Wisdom of the crowds in forecasting COVID-19 spreading severity

Jeremy Turiel and Tomaso Aste

Department of Computer Science, UCL, Gower Street, WC1E6BT London, UK

April 20, 2020

In this work we report that the public reacted on social media at an early stage of the COVID-19 pandemic in a surprisingly accurate way, with activity levels reflecting the severity of the contagion figures registered almost a month later. Specifically, the intensity of COVID-related social media activity from different Italian regions at the beginning of the epidemic (21-24/2/2020), predicts well the total number of deaths reached almost a month later (7/4/2020) in each region. It should be noted that at the time of the initial twitter reaction no tabled regional data on the epidemic was readily available. By the 24th February 2020 only two regions reported death cases and only three reported infected subjects.

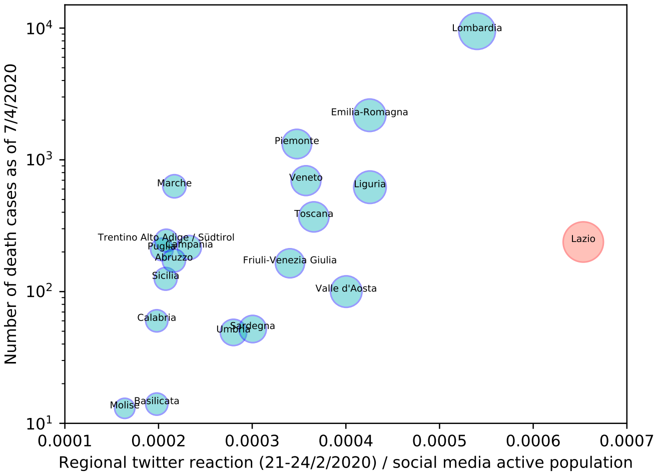

Fig. 1. Demonstration that the cumulative number of deaths in each Italian region on the 7th April 2020 are related to the tweet activity registered a moth earlier. The horizontal axis represents the mean adjusted twitter volume between the 21st February 2020 and the 24th February 2020. The date range corresponds to the peak in social media tension and the beginning of the endogenous countrywide spreading being detected. The vertical axis represent the cumulative number of deaths on the 7th April 2020. This is log-scaled to adjust for the exponential growth of the epidemic. The "Lazio" region, is a clear outlier as most politicians and institutions tweet from the capital, Rome and tweets geolocated to country level default to the capital,

Predicting the spread of COVID-19 has become the focus of many academics and amateurs across the globe. There have been proposed several different modeling tools and intuitions for the forecasting of the severity of the infection (1–4) and, despite some success, there is a shared understanding that forecasting the spread and growth of the epidemic is a challenging task. As the spreading mechanism is not yet fully understood and modelled, predicting the contagion and growth within countries and the regions in each country, before data is available, is essentially impossible. However, this task is extremely useful in order to establish targeted confinement areas, hence containing the virus more effectively while reducing the economic and social disruptions due to the lockdown. The knowledge of this would also allow to allocate resources efficiently across regions. In the present work we use data from twitter activity in different Italian regions to estimate crowd perception of the severity of the event. We then relate the intensity of social media interest with the severity of the infection in the same region in terms of the number of deaths registered the following month. Social sciences often used to forecast product sales by resorting to the “wisdom of the crowds”. These methods works well especially when groups are large and connected opinion dynamics and communication allows crowds to process information (5). In this work we show that such “wisdom” turns out to be accurate also in the prediction of COVID-19 infection severity.

We consider the case of Italy, as twitter activity data is readily available. In Italy the epidemic has now developed to a point where clear distinctions between regions can be made and data at reasonable forecast horizons has been observed.

We analyse tweets from (6), which report COVID-19 related tweets since the 22nd January 2020 . We have geolocated the most popular user locations, covering the vast majority of the dataset, and aggregated the number of unique users discussing coronavirus each day, per Italian region. For simplicity, we will refer to this as tweet volume. We then adjust tweet volume by the population active on social media per region, according to ISTAT[*] data (7, 8).

The main result is reported in Figure 1 where the cumulative number of deaths in each region on 7/4/2020 is plotted in log-scale against the mean adjusted tweet volume registered between 21-24/2/2020. Social media reaction has been adjusted by dividing by the population active on social media in each region, as per the data reported by ISTAT (7, 8). This controls for a bias towards larger regions, although one should be mindful of the higher variance expected in regions with lower tweet volume. We note that regional data for the epidemic was first available on the 24th February 2020, hence after the social media reaction, and at that time most regions still reported no cases, hence not allowing for statistically robust forecasting. The crowds therefore reacted on partial information that was not trivially obtainable from publicly available data. We point out that we used the number of deaths in our model instead of confirmed cases, as we have noticed these to be highly dependent on the number of samples taken which would require a non-trivial rescaling. The dependence on samples strengthens the relation with regional population spuriously.

The evolution of adjusted tweet volume across Italian regions for the period, as well as the growth of reported nationwide positive cases and deaths are reported in Figure 2. We

[*] Istituto Nazionale di Statistica

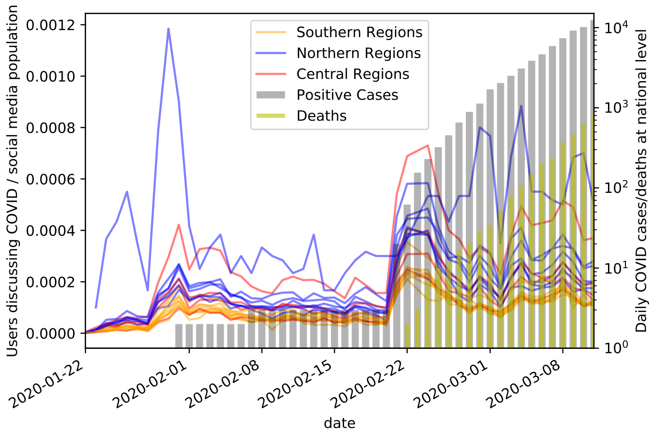

Fig. 2. Representation of the number of active tweet users posting on COVID-19 per day, geolocated and aggregated by Italian region. Regions are coloured according to their classification into Northern, Central or Southern. The plot also Displays the growth of positive COVID cases nationwide as well as the cumulative number of deaths due to the virus. We notice a hierarchy and clustering between different regional areas as the pandemic beings to spread and social media attention peaks (21-24/2/2020). The clusters in particular reflect how areas will be hit by the pandemic.

observe an initial peak in late January, perhaps due to the epidemic in China, but with little differentiation between regions. We then observe a second peak of interest from social media in late February. This appears to be sparked by the endogenous growth of the infection in Italy being measured and reported. At the time (21-24/2/2020) only nationwide epidemic data were available and regional or province breakdowns were only scattered across the news. In Figure 2 we colour-code Italian regions according to the ISTAT[†] characterisation of Northern, Central and Southern. We observe how northern, central and southern regions cluster in order, with regions most hit by the epidemic ranking higher. This seems to suggest that the initial reaction of users on social media had efficiently processed data scattered throughout news channels and performed an accurate risk assessment which is observable in the adjusted social media reaction.

To check that the values of the epidemic are not trivially related to the size of the population in each region (7) and that our analysis adds to this we perform and compare three regression models:

1. adjusted tweets vs. log death cases;
2. log population vs. log death cases;
3. adjusted tweets and log population vs. log death cases.

As it can be noticed from Figure 1, "Lazio" is an outlier due to politicians and central bodies tweeting from it as well as national geolocation defaulting to the capital. This region has been therefore removed from the regression. We also log scale the population to allow for a fair comparison as we notice a sub-linear relation to the number of deaths. We report in Table 1 p-values for the coefficients as well as R^2 for the three regression models.

It can be inferred from Table 1 that adjusted tweet volume is a better regressor than log population. This is shown by both the higher R^2 value in regression 1 with respect to regression 2, and by the significant p-value for tweets in regression 3.

Regression index	$p_{adjusted\,tweets}$	$p_{log\,population}$	R^2
1	$5 \cdot 10^{-4}$	-	0.489
2	-	$9 \cdot 10^{-3}$	0.431
3	$3 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	0.737

Table 1. Tabular summary of coefficient p-values and R^2 values for the regressions discussed above. We confirm that adjusted tweets are a better regressor for the future number of deaths. The regression is performed using the `statsmodels.api.OLS` class from (9).

We conclude that this is an important example of crowd wisdom in a phenomenon which is not directly controlled by the population or its opinion. These results indicate that social media activity may be used to forecast the severity of the spreading of COVID-19 in different countries at an early stage when data from the effect of the disease are not available yet.

Acknowledgments

The authors acknowledges Fabio Caccioli, Licia Capra, Delmiro Fernandez-Reyes, Giacomo Livan and Simone Righi for useful feedback and discussions. The authors acknowledge the EC Horizon 2020 FIN-Tech project for partial support and useful opportunities for discussion. JT acknowledges support from EPSRC (EP/L015129/1). TA acknowledges support from ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018-2 825215).

References

1. Zhang J, et al. (2020) Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside hubei province, china: a descriptive and modelling study. *The Lancet Infectious Diseases.*
2. Chinazzi M, et al. (2020) Preliminary assessment of the international spreading risk associated with the 2019 novel coronavirus (2019-ncov) outbreak in wuhan city. *Lab. Model. Biol. Soc.–Techn. Syst.*
3. Roosa K, et al. (2020) Real-time forecasts of the covid-19 epidemic in china from february 5th to february 24th, 2020. *Infectious Disease Modelling* 5:256–263.
4. Grasselli G, Pesenti A, Cecconi M (2020) Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response. *JAMA.*
5. Bassamboo A, Cui R, Moreno A (2015) Wisdom of crowds in operations: Forecasting using prediction markets. *Available at SSRN 2679663.*
6. Chen E, Lerman K, Ferrara E (2020) Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372.*
7. ISTAT INdS (2020) Popolazione residente al 1° gennaio (http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1).
8. ISTAT INdS (2020) Internet: accesso e tipo di utilizzo: Attività svolte su internet - reg. e tipo di comune (http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPRES1).
9. Seabold S, Perktold J (2010) statsmodels: Econometric and statistical modeling with python in *9th Python in Science Conference.*

[†] Istituto Nazionale di Statistica