COVID-19 Sentiment Analysis Based on Tweets

Valerio La Gatta
University of Naples Federico II

Vincenzo Moscato
University of Naples Federico II

Marco Postiglione
University of Naples Federico II

Giancarlo Sperlí
University of Naples Federico II

Abstract—**This work investigates how individuals in Italy perceived the COVID outbreak and all its implications in real-life. We unveil the most discussed narratives on Twitter and measure how users' interests, sentiment and emotions have evolved over time and across the several aspects of the pandemic. Our analysis shows that while the overall sentiment is negative, Italians have shown upbeat responses to the pandemic, especially in regards to the vaccination campaign. The emotion analysis reveals that while *fear* progressively decreased after the first wave of the pandemic, the overall *anger* has remained constant but gradually turned into various narratives.**

■ **AS OF 31 JANUARY 2020**, the World Health Organisation (WHO) declared a state of world-wide emergency following the spread, in a short space of time, of a new infection caused by a new coronavirus as an etiological agent, scientifically identified as SARS-CoV-2 [1]. The resulting pandemic affected all aspects of our lives and Italy has been one of the first countries in Europe which dealt with the virus. This work analyses the change in perception of COVID-19 in Italy to provide answers to various societal concerns, such as how the perception of the pandemic and the opinions towards vaccines, governments'

measures and lockdown policies have evolved during the pandemic. In particular, we analyze the reactions expressed by users on social networks towards the virus and all its related events, with the aim of understanding what people's real perception was about the virus and all its implications. Unlike recent studies [2], [3], [4], whose primary objective is to classify English tweets into three sentiment classes (i.e. positive, neutral and negative), we focus on data from Italian users and enrich our analysis with a clustering phase which allows us to separately focus on different topics. Furthermore, we provide an analysis of trends in emotions throughout the pandemic.

Our main contributions are as follows:

- We collect a million-scale Italian Twitter dataset regarding the COVID-19 pandemic
- Our analysis covers the whole period of the pandemic, also unveiling how users' interests and sentiment have evolved, and sometimes changed, in time and across different narratives (e.g. economy, vaccine, restriction measures);
- We enhance the sentiment analysis with emotion classification to understand the specific moods and concerns users exhibit towards the several decisions taken to shrink down the sanitary and economic consequences of the pandemic.

Methodology

In this work, we first gathered Italian tweets over the entire global pandemic horizon and subsequently grouped them based on their embedded representations. Finally, for each cluster, we conduct sentiment analysis as well as emotion analysis [5], [6].

We retrieved the tweets used in our analyses with *snscrape*[1] service scraper. We filtered Italian tweets posted from January 2020 to February 2022 containing at least one of these words: "covid", "covid-19", "covid19", "coronavirus", "pandemia" and "nuovo_coronavirus". The selection of the hashtags follows the collection procedure shown in [2]. We removed tweets without text values and duplicate texts. Tweets have undergone additional preprocessing to produce lowercase text devoid of outside links and other special characters — the hash character being the only exception.

We used the resulting tweets corpus to train a Word2Vec model, which maps words in a space where semantically-similar words — which often appear in the same contexts — are close to each other. Specifically, we trained a Skip-Gram model, which is specifically designed to take into account the context in which each word appears.

After having computed tweets embeddings, we applied the KMeans algorithm to find 10 dimensions of analysis according to different (sub-)topics related to the pandemic (e.g. vaccines, government interventions).

[1] https://github.com/JustAnotherArchivist/snscrape

For each cluster, we manually labelled 1000 randomly sampled tweets for sentiment and emotions analysis and then fine-tuned a BERT-based transformer architecture (i.e. Feel-it [7]) on the resulting corpus. Finally, we used the resulting model to provide labels to the entire corpus.

Results

In this section we discuss about the results of our methodological flow on a tweets dataset, collected during the Covid pandemia, whose details are discussed in the next section. In particular, we, firstly, perform a clustering strategy on all tweets in order to unveil the main topics over the entire analysis period, and we, successively, investigate people reaction in terms of sentiment and emotion.

COVID-19 Italian tweets dataset

The collected dataset consists of 2.47 million tweets posted by 183,097 unique accounts between January 9, 2020, and February 1, 2022. In addition, 471k (19.1%) tweets are geolocalized and most of them were published in Lazio (78 810), in Lombardy (60 931) and in Campania (28 010). As shown in Figure 1, the number of relevant tweets explodes with the first cases of infection in Italy and gradually decreases during the first lock-down period. Then, the trend is relatively stable, with small spikes that correspond to major announcements regarding vaccine research, vaccine authorization and government decrees.

In addition to the most common general hashtags that we used to retrieve tweets, such as #CORONAVIRUS and #COVID19, we observed a high proportion of hashtags related to the first wave (#IORESTOACASA, #LOCKDOWN, #PANDEMIA, #DPCM), Green Pass (#GREENPASS), vaccines (#VACCINO, #NOVAX) and Italian politics (#CONTE, #GOVERNO, #SALVINI).

Clusters: analysis dimensions

By examining the most popular terms and hashtags used by the users in each of the 10 clusters retrieved with KMeans, we were able to determine the semantics of each cluster. Specifically, clusters refer to the *First wave* (1.3M), *Vaccination campaign* (80.4k), *National news* (58.1k), *Collateral effects* (63.8k), *Sports news* (34.5k), *Restrictive measures* (76.4k), *Green pass*
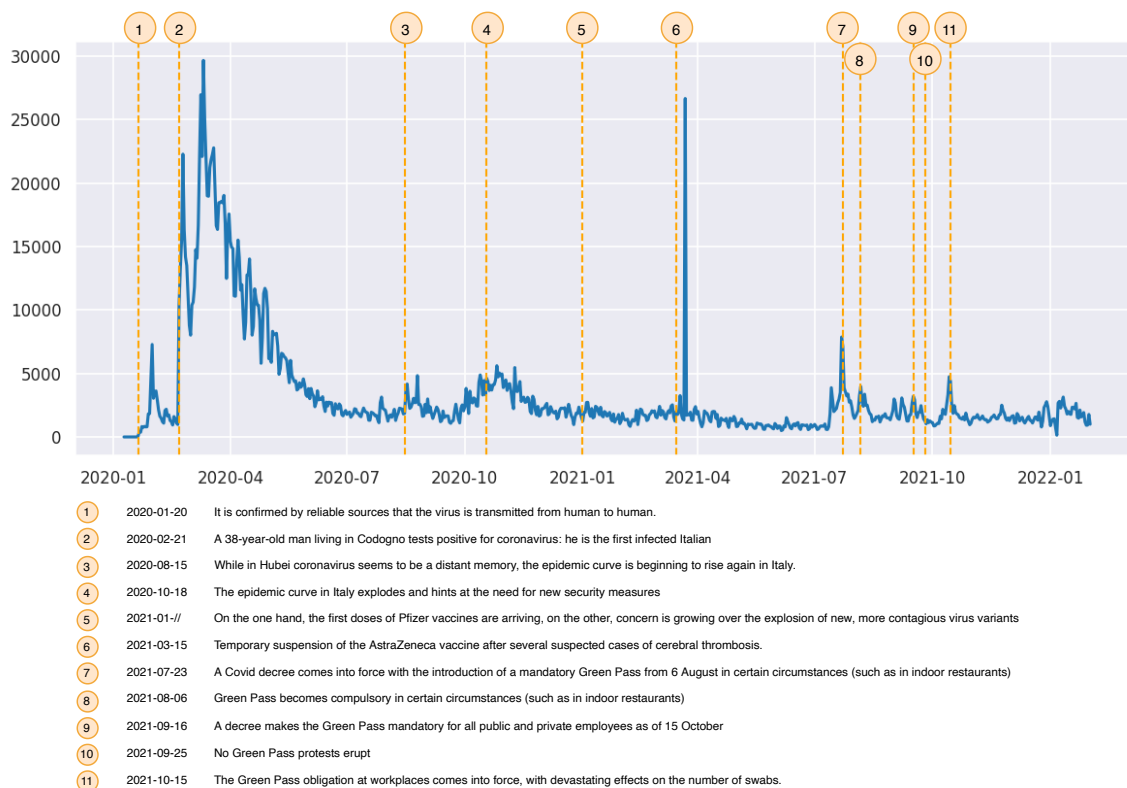
Figure 1: Number of tweets over time in the collected dataset. The times of events related to coronavirus in Italy are marked by dashed vertical lines. Further descriptions of the items are provided in the legend below the chart.

1	2020-01-20	It is confirmed by reliable sources that the virus is transmitted from human to human.
2	2020-02-21	A 38-year-old man living in Codogno tests positive for coronavirus: he is the first infected Italian
3	2020-08-15	While in Hubei coronavirus seems to be a distant memory, the epidemic curve is beginning to rise again in Italy.
4	2020-10-18	The epidemic curve in Italy explodes and hints at the need for new security measures
5	2021-01-//	On the one hand, the first doses of Pfizer vaccines are arriving, on the other, concern is growing over the explosion of new, more contagious virus variants
6	2021-03-15	Temporary suspension of the AstraZeneca vaccine after several suspected cases of cerebral thrombosis.
7	2021-07-23	A Covid decree comes into force with the introduction of a mandatory Green Pass from 6 August in certain circumstances (such as in indoor restaurants)
8	2021-08-06	Green Pass becomes compulsory in certain circumstances (such as in indoor restaurants)
9	2021-09-16	A decree makes the Green Pass mandatory for all public and private employees as of 15 October
10	2021-09-25	No Green Pass protests erupt
11	2021-10-15	The Green Pass obligation at workplaces comes into force, with devastating effects on the number of swabs.

& no-vax (318k), *Government* (138k), *Regional news* (153.9k) and *International news* (84.3k).

People reactions

Sentiment analysis   Fig. 2 shows the trends of percentage of positive and negative sentiments within each cluster. Overall, we can observe that negative sentiment is dominant in most of the clusters, with the exceptions of the "Vaccination Campaign" and the " Sports" clusters. Indeed, Italian people have always favoured the vaccines and the vaccination campaign. This result is also supported by the positive trend of negative sentiment from the second quarter of 2021 in the "Green pass and no-vax" cluster, i.e. there has been increasing skepticism towards people supporting the no-vax movement.

Additionally, we can see that after the initial negative peak, positive and negative emotions in the *First Wave* and *Restrictive measures* clusters align. The negative peak corresponds to the period when the WHO had not yet proclaimed the out-break to be a pandemic. People then recognized (and implicitly accepted) the reasons behind those measures as soon as the virus began to spread over the country, resulting in the first hundreds of deaths.

Emotions analysis   In Figure 3 we show the trends of fear, anger, joy and sadness emotions over the set of clusters under analysis. Plots have been generated by considering, for each emotion, the percentage of tweets where it is predominant over each day during the pandemic. As expected, the prevailing emotions experienced by people tweeting during the pandemic were *fear* and *anger*. During the first days, we can observe a peak in *fear* in all the clusters related to the first wave and Covid-19 information (i.e. First Wave, Collateral Effects, Regional News, International News). After that, we see a significant predominance of *anger* towards Restrictive Measures, Government and National News. While we can observe a constant decreasing trend

Figure 2: Sentiment trends across clusters

of *fear*, suggesting that the illness started to not be considered as a major threat anymore, the same cannot be said of *anger*, since we see a gradual increment in all the clusters regarding the government, news and restrictive measures. It is interesting to notice a major change in the "Green pass and no-vax" clusters, where we can observe a predominance of *anger* over *fear* from July 2021 on, which corresponds to the period when the decree from the government came into force with the introduction of the mandatory Green pass. Interestingly, the National News *fear/anger* trend is opposite to International and Regional news, where the prevailing emotion is *fear*. This is probably due to the fact that national news usually is about government choices, which have usually been strongly discussed by Twitter users. As expected, *joy* is the less prevalent emotion across all the clusters with the exceptions of Sports, where emotions are overlapped, and Vac-

cination Campaign, where we can observe an increase and a peak from the 2021 summer.

Conclusion

In this work, we presented a million-scale Twitter dataset covering the entire period during which COVID-19 spread in Italy. Our analyses revealed people's attitudes and moods towards all narratives related to the pandemic. Our analysis shows Italian people have always looked favourably on vaccinations, despite showing a negative sentiment towards restrictive measures. While fear was the predominant emotion during the first months of the pandemic, it has been then gradually replaced by the anger towards the no-vax campaigns, green pass and other overextended restrictive measures.

◼ REFERENCES

1. M. Cevik et al., "Virology, transmission, and pathogenesis of SARS-CoV-2," *BMJ*, vol. 371, 2020.

Figure 3: Emotion trends across clusters

2. U. Naseem et al., "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, 2021, pp. 1003–1015.

3. M. Qorib et al., "Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset," *Expert Systems with Applications*, vol. 212, 2023, p. 118715, ISSN 0957-4174.

4. D. Mahata et al., "Detecting Personal Intake of Medicine from Twitter," *IEEE Intelligent Systems*, vol. 33, no. 4, 2018, pp. 87–95.

5. R. Mao et al., "The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection," *IEEE Transactions on Affective Computing*, 2022, pp. 1–11.

6. E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, 2016, pp. 102–107.

7. F. Bianchi et al., "FEEL-IT: Emotion and Sentiment Classification for the Italian Language," *WASSA*, 2021.

Valerio La Gatta is a PhD student at the University of Naples Federico II. His research interests are focused on Social Network Analysis and Graph Data Mining. Contact him at valerio.lagatta@unina.it.

Vincenzo Moscato is an Associate Professor at the University of Naples "Federico II". His research activities lay in the area of Multimedia, AI and Social Network Analysis. Contact him at vincenzo.moscato@unina.it.

Marco Postiglione is a PhD student at the University of Naples Federico II. His research interests are focused on Graph Data Mining and Natural Language Processing. Contact him at marco.postiglione@unina.it.

Giancarlo Sperlí is an assistant professor at the University of Naples Federico II. His main research interests are in the area of Multimedia and Social Networks Analysis. Contact him at giancarlo.sperli@unina.it.