BANCA D'ITALIA

EUROSISTEMA

# Questioni di Economia e Finanza

(Occasional Papers)

## Textual analysis of a Twitter corpus during the Covid-19 pandemics

by Valerio Astuti, Marta Crispino, Marco Langiulli and Juri Marcucci

June 2022

Number 692

BANCA D'ITALIA
EUROSISTEMA

# Questioni di Economia e Finanza

(Occasional Papers)

## Textual analysis of a Twitter corpus during the Covid-19 pandemics

by Valerio Astuti, Marta Crispino, Marco Langiulli and Juri Marcucci

Number 692 – June 2022

*The series* Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

*The* Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

*The series is available online at www.bancaditalia.it .*

ISSN 1972-6627 (print)
ISSN 1972-6643 (online)

*Printed by the Printing and Publishing Division of the Bank of Italy*

# TEXTUAL ANALYSIS OF A TWITTER CORPUS DURING THE COVID-19 PANDEMICS

by Valerio Astuti*, Marta Crispino*, Marco Langiulli* and Juri Marcucci*

## Abstract

Text data gathered from social media are extremely up-to-date and have a great potential value for economic research. At the same time, they pose some challenges, as they require different statistical methods from the ones used for traditional data. The aim of this paper is to give a critical overview of three of the most common techniques used to extract information from text data: topic modelling, word embedding and sentiment analysis. We apply these methodologies to data collected from Twitter during the COVID-19 pandemic to investigate the influence the pandemic had on the Italian Twitter community and to discover the topics most actively discussed on the platform. Using these techniques of automated textual analysis, we are able to make inferences about the most important subjects covered over time and build real-time daily indicators of the sentiment expressed on this platform.

**JEL Classification**: C55, C14, C81, L82.
**Keywords**: text as data, Twitter, big data, sentiment, Covid-19, topic analysis, word embedding.
**DOI:** 10.32057/0.QEF.2022.0692

## Contents



* Bank of Italy, Directorate General for Economics, Statistics and Research.

# 1 Introduction

This paper provides a critical overview of some of the most common techniques for the analysis of textual data such as those sourced from social media. The goal is to equip the reader with useful notions on such techniques, readily usable without a background in Big data analytics, rather than to cover the vast literature on Natural Language Processing (NLP) methods. In addition, we provide a demonstration of the potential value of social media data and of these methods for economic research using an extensive dataset collected from Twitter starting in December 2019, containing tweets related to the COVID-19 pandemic.[1]

In comparison to traditional sources - more accurate but available only with a significant delay and at higher costs - data gathered from social media are more timely and much more extensive. For these reasons they can be a useful integration to statistics based on traditional data to assess the impact of sudden changes to the economic situation or external shocks on macroeconomic variables.[2]

Textual data is one of the most multifaceted forms of information available, and as such there are several types of analysis we can perform on a given set of documents: from the purely linguistic structure of the text, to the summary of its content, to the correlation between the analyzed texts and the happening of a given event. In this paper we are not interested in the prediction of any particular phenomenon, so we will focus on *unsupervised* techniques, that is, descriptions of the features of the texts without reference to external information.[3]

Almost all these techniques consist in algorithms for dimensional reduction, which imply projecting each document into a space with a limited number of dimensions. The algorithms are characterized by the detail of the description they attain, the different aspects of the texts they analyze, and their range of applicability.

Whenever the content interpretability is important, each document has to be summarized in a few features (between tens and hundreds), each of which can be

[1]The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy and/or the Eurosystem. We thank Marco De Leonardis, Sabina Marchetti, Riccardo Maria Nusca and Filippo Quarta for the assistance throughout the process of data collection and analysis. We are also grateful to all the participants to an internal Bank of Italy seminar for fruitful discussions and suggestions.

[2]It is important to assess whether data gathered from social media are representative of the population of interest (which, depending on the application, could be for instance the overall country population or the population of users of a given social media).

[3]In the unsupervised learning setting we are just given output data and the goal is to discover *interesting structure* in the data. This is sometimes called knowledge discovery (Murphy, 2012).

assigned a definite meaning. These techniques are usually grouped under the name of *topic analysis*.

However, if we are not interested in the specific contents of our documents, we can reduce the number of features to just one, and reveal the intensity of expression of this feature in each document with a single index. An example is the tone of the documents which might be positive, negative or neutral as in the so-called *sentiment analysis*.

A different class of models, useful when human interpretability is less important, is capable of capturing the context and meaning of words in the documents, by accommodating many more aspects of a text (at the cost of a greater computational burden). These models are called *word embeddings*, because they embed every word in a space with hundreds of dimensions, each dimension encoding some aspect or meaning (*a priori* unknown to the researcher) of that word. Each of these groups of techniques has its peculiarities, advantages and issues, and each of them can be the most suitable for a given task.

After describing the Twitter data and their related issues in Section 2, in Section 3 we will describe the models and techniques of NLP which we employ in the empirical analysis presented in Section 4.

In Section 3.1 we will describe a specific technique to perform topic analysis. In general, topic analysis algorithms are created to extract information from a collection of documents (corpus) by identifying common topics. Their main task is therefore to discover a set of latent topics that best describe the corpus at hand. The topic modelling algorithm we describe, called Structural Topic Model, STM (Roberts et al., 2016), exploits a generative model for every document, assuming that each of them is described by few topics, and that each topic is a probability distribution over a set of words. The intuition behind the STM representation is that (i) documents can be considered similar if they have similar content (in terms of words' counts), and that (ii) from the content we can learn something about the subject(s) contained in the document. Differently from simpler algorithms[4] STM has the ability to exploit metadata about documents (for instance, their date of publication or the geographical location of the author) to improve the assignment of words to the latent topics.

Topic modeling techniques usually rely on representations that do not take into

[4]Such as the Latent Dirichlet Allocation, LDA (Blei et al., 2003).

account the relations between words. For this reason, they are unable to faithfully capture information deriving from the context of a word, or to fully reproduce the semantics of the corpus (they would consider the two sentences "I like Italian food, not Chinese one" and "I like Chinese food, not Italian one" as identical, even if the meaning is clearly different for a human reader). To cover this blind spot we will describe in Section 3.2 a *word embedding* algorithm, which sacrifices simplicity to capture deeper aspects of the texts under study. The scope of a word embedding is to reliably capture the semantic structure of the texts, rather than to synthesize a document or to describe it in terms of a small number of variables. It is still a dimensionality reduction algorithm, but every word is represented as a vector with hundreds of dimensions. The word vectors are oriented in their space such that two vectors are close if the meanings of the associated words are similar. While this representation is not optimized for a particular task, we will see that it allows the extraction of many features of interest from the texts.

In Section 3.3 we will describe two methods to perform sentiment analysis, one of the most straightforward NLP techniques, whose aim is to automatically extract the mood, or tone, expressed in a document. This operation is (usually) easy for a human reader, but less trivial for an algorithm. The simplest approach we adopt is the so-called vocabulary-based (or rule-based) method, which first exploits a rule assigning a sentiment score to every word in a pre-assigned vocabulary, and second obtains the sentiment index of each document by simply aggregating the scores of every word in it. The second approach is based on the word embedding representation. By taking into account the context of words appearing in each document, this method aims at improving the accuracy of the sentiment index associated to each document.

Equipped with the above techniques of text analysis, in Section 4 we will use data collected from Twitter during the COVID-19 pandemic to make inferences about the most relevant subjects covered over time (topic analysis), and to build real-time daily indicators of the sentiment expressed on this platform (sentiment analysis).

Social media data are extensively used for academic research in many different fields, for instance Environmental Science (Moore et al., 2019), Political Science (Beauchamp, 2017), Sociology (Ahmed et al., 2020), Finance (Renault, 2017) and Economics (Angelico et al., 2022, Levy, 2021). On Twitter, registered users can interact posting messages, called "*tweets*", containing a short text, photos, links and

videos. Differently from other social networks, like Facebook, academic researchers can massively download Twitter data using Application Programming Interfaces (API).[5] Hence, since the outbreak of COVID-19, Twitter data were widely exploited to investigate a variety of research issues. For instance, Xue et al. (2020) focus on tweets posted in the early stages of the outbreak to investigate topics and sentiments expressed by the users. Sciandra (2020) analyses the Italian social media communication about COVID-19 through a Twitter dataset collected over two months. Porcher and Renault (2021) use geotagged Twitter data and mobility data to create a daily index of social distancing in the US. Altig et al. (2020) inspect the changes in economic uncertainty for the US and UK before and during the COVID-19 pandemic exploiting - among others - data collected on Twitter. Yaqub (2020) studies the impact of COVID-19 crisis on tweets posted by politicians evaluating the correlation between the number of COVID-19 daily cases in the United States and the sentiment of President Trump tweets.

In comparison to the aforementioned papers, we focus on a longer time period and study a bigger number of tweets, drawing from an extensive dataset collected by the Bank of Italy from December 2019, and containing around 9 million tweets. We investigate how Twitter users reacted to the sequence of events characterizing the period under study, such as the different phases of the pandemic and the evolution of the containment measures enforced by the authorities.

Briefly summarising our main results, we find marked spikes in the number of tweets related to the pandemic, corresponding to the outbreak in Italy and the worsening of the public health situation later in 2020; these spikes confirm that users' mood is extremely responsive to changes in the overall situation and to updates on the evolution of the disease. We find 15 relevant topics in the Twitter discourse over the period under study, describing the public health situation, the restriction policies used to reduce the spread of the virus, economic issues and medical details related to the illness. In addition we extract a daily series for the general sentiment with the two independent methods described above, conveying the same conclusions: the mood of the conversation underwent a sudden dip corresponding to the outbreak of the pandemic and the increasing evidence of the necessity of lockdown measures. Af-

[5]As discussed in Hino and Fahey (2019) quality and representativeness of the collected data are strictly connected with the implemented sampling strategy and the limitations of the available API, thus the data collection process is a crucial step of the analysis.

ter the initial massive plunge, the mood partially recovered, to decrease again in a milder way with the worsening of the situation in late 2020. Finally we construct daily indicators for the attention toward particular themes like "jobs" and "vaccine".

# 2 Data

A description of the data gathering procedure is included in Section 2.1. We show some descriptive statistics on our corpus of tweets and we discuss about Twitter data drawbacks and possible solutions in Section 2.2.

## 2.1 Data collection

The COVID-19 pandemic has been in the limelight on TV talks, newspapers and other media since its first outbreak, with discussions ranging from the tragic death toll and the impact on national economy, to the skepticism of some people questioning the real menace posed by the virus. COVID-19 has been a hot topic also in the social media debate, where users posted their comments and criticisms of news about this issue. We use tweets collected from Twitter social media to assess the sentiment and the topics related to the COVID-19 discussion. Using a private Application Programming Interface (API) we filter all tweets in Italian language containing one or more of the following Italian keywords (English translation in parentheses) related to the COVID-19 pandemic and to the symptoms specific for this disease:[6]

*coronavirus, covid-19, covid19, covid2019, febbre (fever), tosse (cough), difficoltà respiratorie (respiratory difficulties), difficoltà respiratoria (respiratory difficulty), crisi respiratoria (respiratory crisis), difficoltà di respiro (difficulty in breathing), mancanza di respiro (shortness of breath), respiro affannoso (wheezing), respiro corto (shortness of breath), crisi respiratorie (respiratory crises), problemi respiratori (breathing problems), problema respiratorio (breathing problem), brividi (chills), polmonite (pneumonia), polmoniti (pneumonia), dolori (pains), dispnea (dyspnea), mal di testa (headache), respirare (to breath), difficoltà a respirare (difficulty in breathing) tossire (to cough), respiro (breath), dolore al petto (chest pain), dolori muscolari*

[6]We exclude tweets in other languages because we want to focus our analysis on Italian users and also avoid complexity arising from having a vocabulary with multiple languages. It's common practice to include as keywords both terms regarding the topic of interest and related words. Our keywords list contains common symptoms in order to collect also tweets referred to COVID-19 but not explicitly mentioning the name of the disease.

*(muscular pain).*

Using a private API service we manage to go beyond the one-week old limit, typical when using a public Twitter API to retrieve tweets. Applying the previously mentioned filter, we gather tweets sent starting from December 2019, in order to study the weeks before the outbreak of the virus, the different waves and phases of the crisis and the beginning of the vaccination campaign which is still ongoing at the time of writing.

The initial dataset is a corpus of about 24 million of tweets sent between 01/12/2019 and 04/04/2021, containing at least one of the aforementioned keywords, posted by 1.37 million unique users (with an average of 18 tweets per user). The sample contains the body of the posted tweet and a broad set of metadata related to the tweet and the author: date, time, retweet identifier, hashtags, and detailed information on the user, including geographic location.

The impact of COVID-19 was extremely heterogeneous over time and space, especially in the first months when cases were extremely concentrated in specific parts of the country. We include the date of the tweet and the geographic location of the author for the purpose of considering these additional features in our analysis. However, only part of the users are willing to share information on their geographic location; accordingly, this metadata is available only for a sub-sample of our dataset.[7]

## 2.2 Data preparation and descriptive statistics

Pre-processing is of utmost importance before textual analysis, especially for researchers working with data gathered from social media. People don't speak on the web as they would in normal life and language used for tweets and posts is noticeably different from the traditional one used in books or newspapers. Users often include hashtags and emojis in their posts to gain visibility and express feelings and they use specific acronyms and abbreviations which are typical of social network language; typos are also quite common. We use a filter to keep tweets in Italian language and exclude any duplicates or retweets. The total volume of data greatly reduces, and we are left with 9 million tweets in total, posted by 768K unique users, distributed in

[7]Approximately 40% of total tweets are shared by users posting certainly from Italy and the largest part of them provides further information on the region.

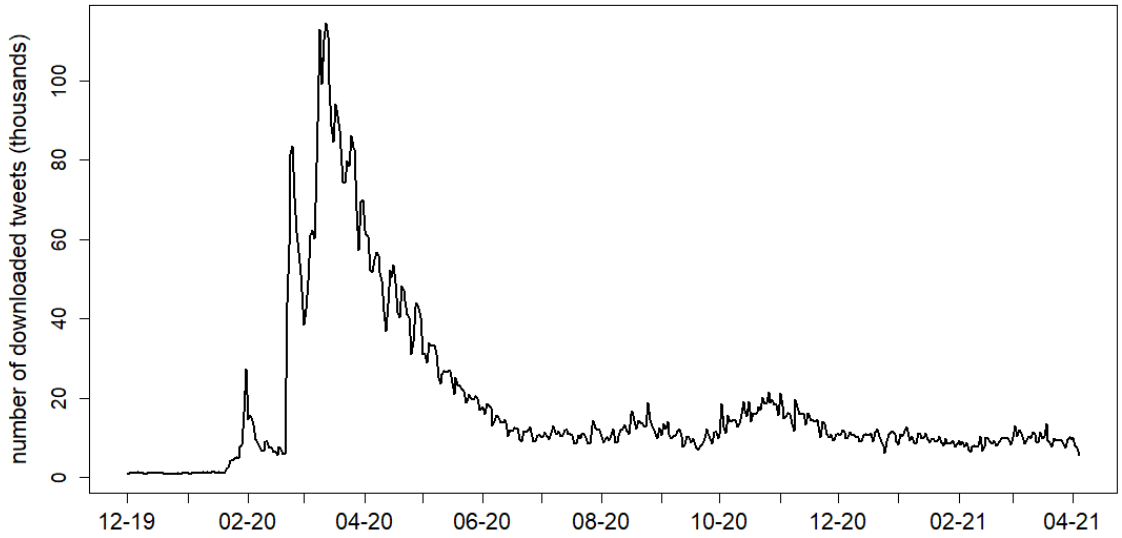


Figure 1: The number of downloaded tweets per day for the period December 2019, April 2021.

time as shown in Figure 1.

Table 1 shows the numbers of tweets containing some of the keywords used to collect them. Most tweets were selected because they contain the name of the disease (*coronavirus, covid19, covid-19, covid2019*) but including symptoms-related keywords we gather over 1 million additional tweets.[8]

[8]We implement a data preparation pipeline in order to achieve an appropriate dataset to apply text mining algorithms. The pipeline is applied on every tweet and works as follows: we convert text to lowercase and remove hashtag symbols (the most frequent ones, excluding those containing the string "covid" or "coronavirus", are reported in Figure 2), numbers, mentions, emojis, URLs, punctuation and special characters.

| **Keyword** | **Volume** (thousands) |
|---|---|
| coronavirus | 4337 |
| covid19 | 1335 |
| covid-19 | 658 |
| brividi (chills) | 274 |
| febbre (fever) | 199 |
| respiro (breath) | 169 |
| covid2019 | 143 |
| mal di testa (headache) | 143 |
| respirare (to breath) | 129 |
| dolori (pains) | 87 |
| tosse (cough) | 63 |
| polmonite (pneumonia) | 53 |

Table 1: The most frequent keywords in the collected data (frequency over 50K).


Figure 2: word cloud of the most frequent hashtags, excluding the ones containing the string "covid" or "coronavirus".

Despite the exceptional size of the dataset, the sample is unlikely to be representative of the Italian population. Data gathered from Twitter and other social media have an intrinsic bias related to the fact that the community on this social network is like a self-selected sample. The issue is then twofold: a sample extracted from Twitter using a filter on keywords can be considered neither representative of Italian population, nor of the topics discussed by the whole national Twitter users' community. The choice of subject-specific keywords unavoidably introduces a *selection bias*.

The sample selection bias with respect to the general population is a common problem and can be explained considering that the Twitter population has a distribution over demographic features (age, gender, education, income etc...) very different from the Italian population, and that not all Twitter users express their opinion on a given subject with the same intensity. According to online sources, the typical Twitter user is male, with a share of 61% (Hootsuite & We Are Social, 2020), he is on average 32 years old (`https://www.oberlo.it/blog/statistiche-twitter`) and well educated. To put it differently: the collected tweets are not posted by individuals randomly chosen from the Italian population, which in principle jeopardizes the possibility of externally validating any inferential conclusion from the sample. To overcome this bias, post-stratification or Bayesian methods may come at hand. However, post-stratification is a difficult task, not only because Twitter accounts cannot

be uniquely associated with individuals - and some accounts are more active than others - but also because the majority of the users do not share relevant personal information. One option is to probabilistically infer some basic information about (a fraction of) the users by applying text analysis techniques on their names, surnames, biographic description (when present), and location (at the municipal level). We may then post-stratify based on surveys' results. This approach is beyond the scope of this study and it will be discussed in future research.

The second aspect of the problem is related to a normalization issue: selecting only tweets speaking about the COVID-19 pandemic, we ignore the information on how prevalent was this topic during the period under study. In principle, the set of tweets containing words related to a given topic could be a negligible part of the total, so that even a perfect knowledge of this set would provide few interesting information (this is less upsetting in terms of the restriction to the Twitter population: the sheer number of Twitter users implies that the sample has *some* relevance in terms of the general population). In other words, the normalization problem arises because we observe a sample of tweets that is not randomly chosen but extracted in order to be related to COVID-19 issues (that is, selected with the keywords listed above). This implies that, in general, any indicator built from these tweets - for instance the sentiment score - is not representative of the overall mood of Italian Twitter community.

In practice, the knowledge of the total daily number of tweets would be sufficient to solve this problem, that is, to assess the relative importance of COVID-19 related tweets (over time) with respect to the Italian Twitter community. Unfortunately, this information is not available from Twitter, nor it is feasible to download all posted tweets. Being aware that there is no trivial solution to this problem, here we propose an estimation of the relevance of COVID-19 subject over time (with respect to the population of general tweets) using the ratio of the number of COVID-19-related tweets to the size of an independent sample, selected with different keywords. The motivation is the following: if we are able to collect all the tweets containing a given word $w$ which is not related to the COVID-19 subject, we can count the tweets containing both $w$ and any of the keywords $k$ used to identify tweets related to COVID-19; denoting by $N(w)$ the number of tweets containing the word $w$, and by $N(\{k\}\ \&\ w)$ the number of tweets containing both the word $w$ and the any of the keywords $k$ used to collect COVID-related tweets, we estimate the proportion of

COVID-related tweets as:[9]

$$R_w = \frac{N(\{k\} \ \& \ w)}{N(w)}.$$

To this aim we employ a dataset of tweets downloaded independently from our data, with keywords related to a different subject.[10] Figure 3 displays the results and plots the weekly average of the estimated ratio $R_w$. The number of tweets related to COVID-19 becomes a significant fraction of the total (more than one sixth, on average, by the end of March) in the first phase of the pandemic, while they constitute a small percentage of the total during the rest of 2020.

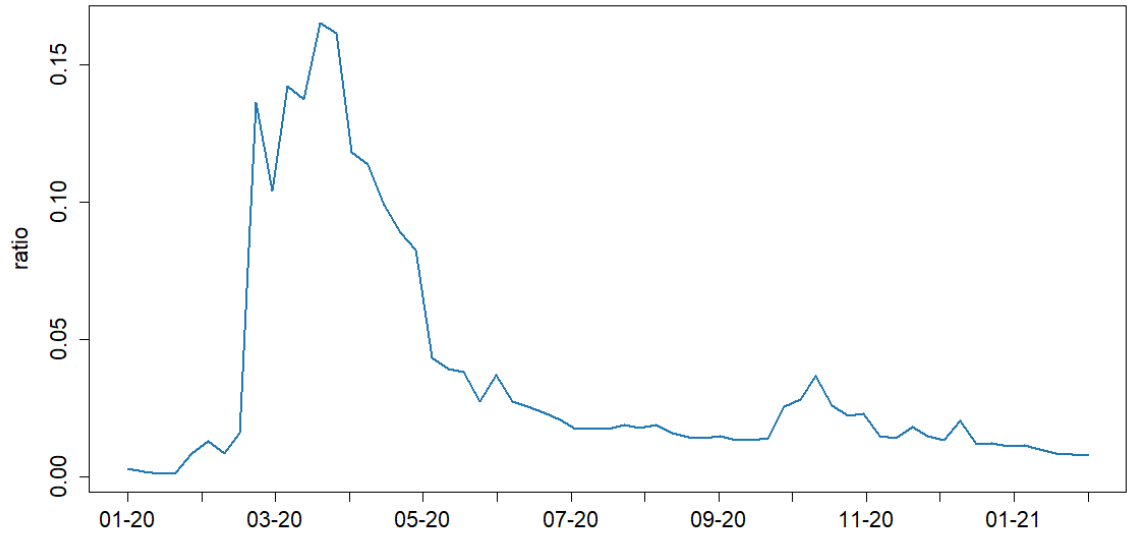


Figure 3: The estimated ratio of COVID-related tweets with respect to the total in 2020 (weekly averages).

As a partial confirmation of this result we can exploit data gathered from Google Trends. Google Trends (`https://trends.google.com/trends`) is a website that provides the time series of the (relative) counts of any search query made on Google that exceeds an unknown threshold based on the geographical location, also giving the possibility to select a time window and a region of interest. The counts, measured with a 0-100 index (normalized on the chosen time window) are available on a weekly basis. Such series offer an overview of how frequent a query is in users' search requests made to Google. It seems reasonable to assume that users try to find on Google

[9] If we can treat the appearance of any of the keywords $\{k\}$ and of $w$ as independent events, $N(\{k\} \ \& \ w)$ can be approximated as $N^T P(\{k\})P(w) \approx N(w)P(\{k\})$, with $N^T$ the total number of tweets and $P(\{k\})$ and $P(w)$ the probabilities of appearance of any of the keywords $\{k\}$ or of $w$.

[10] These keywords contains a selection among 200 terms related to the insurance world; some examples are the words *"beneficiario"* ("beneficiary"), *"broker"* ("broker"), *"cauzione"* ("deposit"), *"fideiussione"* ("surety"), *"liquidazione"* ("liquidation"). Some of the terms may be correlated with COVID, but we assume that on average these correlations will cancel out.

information on topics they care about, and that Google Trends data can be a good external benchmark to make a comparison with Twitter data. While Google Trends data by themselves are not sufficient to provide an estimate of the relative importance of COVID-19 tweets at any given time due to the lack of information about the total number of queries on Google, the time variation of the series can represent an independent confirmation of the behaviour of the series represented in Figure 3. The idea is to select some of the most frequent hashtags of our tweets, and download the corresponding queries from Google Trends, limiting them to lie in the same time period and searched from Italy.

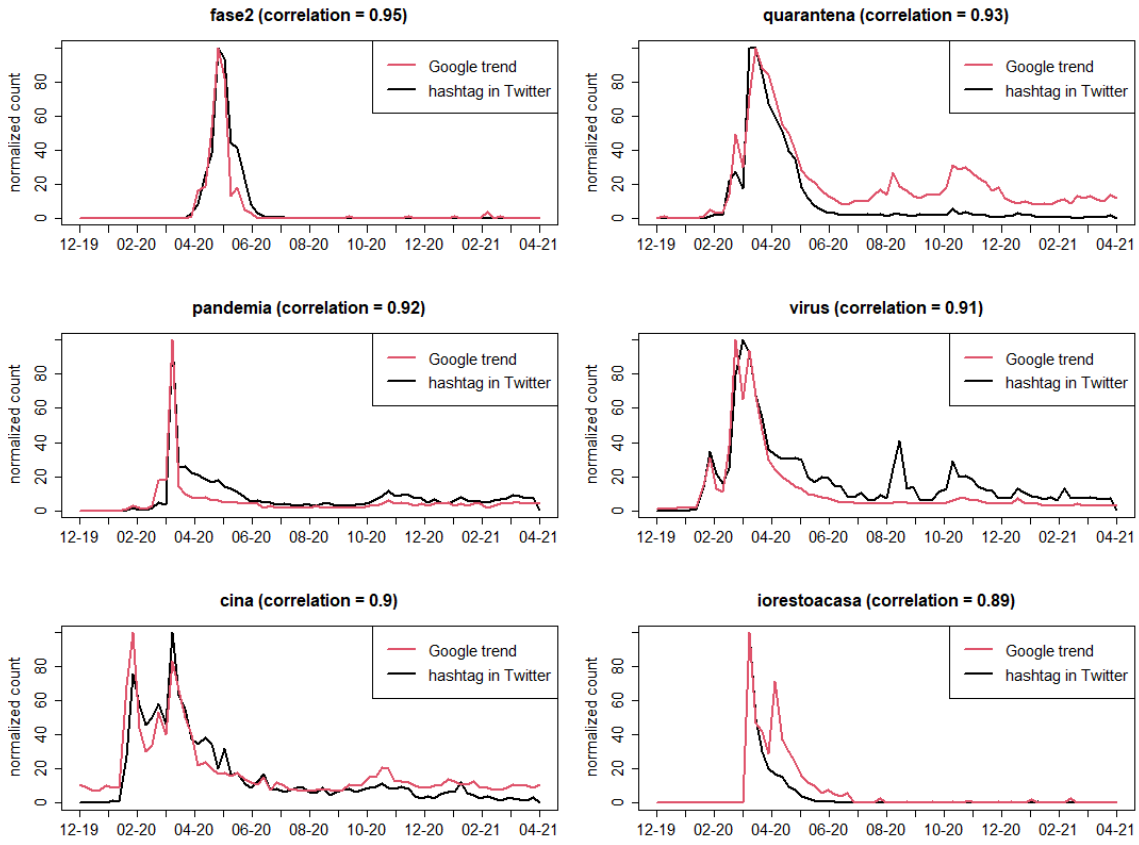


Figure 4: Google trends normalized time series compared to the normalized counts of the corresponding hashtags.

Figure 4 displays the time series with the (normalized) counts of some COVID-19 related hashtags included in our tweets (black lines) and the corresponding Google Trends series (red lines). For these words the two series are highly correlated (correlations reported in the titles of the plots) and show a similar behaviour over time. Assuming this correlation remains valid for other types of searches and hashtags, and exploiting the fact that we can extract from Google Trends ratios between different

searches, we can use the increase of searches related to COVID-19 with respect to some baseline search term as an estimate of the change in the relative amount of COVID-19 tweets, and tweets containing words whose frequency of use is not expected to change over time. More formally, denoting by $r_i^{\mathrm{GT}}$ the ratio between the counts $c(\cdot)$ of Google queries containing the word "coronavirus" and those containing the stopword[11] $i \in \{$"e", "con", "la", "un", "che"... $\}$,

$$r_i^{GT} = \frac{c(\text{“coronavirus”})}{c(i)}.$$

Figure 5 illustrates the described ratio for a selection of words.

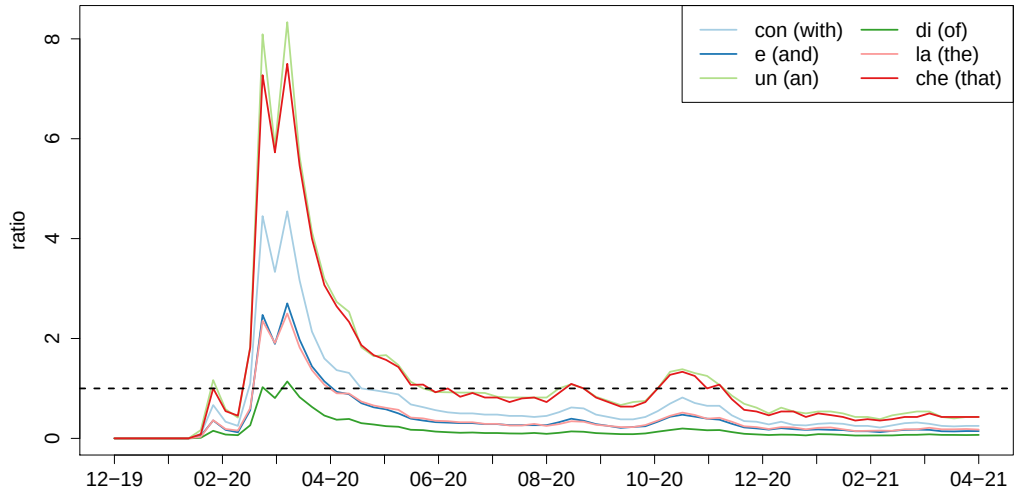


Figure 5: The ratio $r_i^{GT}$ between the counts of the Google queries containing the word "coronavirus" and those containing the words in the legend.

As expected the ratios built with Google Trends data show a similar behavior to the ratio obtained using as benchmark the number of tweets containing uncorrelated words (Figure 3); accordingly the average correlation is 0.85.

# 3 Models

In this section we describe the models we used to analyze the data at hand. Section 3.1 deals with topic modelling, while Section 3.2 explains the word embedding technique. Finally, Section 3.3 describes sentiment analysis presenting two different techniques: vocabulary-based approach (Subsection 3.3.1) and word embedding (Subsection 3.3.2).

[11] Equivalent in English to $i \in \{$"and", "with", "the", "an", "that"... $\}$.

## 3.1 Topic analysis

Topic analysis is a Natural Language Processing (NLP) technique that permits to extract information from a collection of texts by identifying common topics. The main task of topic modeling is therefore to identify the topics that best describe a set of given documents. Generally speaking, a topic modelling algorithm is a generative model describing word counts. Most common probabilistic topic models rely on Bag-of-Words (BoW) representation, that is, they assume that the order of words does not matter. According to this assumption, a document can be represented by a vector of numbers counting the occurrences of a term in the text and neglecting any additional information on words order or co-occurrence.

The most popular probabilistic topic modelling technique is known as Latent Dirichlet Allocation (LDA), developed by Blei et al. (2003). Since its first appearance in 2003, LDA gained a lot of popularity (also testified by the huge number of citations of the original paper - more than 35K), and was exploited in many different contexts.

The main idea behind LDA is the assumption that a set of documents can be described by a distribution of (latent) topics, which can in turn be described by different sets of characterizing words. More formally, each document $d \in \{1, ..., D\}$ is represented as a mixture of $K$ topics. In these mixtures, each word within a given document, $w_{d,n}$, $n = 1, ..., N_d$ belongs to exactly one topic. As a consequence, single documents can be considered as vectors of topic proportions, $\boldsymbol{\theta}_d$, $d = 1, ..., D$, which indicate the percentage of the document belonging to each topic (known as *topic prevalence*), and single topics can be considered as vectors of word proportions $\boldsymbol{\phi}_k$, $k = 1, ..., K$, which indicate the weight of any word in each topic (known as *topic content*).

The likelihood of a given word $w_{d,n}$ in LDA is given by

$$p(w_{d,n}|d) = \sum_{k=1}^{K} p(w_{d,n}|z = k)p(z = k|d),$$

where the distribution of topics within document $d$, $p(z = k|d)$, is modelled as a multinomial distribution with parameter $\boldsymbol{\theta}_d$, $Z|d \sim \text{Mult}(\boldsymbol{\theta}_d)$, and the distribution of words within document $d$ conditioned on belonging to topic $k$, $p(w_{d,n}|z = k)$, is a multinomial distribution with parameter $\boldsymbol{\phi}_k$, $W|z = k \sim \text{Mult}(\boldsymbol{\phi}_k)$.

In LDA, it is further assumed that the prior densities for the random vectors $\boldsymbol{\theta}_d$ and

$\boldsymbol{\phi}_k$ are given by Dirichlet distributions with hyper-parameters $\alpha$ and $\beta$, which, being conjugate to the multinomial distribution, imply that convenient inference techniques can be applied for the estimation procedure (e.g. the Gibbs sampling).

In this paper, we employ a generalization of LDA, namely the Structural Topic Model, STM, (Roberts et al., 2016), which is able to include into the statistical analysis document-level covariate information. This ability aims at improving the assignment of the words to the latent topics in the corpus. More specifically, this model assumes that topic prevalence and topic content can be specified as generalized linear models depending on specific document covariates. In particular, topic prevalence is assumed distributed according to a multivariate logistic normal density having the mean vector parametrized as a function of observed document-level covariates, and topical content is modelled according to an exponential density (similar to a multinomial logistic regression whose covariates are i) the document-level covariates, ii) the world-level latent assigned topic variables and iii) their interactions). We include covariates in the model for topic prevalence as we want to account for the fact that the distribution of topics can be influenced by document-level covariates (for instance, time and space), while the distribution of words within each topic is shared by all documents and do not vary with document specific covariates. Formally, the model for topic prevalence is:

$$\boldsymbol{\theta}_d | \boldsymbol{x}_d, \Gamma, \Sigma \sim \text{LogisticNormal}(\boldsymbol{x}_d \cdot \Gamma, \Sigma),$$

where $\boldsymbol{x}_d$ is the vector of document $d$ covariates, $\Gamma$ is a sparse matrix of coefficients for the topic prevalence model (included to avoid over-fitting), and $\Sigma$ is the covariance matrix, shared by all documents. For additional details on STM model specification, including the choice of prior densities and the precise formulation of the dependence on covariates of topic content which we do not employ here, we refer to Roberts et al. (2016).

The STM model has the major drawback that it looses the convenient conjugacy property of LDA, thus resulting in mathematically intractable posterior distributions. Inference is then carried on with a variational expectation–maximization algorithm that, upon convergence, gives estimates of the model parameters.

The choice of $K$, that is, determining the optimal number of topics, is a fundamental, and generally difficult, task in probabilistic topic modelling. As a matter

of fact, the STM requires to specify $K$ prior to the analysis, and then estimate the model conditionally on it. The usual practice in statistics amounts to fit the model for a range of values, say $k = 1, ..., K$, and then inspect some goodness of fit measures which can drive the choice of $K$.

STMs are implemented in a `R` package, called `stm`, (Roberts et al., 2019), which we employ for the analysis. This package comes with many functionalities, including text-processing functions, convenient tools for summarizing and visualizing the posterior distributions and tools for the choice of $K$. In particular, the `stm` package has a useful function which fits the model for $k = 1, ..., K$, and outputs the following goodness-of-fit measures:

1. The **held-out likelihood** (Wallach et al., 2009), built by keeping out some portion of the words (the test set) in the set of documents, train the model and use the document-level latent variables to evaluate the probability of the held-out test set (high values are best);

2. The multinomial dispersion of the STM **residuals** (Taddy, 2012) (small values are best);

3. The **Semantic coherence** (Mimno et al., 2011) which measures the frequency with which high probability topic words tend to co-occur in documents (high values are best);

4. The **Exclusivity** (Airoldi and Bischof, 2014), which measures the share of top topic words which are distinct to a given topic (high values are best).

There is a trade-off between semantic coherence and exclusivity: if all topics have the same top words we would have extremely high coherence, while by picking completely disjoint topics which do not co-occur in the documents we would obtain high exclusivity. It is therefore important to examine both criteria together, in order to find a value of $K$ for which both measures are reasonably high.

After estimating the model for a chosen $K$, checking relevant words for each topic is an easy way to understand the output and to label topics. This labelling step can be done considering two desirable features of the words that characterize the topics: frequency and exclusivity. A word is frequent if it occurs with high probability when discussing about a topic, while it is exclusive if it appears almost only in a specific

topic. We aim at considering both these features in order to find the most distinctive words for each topic. Along with the highest posterior (`HP`) probability words, `stm` reports for each word the `FREX` metric, proposed by Airoldi and Bischof (2016), which combines semantic coherence and exclusivity using a weighted harmonic mean of the word rank in term of frequency and exclusivity.

## 3.2 Word Embedding

As already mentioned, topic models like STM employ a representation of the analyzed documents which treats every word as an independent object, thus ignoring their order of appearance in a sentence or a document. This representation is detailed enough to perform a topic analysis that relies on the number of occurrences of single words, ignoring the meaning of full sentences or any other relation between words. However this representation cannot take into account similarity between words and this might be a limitation for some tasks.

In the most common form a Bag-of-Words representation identifies a document with a vector having as entries the number of times each word of the vocabulary appears in the document. This document-vectors have the dimension of the full vocabulary and are filled mostly by null values (for a large enough corpus it is very unlikely to find a document containing a large portion of the words in the vocabulary). In the vector space hence generated each word defines a different - orthogonal - direction, such that words like “beauty” and “beautiful” have the same relation as, for example, “beauty” and “ugly”. To overcome both these limitations (that is, the high-dimensionality and sparseness of the vector space and the absence of any information on the relations between words) a method of unsupervised dimensionality reduction can be applied, called *word embedding*. The idea of identifying words by their context dates back at least to Firth (1957) and was later adapted to take advantage of the rapid evolution of computational methods. We used the so-called `word2vec` algorithm, introduced in Mikolov et al. (2013). `Word2vec` is an unsupervised learning algorithm, by which a neural network predicts a word from its context, or - in a variation of the method - the context in which a word appears from the word itself. It is an unsupervised method in that no further information is given to the model in addition to the sequence of words composing the documents, and the algorithm learns to assign a “meaning” to every word (appearing a sufficient number

of times) simply by association with the context in which it frequently appears. In this way every word is mapped on the parameter space of the neural network once it is trained on the corpus, and similar words (i.e. words appearing frequently in similar contexts) will be close in this parameter space. The dimension of the parameter space of the model is chosen by hand, and does not scale with the extension of the corpus or the vocabulary.[12] In common applications `word2vec` models with few hundreds of parameters show good performances in describing corpora composed of millions of different words. We have a (filtered) vocabulary of order $10^5$ words, and used a model with 100 parameters. Both the mentioned problems of the large dimension of the word representation space and the lack of a similarity concept for different words are overcome by this method. Forcing an unsupervised learning algorithm to collect the words in a much smaller space than the one we started with, it organizes them introducing a notion of distance, and describing semantically similar words as close together. An important byproduct of this representation, which we will exploit in our analysis, is the introduction of a linear structure in the representation of words. Loosely speaking a word can be represented as a linear combination of its "component meanings" (Pennington et al., 2014); a classic example of this structure is given in terms of the approximate equation:

$$\text{king} - \text{man} + \text{woman} \approx \text{queen},$$

meaning that in the embedding representation space the closest vector to the linear combination of the word vectors on the left hand side of the equation is usually the one associated to the word we would have picked. This allows us to use in this space the concept of analogies between words, and derive which are the most common associations in the corpus under study. Going a step further in this direction, the authors of Kozlowski et al. (2019), Swinger et al. (2019) characterize words belonging to given fields projecting them on some interesting directions of this "analogy space". With this analysis they are able to represent popular beliefs and prejudices about, for example, the relations between sports, socioeconomic classes and other "cultural dimensions". We will exploit the same principle to show an application of word embedding to sentiment analysis in Section 3.3.2, and to the extraction of other

[12]More precisely: it can grow much more slowly than the dimension of the vocabulary, and still allows for the model to have good performances.

indices related to the COVID-19 pandemic in Section 4.3.

## 3.3 Sentiment analysis

An essential part of our analysis is focused on the mood expressed in the tweets about COVID-19. Various automatic techniques exist to extract the tone from a document; Subsection 3.3.1 describes the *vocabulary-based* approach, which makes use of some human knowledge of the meaning of words. Subsection 3.3.2 explains how to obtain a sentiment index with an alternative method, derived from the word embedding technique of Section 3.2. Later, in Section 4.2, we will show that the results from both techniques coincide over the period analyzed.

### 3.3.1 Vocabulary-based sentiment analysis

The so-called *vocabulary-based* (or *rule-based*) sentiment analysis is the simplest approach to analyze the tone of a document. As the name suggests, this method exploits a rule assigning to every word in the vocabulary a sentiment score. This score is usually decided by human readers, or from previous knowledge of the word meaning in the context studied. Once a sentiment score is assigned to every word, we can evaluate the sentiment of a document summing the values of every word it contains. As an example, we can evaluate the tone of the sentence "I am very happy": the words "I", "am" and "very" are tone-neutral in most context (although the adverb "very" can be used to amplify the tone associated to the adjective "happy"), so they will usually carry a null sentiment score. The word "happy", on the other hand, carries a strong positive meaning. Taking the sum of every individual score we obtain a positive sentiment for the whole sentence. A less trivial example is the sentence "I am sad, but the gift was beautiful": here we have the three tone-carrying words "sad", "gift" and "beautiful". While the first has a strong negative connotation the other two are positive, so the total score of the sentence will depend on the intensity of the tone associated to each word - and ultimately will be dependent on the context and on a human reader's decision.

It is clear that considering just the sum of the word scores, longer documents would have highest scores in absolute value, given that usually they will contain more sentiment-carrying words. To avoid this bias a normalization is necessary in the document score definition. The usual definition considered in the literature normalizes

the sum with the number of sentiment-carrying words present in the given document, or with the sum of the scores absolute values. The simplest non-trivial score association consists in assigning the score 1 to words considered as positive and $-1$ to words considered as negative. With such an assignment the score of a document will be proportional to the difference between the number of positive words and the number of negative words contained in the document. In this case the normalization is usually considered to be the sum of the numbers of positive and negative words found in the document. This translates in the rule:

$$S_i = \frac{P_i - N_i}{P_i + N_i} \tag{1}$$

where $P_i$ is the number of positive words found in the document $i$, and $N_i$ the number of negative words. With such a definition the sentiment score of a document is always in the interval $[-1, 1]$ and the extremes are attained every time only positive or only negative words are found. The same properties can be obtained when word scores are allowed to range in a continuous interval. In this case normalizing with the total number of words found would imply that the extreme document-score values are reached when only maximally scored positive or negative words are found in the document. A more suitable definition of the sentiment score in this case would be:

$$S_i = \frac{\sum_{w_i} S_{w_i}}{\sum_{w_i} |S_{w_i}|} \tag{2}$$

where $S_{w_i}$ is the sentiment score of every individual word of the document $i$.

In our analysis we used as a rule to assign sentiment scores the vocabulary introduced in Bruno et al. (2018). A tone is assigned to almost 19K (18944) words, and the scores range from a minimum value of $-1$ for "*calamità*" ( *"calamity"*), "*malfunzionamento*" ( *"disruption"*) or "*paurosamente*" ( *"fearfully"*) to a maximum of 1 for the words "*felicissimo*" ( *"very happy"*), "*deliziosamente*" ( *"delightfully"*) and "*confidenza*" ( *"confidence"*). The sentiment scores in the vocabulary are decided starting from the ones of the well established "*OpeNER*"[13] dictionary, and enhancing them in order to take into consideration synonyms and antonyms. The final values are obtained from a self-consistent rule in order to maximize the coherence of word scores,

[13]OpeNER stands for Open Polarity Enhanced Name Entity Recognition. This project was funded by the European Commission under the FP7 (7th Framework Program).

and the result is a smoother distribution of the tone than the one from the starting vocabulary. As anticipated we constructed the sentiment score of every tweet as the sum of the individual words score. We adopted however a slightly different normalization than the ones presented above. The former score definitions are well suited to quantify the tone expressed in a long document, but they can become excessively noisy in short text like tweets. As we explained, they reach maximal value when only positive or only negative words are found. In a short document, however, is not unlikely to find by chance a word deemed as positive (or negative), even if the tweet tone is neutral. In this case the tweet tone would be given as maximally positive (or maximally negative) even though its real tone is more neutral.[14] To overcome this inconvenience, we prefer to normalize the index with the length of the document analyzed, to obtain:

$$S_i = \frac{\sum_{w_i} S_{w_i}}{\sum_{w_i} 1} \tag{3}$$

From this definition it follows that to a tweet composed by 9 neutral words (i.e. words with $S_{w_i} = 0$) and 1 positive word with $S_{w_i} = 1$ we assign a score $S_i = 0.1$ instead of $S_i = 1$ as we would have obtained with the other definitions. This seems closer to the real tone of the tweet, in which only one tenth of the words carries a positive tone.[15]

### 3.3.2 Embedding-based sentiment analysis

In the same spirit of Kozlowski et al. (2019), Swinger et al. (2019), we here analyse the mood of the population of Twitter users about themes connected to the COVID-19 pandemic, mimicking the results of a more traditional sentiment analysis. The idea is to exploit the results of the word embedding, in order to provide an unsupervised sentiment index which we will compare to the more traditional vocabulary-based index obtained with the procedure detailed in 3.3.1. This will serve as a consistency check

[14] This can be important when the sentiment is evaluated at the tweet level, but if one considers - as we did - the sentiment aggregated over a large number of tweets the noisy results will simply be averaged to zero.

[15] Before applying vocabulary-based sentiment analysis we perform an additional pre-processing step called "*stemming*". Stemming is a text normalization technique that reduces the dimension of the vocabulary and cuts suffixes from the words keeping just their root. Unlike English language, Italian words have different suffixes to distinguish gender and many exceptions for plurals; verbs are even more irregular, having specific suffixes for different moods and tenses. Many vocabularies, included the one we are using - include only the singular masculine version for names and adjectives and the infinitive for verbs. If the vocabulary includes only the word *"bello"* ("nice", masculine and singular), no scores will be assigned to tweets including words like *"bella"* ("nice" feminine and singular) or *"belli"* ("nice" masculine and plural) or *"belle"* ("nice" feminine and plural). In order to avoid this issue, we apply stemming on tweets and also on the vocabulary.

both for the vocabulary-based sentiment index, and for the word embedding method, which will prove able to extract independently information on what is positive and what is negative in a text.

To evaluate the sentiment expressed by any tweet in this unsupervised setting we take the projection of the vector associated to the tweet with the normalized sum of the vectors associated to words like *"buona"*- *"buono"* ("good"), *"bene"*("well"), *"bella"*- *"bello"* ("beautiful"), *"meravigliosa"*- *"meraviglioso"* ("wonderful"), minus their contraries.[16] In this model the vector representing the tweet is the sum of the vectors associated to its component words, so the sentiment of the tweet will be - similar to the case of the vocabulary-based analysis - the sum of the sentiment associated to each word. In addition to the sentiment direction we can evaluate the projection of the tweets on dimensions like "hope", or on particular issues of interest (in the following we will focus for example on the amount of "job"- and "vaccine"-related tweets). Once an index is associated to every tweet, we consider the evolution of the sentiment by taking the daily aggregate of the results for every different index. The results of the above techniques are shown in Sections 4.2 and 4.3.

# 4 Results of the analysis

This section reports the results of the analyses described in Section 3 applied to the large dataset of COVID-19 related Italian tweets. Section 4.1 deals with the results of topic analysis. Section 4.2 reports the results of sentiment analysis, comparing the indices obtained with the vocabulary-based approach and the word embedding method. Section 4.3 concludes with other results obtained from the word embedding.

[16]We did not include the words *"positiva"*- *"positivo"* ("positive"), because of the double meaning they have in the context of the pandemics: being positive to a test for COVID-19 is not associated to the normal meaning of the word "positive". A reassuring feature of the model is that - defining the direction for positive sentiment as above - the word "positive" has a negative projection on it.

## 4.1 Topic analysis

As previously mentioned, we perform topic analysis with the STM[17] (Roberts et al., 2016), whose main feature is the ability to include in the modelling procedure relevant document-level metadata. In our application we add metadata regarding the geographical location of the user (at the regional level) and the week the tweet was posted. [18]

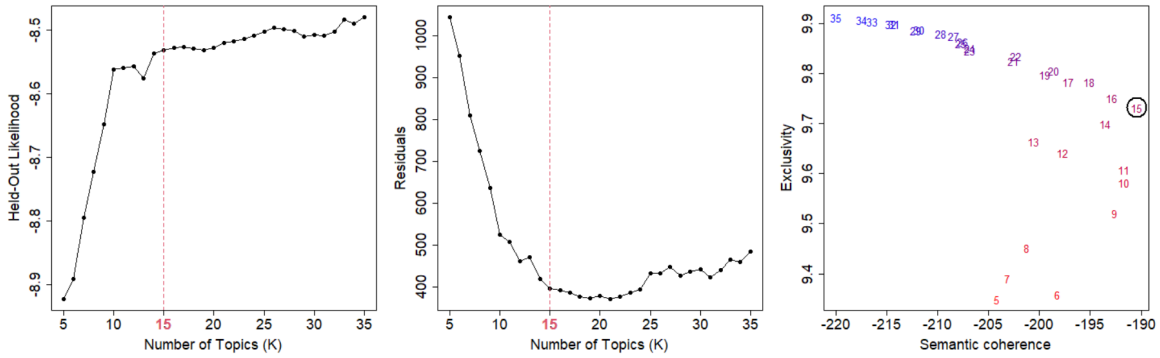


Figure 6: Diagnostic values by number of topics, $k = 5, ..., 35$. Left and middle panels: Held-out likelihood and residuals respectively. The vertical red dashed line marks the values corresponding to the chosen number of topics, $K = 15$. Right panel: exclusivity versus semantic coherence. The points are colored with palette from red (corresponding to small values of $K$) to blue (high values of $K$). Models with fewer topics have generally higher semantic coherence but lower exclusivity, and vice-versa. The black circle marks the value corresponding to $K = 15$.

Estimating a STM on a corpus containing millions of documents is extremely time consuming. In order to reduce the computational burden, we therefore randomly select a sample including a fraction (10%) of the tweets and we use this subset to estimate the model for a range of $K$ values.[19]

Figure 6 shows the plots of the indices introduced in Section 3.1 to choose the number of topics, for $K = 5, .., 35$. We see that, based on these, a good choice for the number of topics is $K = 15$. In fact, both the held-out likelihood (first panel) and the residuals (middle panel) do not improve after $K = 15$, meaning that adding another topic doesn't give much better modeling of the data. Moreover, there is a

[17]Before applying STM, additional data cleaning was implemented: first, we removed words shorter than 3 letters (that are mainly articles or typos, and are therefore not useful for topic analysis) and a standard set of Italian stop-words; In a second moment, we excluded from the vocabulary all the words which appeared less than 10 times and more than 350K times in the corpus, in order to reduce the noise in the estimates. As a matter of fact, the top-frequent words, which are common to most documents, would appear in many topics (possibly, all), making it difficult their interpretation. This cleaning greatly reduces the vocabulary, from 600K terms to 100K.

[18]In particular, to estimate the effect of time (weeks) on topic prevalence, a time covariate is included in the model using a B-spline with 10 degrees of freedom.

[19]The results are stable for different subsamples.

good trade-off between semantic coherence and exclusivity, which are both reasonably high (right panel) for this number of topics.

In light of the above, and being aware that there's not a correct and true value for the number of topics, we run the model on the full corpus of tweets fixing $K = 15$. The algorithm converges after 12 iterations, in approximately 6 hours.

| Topic | Expected topic proportion | Top 8 terms |
|---|---|---|
| Topic 12 | 0.134 | casi, positivi, morti, contagi, italia, dati, bollettino, decessi |
| Topic 9 | 0.077 | italia, coronavirusitalia, lockdown, conte, fase, scuole, zona, marzo |
| Topic 11 | 0.075 | brividi, respiro, mal, casa, respirare, andare, amici |
| Topic 5 | 0.069 | emergenza, salute, news, misure, attività, sicurezza, controlli, ordinanza |
| Topic 10 | 0.069 | positivo, test, lombardia, veneto, napoli, sindaco, quarantena, isolamento |
| Topic 1 | 0.067 | tempi, iorestoacasa, scuola, post, online, spesa, emergenza, distanza |
| Topic 6 | 0.067 | parole, video, storia, vero, foto, leggere, parlare, guarda |
| Topic 14 | 0.065 | via, vaccino, mascherine, repubblica, cina, usa, vaccini, oms |
| Topic 3 | 0.062 | virus, febbre, mascherina, influenza, sintomi, casa, tosse, polmonite |
| Topic 13 | 0.060 | dolori, tanti, dobbiamo, problema, presto, morte, possiamo, problemi |
| Topic 7 | 0.059 | crisi, europa, euro, emergenza, economia, imprese, piano, lavoro |
| Topic 8 | 0.058 | governo, salvini, conte, italiani, giuseppeconteit, lega, fontana, vuole |
| Topic 2 | 0.055 | ospedale, medici, anni, pazienti, morto, ospedali, bergamo, medico |
| Topic 4 | 0.054 | sanità, pandemia, appello, sistema, diffusione, importante, rischio, informazione |
| Topic 15 | 0.030 | testa, situazione, casa, punto, settimana, possibile, causa, mese |

Figure 7: Results for $K = 15$. The discovered topics, ordered by prevalence, along with their 8 most likely characterizing words.

Figure 7 displays the estimated topics, ordered by proportion in the corpus, along with the high probability (`HP`) words that contribute most to each topic (see also the Appendix for a longer list, containing also the most characterizing, `FREX`, words).

Assessing the distinctive words of the estimated topics, we see that a broad set of themes is captured; in the next paragraphs we analyze the most interesting ones for our analysis, which we label, on the basis of our own interpretation, as follows.[20] Topic 12, which is the one with higher prevalence (0.134), can be interpreted as dealing with "*pandemic monitoring*" of COVID-19 official statistics. Topic 1 can be labeled by "*life restructuring*": the top-8 terms refer to internet-based activities which were highly popular during the lockdown, such as distance-learning and smart-working. Topic 3 describes "*health concerns*" with a particular focus on COVID-19 symptoms, and Topic 7 discusses about "*economic worries*".

Figure 8 reports some tweets selected among the most representative for the topics above.

As predictable, these topics assume a different relevance over time (Figure 9) and space: the expected proportion of tweets dealing with these topics changes during the crisis and for users posting from different regions. Every expected topic proportion

[20]Labeling topics learned by topic models is in general a challenging issue. Sometimes the meaning of the topics learned is quite intuitive, but it is often difficult to accurately interpret the meaning of each topic, especially when the total number of topics, $K$, is high. We refer to Ramage et al. (2009), Wan and Wang (2016), for examples of papers dealing with this matter.

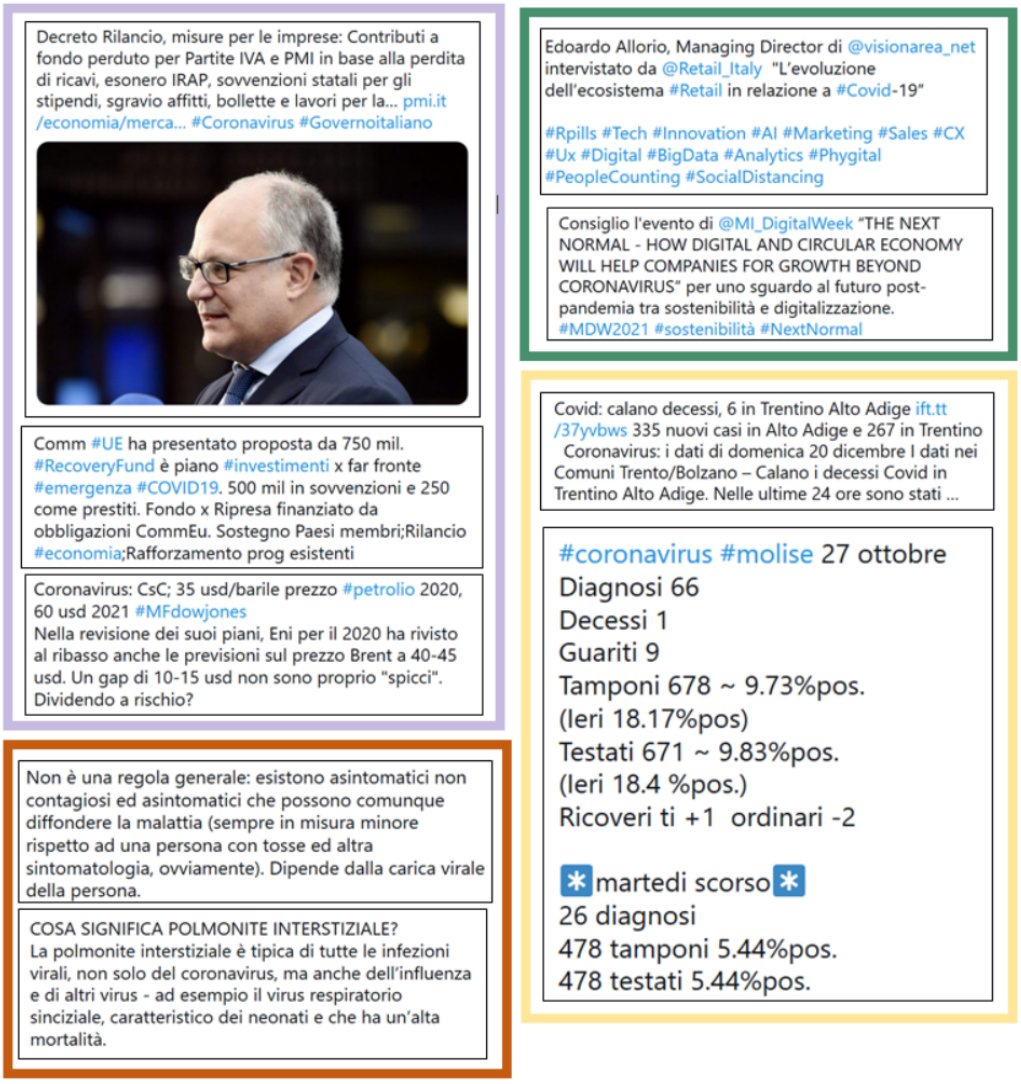


Figure 8: A selection of tweets highly representative of some topics: Topic 7 (violet), topic 1 (green), topic 3 (orange), and topic 12 (yellow).

over time should be considered as the average national frequency and is plotted as a smoothed function of time; parameter estimates of regional effects are topic-specific (can be positive or negative) and should be treated as additive to the frequencies for the overall country. Pandemic monitoring becomes more and more important over time and is the most relevant topic starting from March 2020. Tweets related to health concerns are closely connected to pandemic waves, while posts dealing with economic worries and life restructuring are especially relevant during the first national lockdown (Figure 9).

Life restructuring connected with changes in educational and working methods from home are especially relevant themes from March to June 2020, during the first and most restrictive lockdown (Figure 10, left panel). Focusing on the geographical distribution of tweets dealing with this topic, their proportion is slightly higher in

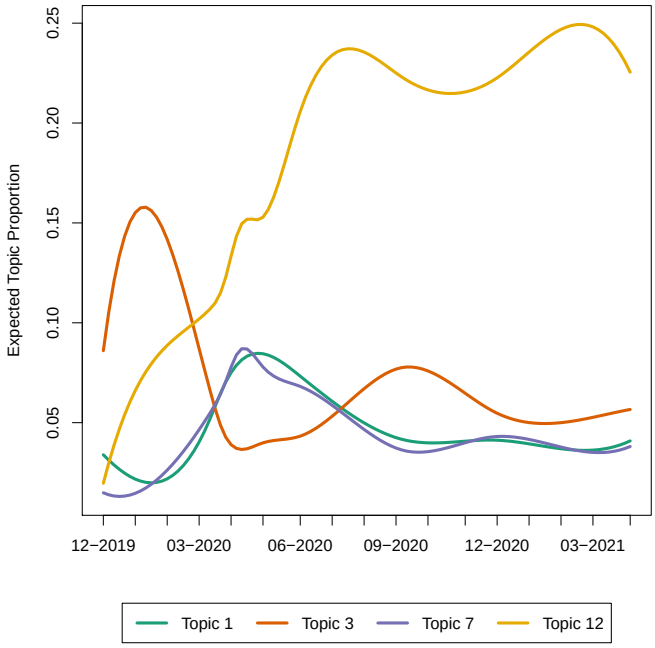


Figure 9: Expected topic proportions in time for selected topics. Topic 1 (life restructuring), topic 3 (health concerns), topic 7 (economic worries), and topic 12 (pandemic monitoring).

more productive regions (Figure 10, right panel).

Tweets about health concerns (COVID-19 symptoms and potential complications and prevention rules) are uncommon during summer 2020 and more popular during the two waves, especially the first one (Figure 11, left panel).[21] Tweets related to this topic are more frequently posted by users living in the regions hit harder by the virus during the first wave (Figure 11, right panel).

[21]Note that the color scales in Figure 10, 11, 12 and 13 are not identical, hence color gradients in each figure correspond to different variations in topic weights.

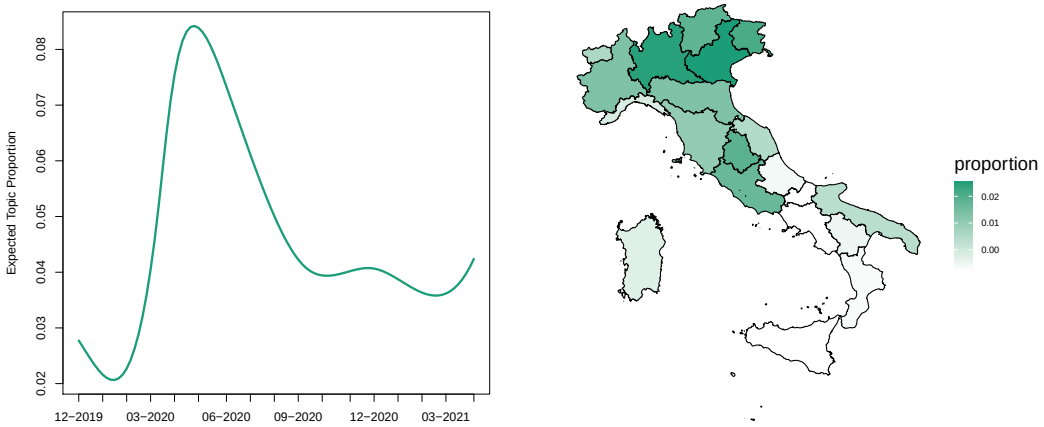


Figure 10: Expected topic proportions in time and space for Topic 1. HP words: `tempi (times)`, `iorestoacasa (Istayathome)`, `scuola (school)`, `post`, `online`; FREX words: `digital`, `working`, `ebook`, `digitaltransformation`, `elearning`.

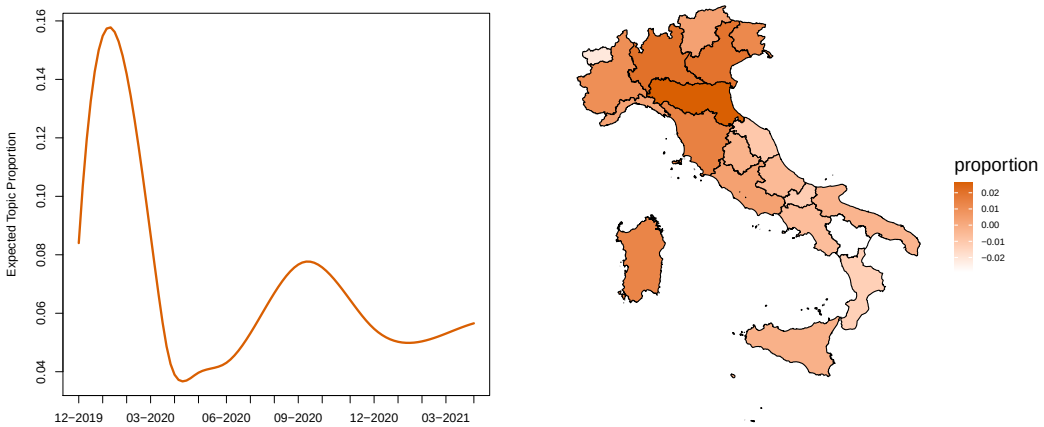


Figure 11: Expected topic proportions in time and space for Topic 3. HP words: `virus`, `febbre (fever)`, `mascherina (face-mask)`, `influenza (flu)`, `sintomi (symptoms)`; FREX words: `stagionale (seasonal)`, `influenza (flu)`, `olfatto (smell)`, `misurare (to measure)`, `pulci (fleas)`.

Posts about economic problems are considerably more frequent during the first lockdown as there was an increasing uncertainty on the length of the economic crisis and negotiations on recovery fund "Next Generation EU" were still ongoing (Figure 12, left panel).

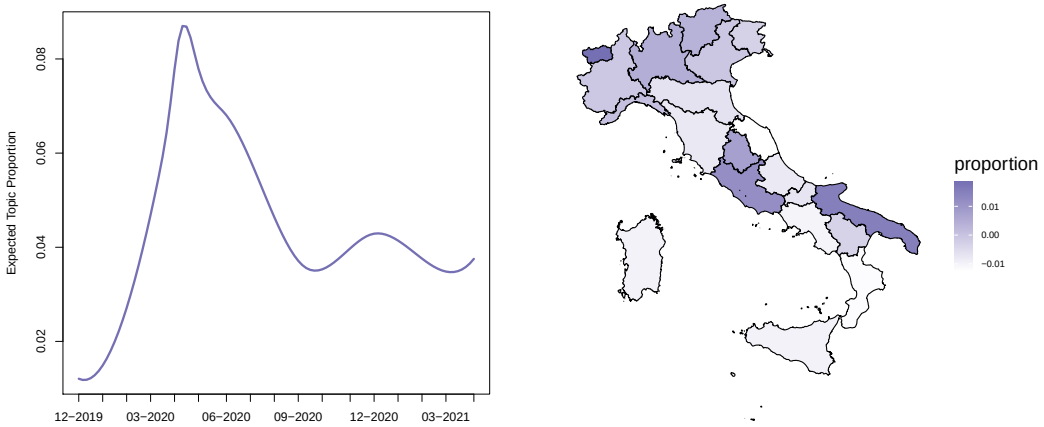


Figure 12: Expected topic proportions in time and space for Topic 7. `HP` words: `crisi (crisis)`, `europa (Europe)`, `euro (Euro)`, `emergenza (emergency)`, `economia (economy)`; `FREX` words: `eep` (Acronym of the Italian online Economic and Political journal, "Economia e Politica", `https://www.economiaepolitica.it/`), `liquidità (liquidity)`, `autonomi (autonomous)`, `eurogruppo (Eurogroup)`, `fiscale (fiscal)`.

Tweets referring to the daily update of pandemic statistics on COVID-19 are the most common: the expected topic proportion increases over the first two quarters of 2020, and continues to be very high during the rest of the period (Figure 13). Data on new daily cases, hospitalisations, intensive care unit admission rates, and current occupancy were a relevant decision-making factor in order to implement and justify the tightening or loosening of restrictive measures.

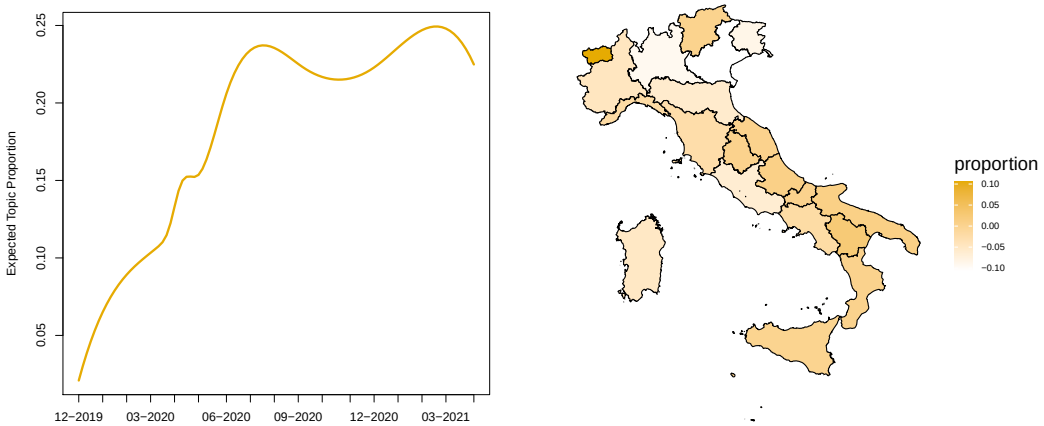


Figure 13: Expected topic proportions in time and space for Topic 12. `HP` words: `casi (cases)`, `positivi (positives)`, `morti (deaths)`, `contagi (contagions)`, `italia (Italy)`; `FREX` words: `bollettino (boullettin)`, `decessi (deaths)`, `situazionecoronavirus (situationCoronavirus)`, `pos`, `grafici (charts)`.

## 4.2 Sentiment analysis

Here we describe the results of the two methodologies outlined in Section 3.3. We start by presenting the results of the vocabulary-based sentiment score detailed in Subsection 3.3.1. After the assignment of a sentiment score to every tweet with the vocabulary-based approach, we considered the daily aggregate, in order to obtain a daily sentiment index. The sudden spike in the number of tweets by the end of February shown in Figure 1 poses the issue of how to normalize this index, in that a large part of the tweets posted after the spike have a neutral sentiment, driving the average index toward zero. The daily index obtained by summing the sentiment of all the tweets is plotted in Figure 14, black line. For the study of the evolution of the sentiment over the course of 2020 it can be useful to fix the sample of users we audit in the analysis. For this purpose we evaluated a different index, taking into consideration only users tweeting before 15 January 2020;[22] this approach aims at limiting the influence of "noisy" users (for instance advertisements, which often include popular hashtags in order to gain visibility on the platform), and gives us a stable benchmark (the sentiment prior to the large media coverage of the pandemic expansion) to evaluate the change over time of the index. The index computed on the selected pool of users corresponds to the blue line in Figure 14.

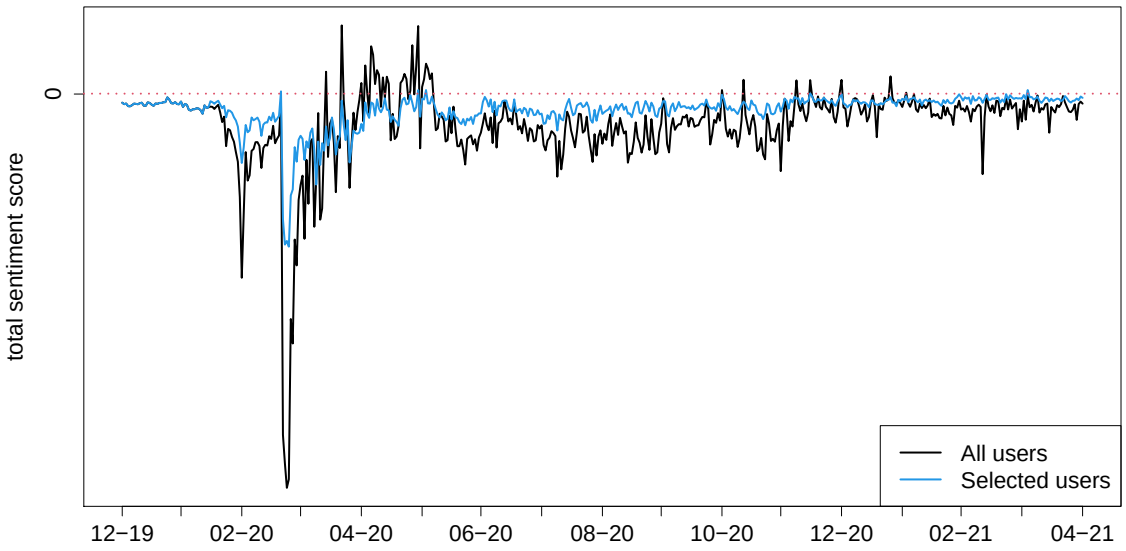


Figure 14: Total daily sentiment index. Black line: all the users in the sample. Blue line: selection of users posting before 15 January 2020.

As we can see from the figure, the sentiment undergoes a sharp (negative) change

[22]From 766K unique users we are left with 25K.

after February 20, 2020, the date of the first quarantine imposed in Italy. The total sentiment becomes more negative mainly because we have a larger number of negative tweets, not because the average sentiment per tweet worsens.

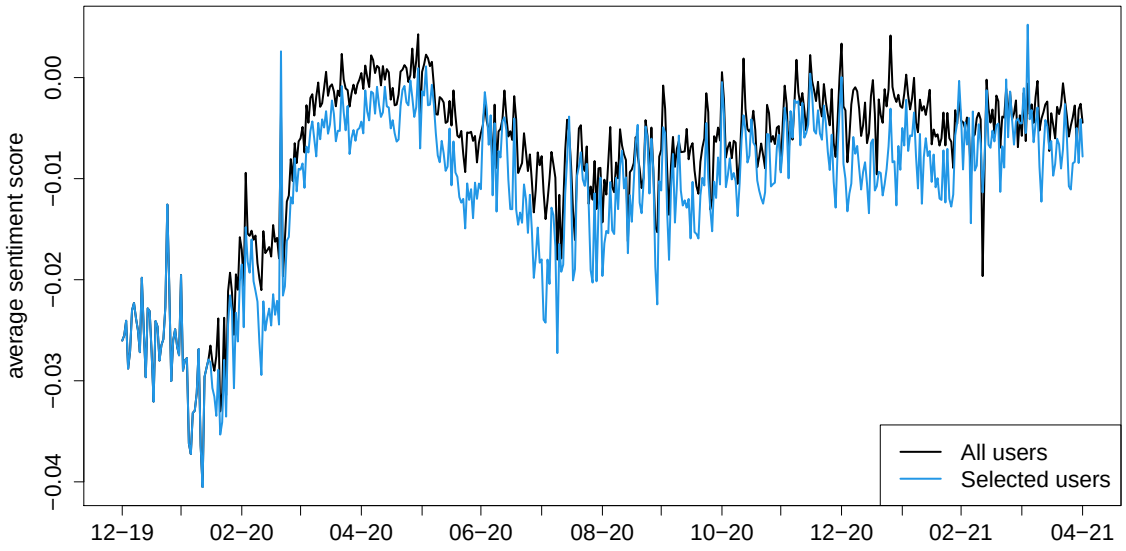


Figure 15: Average daily sentiment index. Black line: all the users in the sample. Blue line: selection of users posting before 15 January 2020.

This conclusion can be verified by looking at the average daily sentiment per tweet reported in Figure 15: the negative peak is no more present. Whenever this adds useful information we try to show both the total and the average daily score, in that they capture in principle very different phenomena: while the former is influenced by a greater number of users tweeting, or habitual users tweeting more, the latter tells us the change in the typical tone of a tweet. This difference is conspicuous when the popularity of a given subject suddenly changes, as in the first period of the COVID pandemic (see also Figure 1).

Next we show the daily sentiment calculated using the results of the trained word embedding model (Subsection 3.3.2) which evaluates the projection of the tweet-vectors on a direction associable with a positive sentiment (see also Section 3.2). The index calculated this way closely resembles the one extracted from the vocabulary-based approach, with a correlation of 0.85 for the sum of the daily sentiment scores and a correlation of 0.89 for the average daily sentiment scores. In Figure 16 we show the comparison of the two indices considering the sum of the daily sentiment scores of every tweet, while in Figure 17 we show the comparison for the average daily sentiment.

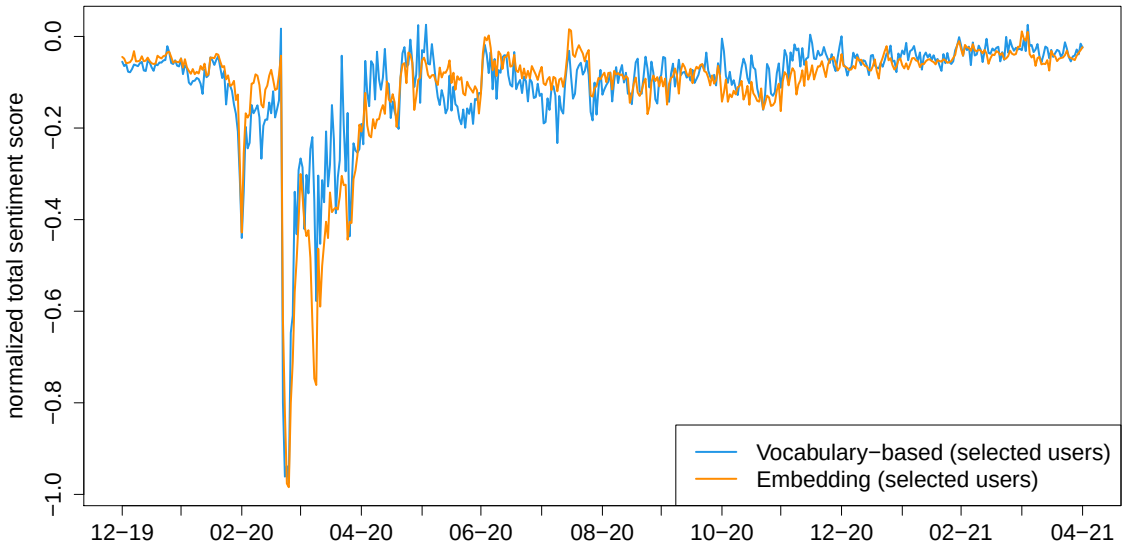


Figure 16: Normalized total daily sentiment index based on the selection of users posting before 15 January 2020. Blue line: index computed with the vocabulary-based technique. Orange line: index computed with the word embedding technique.

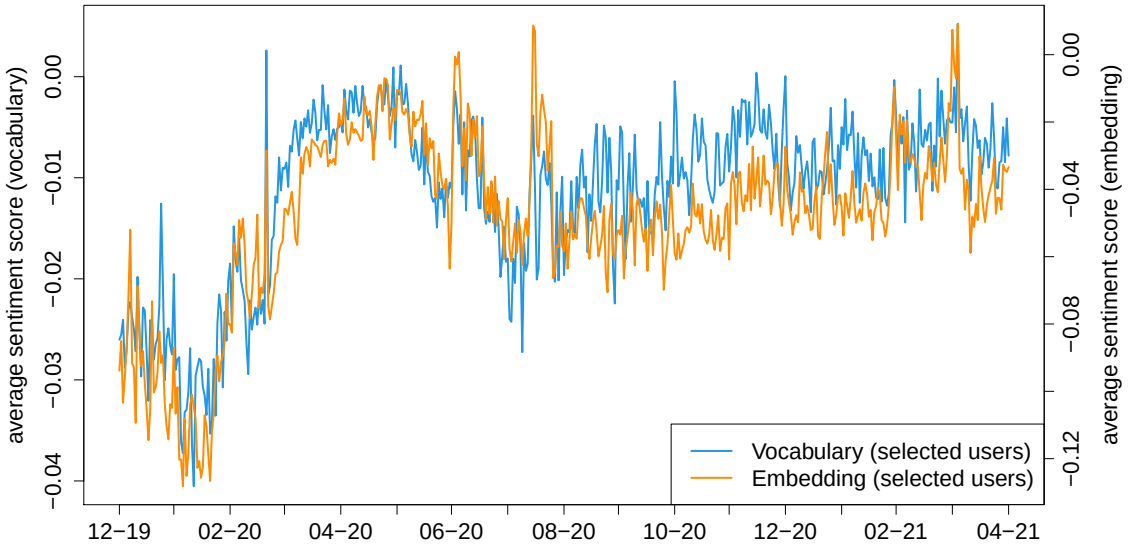


Figure 17: Average daily sentiment index based on the selection of users posting before 15 January 2020. Blue line: index computed with the vocabulary-based technique. Orange line: index computed with the word embedding technique (scale on the right-y-axis).

The sentiment index derived from the word embedding, though being strongly correlated with the standard vocabulary-based one, has a slight advantage in terms of stability. It presents less statistical noise, as captured for example in the signal-to-noise ratio (that is, the ratio between the average signal and its standard deviation): the index derived from the word embedding has an average signal-to-noise ratio of 0.36 (with a daily maximum of 0.87); the vocabulary-based sentiment index, on the

other hand, has a signal-to-noise ratio of 0.10 (with a daily maximum of 0.48). To sum up, even if the two indices have comparable averages, the one extracted from the word embedding has much less uncertainty around its mean value, signaling a greater ability to assign coherent scores to the tweets.

## 4.3 Word embedding: other indices

In addition to the sentiment index, from the word embedding technique we can derive a "hope" index, projecting the tweets on the direction defined by the set of words *"speranza"* ("hope"), *"fiducia"* ("confidence"), *"auspicio"* ("wish"), *"ottimismo"* ("optimism"), *"fede"* ("faith") and their contraries *"disperazione"* ("despair"), *"sconforto"* ("droop"), *"scoramento"* ("discouragement"), *"sfiducia"* ("distrust"), *"delusione"* ("disappointment").[23]

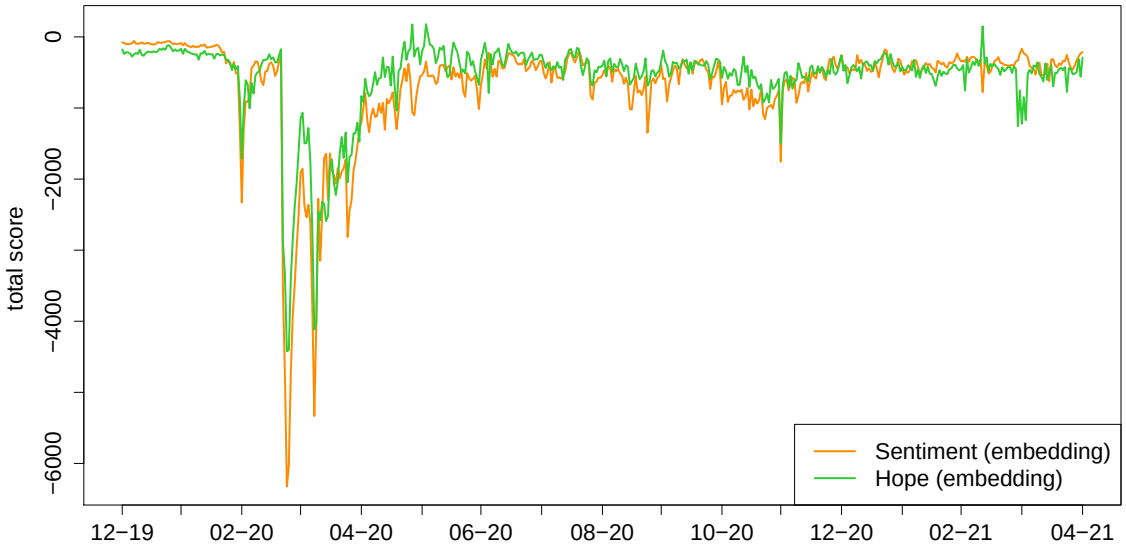


Figure 18: Green line: total "hope" index. Orange line: total sentiment index computed with the word embedding, based on all users.

We can see from Figures 18 and 19 that the behaviour of this index closely resembles the one of the sentiment score for the entire 2020 (correlation = 0.95 before 01/01/2021), while the two indices start to move apart in 2021 (correlation = −0.35 after 01/01/2021). For the average scores we obtain a similar result, with a correlation of 0.85 before 01/01/2021 and −0.52 after 01/01/2021. In particular, there are clear differences in two recent dates: the first one is the installation of the new

[23]It is useful to note that "Speranza" is also the surname of the Italian Minister of Health. This could in principle distort the "hope" index, so we verified the performance of the indicator with and without the word "speranza", obtaining no significant differences.

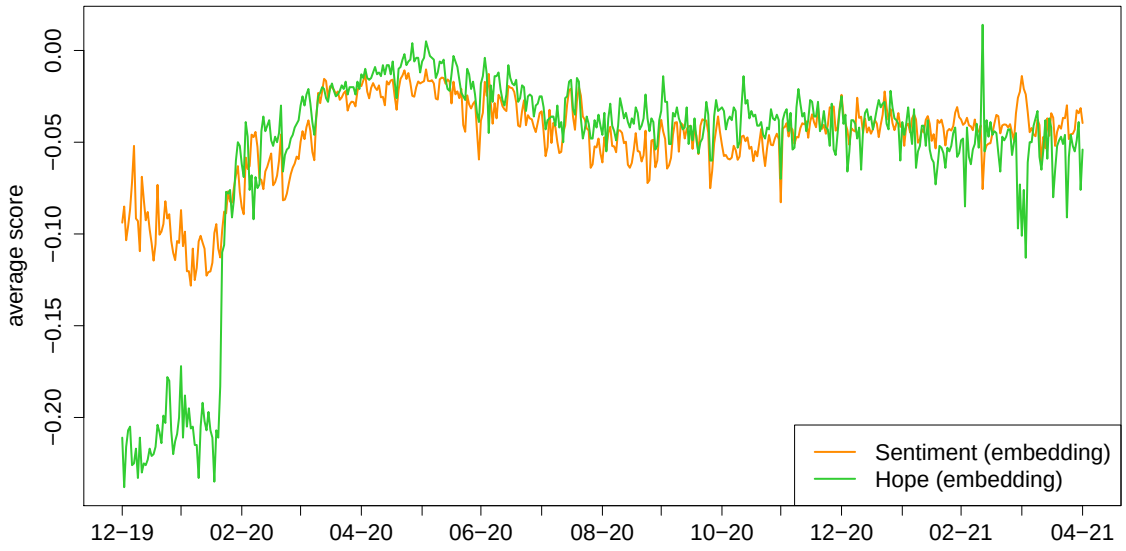


Figure 19: Green line: average "hope" index. Orange line: average sentiment index computed with the word embedding, based on all users.

government with Mario Draghi as president on February 12, 2021: in this occasion we had a spike in the "hope" index but no corresponding variations in the sentiment score. The second one was at the beginning of March; this is particularly interesting, in that the sentiment index has a relative growth in these days, while the "hope" index a sharp decrease. This is a clear manifestation of the difference in the expressions to which the two indicators are sensitive: during the first week of March the worsening epidemiological situation was pointing toward a tougher national lockdown, but at the same time a music festival with a wide audience (*Festival di Sanremo*) was aired every night on national television. While the sentiment index was influenced by the latter element and increased during the week, the "hope" index had an opposite reaction and was more adherent to the worsening of the situation and the probable restoration of a full lockdown.

In a similar fashion we can study the intensity with which a given argument is treated over time. For example we analyzed how much the daily Twitter content is related to words connected to "job" and "vaccine"[24]:

[24]To select the "job" direction we chose the words *"lavoro"* ("job"), *"occupazione"*-*"impiego"* ("employment"), *"salario"* ("salary"), *"stipendio"* ("paycheck"), while for the "vaccine" direction we picked *"vaccino"* ("vaccine"), *"vaccinazione"* ("vaccination"), *"vaccinare"* ("to vaccinate").

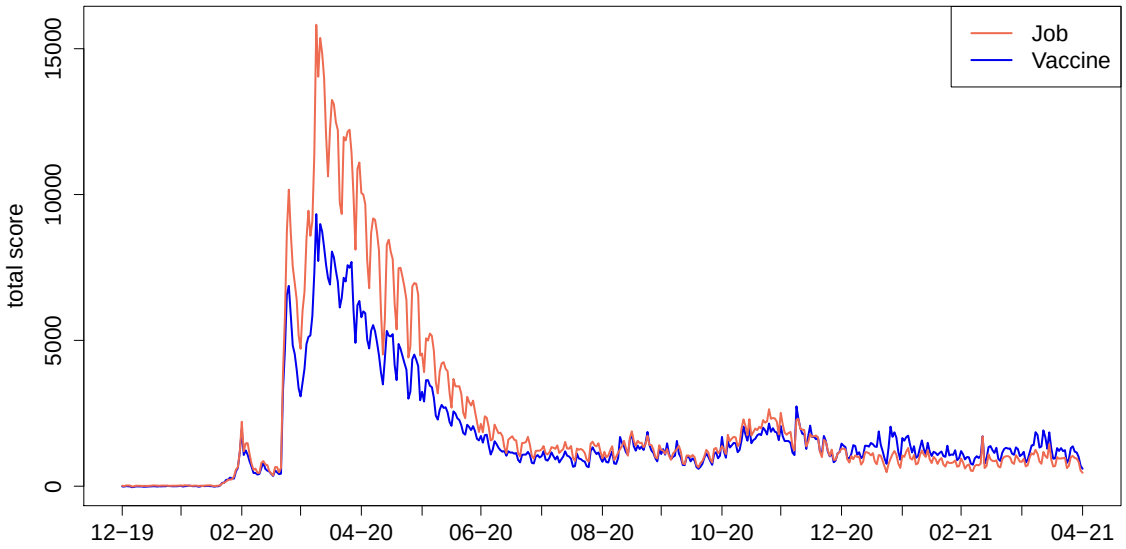


Figure 20: Coral line: total "job" index. Blue line: toal "vaccine" index.

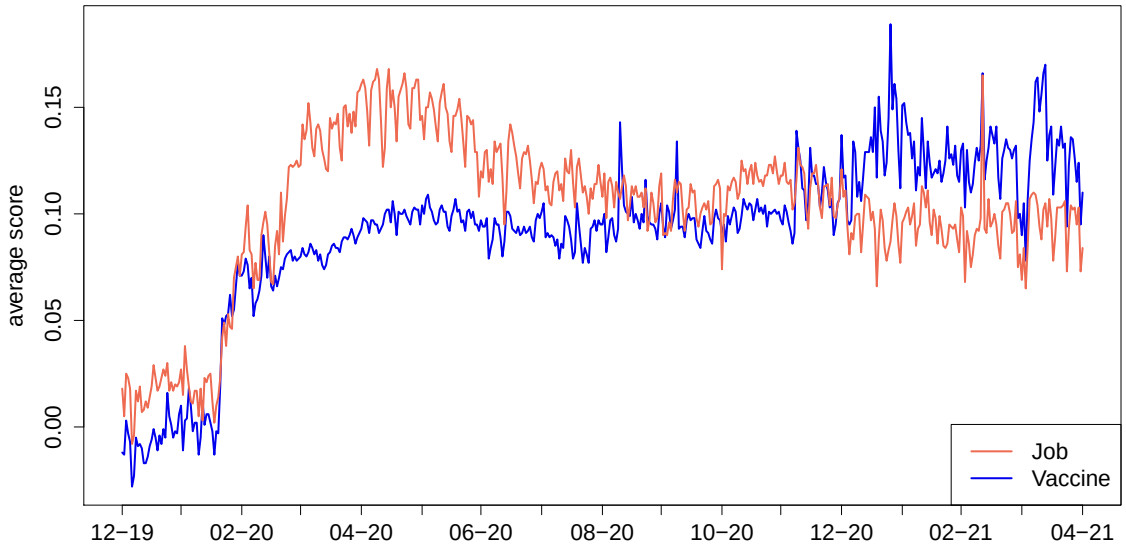


Figure 21: Coral line: average "job" index. Blue line: average "vaccine" index.

We see from Figures 20 and 21 that the attention towards job-related subjects was most prominent in the first part of 2020 (when the sentiment was lower), and vaccine-related conversation took the lead from the end of 2020 to the present days, strictly following the media coverage of the two issues. This is coherent with the results of topic analysis, which showed a spike in the interest toward economic issues shortly after the first lockdown, and a stronger emphasis on the medical aspect of the emergency in more recent times (see Figures 12 and 13).

Finally we show an example of the ability of the model to capture connections between significant words in our set of tweets. In Table 2 we report the associations -

in terms of closest words in the distance defined by the model - with some meaningful words in relation to the pandemic. We can see that the unsupervised model is able

| "galli" | Similarity |
|---|---|
| "clementi" | 0.82 |
| "andreoni" | 0.80 |
| "crisanti" | 0.78 |
| "virologo" | 0.77 |
| "bassetti" | 0.76 |

| "lombardia" | Similarity |
|---|---|
| "veneto" | 0.74 |
| "gallera" | 0.74 |
| "fontana" | 0.74 |
| "regionelombardia" | 0.73 |
| "regione" | 0.68 |

| "vaccini" | Similarity |
|---|---|
| "sputnikv" | 0.82 |
| "astrazeneca" | 0.82 |
| "pfizerbiontech" | 0.82 |
| "pfizer" | 0.82 |
| "genico" | 0.80 |

Table 2: Results from the word embedding: associations with some meaningful words in relation to the pandemic. Left: "galli " (the name of a famous Italian virologist); Middle: "lombardia"; Right: "vaccines".

to capture the similarity between words describing various features of the pandemics, identifying for instance the concepts of "virologist"[25] (Table 2, left), "vaccine" (Table 2, right), and features of geographical regions strongly related in the pandemics evolution[26] (Table 2, middle).

# 5 Concluding remarks

In this paper we proposed a number of automated techniques to analyze textual data, in order to perform a quantitative diagnostic of the Twitter mood over more than a year of COVID-19 pandemic.

We described a tool to perform Topic analysis, namely the Structural Topic Model, and two methods to perform Sentiment Analyses: a simpler one based on a predetermined vocabulary, and a second relying on the word embedding technique.

Our aim was to highlight the usefulness of automated textual analysis to extract quantitative indicators about the public mood, which is usually described only qualitatively and with incomplete data. A crucial advantage in the use of this kind of data and techniques is their timely availability, which makes them a valuable complement for fast predictions or now-casting analysis. Our main results obtained from the sentiment analysis picture a public opinion more negatively influenced at the beginning of the pandemic breakout in Italy, a partial alleviation of the initial negative shock after the first months of lockdown, and a milder worsening of the public mood in correspondence of the relapse of the health conditions after the summer.

[25]Galli, Clementi, Andreoni, Crisanti and Bassetti are among the most famous Italian virologists.

[26]Fontana and Gallera are respectively the Head and Health minister of the regional government of Lombardy. Veneto is the second most hit Italian region during the first wave of the pandemic.

The Structural Topic Model analysis points out a focus towards the medical aspects of the pandemics, the governmental measures to limit the spread of the virus and the economic consequences of these measures. We found 15 main topics in which the public discourse can be decomposed over the course of the period analyzed. For each of these topics we extracted the relative daily weight in the public discourse, showing, for example, a greater interest in the possible health consequences of the contagion at the beginning of 2020, a progressive intensification of the interest for the private consequences of the lockdown soon after, and, after the first few months, a constant component of the conversation focusing on the pandemic monitoring.

Word embedding confirmed the results of the vocabulary-based sentiment analysis, and allowed us to monitor nuances of the public discourse not noticeable with the other techniques (for example topics too weakly represented to be exposed with the other methods, or relations between different words frequently used in the tweets). For example we managed to measure the increase of interest towards the vaccination campaign, or how hopeful tweets were about the future.

This study poses the basis for interesting future developments. The first concerns the possibility of constructing indices which may relate to relevant economic variables. This aspect is particularly relevant in light of the fact that Twitter data come with partial information on the geographical location of the tweeter. The timeliness of the data collecting process and the fine-grained nature of the dataset could allow for the definition of high frequency indices tracking relevant macro-economic variables like propensity to spend, or which sectors are most influenced by current events. For instance, we may look at a regional-based sentiment, and relate it to the evolution of the epidemiological conditions throughout the country in order to exploit its potential predictive power. Another interesting research direction could be to estimate the impact of lockdown measures on topics related to well-being and psychological conditions, in the same spirit of Brodeur et al. (2021). A more methodological direction for future works could be to address the potential selection bias problems: any index built from a social media dataset will have to be externally validated or integrated due to the non-random nature of the selection process. For this reason, post-stratification strategies could be applied by exploiting the meta-information carried by the tweets. In addition, survey data gathered by the Bank of Italy[27] could contribute to precisely

[27]For instance the Special Survey of Italian Households.

delineate a demographic of the users of social networks, studying a representative sample of Italian families.

# References


Ahmed, W., Vidal-Alaball, J., Downing, J., and Seguí, F. L. (2020). COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. Journal of medical internet research, 22(5):e19458.

Airoldi, E. M. and Bischof, J. M. (2014). A Poisson convolution model for characterizing topical content with word frequency and exclusivity. arXiv:1206.4631[CS].

Airoldi, E. M. and Bischof, J. M. (2016). Improving and evaluating topic models and other models of text. Journal of the American Statistical Association, 111(516):1381–1403.

Altig, D., Baker, S., Barrero, J. M., Bloom, N., Bunn, P., Chen, S., Davis, S. J., Leather, J., Meyer, B., Mihaylov, E., Mizen, P., Parker, N., Renault, T., Smietanka, P., and Thwaites, G. (2020). Economic uncertainty before and during the COVID-19 pandemic. Journal of Public Economics, 191:104274.

Angelico, C., Marcucci, J., Miccoli, M., and Quarta, F. (2022). Can we measure inflation expectations using Twitter? Journal of Econometrics, 228(2):259–277.

Beauchamp, N. (2017). Predicting and Interpolating State-Level Polls Using Twitter Textual Data. American Journal of Political Science, 61(2):490–503.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022.

Brodeur, A., Clark, A. E., Fleche, S., and Powdthavee, N. (2021). COVID-19, lockdowns and well-being: Evidence from Google Trends. Journal of Public Economics, 193:104346.

Bruno, G., Marcucci, J., Mattiocco, A., Scarnò, M., and Sforzini, D. (2018). The Sentiment Hidden in Italian Texts Through the Lens of A New Dictionary. Mimeo, Bank of Italy.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis.

Hino, A. and Fahey, R. A. (2019). Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints. International Journal of Information Management, 48:175–184.

Hootsuite & We Are Social (2020). Digital 2020 Global Digital Overview.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. American Sociological Review, 84(5):905–949.

Levy, R. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. American Economic Review, 111(3):831–70.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In Proceedings of the 2011 conference on empirical methods in natural language processing, pages 262–272.

Moore, F. C., Obradovich, N., Lehner, F., and Baylis, P. (2019). Rapidly declining remarkability of temperature anomalies may obscure public perception of climate change. Proceedings of the National Academy of Sciences, 116(11):4905–4910.

Murphy, K. P. (2012). Machine Learning: A probabilistic perspective. The MIT Press.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Porcher, S. and Renault, T. (2021). Social distancing beliefs and human mobility: Evidence from Twitter. PLoS ONE 16(3): e0246949.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings

of the 2009 conference on empirical methods in natural language processing, pages 248–256.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. Journal of Banking & Finance, 84.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. Journal of the American Statistical Association, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2019). Stm: An R package for structural topic models. Journal of Statistical Software, 91(1):1–40.

Sciandra, A. (2020). COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings. In 2020 IEEE Symposium on Computers and Communications (ISCC), pages 1–6.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., and Kalai, A. T. (2019). What are the biases in my word embedding? In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 305–311.

Taddy, M. (2012). On estimation and selection for topic models. In Lawrence, N. D. and Girolami, M., editors, Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of Proceedings of Machine Learning Research, pages 1184–1193, La Palma, Canary Islands. PMLR.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. Proceedings of the 26th International Conference On Machine Learning, ICML 2009, 382.

Wan, X. and Wang, T. (2016). Automatic labeling of topic models using text summaries. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2297–2305.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., and Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. PloS one, 15:e0239441.

Yaqub, U. (2020). Tweeting during Covid-19 Pandemic: Sentiment Analysis of Twitter Messages by President Trump. Digital Government: Research and Practice, 2.

# A Additional material: stm

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| HP | HP | HP | HP | HP |
| tempi | ospedale | virus | sanità | emergenza |
| iorestoacasa | medici | febbre | pandemia | salute |
| scuola | anni | mascherina | appello | news |
| post | pazienti | influenza | sistema | misure |
| online | morto | sintomi | diffusione | attività |
| spesa | ospedali | casa | importante | sicurezza |
| emergenza | bergamo | tosse | rischio | controlli |
| distanza | medico | polmonite | informazione | ordinanza |
| intervista | san | dicono | cittadini | lavoro |
| bambini | infermieri | mani | libertà | ministero |
| FREX | FREX | FREX | FREX | FREX |
| digital | spallanzani | stagionale | diritti | socialnetwork |
| working | ricoverato | influenza | visoni | agenparlitalia |
| ebook | boris | olfatto | allevamenti | force |
| digitaltransformation | martino | misurare | animali | task |
| elearning | cotugno | pulci | umani | disposizioni |
| marketing | dimesso | influenzali | affidabile | castellammare |
| virtuali | johnson | lavarsi | scientifica | see |
| evitiamolo | iperimmune | temperatura | visoniliberi | epidemiologica |
| webinar | policlinico | sintomo | esseri | investiamo |
| dfmlab | antiartrite | termometro | atmosferico | euronews |

| Topic 6 | Topic 7 | Topic 8 | Topic 9 | Topic 10 |
|---|---|---|---|---|
| HP | HP | HP | HP | HP |
| parole | crisi | governo | italia | positivo |
| video | europa | salvini | coronavirusitalia | test |
| storia | euro | conte | lockdown | lombardia |
| vero | emergenza | italiani | conte | veneto |
| foto | economia | giuseppeconteit | fase | napoli |
| leggere | imprese | lega | scuole | sindaco |
| parlare | piano | fontana | zona | quarantena |
| guarda | lavoro | vuole | marzo | isolamento |
| bella | aziende | presidente | dpcm | marche |
| tweet | famiglie | lombardia | maggio | tampone |
| FREX | FREX | FREX | FREX | FREX |
| juventini | eep | fontanapres | voli | crotone |
| finoallafineforzajuventus | liquidità | olimpiadi | rosse | cilentonotizie |
| juventustv | autonomi | fontana | ristoranti | romaforever |
| forzajuventus | eurogruppo | collezionegiorno | coprifuoco | cilento |
| instajuve | fiscale | grimoldipaolo | corea | newscalabria |
| allianzstadium | prestiti | cotone | autocertificazione | calabrianotizie |
| forzajuve | cassa | maxferrari | pallavolo | vallodidiano |
| finoallafine | mutui | boss | codvid | allnews |
| continassa | fiscali | lavaggi | irlanda | vallodidianonotizie |
| alex | economica | beppesala | spaziotransnazionale | pesaro |

| Topic 11 | Topic 12 | Topic 13 | Topic 14 | Topic 15 |
|---|---|---|---|---|
| HP | HP | HP | HP | HP |
| brividi | casi | dolori | via | testa |
| respiro | positivi | tanti | vaccino | situazione |
| mal | morti | dobbiamo | mascherine | casa |
| casa | contagi | problema | repubblica | punto |
| respirare | italia | presto | cina | settimana |
| andare | dati | morte | usa | possibile |
| amici | bollettino | possiamo | vaccini | causa |
| viene | decessi | problemi | oms | mese |
| davvero | tamponi | morire | trump | anno |
|  | guariti | spero | video | italiani |
| FREX | FREX | FREX | FREX | FREX |
| mal | bollettino | gioie | changeitalia | testa |
| tzvip | decessi | forti | bill | assurdo |
| raga | situazionecoronavirus | uomini | gates | restate |
| porca | pos | ovunque | avigan | succede |
| gfvip | grafici | gioia | ilmeteoit | venendo |
| piango | odierni | cbd | randieri | inizia |
| svegliata | calano | mestruali | fakenews | prossima |
| venire | aumentano | atroci | pfizer | punto |
| prelemi | guarigioni | rimedio | lobby | finita |
| sveglio | guariti | cannabis | meteo | matematica |

Table 3: stm results. Highest probability (HP) and most characterizing (FREX) words for the 15 Topics.

# B Timeline of the containment/economic government measures

DL 23/02/20 ⟶ lockdown of 11 northern municipalities

DPCM 01/03/20

DL 02/03/20

DPCM 04/03/20

DL and DPCM 08/03/20 ⟶ lockdown of extended northern provinces (substitutes the former DPCMs of 01 and 04 March)

DL and DPCM 09/03/20 ⟶ national lockdown

DPCM 11/03/20 ⟶ "#IoRestoACasa"

DL 17/03/20 ⟶ "Cura Italia"

DPCM 22/03/20 ⟶ stop to all non-necessary businesses and industries + prohibition to travel outside the region of residence

DL 25/03/20

DPCM 01/04/20

DL 08/04/20 (n. 22 and 23) ⟶ "Liquidità"

DPCM 10/04/20

DPCM 26/04/20

DL 30/04/20

DPCM and DL 10/05/20

DPCM 12/05/20

DL 16/05/20 ⟶ Start Phase 2

DPCM 17/05/20

DL 19/05/20

DPCM 11/06/20

DL 16/06/20

DPCM 14/07/20

DPCM 23/07/20

DL 30/07/20

DPCM 07/08/20

DL 14/08/20

DPCM 07/09/20

DL 08/09/20

DL 11/09/20

DL 07/10/20 ⟶ New restrictions (starting the 8/10/20)

DPCM 13/10/20

DPCM 18/10/20

DL 20/10/20

DPCM 24/10/20

DL 28/10/20

DPCM 03/11/20 ⟶ Zone system starts (in force from the 6th November)

DL 09/11/20

DL 23/11/20

DL 30/11/20

DL 02/12/20

DPCM 03/12/20

DL 18/12/20

DL 05/01/21

DPCM and DL 14/01/21
DL 12/02/21
DL 23/02/21
DPCM 02/03/21 (in force from the 6th March) $\longrightarrow$ tightening of the stringency measures
DL 13/03/21 (n. 30 and 31) $\longrightarrow$ Red zone for the entire country during Easter holidays
DL 22/03/21
DL 01/04/21