2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Online feelings and sentiments across Italy during pandemic: investigating the influence of socio-economic and epidemiological variables

Francesco Scotti, Davide Magnanimi, Valeria Maria Urbano, Francesco Pierri
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Via Giuseppe Ponzio 34/5 20133 Milano, Italia
E-mail: {name.surname}@polimi.it

***Abstract*—During the on-going COVID-19 pandemic, online social media have been extensively used by policy makers and health authorities to quickly disseminate useful information and respond to public concerns in a timely fashion. Notwithstanding the huge amount of literature on analyzing positive and negative emotions conveyed by social media users, researchers have not widely investigated the main determinants of online sentiment during crises.**

To fill this gap, in this paper we analyse a large-scale dataset of over 1.7 M tweets in order to understand whether online feelings, expressed by Italian individuals on Twitter during the pandemic, have been affected by socio-economic and epidemiological variables. Leveraging both panel models and cross-section regressions at different geographical levels, we find that more pessimistic feelings are communicated by users located in areas where the virus hit more severely, with a higher mortality rate and a larger fraction of infected individuals with respect to the local population. Finally, we show that administrative units exhibiting the most positive emotions are those characterized by lower income per capita and larger socio-economic deprivation, suggesting that sentiments in online conversations could be driven by epidemiological factors and by the fear of economic backlashes in wealthier areas of Italy.

Keywords: pandemics; sentiment analysis; social networks; Twitter

INTRODUCTION

As the pandemic of SARS-COV-2 was spreading around the globe during 2020, we witnessed a surge of public interest around social media data as vital sources of information to fight against the virus [6]. As a matter of fact, monitoring human activity may enable authorities and policy makers to detect outbreaks at an early stage, minimise their spread [6] [12] and understand the evolution of individuals feelings during such an unprecedented crisis [3] [1].

COVID-19 is not the first epidemic whose effects on people opinions and emotions were analysed on social media data. During the 2009 H1N1 epidemic, a content analysis of Twitter data revealed the importance of such a platform as a rich source of opinions and shared experiences, which could be used to support health authorities to better respond to public concerns [9]. In a similar analysis during the Ebola outbreak [10], authors showed that the epidemic provoked mainly negative emotions such as anxiety, anger, swearing and death, while in [18] it was confirmed that panic was the prevalent sentiment in the general landscape of social media.

A similar line of research has been followed during the recent pandemic of the novel COVID-19 disease. For instance, governments and sanitary institutions worldwide have used social networks to spread information across the civil community, generating trust and fear as the main reactions. According to the authors of [14], the presence of contrasting emotions among citizens as joy and sadness, disgust and surprise may also reveal the absence of a homogeneous perception of the severity of the situation and the difficulties experienced by policy makers and health authorities to transmit a harmonized and coherent message to society. This issue was corroborated by some evidences exhibiting that individuals on Twitter relied more upon common user generated content rather than on official government and other related institutional communications [17]. Finally, the analysis of the use of language across 39M tweets shared in the United States, allowed the identification of three different phases characterizing online reactions to the pandemic: a refusal stage, while only other countries were experiencing deaths, a suspended reality period, which started after the first COVID-19 victim was announced in each state, and, finally, an acceptance phase, when authorities imposed social distancing measures, in which individuals found a "new normality" for their daily activities, conveying more positive sentiments [1].

Despite the large amount of scientific works which covered the impact of epidemics on individuals' feelings, to the best of our knowledge there is a handful of recent contributions that analyzed to what extent the swing of emotions is influenced by the socio-economic conditions of the population across different geographical areas. In 2017, [7] evaluated the spread of colera in Uganda, investigating socio-economic characteristics of population to understand whether the promotion of social services, such as education, could contribute to control the virus. More recently, [15] carried out a study on the association between socio-economic variables and sentiment of US citizens about lifting mobility restriction put in place during the COVID-19 crisis. They showed that family households, individuals with limited education levels, low-income and higher house rent are more interested in restarting the

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

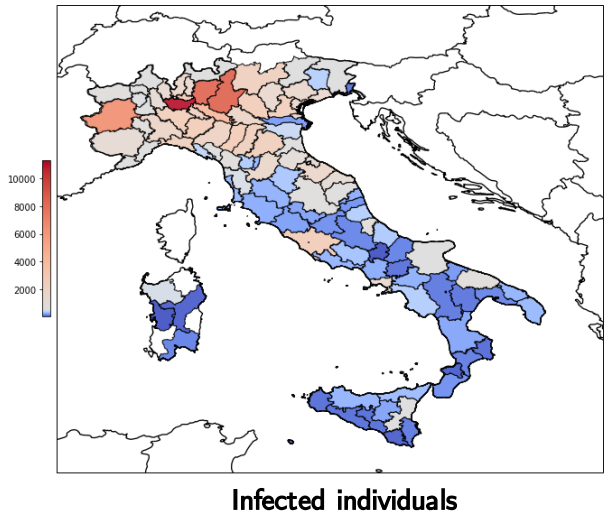

Fig. 1. Number of infected individuals for each province over the entire period of observation. The colorbar is centered at the median value of the distribution. According to official data by Italian Civil Protection, more than 70% of overall number of infected individuals was reported in Lombardia, Piemonte, Emilia Romagna and Veneto, while Campania, Basilicata, Molise, Puglia, Calabria and Sicilia accounted for less than 10% of cases

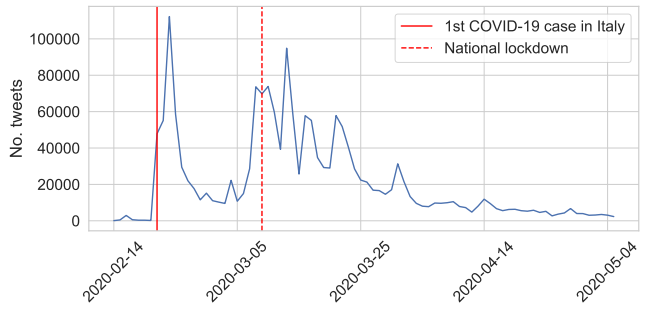

Fig. 2. Number of daily tweets in the period of observation. Red lines indicate respectively the 1st diagnosed COVID-19 case in Italy (solid) and the beginning of national lockdown (dashed).

economy, thus providing relevant signals to policymakers that seek to reduce the impact of the pandemic.

Against this background, in this work we analyse the Italian case and investigate whether the sentiments and the emotions expressed by Twitter users were influenced by their geographical location, by socio-economic factors and epidemiological variables. In particular, we focus on Italy as it was the first European country to be hit by the virus spreading outside China borders, and the first Western country to apply a full lockdown on national scale. The pandemic did not have a homogeneous impact on the Italian peninsula: as the outbreak was located in Lombardy, nearly all northern regions were characterized by large amount of cases with a high risk to overburden the treatment capacity of healthcare infrastructures, whereas southern regions managed to prevent the contagion (see Figure 1).

Recent studies showed that the lockdown had an asymmetric socio-economic impact on municipalities [4], and that variables such as geographical provenance and professional status of individuals had a significant effect on the perceived severity of COVID-19 for health [8]. In this spirit, we study whether this heterogeneity is experienced also in terms of feeling and opinions communicated by Twitter users and we scrutinize the main determinants of these emotions.

To pursue our research goal, we analyse a dataset containing 1.7 million of tweets related to the pandemic and collected between mid-February and beginning of May. We perform an automatic sentiment extraction of tweets and deployed several panel and cross-section regression models to understand the interplay between feelings expressed by online users and a set of socio-economic and epidemiological variables, computed at municipality and province level.

The outline of remaining sections is as follows: in the next section we describe the data used in our analysis; then we provide the methodology employed; next we show experimental results and finally we draw conclusions and future work.

Data collection and description

Twitter

In our study, we leverage a large-scale dataset of tweets obtained using Twitter Streaming API[1] from February 17th to May 5th. We specified the filter query using a non-exhaustive list of Italian keywords related to the pandemic (e.g. "coronavirusitalia", "coronavirusitaly", "covid19italy"), resulting in approximately 1.7 M tweets. We provide in Figure 2 a time series for the daily number of tweets. We can notice a considerable increase in online activity in correspondence of the outbreak in Italy (February 21st) and the resulting lockdown (March 9th). We matched the geo-localization of each tweet, when present, against Italian municipalities, resulting in a total number of approximately 400 k tweets and 1,855 municipalities covered (out of a total of 7,903 municipalities).

Socio-economic and epidemiological variables

To build our regression model we rely on the following *socio-economic* variables, which are available at municipality level:

- **Average income per capita**: it indicates the average income declared by taxpayers in 2018 (thus corresponding to the financial year 2017) in a given area.
- **Inequality**: it is measured as the ratio between mean and median values of the distribution of declared income by taxpayers in 2017. Territories with values larger than 1 are those where income is less equally distributed, as the distribution of income presents a positive skewness which is a symptom of few individuals with large income and a concentration of citizens with lower wages.
- **Fiscal capacity**: it represents administrative revenues based on three different sources: property tax, local income tax, and local fees (measured in 2016). It represents, together with standard expenditure needs, the main building block of the Italian system of fiscal equalization.

[1]https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/api-reference/post-statuses-filter

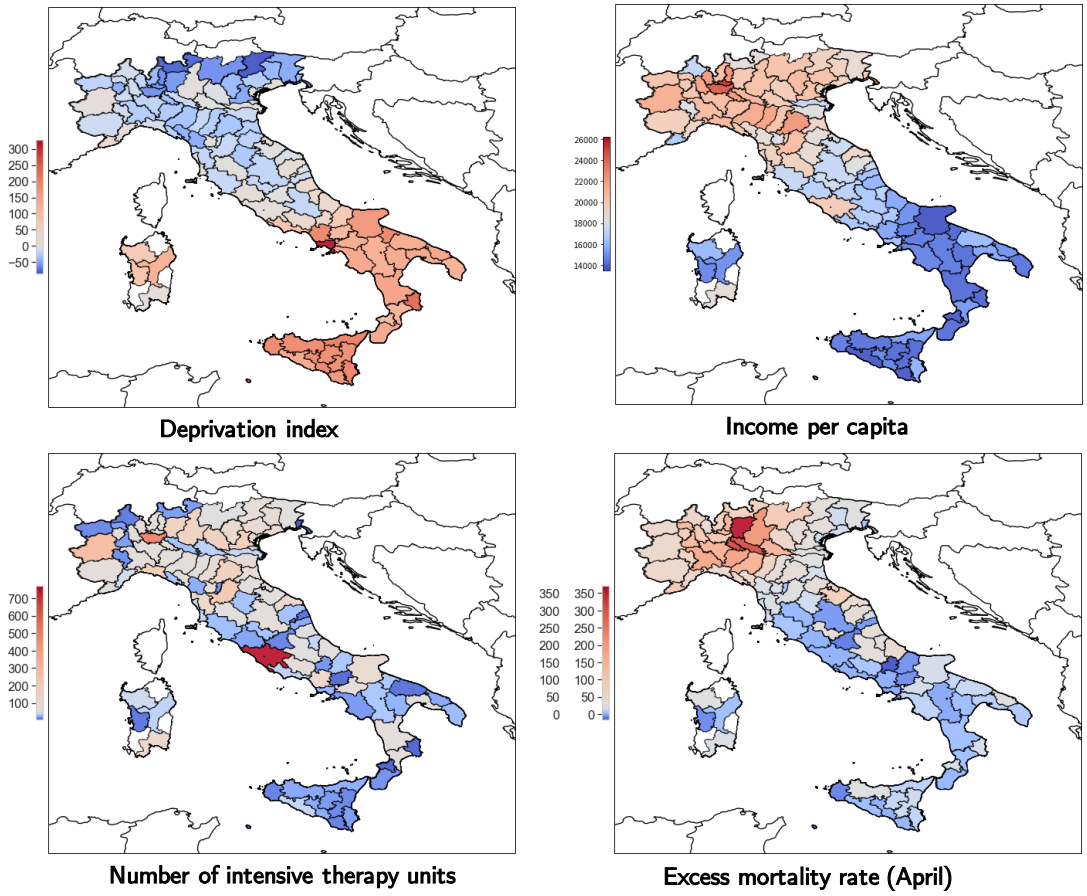

Fig. 3. Geographical distribution of several regression variables measured at province level. In clockwise order: Deprivation index, Income per capita, Excess mortality rate in April and Number of intensive therapy units. For each plot the colorbar is centered to the median value of the distribution.

- **Deprivation index**: This variable is a composite index which covers the following dimensions: education, unemployment, housing, population density and economic poverty. Indicators for each dimension are computed and then transformed in percentage deviation from the national mean, and finally aggregated together with equal weights.

All these measurements are publicly provided by the National Institute of Statistics (ISTAT) and the Ministry of Economics and Finance (MEF).

We further include the following *epidemiological* variables which provide information on the on-going pandemic:

- **Number of intensive/sub-intensive therapy units** available in a given area and which represents a proxy of the capacity of local healthcare infrastructure to provide adequate medical cares to patients with serious respiratory diseases.
- **Extra mortality rate**: this variable was computed by ISTAT by comparing the mortality rate of 2020 w.r.t to the mean mortality of previous 5 years, through the integration of data disclosed by the Italian Superior Sanitary Institute (ISS) and the National Register of the Resident Population Agency (ANPR).
- **Infected individuals (%)**: the fraction of individuals hit by the virus w.r.t to the entire population.
- **Mean basic reproduction number**: the average, over the period of observation, of the expected number of cases generated by an infected individual in a population where all individuals are susceptible.

In particular, all except the extra mortality rate are available only at province level. Measurements are publicly provided by the Italian Civil Protection. As we describe next, we accordingly take them into account only in the regression model at province level.

In Figure 3 we provide the geographical distribution of some variables at province level.

Methodology

Extracting online sentiment

In this work we used the SentITA tool [13], a lexicon-based approach to compute the polarity of each tweet. The SentITA model was trained using two sets of texts in Italian language, one from Twitter and one from Booking reviews. We briefly describe the technique employed in the following.

The combination of the two datasets provides 15.000 positively or negatively labelled sentences, which were then combined with 90.000 Wikipedia neutral sentences to train the model properly. The overall dataset, containing around

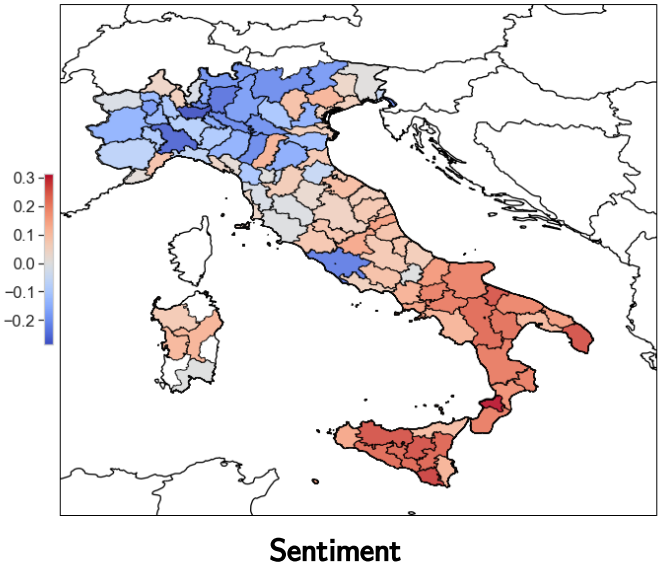

Fig. 4. Geographical distribution of sentiment at province level. The colorbar is centered on zero.

102.000 sentences was used to train a deep learning model, namely an Attentional Bidirectional Recurrent Neural Network with LSTM cells, that operates at word level.

Specifically, each word is represented by five vectors that corresponds to five different features: high dimensional word embedding, word polarity, word NER tag, word POS tag, custom low dimensional word embedding. The high dimensional word embedding are obtained through the Fastex embeddings for Italian. The word polarity is obtained from the OpeNER Sentiment Lexicon Italian. The NEG and POS tags are obtained from the `Spacy` library Tagger model for Italian language, and finally the custom low dimensional word embeddings are generated by random initialization. In its initial layer, the model first performs a dimensionality reduction on the Fastex embeddings and then concatenates them with the rest of the embeddings for each word of the sequence. This concatenation activity is fed in a sequence of two bidirectional recurrent layers with LSTM cells. Results are then passed to the attention mechanism and lastly to the dense layers. This provides the polarity signals. Taking sentences as an input (the maximum sequence length has been set to 35 words) the SentITA tool provides two polarity signals ranging from 0 to 1, one for positive sentiment and one for negative sentiment.

We used the model to assign to each geo-localized tweet two polarity values, and then we computed the average positive (negative) sentiment for each municipality/district in a given period of time by computing the mean positive (negative) signal of tweets originated from a given municipality/district and published in the period of interest.

We show in Figure 4 the geographical distribution of sentiment across Italian provinces.

Regression model

In order to investigate whether the sentiment of different administrative units was affected by socio-economic factors and epidemiological variables, we apply a panel regression analysis with the following specification equation:

$$Y_{i,t} = \beta_0 + \gamma X_i + \delta Z_{i,t} + u_i + \epsilon_{i,t} \quad (1)$$

where index i refers to the underlying administrative unit (either municipalities or provinces) and subscript t indicates the considered unit of time, i.e. four non overlapping windows of 14 days starting from March 2nd.

More in detail, the dependent variable represents the average sentiment at municipality (province) level, computed as the mean net sentiment of all tweets geo-localized in a certain municipality (province) in the period t. In particular, as the SentIta library associates to each tweet a positive and negative score, the net sentiment score is computed as the difference between the positive and the negative value.

For what concerns the right-hand side of the equation, X_i is a matrix of time fixed regressors covering socio-economic and health infrastructure capacities dimensions described in the previous section.

By plugging the set of *socio-economic* variables (plus the number of intensive/sub-intensive therapy units) we control for the fact that the heterogeneity in the sentiment across different provinces might be driven by wealth, economic disparities and capacities of the local health infrastructure. Indeed, as the restrictive measures adopted during the lockdown severely hit several production sectors, value chains, trade exchange and the sanitary system risked to be overburden by the number of patients, we suspect that these variables might be relevant determinants to explain the sentiment of Twitter users.

As the COVID-19 pandemic was primarily a sanitary emergency, we include into the model the matrix $Z_{i,t}$ which encompasses time variant regressors addressing the *epidemiological* dimension. For these factors we take into account also the previous lag, as we suspect that the sentiment may show a certain delay with respect to the pandemic pattern and current feelings might be affected by events occurred in previous time units.

To provide robustness to our estimates we carry out our analysis at two levels of granularity, combining a panel and cross section regression analysis, using first provinces and then municipalities as reference administrative units. Insofar we assess whether our estimates are affected by the aggregation process which might shrink some relevant effects once moving to the province perspective, or if instead results are stable at different geographical scales.

Overall, our dataset comprises 1,855 out of 7,903 Italian municipalities, covering more than 70% of overall Italian population.

Results and Discussion

Regression at province level

In this section we describe results of our regression model when considering provinces as administrative units, and thus plugging into the model *socio-economic* and *epidemiological* variables at province level. Results are provided in Table I.

	Dependent variable:			
	Province Sentiment			
	(FE)	(RE)	(GMM)	(OLS)
Intercept		0.089	0.076	0.018
		(0.075)	(0.245)	(0.472)
Income pc		-0.231^{**}	-0.274^{**}	-0.340^{***}
		(0.110)	(0.114)	(0.108)
Inequality		-0.071	-0.048	0.036
		(0.093)	(0.082)	(0.086)
Deprivation		0.222^{*}	0.289^{***}	0.220^{*}
		(0.127)	(0.107)	(0.117)
Fiscal Capacity		-0.047	-0.020	-0.122^{*}
		(0.079)	(0.062)	(0.072)
Extra Mortality t	-0.013^{*}	-0.082	-0.107^{*}	-0.373^{***}
	(0.0074)	(0.061)	(0.055)	(0.088)
Extra Mortality t-2	0.105	0.080	-0.036	-0.015
	(0.092)	(0.055)	(0.027)	(0.049)
Extra Mortality t-4	-0.027	-0.020	-0.028	
	(0.047)	(0.040)	(0.047)	
Infected Individuals t	-0.167^{**}	-0.124^{*}	-0.127^{*}	-0.827^{*}
	(0.073)	(0.066)	(0.076)	(0.521)
Infected Individuals t-1	-0.152^{*}	-0.115	-0.028	-0.849^{*}
	(0.079)	(0.073)	(0.078)	(0.534)
Intensive Therapy		-0.136^{**}	-0.093	-0.152^{**}
		(0.064)	(0.073)	(0.071)
Reproduction Number t	-0.017	-0.011	0.105	-0.025
	(0.035)	(0.034)	(0.226)	(0.060)
Reproduction Number t-1	0.064	0.080^{**}	0.132	0.005
	(0.039)	(0.038)	(0.088)	(0.055)
Observations	420	420	420	105
Adjusted R^2	0.109	0.283	0.135	0.763
Hausman test	5.78^{-11}			
Hansen test			0.94	
Arellano Bond test order 1			3.54^{-8}	
Arellano Bond test order 2			0.34	
Breusch-Pagan Test	3.95^{-8}			
Note:		*p<0.1; **p<0.05; ***p<0.01		

TABLE I

RESULTS OF REGRESSION MODEL(S) APPLIED AT PROVINCE LEVEL. IN THE OLS REGRESSION MORTALITY T, CONTAGIONS T, AND REPRODUCTION COEFFICIENT T REFERS TO THE CORRESPONDNET VARIABLES MEASURED IN THE MONTH OF APRIL, WHILE THE PREVIOUS LAG REFERS TO THE SAME FACTORS MEASURED IN MARCH

In a first step, we estimate a random effects panel model based on White heteroskedasticity robust standard errors [19]. Indeed, the Breusch Pagan test clearly rejects the hypothesis of constant variance across the error terms. Moreover, as the individual component of the error term u_i cannot be a-priori considered uncorrelated with the time varying regressors [11], we rely on the formal Hausman test to distinguish between a random and a fixed effect model. A p-value < 0.05 strongly rejects the null hypothesis and is clearly in favour of the fixed effects specification. In this case, as the fixed effect model is based on the within transformation which subtracts to each observation the individual mean from each variable, the time invariant regressors cannot be estimated.

However, these results might still be biased, due to the fact that the idiosyncratic component of the error term $\epsilon_{i,t}$ might still be correlated with some of the explanatory variables. As a consequence, notwithstanding the exploitation of a fixed effects model, the regressors might still be endogenous. To deal with this issue we rely on a Generalized Methods of Moments (GMM) estimator, based on the [2] methodology. We limit the number of instruments to two per each regressor, exploiting lags 2 and 3, to guarantee a parsimonious usage [16]. The reason for this is that adopting too many instruments, reduce the power properties of the Hansen test [5] and lead to a downward-bias standard error [20]. In this case we perform a Hansen test to verify that the identified instruments are valid and robust. A p-value $= 0.94$ is in favour of the null hypothesis, which states the exogeneity of the instruments. Moreover, the GMM estimator provides reliable results in case the first differenced error term does not display second order autocorrelation and for this reason, we perform the Arellano-Bond test. In particular, we show evidence of first order autocorrelation, but absence of statistically significant autocorrelation of order 2 (ACF(2) p-value $= 0.34$), contributing to the validity of the results.

The results are mainly coherent among the different estimated models. Indeed the Random Effects and the GMM clearly suggests a relevant impact on the province sentiments of the income per capita and of the deprivation index. In particular, individual wealth is negatively correlated with Twitter users feelings and therefore, a lower income per capita on average is associated with more positive sentiments. Conversely, deprivation positively affects the dependent variable suggesting that more optimistic feelings are expressed in areas with higher socio-economic imbalance. This means that regions with the lower sentiment values are not those characterized by higher levels of poverty and disparities, but the wealthier and richer areas of the Italian peninsula. This is confirmed by the geographical distribution of sentiment, which exhibits a strong polarization between North (negative) and South (positive).

Furthermore, we find a strong significance also among epidemiological variables. We can observe how both extra mortality rate and fraction of infected individuals have a negative impact on the local sentiment. Indeed, areas which have been more severely hit by the virus in terms of deaths and number of contagions are also those displaying the most pessimistic feelings. Interestingly, the mortality is not significant in its previous lags, and this might be due to the average duration of COVID-19 disease. This pattern is observed also with the number of infected individuals, even if in a less evident way as in case of the Fixed Effects model the coefficient is still negative and statistically significant.

Finally, we find that the number of intensive and sub-intensive therapy units at hospitals are negatively correlated with the provincial sentiment. This means that areas with a higher capacity in the healthcare infrastructures have been characterized by less optimistic feelings, suggesting that in the North of Italy the local population has perceived the overburden of the sanitary system. On the other hand, inequality, fiscal capacity and the basic reproduction number and its previous lag seem not to be key determinants of Twitter users sentiments.

In order to verify whether the results were mainly driven

by the temporal perspective or instead they were stable across the analysed period, we perform a cross section analysis based on Ordinary Least Squares (OLS). In this case, we average all the observations related to each province for every variable. The results corroborates our previous estimates as again provinces with more pessimistic sentiments are those with higher income per capita and lower deprivation. The only exception is characterized by the significance of the fiscal capacity coefficient, which , however, confirms that the feelings were more negative in richer areas where individuals declare higher income.

In this case, the extra mortality, the contagions and basic reproduction numbers are measured for the month of April and March. Interestingly we observe that while the fraction of infected individuals is significant in both months, the extra mortality rate is statistically relevant only in April, confirming our hypothesis on the delayed impact of the infection when it causes death. Finally, we confirm the negative impact of intensive therapy units on province sentiments.

Overall we find a strong evidence of the relevance of both *socio-economic* and *epidemiological* dimensions, and we observe that feelings are not homogeneous across the Italian peninsula, as if the whole country was exposed to the same emergency, but that results are mainly driven by the local intensity of the COVID-19 epidemic. Moreover, Twitter sentiments are not subject to the fear of enlargement of disparities and poverty as the areas characterized by lower income per capita and deprivation are those externalizing more positive feelings.

Regression at municipality level

As a second validation step, we replicate both panel and cross-section analyses at municipality level. We aim to investigate the robustness of the results through the introduction of additional sources of heterogeneity, and by significantly enlarging the sample size. Results are provided in Table II.

Indeed, results at province level might be influenced by the fact that combining together many municipalities, the differences among Northern and Southern areas might be amplified, and thus driving our results. On the other hand, by leveraging the municipality level, we can observe a higher variability across areas in the same province, as there might be some administrative units having different socio-economic and epidemiological characteristics with respect to other municipalities in the same province. Therefore, such higher level of detail allows to test the stability and validity of our previous considerations. However, for what concerns *epidemiological* variables we only take into account the extra mortality rate, as the other variables are only available at province level.

Overall, the results are coherent with our previous analysis. Indeed, the estimated models show a negative significant impact of income per capita and a positive relationship between deprivation and average municipality sentiment. Moreover, we confirm the relevance of the contemporary extra mortality rate, that has a negative effect on municipality feelings. In this case, we do not estimate the fixed effect model, as the Hausman

	Dependent variable:			
	Municipality Sentiment			
	RE	GMM	(OLS)	(OLS + FE)
Intercept	0.007	−0.009	−0.000	−0.123
	(0.015)	(0.027)	(0.023)	(0.108)
Income pc	−0.031*	−0.304	−0.055**	−0.057*
	(0.018)	(0.244)	(0.028)	(0.034)
Inequality	−0.011	0.054	−0.005	−0.120*
	(0.018)	(0.053)	(0.028)	(0.067)
Deprivation	0.034**	0.156*	0.046*	0.054
	(0.017)	(0.095)	(0.026)	(0.037)
Fiscal Capacity	−0.033*	−0.172*	−0.060**	−0.061*
	(0.019)	(0.106)	(0.029)	(0.035)
Extra Mortality t	−0.032**	−0.064*	−0.040	0.045*
	(0.014)	(0.035)	(0.024)	(0.027)
Extra Mortality t-2	0.002	−0.013	−0.043*	−0.037
	(0.015)	(0.021)	(0.025)	(0.026)
Extra Mortality t-4	0.033**	−0.033*		
	(0.014)	(0.017)		
Observations	5566	5566	1855	1855
Adjusted R^2	0.134	0.275	0.216	0.396
Hausman Test	0.18			
Hansen Test		0.98		
Arellano Bond test order 1		5.92^{-9}		
Arellano Bond test order 2		0.47		
Breusch-Pagan test	2.05^{-12}			
Note:	*p<0.1; **p<0.05; ***p<0.01			

TABLE II
RESULTS OF REGRESSION MODEL(S) APPLIED AT MUNICIPALITY LEVEL. IN THE OLS REGRESSION MORTALITY T REFERS TO THE EXTRA MORTALITY IN APRIL, WHILE THE PREVIOUS LAG TO THE CORRESPONDENT VARIABLE IN MARCH

test accepts the null hypothesis that the random effect is the preferred specification (p-value $= 0.18$), while the Hansen test is in favour of the validity of instruments. In this case, for the cross-section analysis, in addition to the standard OLS regression we perform an OLS with regional fixed effects to avoid the risk that results are driven only by data measured in a specific area of Italy. We confirm previous results, and we find again that the mortality is more significant in April, rather than in March, suggesting a delayed effect on Twitter online conversations. At this level of analysis, the main exception is constituted by the fiscal capacity variable, which becomes significant in all the estimated models reinforcing the idea that negative messages were widespread in wealthier areas.

This analysis corroborates our previous findings and confirms the relevance of both socio-economic and epidemiological variables at driving Twitter emotions and feelings.

CONCLUSIONS

In this paper we proposed an analysis of the determinants of online feelings and sentiments expressed by Italian users on Twitter during the COVID-19 pandemic. Leveraging over 1.7 M tweets related to the virus and collected during the crisis period, we applied a set of panel and cross-section regressions to understand whether epidemiological and socio-economic variables had an impact on the the feelings conveyed on the platform.

In particular, we find evidence of a strong geographical heterogeneity, which reflects different emotions communicated by citizens in the North and South of Italy. More in detail, a severe pessimism is experienced in areas where the virus had

a stronger impact in terms of contagions, and extra number of deaths with respect to past years. Furthermore, we find a significant negative relationship between the local feelings and the number of intensive and sub-intensive therapy units in Italy, suggesting that individuals living in most affected regions, such as Lombardy and Piemonte, have perceived an overburden of healthcare infrastructures during the critical phases of the pandemic despite the higher number of units available. On the other hand, areas exhibiting the most positive emotions are those characterized by lower wealth and higher levels of deprivation.

Overall, our estimates, which are very stable among different models and both at province and municipality level, show that negative sentiments were mainly driven by the ongoing sanitary emergency. Furthermore, the analysis of socio-economic variables seems to suggest that feelings could be negatively influenced by the fear of deterioration of the socio-economic position of the wealthier areas.

Moreover, our results, provide relevant contribution from a methodological point of view providing relevant insights for further research. Indeed, the field of the analysis of socio-economic variables with respect to pandemics is at an embryonic stage. Lastly, this work could provide significant clues to policy makers on the impact of the pandemic on citizens' sentiments. These can be exploited in order to design targeted and effective communication preventing the heterogeneous spread of feelings across Italy.

As a future research direction, we plan to perform a comparative analysis of online sentiment across multiple countries, and extending the period of observation in order to capture the impact of the release of mobility restrictions on online conversations and to strengthen the representativeness of our analysis.

REFERENCES

[1] L. M. Aiello, D. Quercia, K. Zhou, M. Constantinides, S. Šćepanović, and S. Joglekar. How epidemic psychology works on social media: Evolution of responses to the covid-19 pandemic. *arXiv preprint arXiv:2007.13169*, 2020.

[2] M. Arellano and S. Bond. Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies*, 58(2):277–297, 1991.

[3] A. I. Bento, T. Nguyen, C. Wing, F. Lozano-Rojas, Y.-Y. Ahn, and K. Simon. Evidence from internet search data shows information-seeking responses to news of local covid-19 cases. *Proceedings of the National Academy of Sciences*, 117(21):11220–11222, 2020.

[4] G. Bonaccorsi, F. Pierri, M. Cinelli, F. Porcelli, A. Galeazzi, A. Flori, A. L. Schmidth, C. M. Valensise, A. Scala, W. Quattrociocchi, and F. Pammolli. Economic and social consequences of human mobility restrictions under covid-19. *Proceedings of the National Academy of Sciences*, 2020.

[5] C. G. Bowsher. On testing overidentifying restrictions in dynamic panel data models. *Economics letters*, 77(2):211–220, 2002.

[6] C. O. Buckee, S. Balsari, J. Chan, M. Crosas, F. Dominici, U. Gasser, Y. H. Grad, B. Grenfell, M. E. Halloran, M. U. G. Kraemer, M. Lipsitch, C. J. E. Metcalf, L. A. Meyers, T. A. Perkins, M. Santillana, S. V. Scarpino, C. Viboud, A. Wesolowski, and A. Schroeder. Aggregated mobility data could help fight covid-19. *Science*, 2020.

[7] G. Bwire, A. Munier, I. Ouedraogo, L. Heyerdahl, H. Komakech, A. Kagirita, R. Wood, R. Mhlanga, B. Njanpop-Lafourcade, M. Malimbo, et al. Epidemiology of cholera outbreaks and socio-economic characteristics of the communities in the fishing villages of uganda: 2011-2015. *PLoS neglected tropical diseases*, 11(3):e0005407, 2017.

[8] C. Cerami, G. C. Santi, C. Galandra, A. Dodich, S. F. Cappa, T. Vecchi, and C. Crespi. Covid-19 outbreak in italy: Are we ready for the psychosocial and economic crisis? baseline findings from the longitudinal psycovid study. *Baseline Findings from the Longitudinal PsyCovid Study (4/2/2020)*, 2020.

[9] C. Chew and G. Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[10] I. C.-H. Fung, Z. T. H. Tse, C.-N. Cheung, A. S. Miu, and K.-W. Fu. Ebola and the social media. 2014.

[11] N. Islam. What have we learnt from the convergence debate? *Journal of economic surveys*, 17(3):309–362, 2003.

[12] C. J. McInnes and H. Hornmoen. 'add twitter and stir': The use of twitter by public authorities in norway and uk during the 2014-15 ebola outbreak. *Observatorio (OBS*)*, 12(2):23–46, 2018.

[13] S. Pelosi. Sentita and doxa: Italian databases and tools for sentiment analysis purposes. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 226–231, 2015.

[14] D. Prabhakar Kaila, D. A. Prasad, et al. Informational flow on twitter–corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(3), 2020.

[15] M. M. Rahman, G. Ali, X. J. Li, K. C. Paul, P. H. Chong, et al. Twitter and census data analytics to explore socioeconomic factors for post-covid-19 reopening sentiment. *Nawaz and Li, Xue Jun and Paul, Kamal Chandra and Chong, Peter HJ, Twitter and Census Data Analytics to Explore Socioeconomic Factors for Post-COVID-19 Reopening Sentiment (June 30, 2020)*, 2020.

[16] D. Roodman. How to do xtabond2: An introduction to difference and system gmm in stata. *The stata journal*, 9(1):86–136, 2009.

[17] J. Samuel, G. Ali, M. Rahman, E. Esawi, Y. Samuel, et al. Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, 11(6):314, 2020.

[18] S. Towers, S. Afzal, G. Bernal, N. Bliss, S. Brown, B. Espinoza, J. Jackson, J. Judson-Garcia, M. Khan, M. Lin, et al. Mass media and the contagion of fear: the case of ebola in america. *PloS one*, 10(6):e0129179, 2015.

[19] H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

[20] F. Windmeijer. A finite sample correction for the variance of linear efficient two-step gmm estimators. *Journal of econometrics*, 126(1):25–51, 2005.