_days is set equal to the length of the train data grouped by date
- N is fixed for each country, i.e. the total population
- L is fixed for each country, i.e. lockdown date
- D is set to vary from 5 to 20 (since it takes on avg. 5 days to show symptoms, at most 14, according to studies)
- CFR is set to vary from 0.1%–10%
- R_0 and R_0,2 are set to vary from 0.1 to 3.5

All assumptions are based on scientific studies.

SIR MODEL

SIR WITH TIME-DEPENDENT R_0 AND CFR

To make better predictions, R_0 and CFR are treated as functions, i.e. the variables change continuously and don't jump at the Lockdown date. Also, the CFR was until now treated as constant, however, with more people infected, treatment becomes less available and the case fatality rate increases. Now, CFR is treated as a function of:

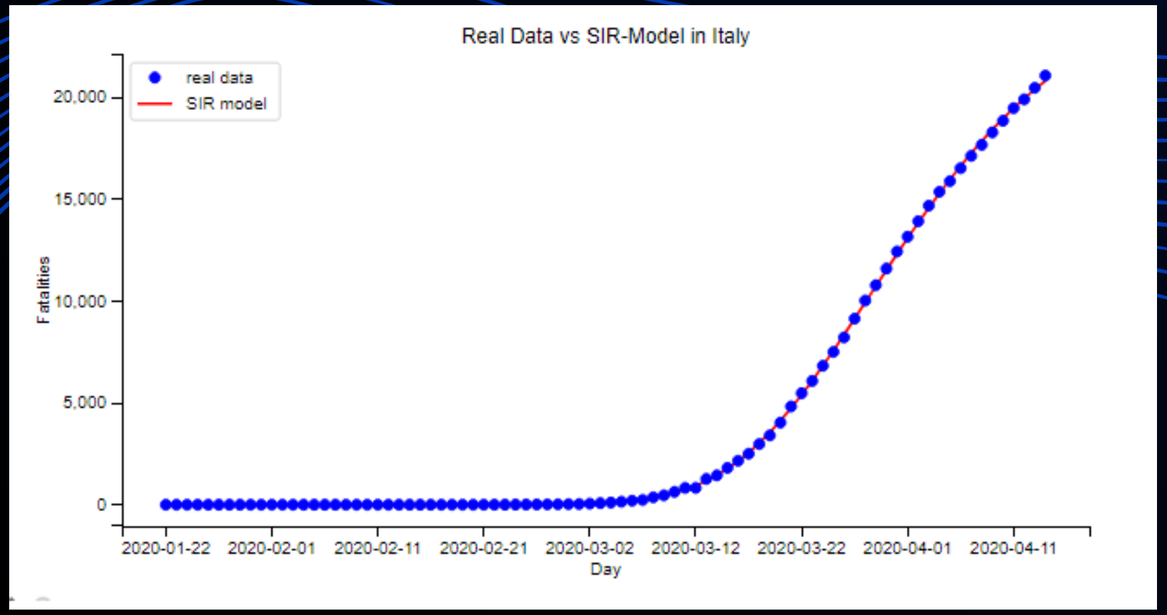
$$CFR(t) = s \cdot \frac{I(t)}{N} + \alpha_{OPT}$$

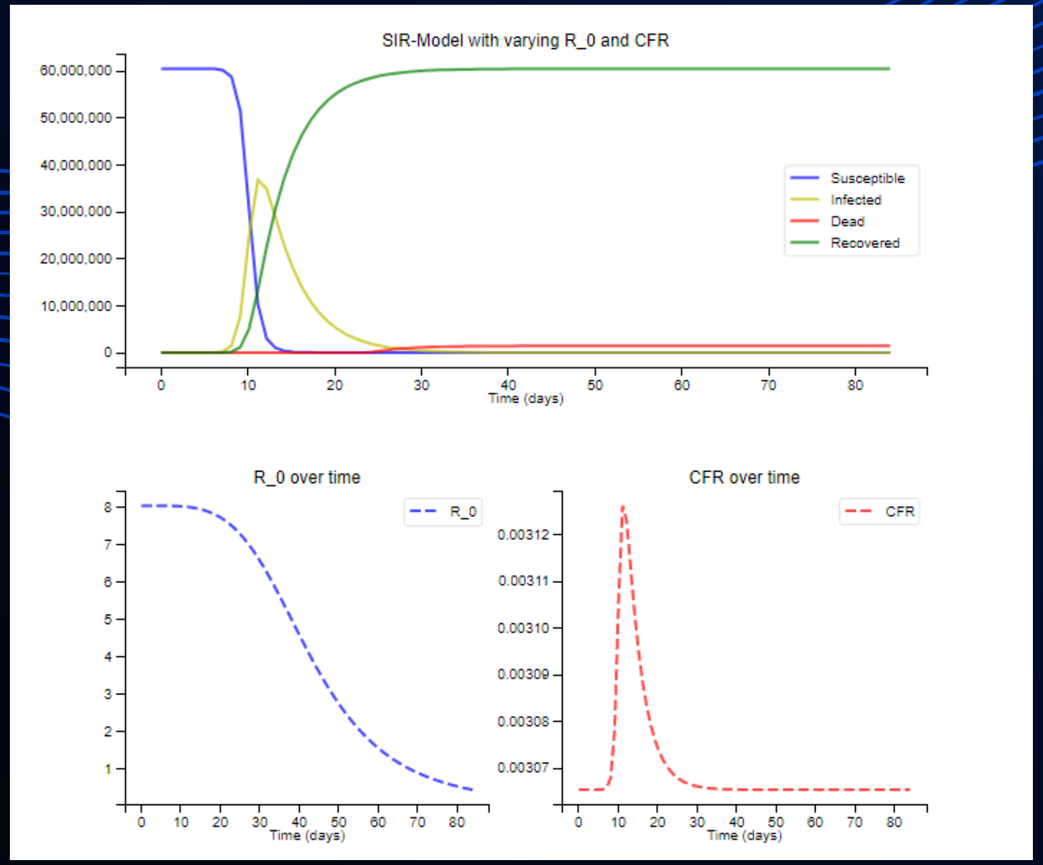
- I(t)/N, the fraction of infected of the total population
- s an arbitrary and fixed scaling factor
- And the last parameter α being the CFR with optimal treatment avaiable.

In comparison with the non-extended SIR Model, the **one with time-varying R0 and CFR seems to fit better the data.**
In the following slides, R_0 and CFR are shown, over time, for each country.

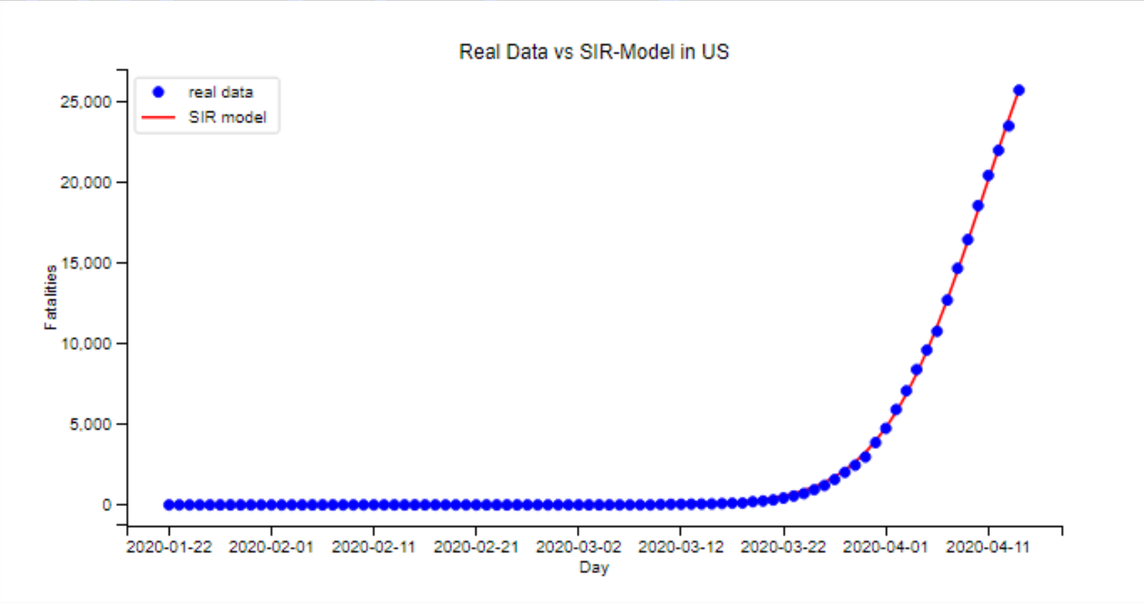
All assumptions are based on scientific studies.

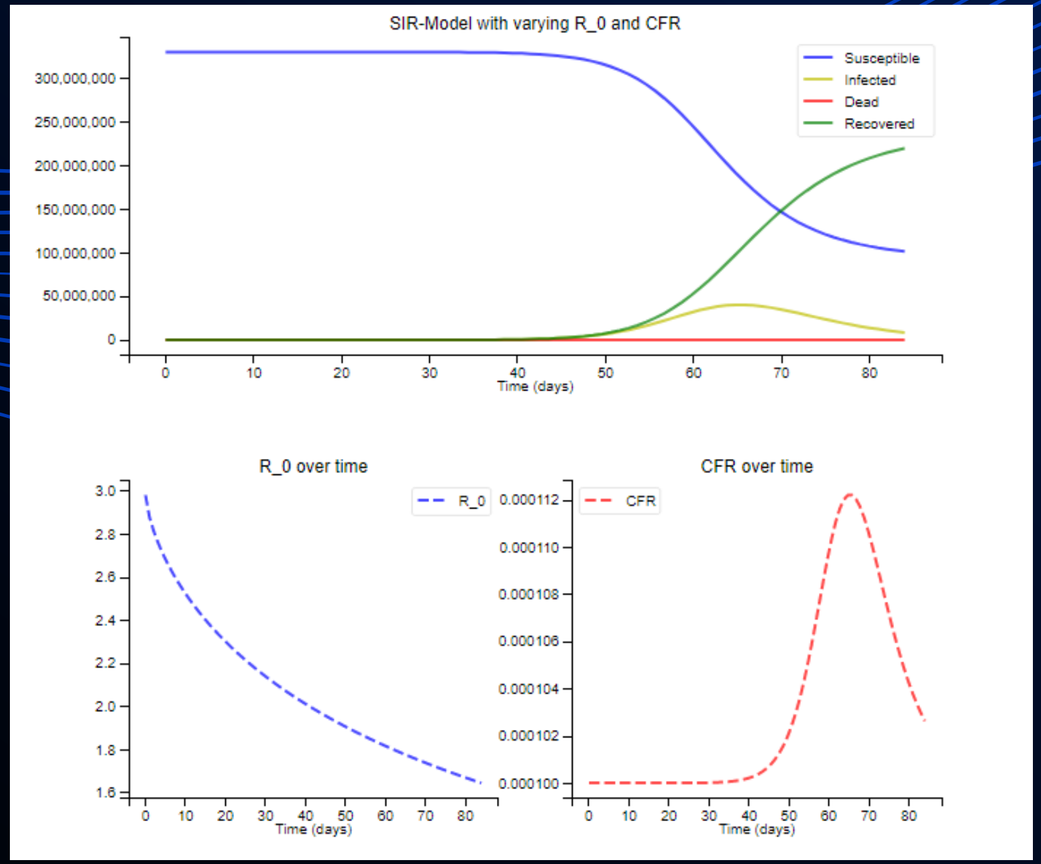
FITTING SIR TO DATA : ITALY



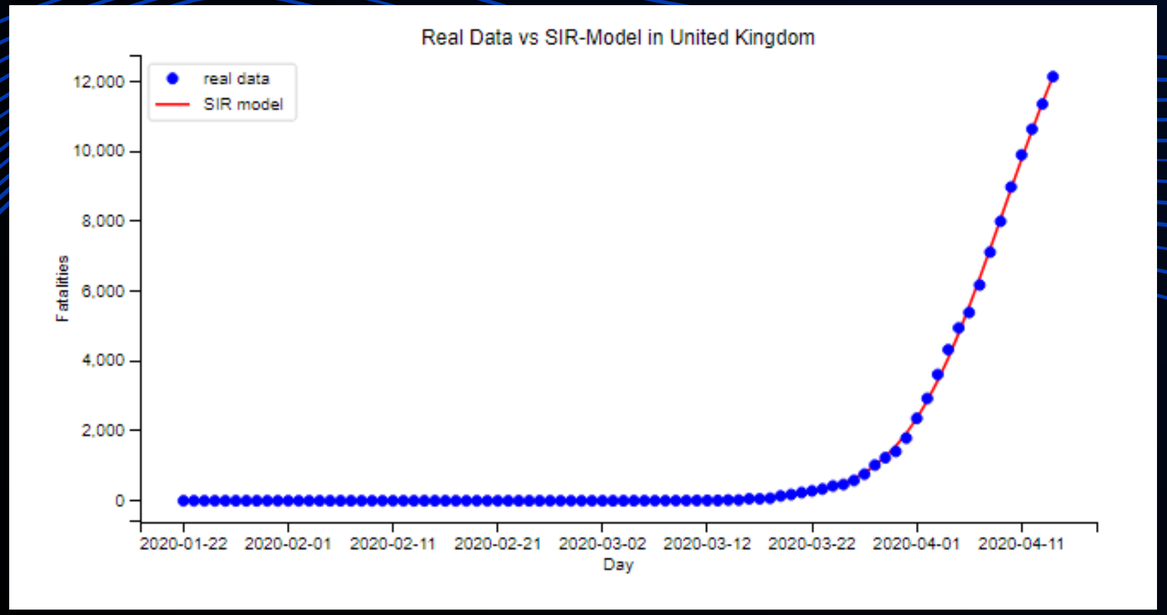


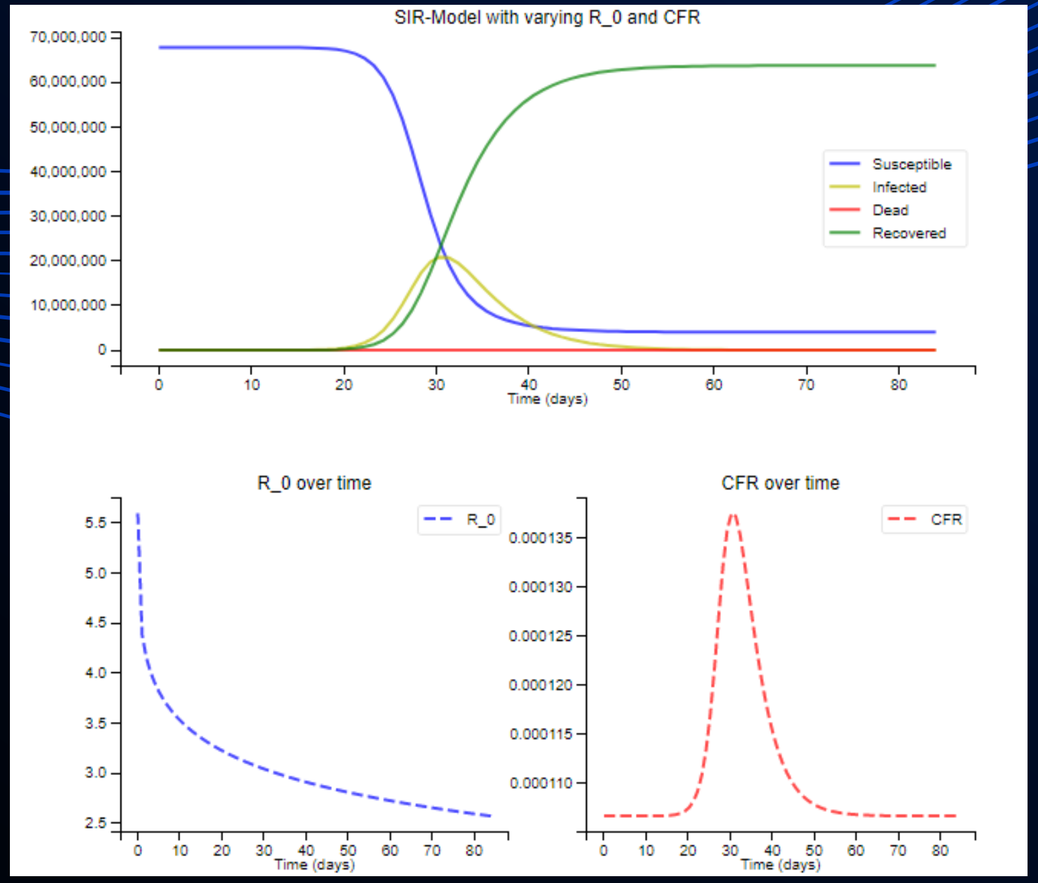
FITTING SIR TO DATA : US





FITTING SIR TO DATA : UK





PREDICTIONS

PREDICTIONS

The models are fitted only to some selected countries because it's difficult to initialize the fit parameters correctly for countries with few statistics, since the results for some of them can be ambiguous. The regression coefficient (R2) is used to evaluate the fitting ability of various methods.

LOGISTIC MODEL

Logistic model is mainly used in epidemiology. It is commonly to explore the risk factors of a certain disease, and predict the probability of occurrence of a certain disease according to the risk factors. The development and transmission law of epidemiology can be predicted through logistic regression analysis.

$$Q_t = \frac{a}{1 + e^{b - c(t - t_0)}}$$

- Qt is the cumulative confirmed cases (deaths)
- a is the predicted maximum of confirmed cases (deaths)
- b and c are fitting coefficients
- t is the number of days since the first case
- t0 is the time when the first case occurred

PREDICTIONS

GOMPERTZ MODEL

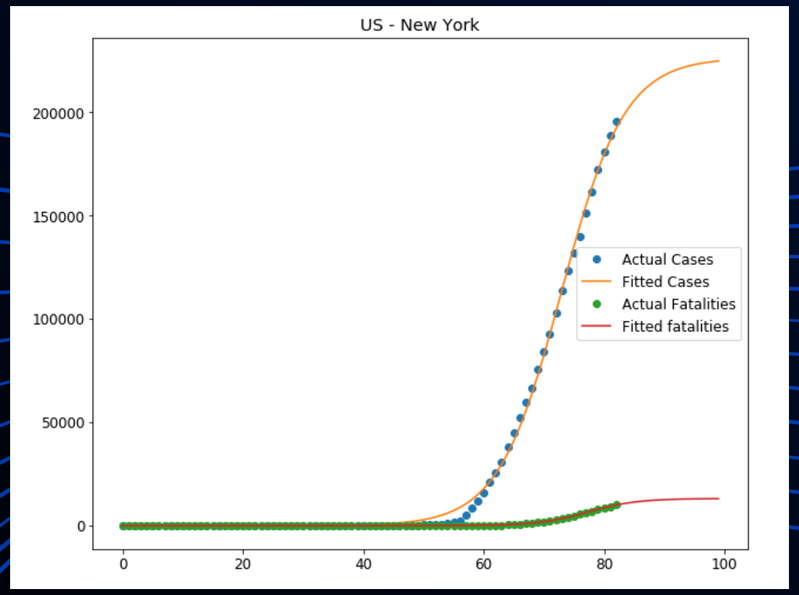
The model was originally proposed by Gomperts (Gompertz,1825) as an animal population growth model to describe the extinction law of the population. The development of infectious diseases is similar to the growth of individuals and populations.

$$Q_t = ae^{-be^{-c(t-t_0)}}$$

- Qt is the cumulative confirmed cases (deaths)
- a is the predicted maximum of confirmed cases (deaths)
- b and c are fitting coefficients
- t is the number of days since the first case
- $t0$ is the time when the first case occurred

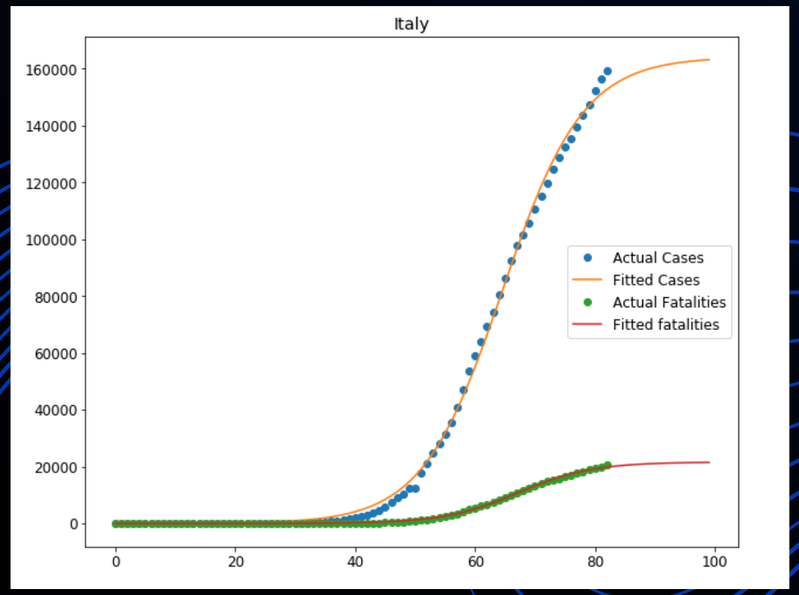
Judging from the fitting precision of the models, the Logistic model is better than the Gompertz one.

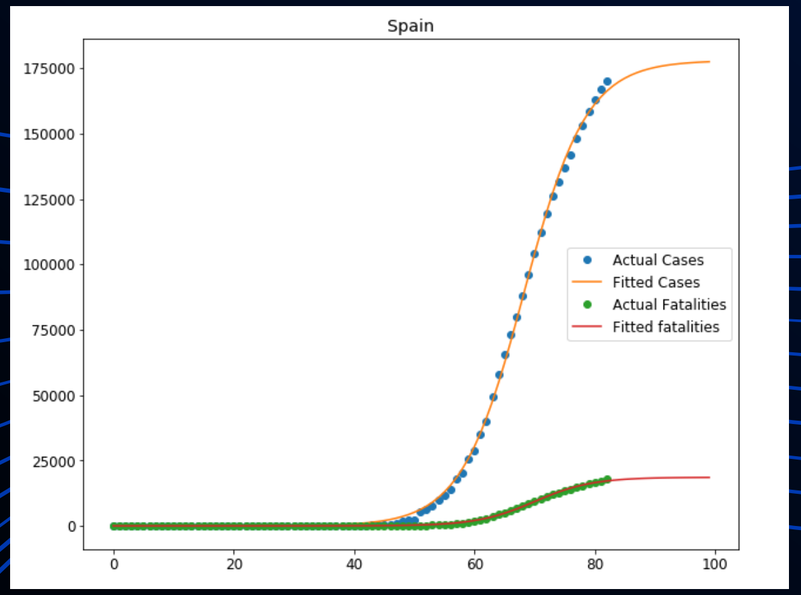
PREDICTION RESULTS: US (NY)

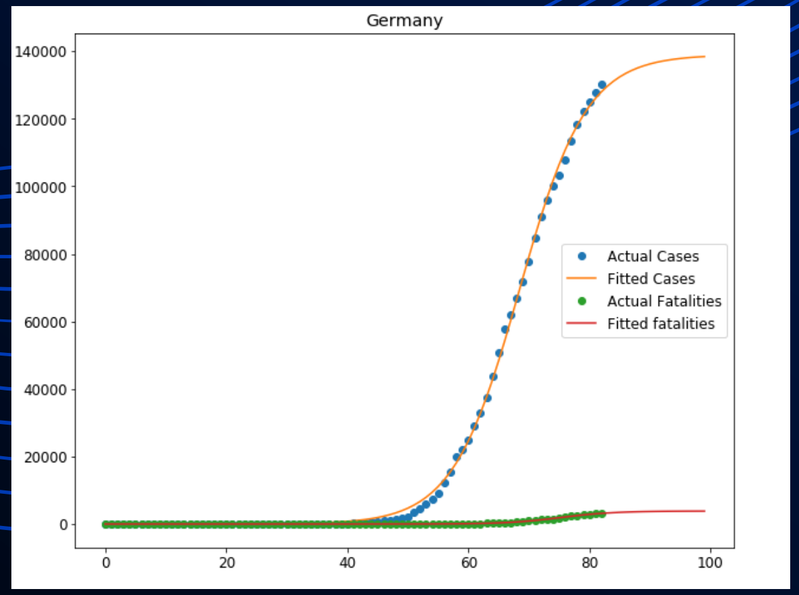


New York's confirmed cases are expected to grow till 230.000 people infected. However, the forecasts, given the contagiousness of the virus and the population density, will certainly be worse in the future.

PREDICTION RESULTS: IT, ES, DE

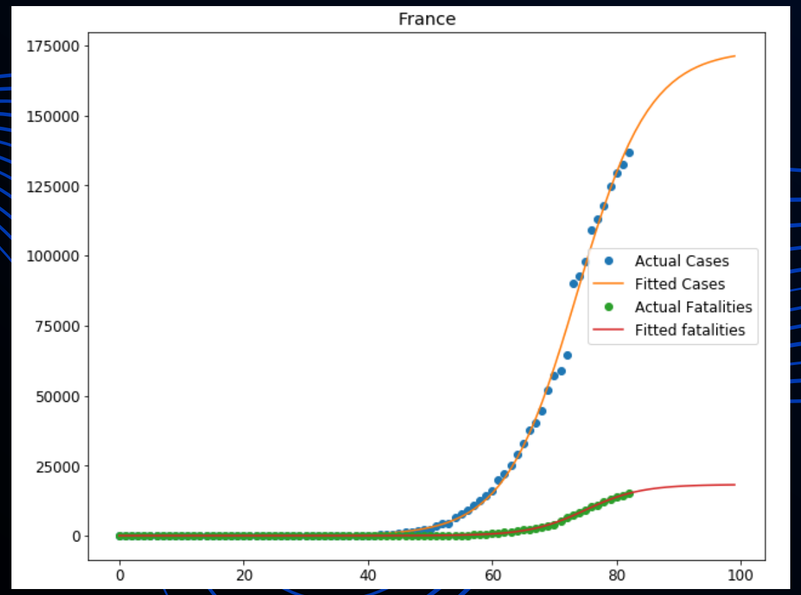


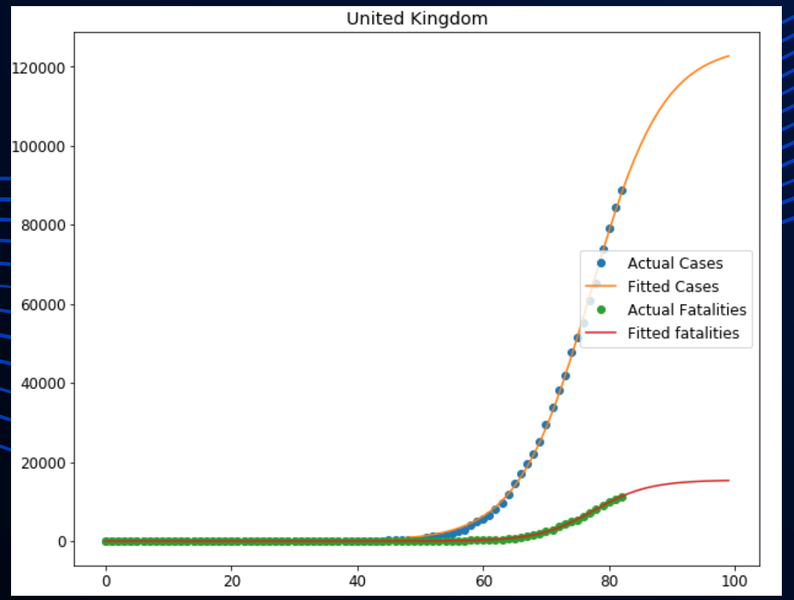




For what concerns Italy, Spain and Germany, thanks to national responses to the virus (lockdowns), it seems that, in the near future, there will be a turning point into a stable direction.

PREDICTION RESULTS: FR, UK





We cannot say the same for France and United Kingdom, where, due to a lack of an immediate national response, the contagion will not reach a stable point within the near future.

SENTIMENT ANALYSIS

SENTIMENT ANALYSIS

Has the news media been overreacting or under-reacting during the development of COVID-19? What are the media's main focuses? How is the news correlated to public reactions or policy changes? We might find many insights with more than 3,500 CBC news articles.

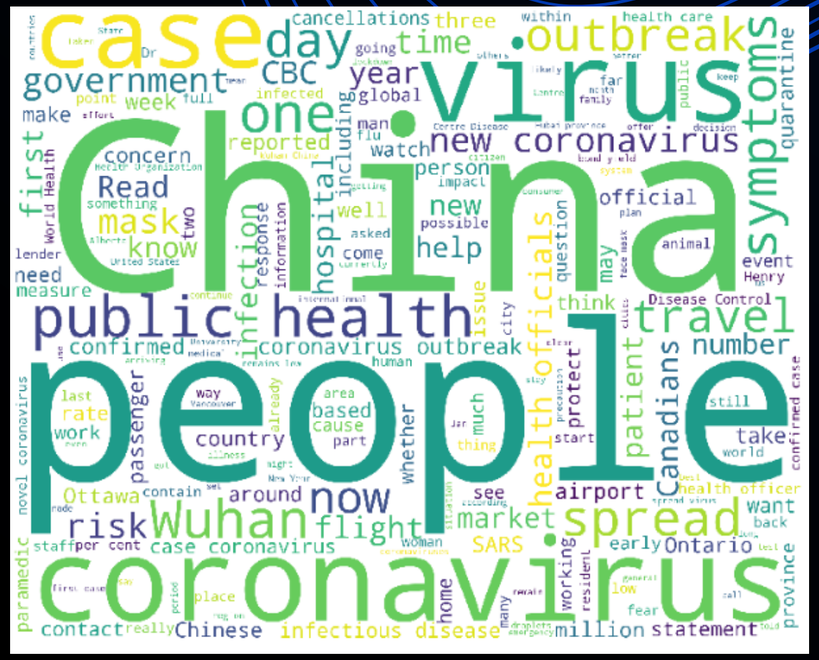
Through the use of web scraping and NLP (sentiment analysis), this analysis shows the media’s reaction to the current Coronavirus situation on a month basis in order to understand the evolution of news articles’ main focus.

SENTIMENT ANALYSIS: JANUARY

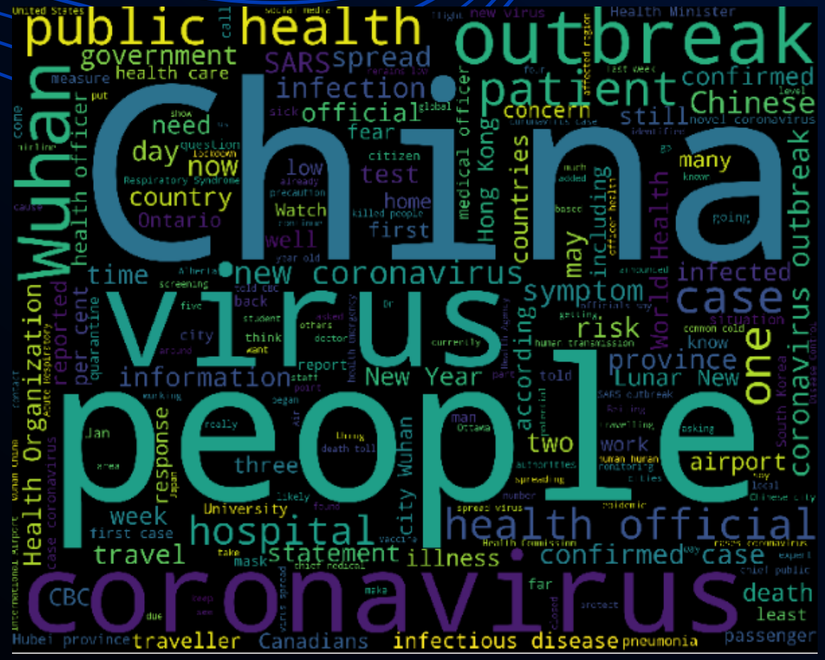
Starting from January, it can be noticed that articles are 73% negative and only 22% positive in polarity. The first split of the analysis is focused on the first-wave of the pandemic, as showed in the worldcloud by the size of the word China.

-1	73.369565
1	21.739130
0	4.891304

Positive Words



Negative Words

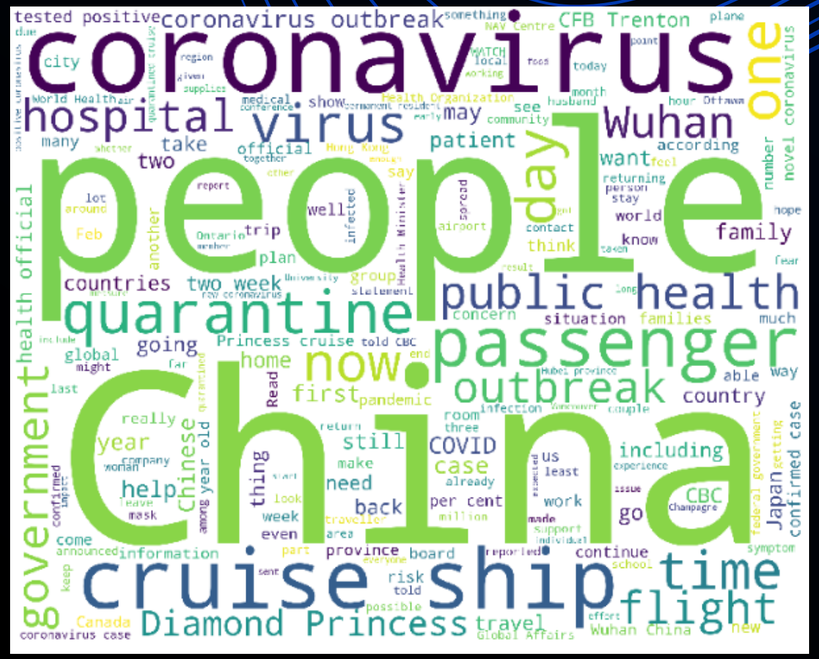


SENTIMENT ANALYSIS: FEBRUARY

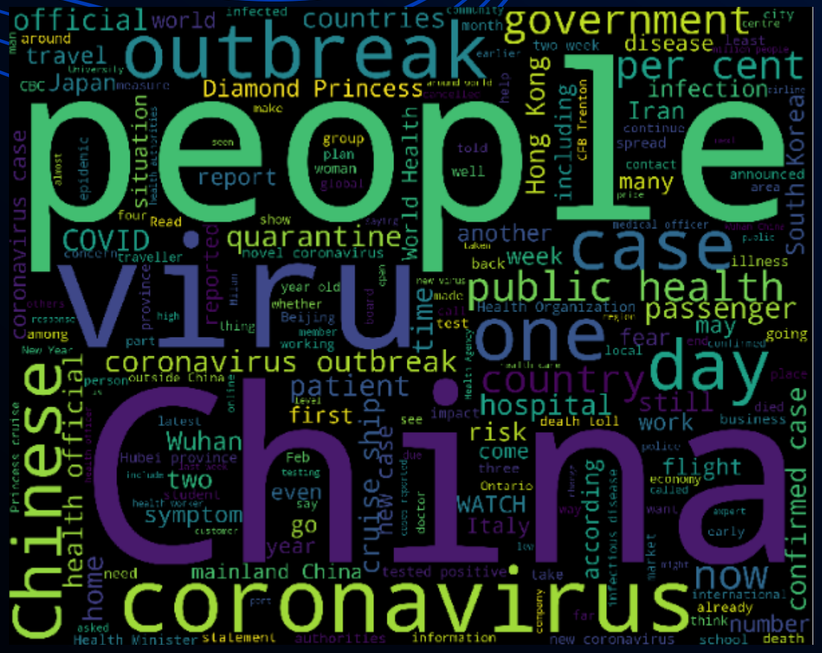
Focusing on February, it can be pointed out that the news are 43% positive in polarity, since this is the period in which scientists started to study coronavirus, symptoms and treatment. The worldcloud shows the high frequency of news about Diamond Princess and China's lockdown.

-1	52.261307
1	43.718593
0	4.020101

Positive Words



Negative Words

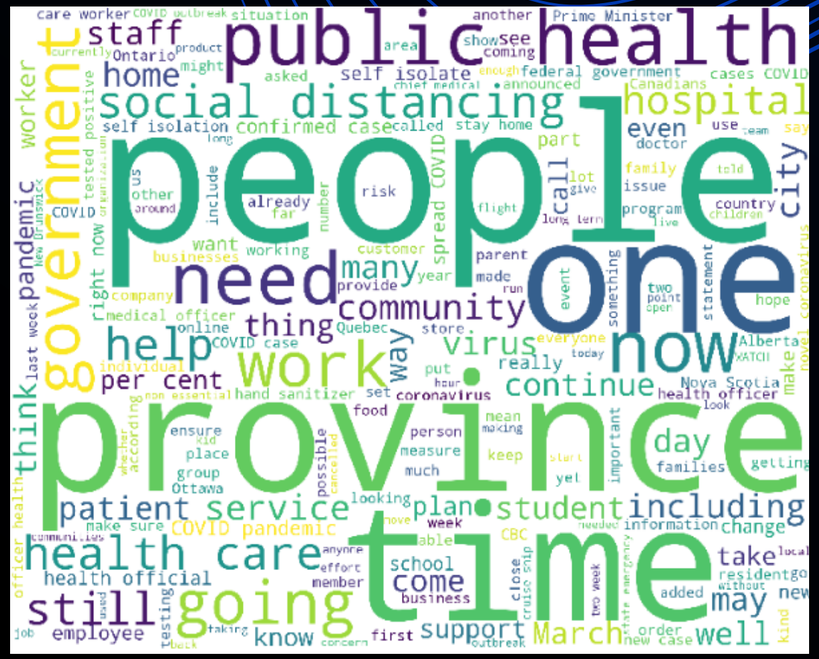


SENTIMENT ANALYSIS: MARCH

At the end, on the last month of the analysis (March), the positive news surpassed the negative ones in polarity and, the worlcloud shows how the main news' features are changed. If before we had China as one of the main words, now the focus is on the possible national responses to coronavirus, as evidenced by words like government, social distancing, self isolate, healthcare, work, help and community.

1	64.363144
-1	31.910569
0	3.726287

Positive Words



Negative Words

For further information about Python codes, we reccomend to visit our repository on GitHub: https://github.com/MarcoDibuono/COVID-19-Global-Outlook