Department of Business and Management

Master Thesis in
Global Management and Politics

Different ways of telling the pandemic: an examination of Covid-19 tweets from major Italian newspapers

Supervisor:
Prof. Marco Mingione

Co-Supervisor:
Prof. Silvia Dello Russo

Candidate:
Raffaele Di Cristo

ACADEMIC YEAR 2022/2023

Contents

List of Figures

List of Tables

R Coding: Scraping Tweets

```
#ID TOKEN
#Packages
packs <- c("tidyverse", "magrittr", "lubridate", "httr", "
    jsonlite", "rtweet", "twitteR")
sapply(packs, require, character.only = T)

#Bearer Token for Twitter Recognition
bearer_token_prem <-  "your_bearer_token"

api_key_prem <- "your_api_key"
api_secret_prem <- "your_api_secret"
bearer_token_prem <- "your_bearer_token"
client_ID_prem <- "your_client_ID"
client_secret_prem <- "your_client_secret"
access_token_prem <- "your_access_token"
access_secret_prem <- "your_access_secret"
twitteR::setup_twitter_oauth(api_key_prem, api_secret_prem,
    access_token_prem, access_secret_prem)

#Creation of Day Sequence: format employed is YYYY-MM-DD
day_seq <- seq(as.Date("2020-02-01"), as.Date("2022-02-28"),
     1)

#Creation of the vector for queries
#from: by which user?
#start_time, end_time: day and time of beginning and end of
    the time window of interest
#max_results: the maximum number of tweets we can download (
    initially set as 150 and doubled afterwords)
#tweet.fields: fields saved

testata <- "corriere" #name of twitter account without @
```

```
query_s <- paste0("from:", testata,"&start_time=",
                  day_seq[-length(day_seq)],
                  "T00:00:00.00Z&end_time=",
                  day_seq[-1],
                  "T00:00:00.00Z&max_results=150&tweet.
                      fields=id,text,author_id,created_at,
                      entities")


query_s <- paste0("from:Corriere", "&start_time=", day_seq[-
    length(day_seq)],"T00:00:00.00Z&end_time=", day_seq[-1],
    "T00:00:00.00Z&max_results=150&tweet.fields=id,text,
    author_id,created_at,entities")

#Output List: named with each italian newspaper
corriere <- list()

#Iteration Process
for(i in 1:length(query_s)){


getTweets <- GET(paste0('https://api.twitter.com/2/tweets/
    search/all?query=', query_s[i]),
                 accept_json(),
                 add_headers('Authorization' = paste0('
                     Bearer ', bearer_token_prem)))

#Reminder: status code = 200 iteration was successful
#getTweets$status_code
#fromJSON function to order downloaded data

json_data <- fromJSON(httr::content(getTweets, as = "text"),
     flatten = TRUE)

foo <- json_data$data %>% as_tibble() %>% mutate(created_at
```

```
    = lubridate::ymd_hms(created_at))

if(any(grepl("^entities.hash",colnames(foo)))){
        which(!map_lgl(foo$entities.hashtags, is.null))

corriere[[i]] <- foo %>% select(id, text, author_id, created
    _at, entities.hashtags)
     }
  }

if(as.numeric((getTweets$all_headers[[1]]$headers$`x-rate-
    limit-remaining`))<50){
     Sys.sleep(60)
  } else{
     Sys.sleep(3)
  }
  print(i)
}
#Combination at the end of the loop
corriere_combined <- reduce(corriere, bind_rows)

#Saving the object
save(corriere_combined, file = paste0("AllTweets_", testata,
     ".RData"))
```

Chapter 1

Introduction

Textual data analysis is now one of the most crucial research topics, given the increasing availability of unstructured data and the potential information that can be retrieved from it. With the increasing availability and popularity of opinion-rich resources such as social media, new opportunities and difficulties emerge as it is possible to employ information technologies to proactively seek for and comprehend the opinions of others.

Data is now generated in a wide range of formats and at a rate that shows no signs of stopping. Finding information that can help to make decisions among hundreds of documents, web pages, and social media feeds is a difficult and time-consuming task. As a result, textual information is frequently neglected or used only in part to support decision making processes.

Natural Language Processing (NLP) approaches, in this perspective, enable to digest text data more effectively by providing the possibility to identify relevant patterns and the capacity to transform textual data into usable knowledge.

One of the most interesting and widely used social platforms for generating a large amount of textual data is Twitter, whose use has both significant potential and limitations as well as methodological and ethical considerations relating to privacy and copyright.

Several variables can be identified as underpinning the proliferation of Twitter usage in the social sciences: firstly, Twitter's search capability and the fact that tweets display in Google search results make it easier to identify and follow conversations. Secondly, Twitter's API (Application Programming Interface) is

more open and accessible compared to other social media platforms, facilitating the ability for programmers to develop data access techniques and, consequently, providing additional resources to scholars. Moreover, Twitter's hashtag regulations make it easy to collect, sort, and expand data collecting searches.

Aside from content analysis, there are several approaches which may be adapted to Twitter features: because data on Twitter is available in real time, time series analysis, which is typically used to study the occurrence of peaks of tweets around a certain topic, is suitable. Furthermore, given the homogeneity of tweet length, the Sentiment Analysis applies well to Twitter.

The main objective of the present study is to address the research question concerning the analysis of textual information provided by tweets of Italian newspapers, in order to determine whether they behaved similarly when reporting on the Covid-19 pandemic in italy or not. Despite providing only a partial picture of all information provided during the pandemic in Italy, the assumption is that Twitter is sufficiently representative of all information provided during the pandemic in Italy by national magazines.

Italy emerged as the first nation in the Western Europe to be afflicted by the COVID-19 pandemic, and it is widely considered as one of the worst-affected countries. As stated in the seventh report produced jointly by Istituto Nazionale di Statistica (ISTAT) and Istituto Superiore di Sanità (ISS), from the outset of the pandemic up to February 2022, 145.334 deaths related with the diagnosis of SARS-cov-2 infection were reported in the ISS's integrated National Surveillance System COVID-19. Precisely, there were 77.778 deaths in 2020, 59.136 in 2021, of which about 8.000 related to diagnosis in 2020, and 8429 deaths in January 2022. Moreover, The COVID-19 standardised annual mortality rate declined from 50.5 deaths per 100.000 inhabitants in 2020 to 40.3 in 2021. This decline is entirely related to the North (73.1 in 2020, 39.4 in 2021), whereas the rate increases in the Centre (31.3 in 2020 and 37.0 in 2021) and the South (28.2 in 2020 and 43.3 in 2021).

Despite being identified in late January the first two cases of coronavirus in Rome, on 21 February is identified what will mistakenly be patient zero, a 38-year-old from Codogno. Several outbreaks are present in some areas of northern Italy such as Vo' Euganeo and in the province of Bergamo.

In light of this, the present study is based on a temporal horizon which goes from 2020-02-01 to 2022-02-28. Since the pandemic represents a world-wide health concern, it cannot be argued that it has stopped in Italy. The World Health Organization will announce the end of the pandemic based on evidence from around the world. In this regard, it is crucial to emphasize that the President of the United States, Joe Biden, signed the measure approved by the Federal Congress that declares the end of the national emergency associated with the Covid-19 pandemic. The measure, which divided the Democrats, was recently approved by the Federal Senate with a bipartisan majority of 68 votes to 23. However, for the implementation of this study, a decision was made to designate 2022-02-28 as an ending date. The underlying assumption is that starting from this date, there has been a shift in terms of communication provided by national newspapers in Italy from the Covid-19 to the outbreak of the war between Russia and Ukraine. Furthermore the SARS-cov-2 virus has mutated over time and new variants, have spread rapidly in Italy. The virus circulating in the first phase of the pandemic was followed by the spread of the alpha variant, then the delta variant from July to November 2021 and finally from the beginning of December the diffusion of the Omicron variant. Despite the fact that the above mentioned variants of concerns show a higher degree of transmissibility, they have become less aggressive, attempting to make the disease less harmful.

1.1 Literature Review

Research in the field shows that Twitter has been previously used, as a social media, for data analysis during the pandemic. In their study, Jimenez-Sotomayor et al. (2020), evaluated tweets about older adults and Covid19 in order to identify ageist or highly offensive content.

Mourad et al. (2020) highlighted the relevance of employing social networks in a worldwide pandemic crisis by relying on trustworthy users from a variety of industries.

Lahuerta-Otero and Cordero-Gutiérrez (2016) presented a quick investigation of the behavior of a specific type of tweeter user and assessed their influence ratio using various data mining approaches.

Xue et al. (2020) demonstrated that Twitter data and machine learning technologies may be used for an infodemiology study, allowing researchers to investigate evolving public debates and sentiments during the COVID-19 pandemic.

Yang et al. (2021) analyzed the degree to which the pandemic has spread on social media platforms like Facebook and Twitter.

Medford et al. (2020) created a set of COVID-19-related hashtags to search for relevant tweets throughout a two-week period from January 14th to January 28th, 2020.

Pastor (2020) depicts Filipinos' feelings about the effects of extreme community quarantine induced by the COVID-19 Pandemic. Based on the users' tweets, the researcher also examines the impact of excessive community quarantine and other Pandemic consequences on personal lifestyle.

Mantas et al. (2020) applied Natural Language Processing in their study. The application of a clustering technique to publically available Tweets posted by African Americans allowed them to discover several topics related to the pandemic.

Furthermore, in a research conducted by Ordun et al. (2020) ,Latent Dirichlet Allocation (LDA) topic modeling was used to produce twenty separate topics about case spread, healthcare staff, and personal protective equipment (PPE). The application of this technique helped them demonstrating how information about the Covid-19 spreads via Twitter.

Despite the existence of several studies on the field, not so many analyses have been conducted in Italy for the reference period. Hence, in contrast with other previous works which mainly focus on English tweets or diverse countries, the main contribution of this study is to provide, through data mined from Italian newspapers, a unique framework gathering information related to the Covid-19 crisis in Italy.

This work not only can be employed for future research, but it also provides a clear picture of how information was managed in Italy throughout the pandemic.

The study is organized as follows: Chapter 2 provides an overview of the methods used in the present analysis, including Natural Language Processing, Text Mining and Topic Modelling. Chapter 3 refers to all the steps related to Data, starting from the selection up to the scarping and keywords filtering. Subsequently, Chapter 4 provides an overview of the main findings of this work, including the description of relevant outcomes. In this section, the above mentioned methods have been implemented. Finally, Chapter 5 concludes the research.

Chapter 2

Methods

Textual data may be situated among a wide range of disciplines such as linguistics, content analysis, information retrieval and artificial intelligence. While several publications and book sections have analyzed the field of textual analysis, much fewer have actually described how to treat and analyze this kind of data (Burnard, 1996).

Textual data is frequently generated in real-world contexts such as social media and customer feedback. This implies that it can reflect real-world trends and concerns, making it applicable to a wide range of applications. Despite the flexibility and accessibility of textual data it shows several limitations: firstly, textual data might be imprecise and hard to comprehend. Indeed, it is not always evident what is the meaning of a single term or phrase in a particular context, rendering valid conclusions from the data challenging.

Secondly, textual data can be controversial since the same text may be perceived in a different way. As a result, contradictions in data analysis and modeling may arise. Moreover, textual data, unlike structured information such as numerical or categorical data, falls into the domain of the so called unstructured data. This implies that this kind of data has not a defined format, making it difficult to handle and analyze.

Finally textual data is frequently heavily influenced by the context in which it was generated. Harnessing this context and using it to gain insights from data can be difficult. In this regard, there have been significant changes in the overall

setting of this field of study, as well as the goals and methodological techniques it employs.

NLP and Text Mining (TM) procedures will be subsequently discussed and adopted in practice for the present work. Despite being related topics, it can be argued that Text Mining aims at identifying useful information in text by translating it into data that can be further analyzed. On the other hand, Natural Language Processing is a TM component that performs a type of linguistic analysis with the aim of facilitating a machine in understanding text.

2.1 Natural Language Processing

NLP is a field of research and application that investigates how computers may be used to comprehend and modify natural language text or speech in order to accomplish meaningful tasks.

The goal of NLP researchers is to learn how humans perceive and use language so that appropriate tools and techniques may be developed to help computer systems understand and manipulate natural languages. As a consequence of NLP as an ongoing area of research, it can be argued that there is no single agreed-upon definition around it.

According to Liddy (2001) NLP can be defined as a theoretically motivated range of computational techniques for evaluating and reproducing naturally occurring texts at one or more linguistic levels in order to achieve human-like language processing for a variety of tasks or applications.

This definition is in line with the one provided by Kapukaranov and Nakov (2015): in accordance with their point of view, Natural Language Processing, as the subfield of computer science, employs computational techniques aimed at learning, understanding, and producing human language content.

Furthermore Natural Language Processing has a mixed influence. Indeed, among the key contributors to NLP are: Linguistics, Computer Science and Cognitive Psychology. The first one is concerned with the development of formal, structural models of language and the discovery of language universals. Computer Science refers to developing internal representations of data and efficient processing of these structures. Last but not least, Cognitive Psychology

perceives language usage as a window into human cognitive processes, with the aim of modeling the use of language in a psychologically reasonable way.

Since its origin in 1950s, research about NLP has been focusing on several topics such as machine translation, information retrieval, text summarization, question answering, information extraction, topic modeling, and more recently, opinion mining.

Machine Translation (MT) is perhaps the most considerable way through which computers could facilitate human-human communication. MT is not essentially a task to be performed solo by machines. Actually, MT can be reconceptualized as an opportunity, for computer-supported cooperative work, to exploit human skills (Green et al., 2014). In such a scheme, machine intelligence aims at human-computer interface characteristics such as formulating good suggestions and responding effectively to human input, rather than completely replacing a human translator's skills and knowledge.

Information retrieval (IR) is finding material, usually documents, of an unstructured nature, usually text, that satisfies an information need from within large collections, usually stored on computers (Schütze et al., 2008).

According to this definition, IR was a niche activity in the past reserved for library professionals, paralegals, and other professional searchers. Now that the world has changed, hundreds of millions of individuals use an online search engine or scan their email every day to retrieve information.

2.2 Text Mining

Text mining is a multidisciplinary study subject that employs approaches from several domains such as computer science, linguistics, and statistics.

As stated by several researchers, Text mining refers to the usage of massive online text collections to identify new evidence and patterns about the world. Traditional text mining methods (Weiss et al., 2010) are inherited from the data mining community, such as document clustering (Zhao and Karypis, 2005) and document categorization (Sebastiani, 2002).

The concept behind both is to convert the text into a systematized format based on phrase frequencies and then employ typical data mining methods. Document clustering is commonly used to group news items or information service documents (Steinbach et al., 2000), while text classification algorithms are employed in applications such as e-mail filters and automatic classification of papers in business archives (Miller, 2014).

More advanced text mining techniques have been adopted in recent years for analyses in a variety of domains, such as linguistic stylometry (Girón et al., 2005), which calculates the probability that a particular author wrote a given text by observing the author's way of writing, or in search results for discovering rankings of documents from web search records of user activity (Radlinski and Joachims, 2007).

Recent improvements in document exchange have provided significant notions for automatic text management. The semantic web (Berners-Lee et al., 2001) disseminates defined frameworks for document sharing, allowing agents to execute linguistic procedures on them. This is accomplished by including metadata and by marking the content with labels. RDF (Manola et al., 2004) is one major format, and attempts to manage it have already been done in R (R Development Core Team 2007) with the Bioconductor initiative (Gentleman et al., 2004). This advancement provides a significant degree of adaptability in document exchange. However, as XML-based formats gain attention, programs must be capable of handling XML documents and metadata.

The advantage of text mining stems from the massive amount of useful data variable in texts that is not accessible in conventional data formats for a variety of explanations: for centuries, text has been the preferred method of keeping

data, and primarily time, individual, and budget restrictions prevent us from converting texts into well-organized structures such as data frames and tables.

Latent semantic research methods in bioinformatics (Dong et al., 2006), the use of statistical techniques for instantly inspecting jurisdictions (Feinerer and Wild, 2007), plagiarism detection in academic institutions and book publishers, machine assisted cross-language information retrieval (Li and Shawe-Taylor, 2005), or responsive web filtering learning through statistical inference are examples of statistical perspectives for text mining methods in study and business intelligence.

Help desk investigations (Sakurai and Suyama, 2005), assessing customers' needs by evaluating qualitative interviews (Feinerer and Wild, 2007) , instant ranking (Wu and Chen, 2005), fraud detection by inspecting claim alert, or decoding social media platforms for specific patterns such as concepts for new goods are other common scenarios.

Text mining capabilities are now available almost in each significant statistical programming package, and several data mining tools supply answers for text mining issues. Overall, the main stages of text mining involve acquiring, preprocessing, representing, extracting, analyzing, and interpreting text data to extract useful insights and knowledge from it.

2.2.1 Preprocessing

To acquire all words used in a particular text, a tokenization method is necessary, which involves splitting a text document into a stream of words by deleting all punctuation marks and replacing tabs and other non-text characters with single white spaces (Hotho et al., 2005). This tokenized representation is then utilized to do additional processing. The dictionary of a document collection is the set of various terms created by combining all text documents in a collection.

Being D the collection of documents and $T = \{t_1, \ldots, t_m\}$ the lexicon, in other words the group of all lexical items present in D, then tf(d,t) represents the absolute frequency of term $t \in T$ in document $d \in D$. Term vectors are denoted by $\vec{t_d} = (tf(d, t_1), \ldots, tf(d, t_m))$.

The set of words characterizing the documents can be decreased by filtering, lemmatization or stemming approaches to lower the size of the dictionary and

hence the dimensionality of the description of documents inside the collection.

Filtering methods remove terms from the lexicon and, as a result, from documents. Stop word filtering is a common filtering approach. Stop word filtering removes words with little or no content information, such as articles, conjunctions, prepositions, and so on. Additionally, words that occur infrequently can be said to have minimal information value to distinguish them across documents, and terms that appear infrequently are expected to be of minor statistical significance and can be eliminated from the lexicon.

Lemmatization attempts to match verbs to the infinite form and nouns to the singular form. Nevertheless, in order to do so, the word form, i.e. the speech component of each term in the text file, must be known. Because this labeling technique is typically time-consuming, stemming approaches are commonly used in practice.

Stemming approaches attempt to reconstruct the essential forms of words. A stem is a naturally occurring set of words that have the same, or very similar, meaning. Every word is represented by its stem after the stemming process.

2.2.2 Text Summarization

Text Summarization (TS) can be defined as the process of generating a shortened form of text document by maintaining relevant information and general significance of source text (Andhale and Bewoor, 2016). In the presence of a large text, *Automatic Text Summarization* (ATS) may turn out to be a significant way of searching relevant information in a short time with tiny efforts.

TS process entails three major phases:analysis, transformation and synthesis. The first one refers to the examination of source text and selection of properties. Transformation refers to the conversion of the result of the analysis and finally, the last step refers to the demonstration of the summary (Jones, 2007).

It is possible to classify TS approaches into two classes: extractive and abstractive. The first case, extractive summarization, is about extrapolating significant sentences or expressions from the source papers and clustering them in order to create a summary without implying a change in the source text. On the other hand, abstractive summarization refers to the comprehension of the source text by adopting the linguistic technique to understand and assess

the text (Khan and Salim, 2014). The aim of this second class is to create a simplified summary, by transfering data in a precise way.

It is possible to mention and adopt several methods when it comes to extractive summarization: *Term Frequency, Inverse Document Frequency, Cluster Based, Text Summarization with Neural Network, Text Summarization with Fuzzy Logic, Graph Based, Latent Semantic Analysis, Machine Learning Approach and Query Based Summarization.* These techniques have been further discussed by Allahyari et al. (2017) in their study.

Term Frequency (TF) and the Inverse Document Frequency (IDF) techniques show the importance of a term in a given document. Specifically, TF refers to the number of times a word occurs in the document. On the other hand, IDF is a measure that reduces the weight of words that appear repeatedly and, at the same time, increases the weight of words that seldom appear. The mathmatical summarization is presented as follow:

$$d^i = TF(t_i, d) \cdot IDF(t_i), \tag{2.1}$$

where:

d^i = TF-IDF of term t_i in document d.

TF = Term Frequency.

IDF = Inverse Document Frequency.

$t_i = i^{th}$ term.

d = document.

It can be argued that the greater the value of a Term Frequency, the greater its significance in a given document.

Cluster Based (CB) method is particularly relevant when the document for which summary is being supplied concerns different topics. Summary sentence is chosen based on the existing association between the sentence and the theme of the cluster. The latter, is represented by high frequency term. Hence, CB method produces highly significant summary, to the given query or document topic.

2.3 Topic Modelling

Topic models refers to Bayesian statistical frameworks in which unstructured information, typically a collection of text materials, are organized according to intrinsic themes named topics with multinomial distributions on strings (Zhao et al., 2015). Topic modeling assumes that the corpus comprises a specific number of underlying topics and that every document includes several topics in variable amounts. Various topic models have been proposed, such as Latent Semantic Indexing (Deerwester et al., 1990), Probabilistic Latent Semantic Analysis (Hofmann, 1999) and the so called Latent Dirichlet Allocation (Blei et al., 2003).

The latter will be further discussed as it will be employed for the present study. The underlying implication of Latent Dirichlet Allocation, the most widely used topic modeling, is that each document may be expressed as a probabilistic distribution spanning hidden topics, with each topic distribution having a similar Dirichlet baseline.

Assuming a corpus D of M documents each with N_d words $d \in (1, ..., M)$ LDA predicts D using the generating procedure outlined below:

1. Choose a multivariate statistical distribution ϕ_t for topic $t \in (1, \ldots, T)$ from a Dirichlet distribution with component β.

2. Choose a multivariate statistical distribution θ_d for document $d \in (1, \ldots, M)$ from a Dirichlet distribution with parameter α.

3. For a word $w_n, (n \in \{1, \ldots, N_d\})$ in a document d:

   (a) choose a topic z_n from θ_d

   (b) choose a word w_n from ϕ_{zn}

The probability of observed data D is determined and maximized as follows in order to derive the unknown variables and hyper parameters:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\sum_{n=1}^{N_d} p(z_{dn}|\theta_d) p(w_{dn}||z_{dn}, \phi) P(\phi|\beta) \right) d\theta_d d\phi \quad (2.2)$$

However, problems emerge for unstructured document sets where the content and amount of specific themes are uncertain at the outset. This implies the ideal number of topics is unclear, and changing amounts of topics will almost certainly result in significantly different corpus architecture (Zhao et al., 2015).

A lack of topics may result in a model that is too narrow to provide precise classifiers. An excessive amount of subjects, on the other hand, may result in a model that is particularly complicated, rendering interpretation and personal evaluation quite hard.

Chapter 3

Data

3.1 Sample Selection and Description

For this study, sample selection was conducted based on Italian newspapers. According to the data provided by "Agenda del Giornalista" [1] , the number of newspapers officially registered in Italy in 2022 amounts to 3.444, with 89 news agencies, 98 newspapers and 745 specialized and technical radio and television newspapers.

Despite the large amount of newspaper listed in Italy, this study only takes as a point of reference the major national newspapers. The list, made available online by the association "Accertamenti Diffusione Stampa" [2] ,included several sections and divided newspapers for category: national newspapers, online newspapers, sport newspapers and financial newspapers.

The list contains a total number of 24 newspapers as a starting point. However, it is crucial to highlight some considerations: among the above-mentioned national newspapers, some of them were excluded from the analysis for several reasons. Firstly, the domestic newspaper Domani, published by Società Editoriale Domani SpA and distributed by RCS mediagroup, was omitted from the list because of its foundation in May 2020. Therefore, Domani had not yet been founded when the pandemic broke out in Italy. Similarly La Ragione, a national newspaper of liberal-democratic and pro-European inspiration founded in June

[1] See https://stampafinanziaria.com/quanti-sono-i-giornalisti-in-italia/
[2] See https://giornali.it/quotidiani-nazionali/

2021, was excluded from the list.

Moreover, since the focus of the analysis is Italy, the international newspaper Metro, founded in 1995 in Sweden, was excluded from the list. Additionally, Il Dubbio, an Italian daily newspaper directed by Davide Varì, founded in December 2015 and published by Edizione Diritti e Ragione, was omitted from the list.

After careful considerations for the implementation of the analysis under study, the list that initially included 24 newspapers has been now reduced to 20 newspapers. Table 3.1 shows the list of selected newspapers and their political orientation.

La Repubblica, La Stampa and Il Secolo Decimo Nono share the same editor: GEDI Gruppo Editoriale. The latter, not only publishes the three newspapers but also several local magazines.

In the list, there are three other newspapers that have in common the same editor: this is the case of Il Messaggero, Il Mattino di Napoli and Leggo edited by Caltagirone Editore. The holding strengthens its position in the Italian publication market by investing in title graphic redesign, launching online editions of all newspapers, and converting Leggo into the first Italian social press. Il Manifesto and Il Foglio are the only two national newspaper to be published by a cooperative society: Il Nuovo Manifesto-Società Cooperativa Editrice and Foglio Quotidiano Società Cooperativa.

After a short description of selected Italian Newspapers, further considerations were made regarding the political orientation of each magazine: because of the scarcity of official sources and explicit statements by the various newspapers, observations were drawn on the basis of the ownership of each magazine.

It is possible to highlight that there are no leftist newspapers at the exception of Il Manifesto. At the opposite, the majority of national Italian newspapers are owned by financial groups with lobbying interests and therefore they tend to be more or less right-oriented conservatives. Moreover, La Notizia and Il Fatto Quotidiano can be classified as liberal.

3.2 Scraping of Tweets

Modern society is increasingly reliant on data, emphasizing the relevance of accurate data collection in order to make important decisions and draw conclusions.

There are several methods of collecting data: surveys, interviews, focus groups, social media and many more. A decision was made to collect data from Twitter. The latter has become, over years, a useful tool for journalists around the world: not by chance Twitter is categorized as "news app" on some app stores rather than a mere social network. In light of this trend, Twitter is a great source for data collection and especially for the present study: according to statistics, more than 80 % of young journalists obtain their news via Twitter[3].

Despite not being the most popular social media in terms of monthly active users, most popular platforms such as Facebook and WhatsApp do not make their data as widely available as Twitter does. It may be claimed that no other social media network has the infrastructure that Twitter has: its singularity is given by the fact that it offers almost 100 % of its data through APIs. [4]

Although being a good source for data collection, Twitter may be considered as a subset of all conceivable information, therefore it only provides a partial picture of the phenomenon; nevertheless, we assume that it is sufficiently representative of all information provided during the pandemic in Italy.

There are several methods to collect data from Twitter: the web client, available without the need of a Twitter account, is the easiest method to gather data from Twitter. However, this method has changed over time. Indeed, in the past the web client method provided tweets from the previous seven days. Starting from July 2016 it is capable of retrieving tweets through the Advanced Search feature. The output can be placed into a word processor for additional treatment. This method of collecting Tweets takes time, and irregular line structure makes the automated analysis challenging.

A more complicated method requires accessing the Twitter APIs via software

[3] See `https://www.websiterating.com/it/research/twitter-statistics/#references`

[4] Application Programming Interface: in basic terms, APIs are a set of functions and procedures that allow for the creation of applications. They access the data and features of other applications, services, or operating systems.

and programming tools. There are two main API models available: REST and Streaming [5] . In the first case, the client will execute one or more queries to the API using a REST architecture, and Twitter's server will react by providing desired information to the topical application. Having specified lexical items or hashtags, from a specific time period, the REST API may be a valuable resource for gathering data with certain, pre-defined criteria for a corpus-based investigation.

On the other side, clients who use the Streaming API, keep the connection open for as long as they want. Huge amounts of data can be collected very rapidly based on the criteria of the API request. In order to access the Twitter API, many packages throughout many computer languages have been created [6]. For instance, twitteR (Gentry et al., 2016) and StreamR (Barberá and Rivero, 2015), can be utilized for data collection in R environment. The first package, TwitteR, queries the REST API many times and gathers information into a hierarchical structure that may be quickly translated to a R dataset for additional analysis, saved as a text document, or stored to a database. The second one , StreamR package, allows the client to use the Streaming API. Before the scraping phase, several packages have been installed and loaded to the library: "tidyverse", "magrittr", "lubridate", "httr", and "jsonlite". We excluded "rtweet" and "twitteR" packages as not necessary for this stage.

The scraping of the tweets, through R software, has been organized as follow: creation of a script for each newspaper. As a result, 20 R scripts were created for this study. Moreover, for each magazine the objective was to download all the tweets made in the time span. Due to constraints imposed by Twitter , tweets were downloaded daily [7]. Since the reference period of this analysis is Covid-19, a daily sequence covering the entire duration of the pandemic in Italy has been created: the sequence goes from February 1, 2020, when the first two cases of infection were confirmed in Italy [8] , to February 28, 2022, which corresponds to the beginning of the war between Ukraine and Russia. The underlying assumption is that after the outbreak of war, the focus on information in newspapers

[5] see https://dev.twitter.com/overview/documentation
[6] see https://dev.twitter.com/overview/api/twitter-libraries
[7] see https://developer.twitter.com/en
[8] The first two cases confirmed are two Chinese tourists hospitalized since January 29 at the National Institute for Infectious Diseases "Lazzaro Spallanzani".

shifted from the pandemic to the war.

Subsequently the vector for queries was created. The latter requires several components to be specified: from which user, start time and end time, max results and tweets field.

User refers to the name adopted by each magazine in Twitter. `Max results` refers to the maximum number of tweets we can download. At a first stage, this parameter was set to 150, but it was subsequently doubled to 300 in the hope that no newspaper posted more than 300 tweets per day.

Tweets field refers to the fields we save: `id, text, authorid , createdat and entities` were saved in the present study.

Consequently, the iteration process was performed for each magazine: the duration of the procedure was about two hours per iteration. At the end of the loop is the final phase of the combination and the saving of the object. This procedure has been applied to every single magazine.

The iteration process was successful for all the newspapers at the exception of two: Il Resto del Carlino and Italia Oggi. In the latter two cases, the number of total tweets downloaded did not coincide with the daily sequence created at first.

As far as il Carlino, the first available tweet was dated "20-05-2021" while the last one was dated "18-02-2022" with a total amount of 267 tweets over the whole period. Similarly, it the case of Italia Oggi, the first tweet downloaded was dated "13-03-2020" whereas the last one was dated "23-02-2022", with a total amount of 1802 tweets over the whole period.

Checks on individual twitter accounts showed that during the reporting period the two newspapers had posted regularly. For this reason they were excluded for further processing in the present analysis. A total number of 793.739 of tweets has been collected.

3.3 Keywords Filtering

As the focus of the present study is Covid-19 in Italy, collected data was filtered with keywords related to the pandemic in order to rely on more accurate and relevant results. Therefore, the list included the following Italian words: "coronavirus, covid, quarantena, pandemia, lockdown, tampon*, assembrament*, restrizion*, isolamento, batterio, contagi*, distanziamento, mascherin*, epidemia". The total number of tweets related to the pandemic is 71.529.

This decision is consistent with past studies in the field, which adopted a similar approach. Lamsal (2020) used the keywords "corona" and "coronavirus" in his research. The latter were used to filter the Twitter stream. Furthermore, as the pandemic progressed, a number of additional possible keywords emerged. Indeed, new keywords were gathered and added to the filtering section. All of this contributed to the development of a large-scale Twitter dataset containing over 310 million COVID-19 specific English language tweets.

Wang et al. (2020) used the Twitter streaming API to collect Dutch tweets, which were then filtered with a set of pandemic-related keywords such as "corona," "covid," "stay home," and "facial mask." This enabled them to analyze Dutch public sentiment on governmental COVID-19 measures using text data gathered from three different online media sources.

Other words related to the pandemic could have been included in our list and this would have led to a greater number of observations per newspaper. However, taking into account the limited number of keywords adopted by similar investigations in the field, a decision was made to stick with the above-mentioned keywords as they are considered exhaustive for the present investigation.

Considering that tweets have been downloaded individually for each newspaper, all data frames have been initially stored in a list and merged by using rbind function afterwards. Hence, joining multiple data.frames is underpinned by the idea of facilitating the next step of the present study: exploratory data analysis.

Chapter 4

Findings and Discussion

4.1 Exploratory Data Analysis

Exploratory Data Analysis is the fundamental method that entails various preliminary investigations on data in order to comprehend the data and get as many insights from it as possible via the use of descriptive statistic and graphical visualizations.

Generally speaking, exploratory data analysis entails the following steps: summarization of the data set via summary statistics, visualization of the data set by using charts, and identification of null data. By performing the above-mentioned steps, it is possible to have a clear idea of how values are distributed in the dataset and detect any issues.

Exploratory data analysis have been conducted employing the tidyverse package from R software.

In the present analysis the data set includes 5 variables displayed as follows: `id, text, author.id, created.at and entities.hashtags.` The first three variables are character, also known as string variables. They contain text-based information that the system recognizes. Letters, special characters and even numerals may be included. Continuing with the description of our data set, the fourth variable is created.at: this is a date-times variable which resulted from the day sequence created during the data collection process.

Last but not least, is the entities.hashtag variable which is a list containing the hashtags used per each tweet. NULL values appeared for tweets with no

hashtags at all.

Among the magazines analyzed in the sample, it is possible to notice that Repubblica is the one with the highest number of observations (172.180). Despite being the second most read newspaper in Italy [1] , it holds the record in the world of Twitter. This may be explained by the fact that Repubblica joined Twitter in January 2009 while Corriere della Sera, the most read newspaper in Italy according to the ADN [2] , joined Twitter two years after in October 2011. Moreover, there is also a significant difference in terms of number of followers: Repubblica has 3.5 million followers while Corriere has 2.7 million followers.

The situation does not change if we take into consideration data after filtering them with keywords related to the pandemic. Repubblica is still the newspaper with the highest number of observations (12.059).

At the opposite, Il Secolo Decimo Nono is the newspaper with the lowest number of observations (4.217). As a result, the newspaper founded in Genova is not even mentioned in the top 20 best-selling newspaper in Italy. It has joined twitter in June 2009 and counts 62.4 thousands followers. However, the situation does change in this case if we have a look at data after filtering: Il Secolo Decimo Nono is not the newspaper with the lowest number of observations. Il Manifesto, with 273 observations, occupies this role. This means that, among the newspapers in the sample, Il Manifesto is the one that tweeted less on Covid-19 in Italy.

At this point it is legitimate to answer a question: to what extent did each Italian newspaper talk about Covid-19 in the reporting period? To answer this question, it is possible to calculate the ratio between the number of tweets filtered with keywords and the total number of tweets without filtering. Values, in percentage terms, are shown in Table 5.2. In the case of Libero newspaper, 25 % of tweets were related to the pandemic. Even though Repubblica is the newspaper with the highest number of tweets before and after filtering with keywords, only 7 % of its tweets were related to Covid-19. Moreover, almost 22 % of tweets were related to the pandemic for Il Tempo. With regard to Il Manifesto, 4 per cent of its tweets were related to the pandemic and this is in

[1] See http://www.data24news.it/media/top20-dei-quotidiani-piu-venduti-al-primo-posto-corriere-della-sera-ultimo-quotidiano/

[2] Accertamenti Diffusione Stampa is the the association that certifies the data of circulation and circulation of the daily and periodic press of any species published in Italy

line with the fact that, among the newspapers, Il Manifesto tweeted less on the pandemic in Italy. Overall, it is possible to state that, in percentage terms, "only" 9 % of Total Tweets posted by italian newspapers are related to the pandemic.

Below a time series of all tweets displayed per day is proposed. As we can observe, the majority of selected Italian newspapers posted less than 300 tweets per day. This is in line with the choice adopted of 300 as a reasonable threshold.

However, at this stage some issues have emerged: if we look at the time series of Repubblica, it is possible to observe that starting from February 2020 up to February 2021 a part of tweets was lost. This result stems from the fact that Repubblica posted more than 300 tweets per day, exceeding the specified limit. Furthermore, as indicated by the fictitious line in the graph, Twitter database for Il Giornale has been erased for the period January 2020 to February 2021. Furthermore, if we observe the time series of all tweets related to the pandemic, it is possible to notice that the majority of selected Italian newspapers follows a common pattern. This tendency is hardly surprising if we take into consideration the trend of the pandemic itself in Italy. It is not surprising as well that the issue of all tweets related to Il Giornale has resurfaced at this point. Another noteworthy point is that Il Tempo appears to have posted more tweets related to the pandemic at the end of 2021. This is in contrast with the common path followed by the majority of selected Italian newspapers which showed a pick of posted tweets during the first wave in Italy.

Another potential aspect to analyze in the present study is the identification of the period when the highest number of tweets was recorded.

To answer this question, a time series has been proposed. Figure 4.3a provides an overview of the total number of tweets displayed per day while Figure 4.3b shows the total number of tweets related to the pandemic, displayed per day.

As we can observe from figure 4.3b the highest number of tweets related to the pandemic was registered in March 2020, when over 7.600 tweets were posted. This may be explained by the beginning of the first devastating wave for Italy with the national lockdown. On March 11, 2020, the World Health Organization declared Sars-Cov-2 infection to be a pandemic, and Italy was the very first European country to be impacted by the virus, necessitating immediate action to control transmission across the country and provide care (Boriani et al., 2023). March 2020 was also the month in which the first actions were taken to tackle the pandemic: the first Covid-19 vaccine trials began on March 16, 2020, just over two months after the viral sequence of Sars-Cov-2 was identified. The first mrna vaccine and cansino-viral vector vaccine were used in modern clinical studies. All of this may explain why the pandemic has received more attention in Italy in March 2020.

The date of March 18, 2020 coincides to the day when the maximum amount of tweets about the pandemic was recorded in Italy. As a result, it is not unexpected that this day occurs during the week with the highest number of tweets (2020-03-15 - 2020-03-22).

Nonetheless, something "exceptional" beyond March 18, 2020 may explain the record. The National Day in commemoration of all the victims of the coronavirus epidemic, also known as the National Day for the Victims of COVID-19, is an Italian national holiday commemorated on March 18 in memory of those who died in Italy as a result of SARS-cov-2 during the COVID-19 pandemic.

The Chamber of Deputies established it on July 23, 2020, and the Senate of the Republic established it on March 17, 2021. The chosen date was determined on the same day in 2020 when the Italian Army's heavy vehicles assisted in the removal of hundreds of coffins deposited at the monumental cemetery of Bergamo, whose column of means triggered a lot of interest in public opinion during the full first wave of the coronavirus.

As a confirmation of what has been mentioned above, Figure 4.4 shows that more tweets related to the pandemic were created during the week (Monday - Tuesday) rather than on the weekend (Friday - Sunday). Daily epidemiological situation updates were issued via news conferences, press releases, and website updates. In general, the days of the week with the most covid communications

in Italy were those when significant events or crucial choices were made by government officials. For example, during the first wave the Italian government issued a decree on March 9, 2020, imposing the closure of schools, universities, museums, theaters, and other public spaces across the country in reaction to the COVID-19 pandemic. Following that, on March 11, 2020, a decree was published that imposed additional limitations, including the closing of non-essential establishments and the requirement for grocery stores and supermarkets to close at 18:00. Moreover, on March 21, 2020, the decree "Io resto a casa" is passed, requiring all Italian citizens to stay at home save for job, health, or need. On May 4, 2020, the Italian government begins the first phase of the country's gradual reopening, with the restart of production activity and the resumption of some commercial activities.

It is crucial to recall that Italy has recorded 5 waves during the pandemic. The latters have been defined as follows: the first wave, goes from March 2020 to May 2020; the second wave goes from October 2020 to January 2021; the third wave goes from February 2021 to May 2021; the fourth wave goes from June 2021 to October 2021; and the fifth wave goes from November 2021 to February 2022 [3]

As observed from Figure 4.3b the first wave was the period with the highest number of tweets. At the same time it is possible to notice a pattern that seems to be consistent with the various waves: although to a lesser extent than the previous wave, the number of tweets related to the pandemic increases at each wave. At the end of October 2020, despite a summer in which cases in Europe were kept to a minimum, France, Spain, Germany, and then Italy registered an increase in the number of infections. In addition to the total number of cases, which is incomparable between the first and second waves, Italy was experiencing widespread infection. This is in contrast to March 2020, when the majority of illnesses and deaths were registered in the north of Italy. The number of tweets related to the pandemic started to decrease from the end of 2020. Indeed, Pfizer BioNTech created the first Covid-19 vaccine in history in mid-December 2020 and consequently first vaccine injections took place simultaneously in all member countries. By the end of the fifth wave it is possible to notice that the

[3]Data from Italian Civil Protection Department. https://opendatadpc.maps.arcgis.com/apps/dashboards/b0c68bce2cce478eaac8 2fe38d4138b1.

number of tweets related to the pandemic is extremely low compared to the previous waves. This coincides with the beginning of the war between Russia and Ukraine, a period in which the main topic of news in Italy was no longer Covid-19 but the struggle between the two nations, having far-reaching consequences for the entire world.

Among the variables included in the data set there is "entities.hashtags" , containing the list of hashtags used per each tweet. Hashtags, denoted by the appropriate mark, are used on Twitter to classify words or subjects. This feature was developed by Twitter and enables users to rapidly explore topics of interest.

At this point we will try to identify the most used hashtag. Fig 4.5 shows the list of top 20 popular hashtags from tweets filtered with keywords related to the pandemic: coronavirus has been used 12.233 times, covid 4.294 times, covid19 has been used 3.199 times, iltempoquotidiano 3.007 times, vaccino 1.883 times , lockdown 1.304 times , conte 1.042 times , coronavirusitalia 957 times, governo 855 times , roma 751 times, greenpass 658 times, lombardia 590 times, edicola 578 times, italia 554 times, draghi 466 times, pandemia 461 times, bollettino 428 times, fattoquotidiano 424, salvini 420 times and Speranza 371 times. Most of the hashtags are strictly related to the pandemic while some others are related to the name of the newspaper or to the name of political figures who played a primary role during the pandemic. Indeed, in the wake of the 2021 government crisis, Giuseppe Conte and Mario Draghi assumed the presidency of the council. A shift in executive leadership that, while not involving a turnover of the health ministry's political leadership, did involve changes in administrative leadership.

4.2 Text Mining Results

This section explores the practical implementation of the methodology and theoretical framework discussed in Section 2. Several packages have been used for Text Mining: `"tm", "Snowball", "worldcloud2" and "ggplot2".`

As a first step, a large list containing 18 elements, each of them corresponding to a single national newspaper, was created. Subsequently collected data was loaded as a Corpus, a huge collection of texts used in data analysis for checking frequencies and verifying grammatical structures within a certain language area.

Characters such as "/" and "@" were replaced with white spaces by using "tmmap" function. Additional steps involved the conversion of the text to lower case, the removal of numbers and the removal of italian common stopwords. The latter, does include general italian words but in order to obtain more robust results, a character vector was created specifying our own stop words. The vector includes words related to the name of the newspapers as well as words like "tco" and "dopo" as non significant for the present investigation. Furthermore, punctuations and extra white spaces were eliminated. Finally a term-document matrix was built after all the previous steps.

As we can observe from Table 4.2 , the term-document matrix is composed of 121.402 terms and 18 documents. A 90 per cent dispersed Term-Document Matrix (TDM) indicates that 90 per cent of the cells in the matrix have a zero value, indicating that most terms appear in only a few documents and most documents contain only a few terms. Just 10 per cent of the cells in the matrix have a value different from zero.

Table 4.3 shows the top ten most frequent words. As exptected, the words coronavirus and covid are the ones with the highest frequency meaning that they appear more frequently in the document collection. Considering that the word coronavirus is the one with the highest frequency, a decision was made to visualize the plot of terms correlated with this word. Hence, results are shown in Figure 4.6.

As expcted, words such as "scoperta" and "wuhan" are highly associated (0.94) with the word "coronavirus". This is in line with the fact that the SARS-cov-2 coronavirus, which causes COVID-19 disease, was discovered in late 2019 in Wuhan, Hubei Province, China. Yet, the precise origin of the virus is still

unknown.

According to some studies, the virus may have originated naturally from an animal, specifically the bat. However, there is also a theory that the virus may have been accidentally released by a Wuhan laboratory studying coronaviruses. This theory has been the subject of debate and investigation by international bodies.

4.3 A focus on Political Orientation

It is critical to emphasize that the Italian press is free and independent, and that diversity of perspectives and voices is a key value of democracy in italy.

Despite this, major national newspapers in Italy are often regarded as inclined to specific political currents. Therefore, as stated in Section 3 it is crucial to remember that the political orientation of newspapers can vary depending on the director or owners, and that columnists frequently share views that differ from the newspaper's editorial line.

In light of this, the present section aims at exploring potential differences which might arise from different political orientation. As a first step, among the italian national newspapers, a decision was made to select one journal per political orientation.

For the leftist part, Il Manifesto is the only one with such orientation in Italy. The latter is a daily newspaper published in Rome that was created in 1969. The newspaper considers itself to be a leftist and anti-capitalist news source.

Il Manifesto's editorial style is characterized by a significant emphasis on social and environmental issues, the battle against inequality, and the protection of workers' rights. The newspaper is critical of the current economic and political system and calls for a fundamental social reform.

In terms of politics, Il Manifesto has historically been associated with the Italian Communist Party and leftist groups opposed to the dominant political and economic system. However, in the years following the fall of the Berlin Wall, the newspaper adopted a more autonomous stance and broadened its scope to include anti-militarism, feminism, and migrant rights.

At the opposite, since the majority of italian newspapers tend to be right-oriented conservatives, Il Secolo Decimo Nono was selected among all of them. The latter is a Genoa-based Italian newspaper that was created in 1886.

In terms of political orientation, the journal considers itself independent and not affiliated with any political party. Yet, the newspaper's editorial line is frequently considered similar to the stances of the Christian Democrats and the center-right, notably on economic and foreign policy issues. Furthermore, the journal has deep roots in the Liguria region and frequently covers problems

pertaining to its social and political realities.

Finally, for the liberal orientation, Il Fatto Quotidiano was selected. This decision stems from the fact that certain politicians and the media have frequently criticized Il Fatto Quotidiano for its partiality, but the newspaper has always claimed to be an independent journalistic institution free to express their thoughts in a critical and objective manner.

Figure 2 shows the worldcloud of terms related to the pandemic which have been employed by the three national newspapers.

At the outset, it is possibile to notice some similarities: the most used term for each one of them is "coronavirus", followed by "covid". This result is not surprising considering the topic of the present analysis. However, it is possible to highlight the presence of slight differences. Indeed, Il Manifesto, in quality of letf-oriented newspaper, tends to adopt a style of communication based on progressive and internationalist ideas, typical values of leftist parties. In light of this, it is not surprising that Il Manifesto decided to use more the word "lockdown" rather than "quarantena" as it was the case for Il Secolo Decimo Nono. According to the above mentioned statement, Il Secolo Decimo Nono has deep roots in the Liguria region: this is in line with the fact that Il Secolo, as a right oriented newspaper, tends to employ a style of communication based on patriotic, conservative and religious values. This can be seen in Fig 2b, where words such as "coronavirusliguria" , "genova" appear in the wordcloud. Furthemore, as we can observe from 2c, objective terms such as "isolamento", "restrizioni" emerge in the wordcloud, as a confirmation to the liberal orientation of Il Fatto Quotidiano.

As shown in Table 3.1, it is possible to observe that several italian newspapers share the same editor. This is the case of Il Mattino, Il Messaggero and Leggo for Caltagirone Editore as well as Il Secolo Decimo Nono, La Repubblica and La Stampa for GEDI Gruppo Editoriale.

4.4 Topic Modelling: Latent Dirichlet Allocation

The process begins with the typical examination of the corpus data. As we are dealing with tweets which are extremely short texts, a decision was made to concatenate single documents in order to obtain longer textual entities for topic modelling.

Italian stopwords were removed during text preparation because they tend to appear as interference in the LDA model's estimated themes. Moreover customized stopwords were subsequently specified as a character vector and removed afterwords.

Moreover, special characters were deleted and words were stemmed and transformed to lowercase characters. In order to compute a topic model on the processed data, a document term matrix was built. Furthemore, as a step to accelerate the model computation, a specific minimal frequency was set equal to 5. In other words, all terms that appear less than 5 times in all documents are deleted from the DTM. The imposition of a minimum frequency helps to reduce noise in the DTM, eliminating words that are too rare and therefore may not be significant for the present analysis.

The number of topics K is the most critical parameter to define in advance for parametric systems such as Latent Dirichlet Allocation (LDA). Four approaches were used in this case in order to identify the optimal number of topics: namely the metrics "CaoJuan" , "Arun", "Griffiths" and "Deveaud" are included in the package `ldatuning` as discussed by Nikita (2016).

Specifically, the first two metrics show a low amount of subjects, while the others have a high number of themes. Optimally, several methods should converge and show peaks and dips respectively for a certain number of topics. Results are shown in Fig 3

Despite a suggestion of 5 as optimal number of "k" topics, a decision was made to vary this parameter k from 2 to 5. As a result, given the small variances in scores, we decided on maintaining $k = 2$ to ensure that the information remains interpretable. Figure 4 shows the 20 terms that are most common within each topic.

Despite the presence of common terms such as "cornavirus" , "italia", "quarantena", "lockdown" and "mascherina" it can be argued that there are some terms which appear only in a topic. This is the case of terms as "casi", "tamponi", "dati", "milioni", as they only appear in Topic 2. At the opposite words as "governo", "aggiornamento", "greenpass", "vaccino" and "novax" only appear in Topic 1. The presence of such specific words in Topic 2 might suggest that it refers to communications provided by the italian government during the pandemic, through the announcement of data related to the development of the pandemic. At the opposite, specific words appearing in Topic 1 might suggest that it refers to communication provided by the italian government with respect to the vaccination campaign during the pandemic in italy.

Another option entails the possibility to look at the words with the highest difference in Beta between topic 1 and 2. Results are shown in Table 4.4 and Fig 4.10. In light of this, it is possible to observe that Topic 1 includes words such as "astrazeneca", "pfizer", "varianteomicron" and "bassetti". At the opposite Topic 2 is characterized by words as "andamento", "aggiornamento", "epidemia" and "news". These results provides an additional understanding of the two topics identified by the algorithm at the outset of the present section. Therefore, topic 2 might refer to the provision of numerical data on the pandemic trend by the central government to prevent the spread of the virus. At the opposite, topic 1 might refer to the communication released during the pandemic by national and regional authorities with respect to the vaccination campaign and the mutation of Covid-19 virus.

Name	Foundation	Editor	Political Orientation
Avvenire	1968	Avvenire Nuova Editoriale s.p.a.	Conservative
Corriere della Sera	1876	RCS MediaGroup	Conservative
Il Fatto Quotidiano	2009	Società Editoriale Il Fatto s.p.a.	Liberal
Il Foglio	1996	Foglio Quotidiano Società Cooperativa	Conservative/Right
Il Giornale	1974	Società Europea di Edizioni s.p.a.	Conservative/Right
Il Manifesto	1969	Il Nuovo Manifesto-Società Cooperativa Editrice	Left
Il Mattino	1892	Caltagirone Editore	Conservative
Il Messaggero	1878	Caltagirone Editore	Conservative
Il Resto Del Carlino	1885	Editoriale Nazionale	Conservative
Il Riformista	2002	Romeo Editore s.r.l.	Conservative
Il Tempo	1944	Gruppo Angelucci	Conservative/Right
Il Secolo XIX	1886	GEDI Gruppo Editoriale	Right
Il Sole 24 ORE	1965	Confindustria	Conservative
Italia Oggi	1986	Gruppo Class	Conservative/Right
La Notizia	2013	La Notizia s.r.l.	Liberal
La Repubblica	1976	GEDI Gruppo Editoriale	Conservative
La Stampa	1867	GEDI Gruppo Editoriale	Conservative
La Verità	2016	La Verità s.r.l.	Conservative/Right
Leggo	2001	Caltagirone Editore	Conservative
Libero Quotidiano	2000	Editoriale Libero s.r.l.	Conservative/Right

Table 3.1: List of selected Italian Newspapers and corresponding political orientation.

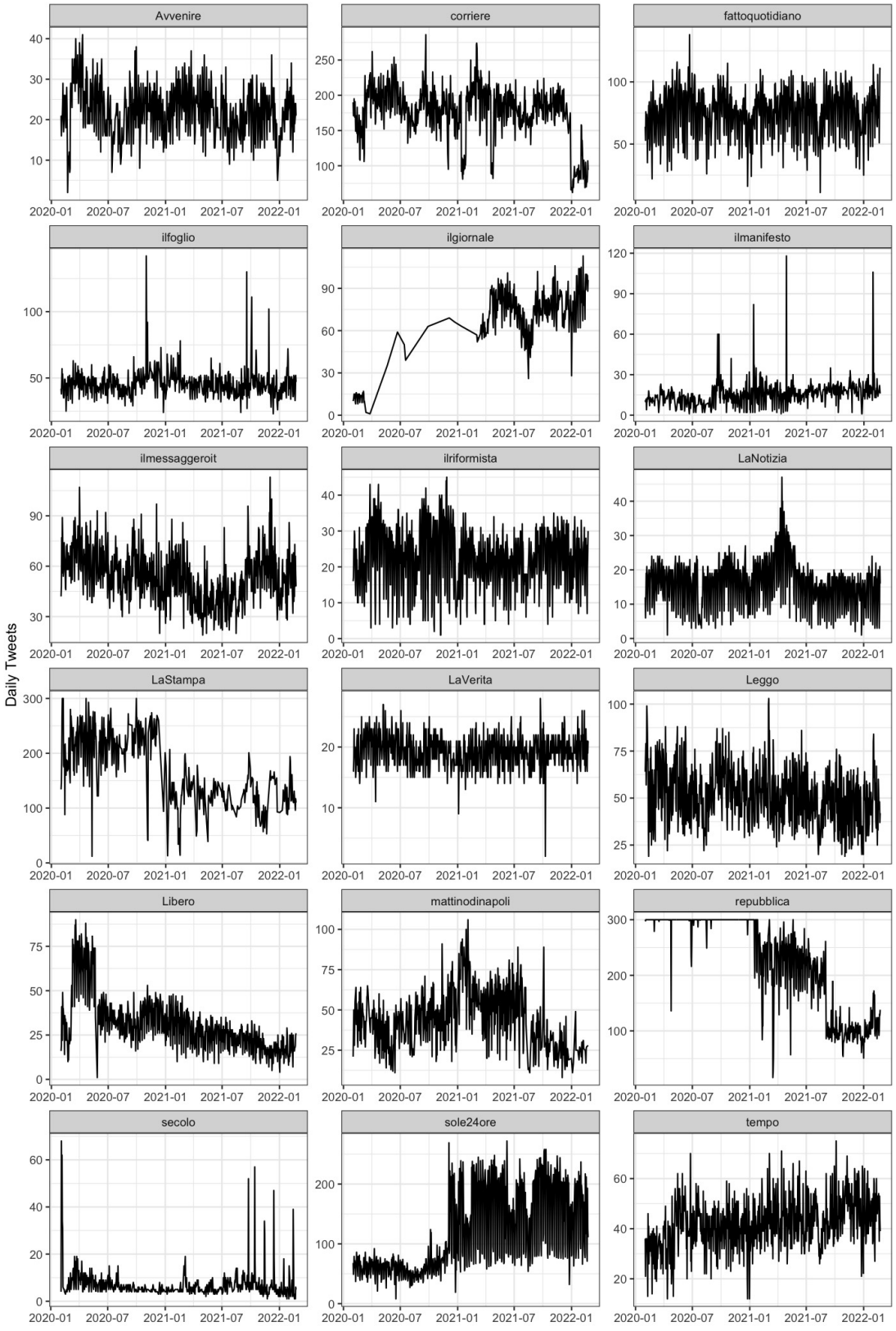


Figure 4.1: Time Series of all Tweets displayed per days

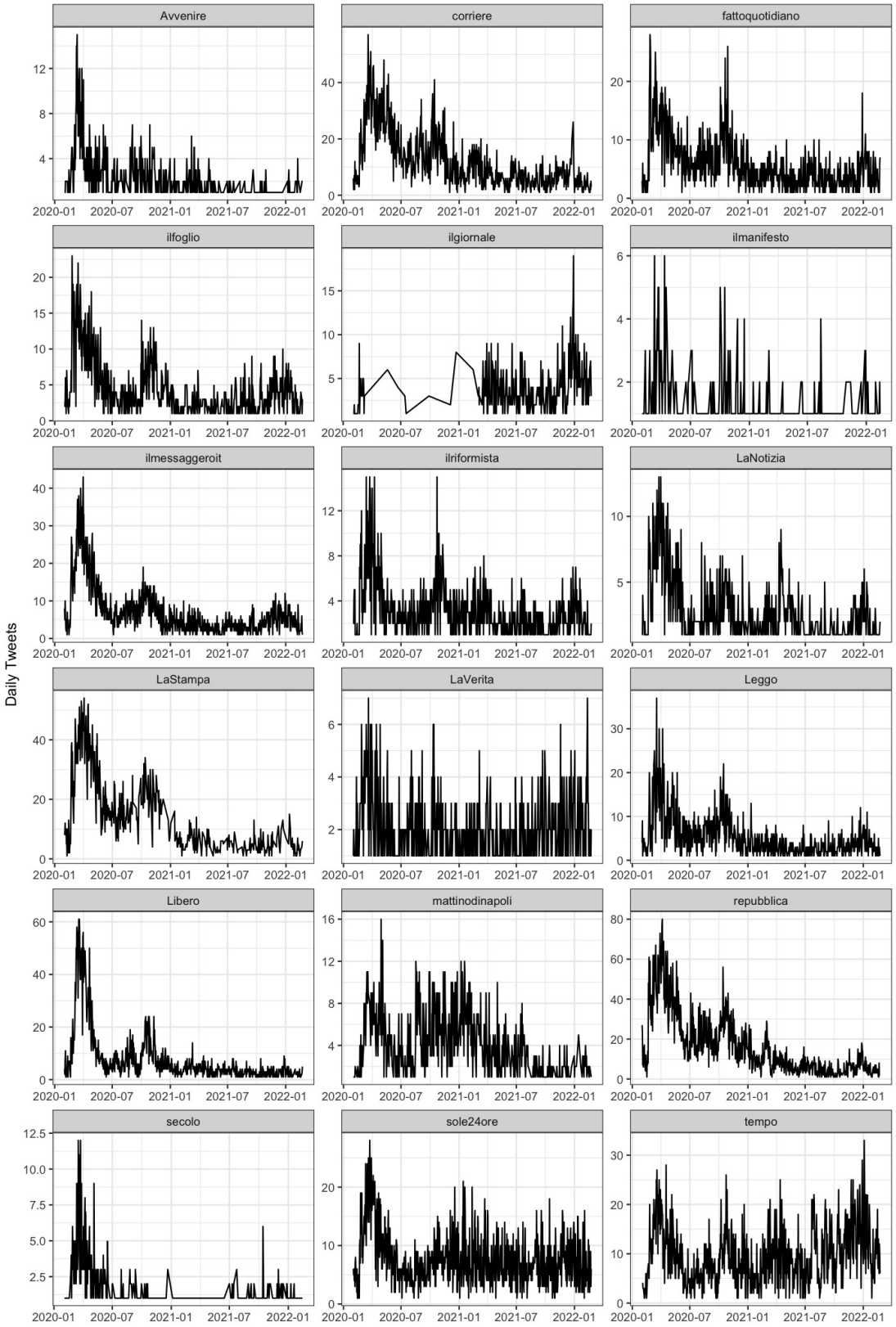
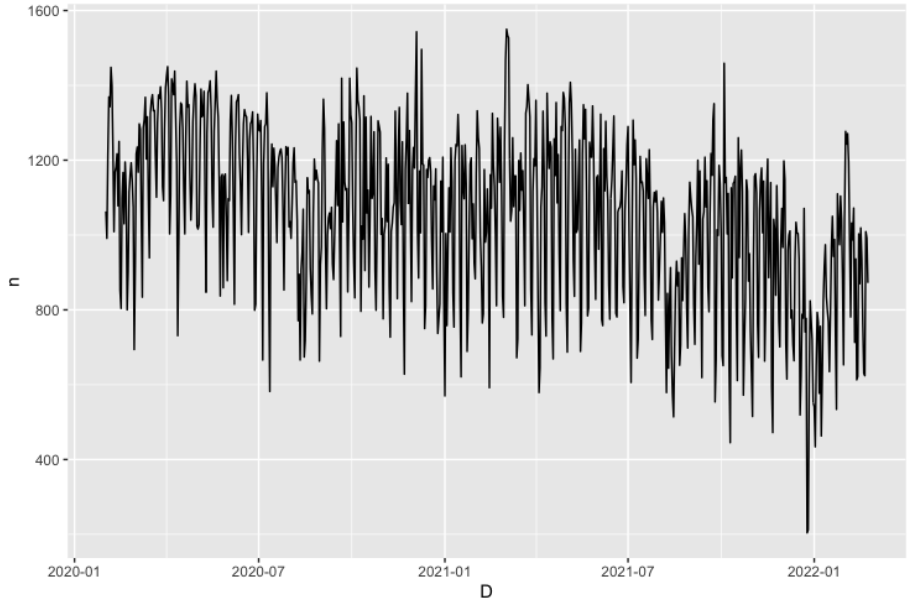


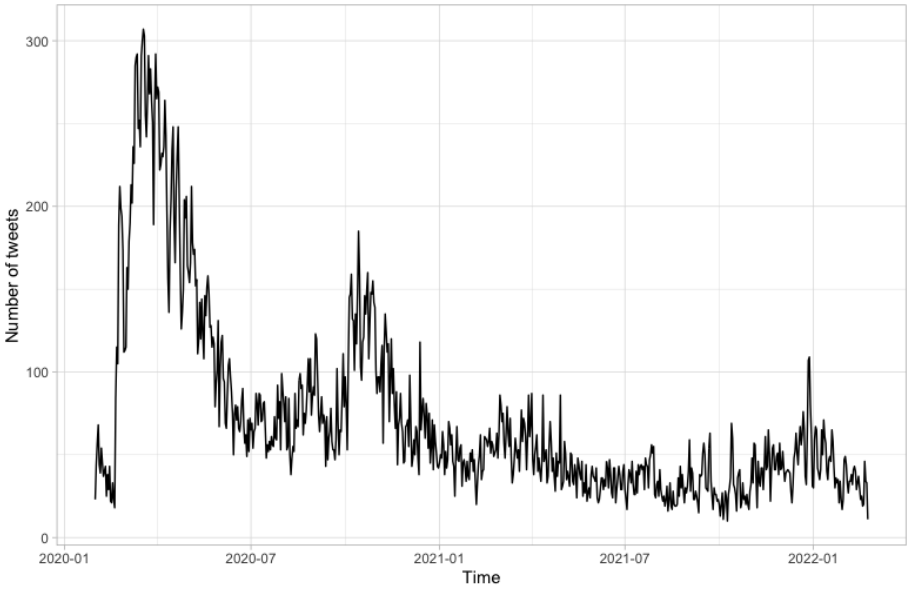
Figure 4.2: Time Series of all Tweets related to the pandemic displayed per days

Name	Tweets before filtering	Tweets after filtering	Ratio(%)	Daily Average of tweets	Monthly Average of tweets
Libero Quotidiano	23.069	5.667	25 %	228	923
Il Tempo	31.678	7.099	22 %	16	1.267
La Notizia	11.791	1.669	14 %	19	472
Il Secolo XIX	4127	583	14 %	6	169
Il Messaggero	40.345	5.239	13 %	54	1.614
Il Riformista	16.710	2.056	12 %	22	668
Leggo	38.309	3.782	10 %	31	1.532
Il Foglio	30.548	2.765	9 %	41	1.222
La Verità	13.979	1.244	9 %	51	559
Il Mattino	26.265	2.254	9 %	35	1.051
La Stampa	74.816	6.408	9 %	99	2.993
Avvenire	14.869	1.125	8 %	20	595
Il Fatto Quotidiano	55.404	4.147	7 %	73	2.216
La Repubblica	172.180	12.059	7 %	117	6.887
Corriere della Sera	119.311	8.212	7 %	158	4.772
Il Sole 24 ORE	88.524	5.668	6 %	42	3.541
Il Giornale	22.802	1.194	5 %	30	912
Il Manifesto	8.922	358	4 %	12	357

Table 4.1: Percentage of tweets related to the pandemic with daily and monthly average.



(a) Total Tweets



(b) Tweets related to the pandemic.

Figure 4.3: Daily Time Series of all tweets and Daily Time Series of tweets related to the pandemic.

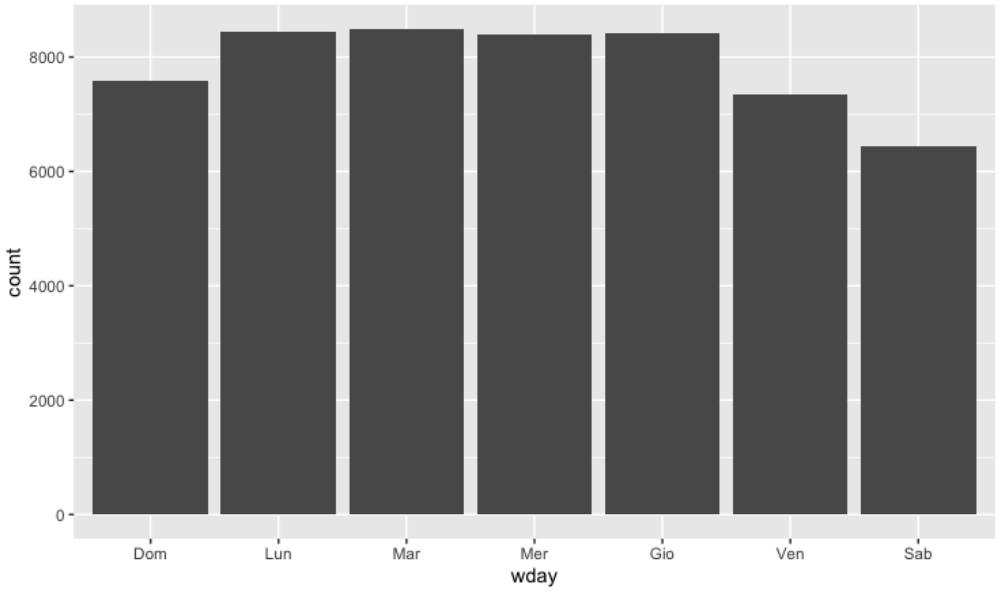


Figure 4.4: Days of week when the highest number of tweets related to the pandemic was registered.

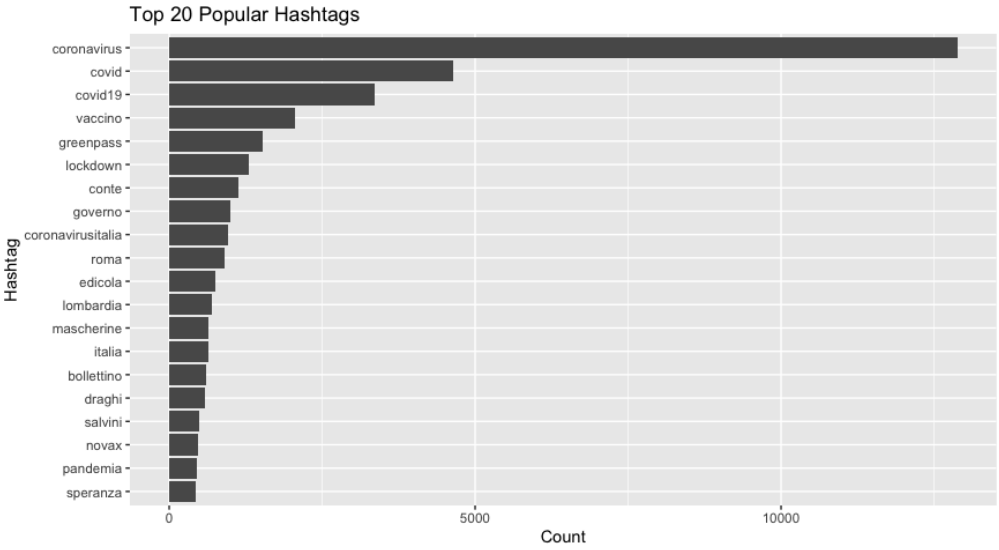


Figure 4.5: Top 20 popular hashtags from tweets related to the pandemic.

Term Document Matrix	(terms: 121402, documents: 18)
Non-/sparse entries:	208950/1976286
Sparsity:	90 %
Maximal term length:	54
Weighting:	term frequency (tf)

Table 4.2: Term Document Matrix

	word	freq
coronavirus	coronavirus	22.114
covid	covid	12.146
pandemia	pandemia	6.861
lockdown	lockdown	6.106
quarantena	quarantena	4.039
tamponi	tamponi	3.855
italia	italia	3.399
mascherina	mascherina	2.977
contagio	contagio	2.856
vaccino	vaccino	2.579

Table 4.3: Top 10 most frequent words

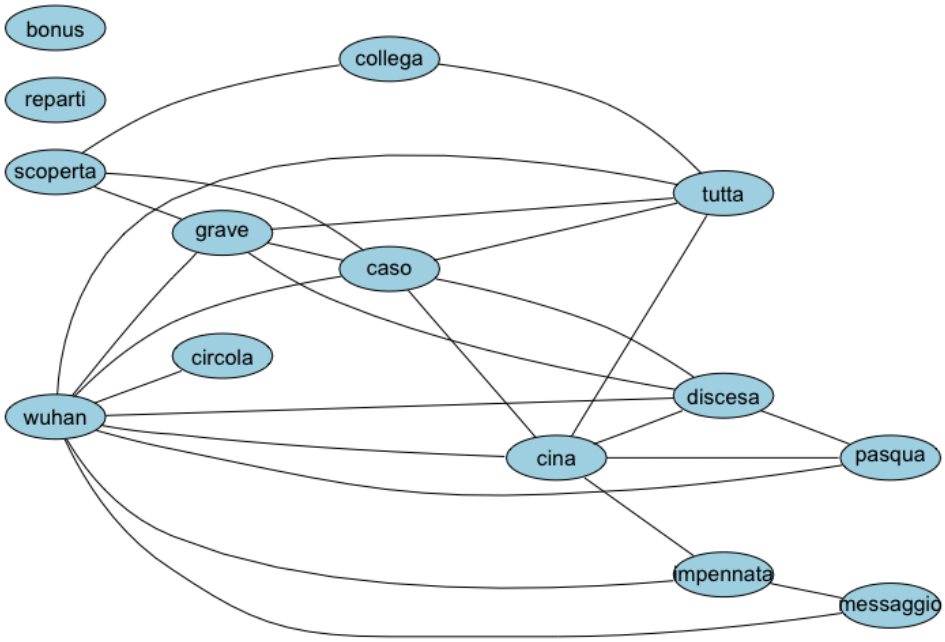
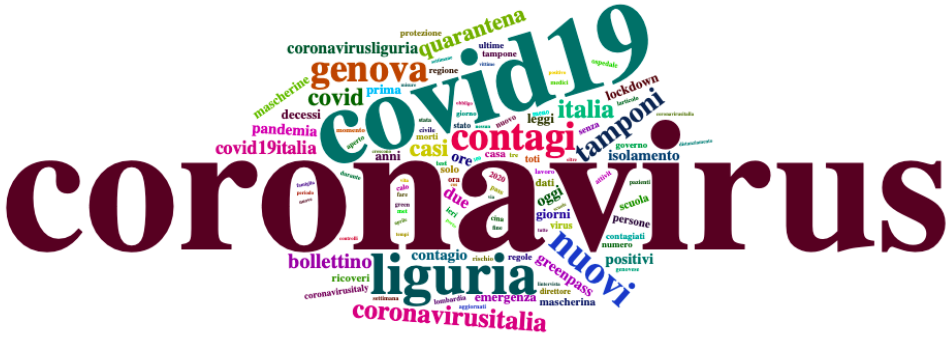


Figure 4.6: Plot of terms correlated with the word "coronavirus".

(a) Il Manifesto



(b) Il Secolo

(c) Il Fatto Quotidiano

Figure 4.7: Wordclouds of terms used by Il Manifesto, Il Secolo and Il Fatto Quotidiano for tweets related to the pandemic

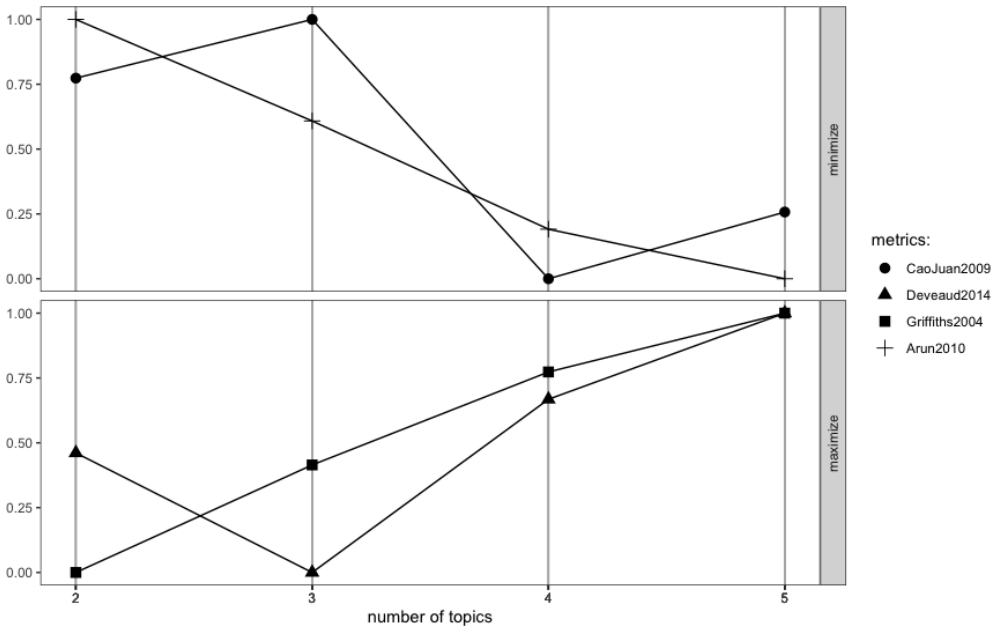


Figure 4.8: Identification of parameter K: optimal number of topics.

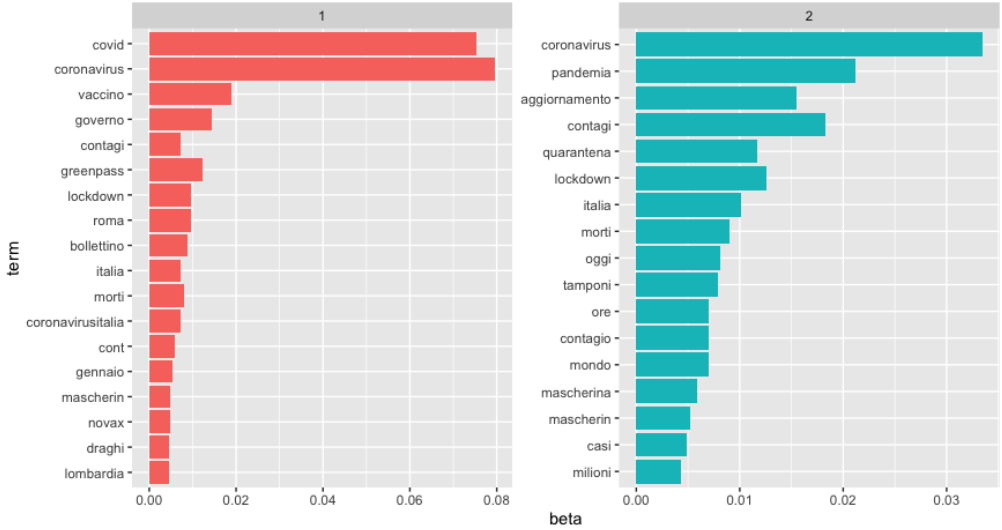


Figure 4.9: Most common terms within each topic.

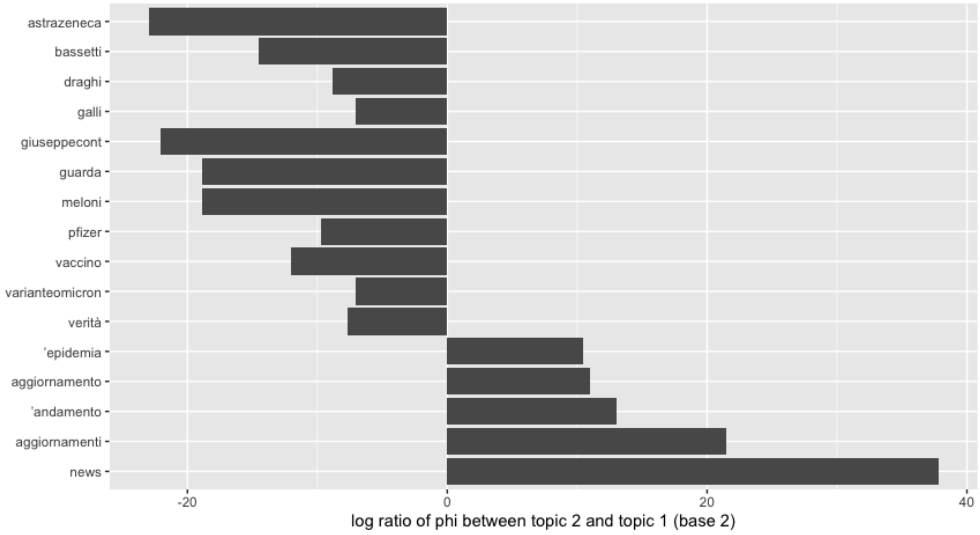


Figure 4.10: Words with the highest difference in β between Topic 1 and 2

term	topic1	topic2	logratio
aggiornamento	2.09e- 5	0.00727	8.44
assembramenti	7.60e- 4	0.00139	0.871
bollettino	3.20e- 3	0.00106	-1.59
contagiati	1.26e- 3	0.000426	-1.57
dpcm	1.33e- 3	0.000498	-1.42
decessi	1.02e- 3	0.000612	-0.738
distanziamento	5.58e- 4	0.00181	1.70
draghi	1.86e- 3	0.000626	-1.57
governo	5.66e- 3	0.00211	-1.42
guariti	1.04e- 3	0.000295	-1.81
greenpass	2.16e- 3	0.0000491	-5.46
isolamento	1.40e- 3	0.00340	1.28
mascherina	3.41e- 3	0.00651	0.932
positivi	3.88e- 3	0.00216	-0.848
novax	1.35e- 3	0.0000584	-4.53
restrizioni	1.65e- 3	0.00283	0.779
ricoveri	1.34e- 3	0.000422	-1.66
vaccino	7.87e- 3	0.000972	-3.02

Table 4.4: Words with the highest difference in β between Topic 1 and 2

Chapter 5

Conclusion

The present study has analyzed textual information provided by tweets of italian newspapers during the pandemic, starting from the first cases in the north of Italy up to the outbreak of the war between Russia and Ukraine. Several outcomes emerged from this investigation: firstly, a unique data set on Covid-19 in italy has been created. The latter, despite the focus on only one country, represents a potential contribution to the expansion of the existing literature. Furthermore, the analysis conducted with Natural Language Processing and Text Mining techniques also provides a clear picture of how information was managed in Italy throughout the pandemic with a focus on the political orientation of the selected newspapers. In light of this, even if Covid-19 may be represented as an objective virus, this work confirmed that there are still differences in the style of communication around it, thus allowing a certain degree of subjectivity. These differences, despite being slight, arise from the different values and ideologies associated to the political orientation of the italian newspapers.

As a result, similarities have emerged with respect to italian national newspapers that were part of the same owner group whereas communicative differences, in terms of frequency and choice of words, have emerged. In addition, results confirmed the assumption that starting from the beginning of the war between Russia and Ukraine, there has been a shift in terms of communication provided by national Italian newspapers. The application of Sentiment Analysis, which

aims at automatically identifying and extracting opinions, emotions, and attitudes expressed in a text, has been excluded in the present work. This decision stems from the potential bias which might have resulted in words specific to the pandemic such as "positivo". The latter term might have been recognized as a positive emotion while it hides a negative sentiment. Moreover, results from Topic Modelling Section suggest that vaccine has been a crucial topic during the pandemic in Italy. At this regard, not only the above mentioned topic involved the mutation of the virus but also the rejection of the vaccine itself.

According to a report released by GIMBE, foundation which aims at promoting the dissemination and application of the best scientific evidence in order to improve people's health, 8.63 million people have not received a single dose of vaccine. This result is not surprising if we consider the high degree of misinformation and the low level of trust that the Italian population attributes to political parties.

Ultimately, it can be argued that the italian government, as well as other european and non european countries, was not prepared for a proper management of the pandemic in Italy. This is in line with a statement released by Frank Snowden during an interview in which it was stressed that the years before the pandemic were marked by cuts in scientific research and expenditure on the health system.

Bibliography

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*, 2017.

Narendra Andhale and Laxmi A Bewoor. An overview of text summarization techniques. In *2016 international conference on computing communication control and automation (ICCUBEA)*, pages 1–7. IEEE, 2016.

Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of twitter users. *Social Science Computer Review*, 33(6):712–729, 2015.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Giuseppe Boriani, Federico Guerra, Roberto De Ponti, Antonio D'Onofrio, Michele Accogli, Matteo Bertini, Giovanni Bisignani, Giovanni Battista Forleo, Maurizio Landolina, Carlo Lavalle, et al. Five waves of covid-19 pandemic in italy: Results of a national survey evaluating the impact on activities related to arrhythmias, pacing, and electrophysiology promoted by aiac (italian association of arrhythmology and cardiac pacing). *Internal and Emergency Medicine*, 18(1):137–149, 2023.

Philip Burnard. Teaching the analysis of textual data: an experiential approach. *Nurse education today*, 16(4):278–281, 1996.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Qi-Wen Dong, Xiao-long Wang, and Lei Lin. Application of latent semantic analysis to protein remote homology detection. *Bioinformatics*, 22(3):285–290, 2006.

Ingo Feinerer and Fridolin Wild. Automated coding of qualitative interviews with latent semantic analysis. In *Information systems technology and its applications–6th international conference–ISTA 2007*. Gesellschaft für Informatik e. V., 2007.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1–16, 2004.

Jeff Gentry, Maintainer Jeff Gentry, Suggests RSQLite, and RMySQL License Artistic. Package 'twitter'. *Cran. r-project*, 2016.

Javier Girón, Josep Ginebra, and Alex Riba. Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 59(1): 19–30, 2005.

Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187, 2014.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. A brief survey of text mining. *Journal for Language Technology and Computational Linguistics*, 20 (1):19–62, 2005.

Maria Renee Jimenez-Sotomayor, Carolina Gomez-Moreno, and Enrique Soto-Perez-de Celis. Coronavirus, ageism, and twitter: An evaluation of tweets about older adults and covid-19. *Journal of the American Geriatrics Society*, 68(8):1661–1665, 2020.

Karen Spärck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481, 2007.

Borislav Kapukaranov and Preslav Nakov. Fine-grained sentiment analysis for movie reviews in bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274, 2015.

Atif Khan and Naomie Salim. A review on abstractive summarization methods. *Journal of Theoretical and Applied Information Technology*, 59(1):64–72, 2014.

Eva Lahuerta-Otero and Rebeca Cordero-Gutiérrez. Looking for the perfect tweet. the use of data mining techniques to find influencers on twitter. *Computers in human behavior*, 64:575–583, 2016.

Rabindra Lamsal. Design and analysis of a large-scale covid-19 tweets dataset. *applied intelligence*, 51:2790–2804, 2021.

Yaoyong Li and John Shawe-Taylor. Using kcca for japanese-english cross-language information retrieval and classification. *Journal of intelligent information systems*, (tba), 2005.

Elizabeth D Liddy. Natural language processing. 2001.

Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107):6, 2004.

J Mantas et al. Application of topic modeling to tweets as the foundation for health disparity research for covid-19. *The Importance of Health Informatics in Public Health during a Pandemic*, 272:24, 2020.

Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl, and Christoph U Lehmann. An "infodemic": leveraging high-volume twitter data

to understand early public sentiment for the coronavirus disease 2019 outbreak. In *Open forum infectious diseases*, volume 7, page ofaa258. Oxford University Press US, 2020.

Thomas W Miller. *Modeling techniques in predictive analytics: business problems and solutions with R*. Pearson Education, 2014.

Azzam Mourad, Ali Srour, Haidar Harmanani, Cathia Jenainati, and Mohamad Arafeh. Critical impact of social networks infodemic on defeating coronavirus covid-19 pandemic: Twitter-based study and research directions. *IEEE Transactions on Network and Service Management*, 17(4):2145–2155, 2020.

Murzintcev Nikita and Maintainer Murzintcev Nikita. Package 'ldatuning', 2016.

Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020.

Cherish Kay Pastor. Sentiment analysis of filipinos and effects of extreme community quarantine due to coronavirus (covid-19) pandemic. *Available at SSRN 3574385*, 2020.

Filip Radlinski and Thorsten Joachims. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 570–579, 2007.

Shigeaki Sakurai and Akihiro Suyama. An e-mail analysis method based on text mining techniques. *Applied Soft Computing*, 6(1):62–71, 2005.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Fabrizio Sebastiani. Acm computer surveys. *Machine learning in automated text categorization*, 3:1–47, 2002.

Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. 2000.

Shihan Wang, Marijn Schraagen, Erik Tjong Kim Sang, and Mehdi Dastani. Public sentiment on governmental covid-19 measures in dutch social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

Sholom M Weiss, Nitin Indurkhya, Tong Zhang, and Fred Damerau. *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media, 2010.

Y-FB Wu and Xin Chen. elearning assessment through textual analysis of class discussions. In *Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05)*, pages 388–390. IEEE, 2005.

Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *Journal of medical Internet research*, 22 (11):e20550, 2020.

Kai-Cheng Yang, Francesco Pierri, Pik-Mai Hui, David Axelrod, Christopher Torres-Lugo, John Bryden, and Filippo Menczer. The covid-19 infodemic: twitter versus facebook. *Big Data & Society*, 8(1):20539517211013861, 2021.

Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, volume 16, pages 1–10. Springer, 2015.

Ying Zhao and George Karypis. Topic-driven clustering for document datasets. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 358–369. SIAM, 2005.

Summary

Introduction

Data is now generated in a wide range of formats and at a rate that shows no signs of stopping. Finding information that can help to make decisions among hundreds of documents, web pages, and social media feeds is a difficult and time-consuming task. In light of this, one of the most interesting and widely used social platforms for generating a large amount of textual data is Twitter: Twitter's API (Application Programming Interface) is more open and accessible compared to other social media platforms, facilitating the ability for programmers to develop data access techniques and, consequently, providing additional resources to scholar.

The main objective of the present study is to address the research question concerning the analysis of textual information provided by tweets of italian newspapers, in order to determine whether they behaved similarly when reporting on the Covid-19 pandemic in italy or not. Despite providing only a partial picture of all information provided during the pandemic in Italy, the assumption is that Twitter is sufficiently representative of all information provided during the pandemic in Italy by national magazines.

The present study is based on a temporal horizon which goes from 2020-02-01, when patient zero was identified in Codogno, to 2022-02-28. The underlying assumption is that starting from this latter date, there has been a shift in terms of communication provided by national newspapers in Italy from the Covid-19 to the outbreak of the war between Russia and Ukraine.

Research in the field shows that Twitter has been previously used, as a social media, for data analysis during the pandemic. In their study Medford et al. [2020] created a set of COVID-19-related hashtags to search for relevant tweets throughout a two-week period from January 14^{th} to January 28^{th}, 2020. Mantas et al. [2020] applied Natural Language Processing in their study. The application of a clustering technique to publically available Tweets posted by African Americans allowed them to discover several topics related to the pandemic. Furthermore, in a research conducted by Ordun et al. [2020] ,Latent Dirichlet Allocation (LDA) topic modeling was used to produce twenty separate topics

about case spread, healthcare staff, and personal protective equipment (PPE). The application of this technique helped them demonstrating how information about the Covid-19 spreads via Twitter.

Despite the existence of several studies in the field, not so many analyses ave been conducted in Italy for the reference period. Hence, in contrast with other previous works which mainly focus on English tweets or diverse countries, the main contribution of this study is to provide, through data mined from Italian newspapers, a unique framework gathering information related to the Covid-19 crisis in Italy.

This work not only can be employed for future research, but it also provides a clear picture of how information was managed in Italy throughout the pandemic.

Data

Despite the large amount of newspaper listed in Italy, this study only takes as a point of reference the major national newspapers. After careful considerations for the implementation of the analysis under study, the list provided by the association "Accertamenti Diffusione Stampa" that initially included 24 newspapers has been now reduced to 20 newspapers. After a short description of selected Italian Newspapers, further considerations were made regarding the political orientation of each magazine: because of the scarcity of official sources and explicit statements by the various newspapers, observations were drawn on the basis of the ownership of each magazine.

Collected data was filtered with keywords related to the pandemic in order to rely on more accurate and relevant results. Therefore, the list included the following Italian words: "coronavirus, covid, quarantena, pandemia, lockdown, tampon*, assembrament*, restrizion*, isolamento, batterio, contagi*, distanziamento, mascherin*, epidemia". The total number of tweets related to the pandemic is 71.529. In order to determine at which percentage each Italian newspaper talked about Covid-19 in the reporting period, it is possible to calculate the ratio between the number of tweets filtered with keywords and the total number of tweets without filtering. Overall, as shown in the table below, it is possible to state that, in percentage terms, "only" 9 % of Total Tweets posted by italian newspapers are related to the pandemic.

Name	Foundation	Editor	Political Orientation
Avvenire	1968	Avvenire Nuova Editoriale s.p.a.	Conservative
Corriere della Sera	1876	RCS MediaGroup	Conservative
Il Fatto Quotidiano	2009	Società Editoriale Il Fatto s.p.a.	Liberal
Il Foglio	1996	Foglio Quotidiano Società Cooperativa	Conservative/Right
Il Giornale	1974	Società Europea di Edizioni s.p.a.	Conservative/Right
Il Manifesto	1969	Il Nuovo Manifesto-Società Cooperativa Editrice	Left
Il Mattino	1892	Caltagirone Editore	Conservative
Il Messaggero	1878	Caltagirone Editore	Conservative
Il Resto Del Carlino	1885	Editoriale Nazionale	Conservative
Il Riformista	2002	Romeo Editore s.r.l.	Conservative
Il Tempo	1944	Gruppo Angelucci	Conservative/Right
Il Secolo XIX	1886	GEDI Gruppo Editoriale	Right
Il Sole 24 ORE	1965	Confindustria	Conservative
Italia Oggi	1986	Gruppo Class	Conservative/Right
La Notizia	2013	La Notizia s.r.l.	Liberal
La Repubblica	1976	GEDI Gruppo Editoriale	Conservative
La Stampa	1867	GEDI Gruppo Editoriale	Conservative
La Verità	2016	La Verità s.r.l.	Conservative/Right
Leggo	2001	Caltagirone Editore	Conservative
Libero Quotidiano	2000	Editoriale Libero s.r.l.	Conservative/Right

Table 5.1: List of selected Italian Newspapers and corresponding political orientation.

Name	Tweets before filtering	Tweets after filtering	Ratio(%)	Daily Average of tweets	Monthly Average of tweets
Libero Quotidiano	23.069	5667	25 %	228	923
Il Tempo	31.678	7.099	22 %	16	1.267
La Notizia	11.791	1.669	14 %	19	472
Il Secolo XIX	4.127	583	14 %	6	169
Il Messaggero	40.345	5.239	13 %	54	1.614
Il Riformista	16.710	2.056	12 %	22	668
Leggo	38.309	3.782	10 %	31	1.532
Il Foglio	30.548	2.765	9 %	41	1.222
La Verità	13.979	1.244	9 %	51	559
Il Mattino	26.265	2.254	9 %	35	1.051
La Stampa	74.816	6.408	9 %	99	2.993
Avvenire	14.869	1.125	8 %	20	595
Il Fatto Quotidiano	55.404	4.147	7 %	73	2.216
La Repubblica	172.180	12.059	7 %	117	6.887
Corriere della Sera	119.311	8.212	7 %	158	4.772
Il Sole 24 ORE	88.524	5.668	6 %	42	3.541
Il Giornale	22.802	1.194	5 %	30	912
Il Manifesto	8.922	358	4 %	12	357

Table 5.2: Percentage of tweets related to the pandemic with daily and monthly average.

Methods

Textual data may be situated among a wide range of disciplines such as linguistics, content analysis, information retrieval and artificial intelligence. Natural Language Processing, Text Mining and Topic Modelling have been implemented for the present study. The scraping of the tweets, through R software, has been organized as follow: first step of the process refers to the creation of a script for each newspaper. As a result, 20 R scripts were created for this study. Moreover, for each magazine the objective was to download all the tweets made in the time span. Due to constraints imposed by Twitter , tweets were downloaded daily. Subsequently the vector for queries was created. The latter requires several components to be specified: from which user, start time and end time, max results and tweets field. Max results refers to the maximum number of tweets we can download. At a first stage, this parameter was set to 150, but it was subsequently doubled to 300 in the hope that no newspaper posted more than 300 tweets per day. Consequently, the iteration process was performed for each magazine: the duration of the procedure was about two hours per iteration. At the end of the loop is the final phase of the combination and the saving of the object. This procedure has been applied to every single magazine. The iteration process was successful for all the newspapers at the exception of two: Il Resto del Carlino and Italia Oggi. In the latter two cases, the number of total tweets downloaded did not coincide with the daily sequence created at first. Therefore, a large list containing 18 elements, each of them corresponding to a single national newspaper, was created. Subsequently collected data was loaded as a Corpus, a huge collection of texts used in data analysis for checking frequencies and verifying grammatical structures within a certain language area. Characters such as "/" and "@" were replaced with white spaces by using "tmmap" function. Additional steps involved the conversion of the text to lower case, the removal of numbers and the removal of italian common stopwords. Finally a term-document matrix was built after all the previous steps.

Results and Discussion

The observation of the time series of all tweets related to the pandemic,allowed us to notice that the majority of selected Italian newspapers follows a common pattern: this tendency is hardly surprising if we take into consideration the trend of the pandemic itself in Italy. Although to a lesser extent than the previous wave, the number of tweets related to the pandemic increases at each wave. Another noteworthy point is that Il Tempo appears to have posted more tweets related to the pandemic at the end of 2021. This is in contrast with the common path followed by the majority of selected Italian newspapers which showed a pick of posted tweets during the first wave in Italy. The date of March 18, 2020 coincides to the day when the maximum amount of tweets about the pandemic was recorded in Italy. As a result, it is not unexpected that this day occurs during the week with the highest number of tweets (2020-03-15 - 2020-03-22). Nonetheless, something "exceptional" beyond March 18, 2020 may explain the record. The National Day in commemoration of all the victims of the coronavirus epidemic, also known as the National Day for the Victims of COVID-19, is an Italian national holiday commemorated on March 18 in memory of those who died in Italy as a result of SARS-cov-2 during the COVID-19 pandemic.

Another section of the present study aims at exploring potential differences which might arise from different political orientation. As a first step, among the Italian national newspapers, a decision was made to select one journal per political orientation. As shown in the word clouds below it is possible to highlight the presence of slight differences. Indeed, Il Manifesto, in quality of left-oriented newspaper, tends to adopt a style of communication based on progressive and internationalist ideas, typical values of leftist parties. In light of this, it is not surprising that Il Manifesto decided to use more the word "lock-down" rather than "quarantena" as it was the case for Il Secolo Decimo Nono. The latter has deep roots in the Liguria region: this is in line with the fact that Il Secolo, as a right-oriented newspaper, tends to employ a style of communication based on patriotic, conservative and religious values. This can be seen by looking at the wordcloud related to the italian newspaper where words such as "coronavirus-liguria" , "genova" appear. Furthemore, as we can observe, objective terms such

as "isolamento", "restrizioni" emerge in the wordcloud, as a confirmation to the liberal orientation of Il Fatto Quotidiano.

Finally, with reference to the Topic Modelling section, in order to determine K, the number of optimal topic, four approaches were used: CaoJuan2009, Arun2010, Griffiths2004 and Deveaud2014. The first two have a low amount of subjects, while the others have a high number of themes. Optimally, several methods should converge and show peaks and dips respectively for a certain number of topics. Despite a suggestion of 5 as optimal number of "k" topics, a decision was made to vary this parameter k from 2 to 5. As a result, given the small variances in scores, we decided on maintaining k = 2 to ensure that the information remains interpretable. If we look at most common terms within each topic it can be argued that there are terms which appear only in a topic. This is the case of terms as "casi", "tamponi", "dati", "milioni", as they only appear in Topic 2. At the opposite words as "governo", "aggiornamento", "greenpass", "vaccino" and "novax" only appear in Topic 1. The presence of such specific words in Topic 2 might suggest that it refers to communications provided by the italian government during the pandemic, through the announcement of data related to the development of the pandemic. At the opposite, specific words appearing in Topic 1 might suggest that it refers to communication provided by the italian government with respect to the vaccination campaign during the pandemic in italy.

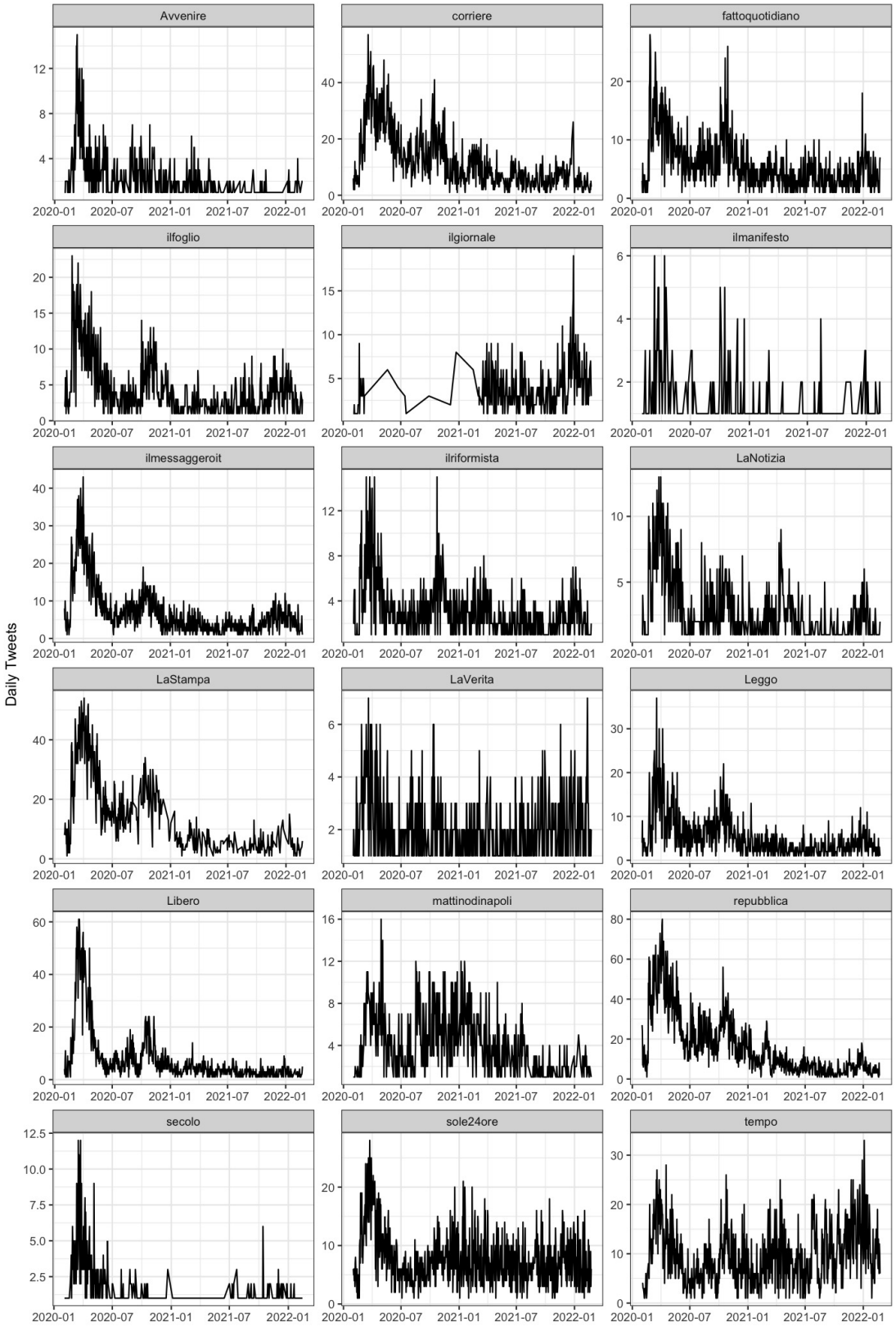
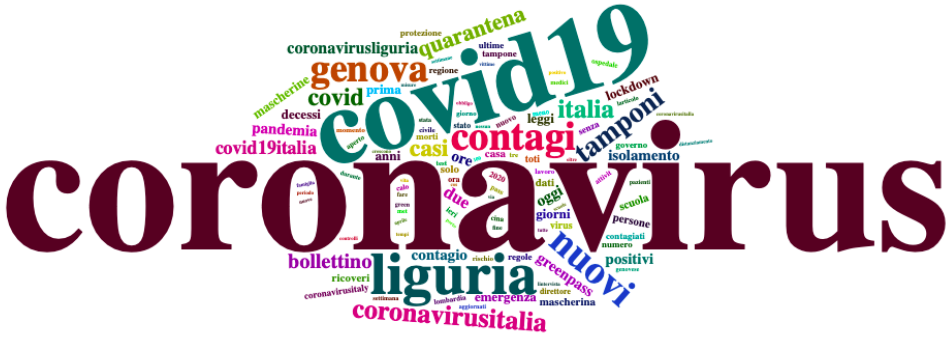


Figure 1: Daily Time Series of all Tweets related to the pandemic displayed per each selected Italian newspaper.

(a) Il Manifesto



(b) Il Secolo

(c) Il Fatto Quotidiano

Figure 2: Comparison between wordclouds of terms used by Il Manifesto, Il Secolo and Il Fatto Quotidiano for tweets related to the pandemic.

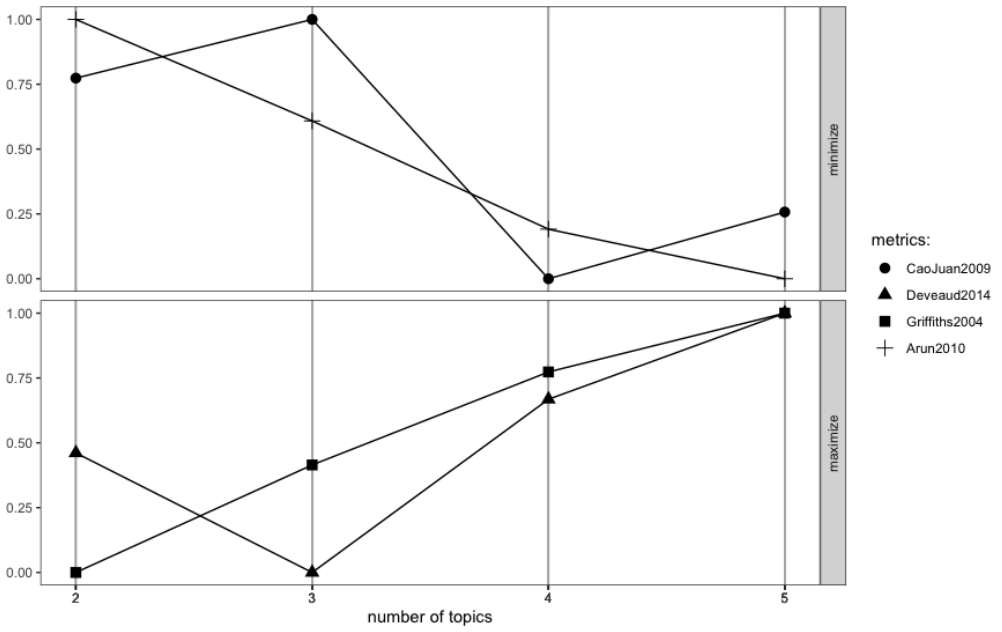


Figure 3: Identification of parameter K: optimal number of topics.

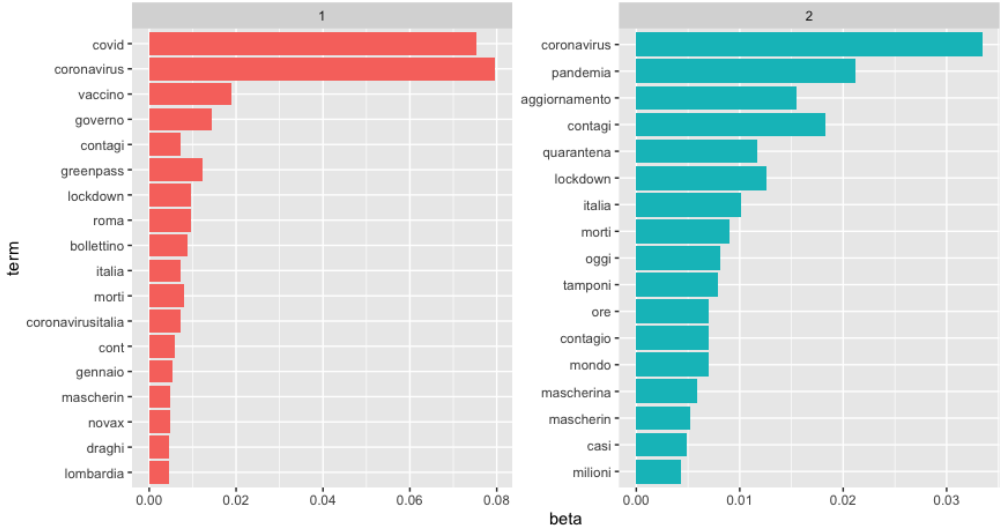


Figure 4: Most common terms within each topic.

Conclusion

Several outcomes emerged from this investigation: firstly, a unique data set on Covid-19 in italy has been created. The latter, despite the focus on only one country, represents a potential contribution to the expansion of the existing literature. Furthermore, the analysis conducted with Natural Language Processing and Text Mining techniques also provides a clear picture of how information was managed in Italy throughout the pandemic with a focus on the political orientation of the selected newspaper. As a result, similarities have emerged with respect to italian national newspapers that were part of the same owner group whereas communicative differences, in terms of frequency and choice of words, have emerged. In addition, results confirmed the assumption that starting from the beginning of the war between Russia and Ukraine, there has been a shift in terms of communication provided by national Italian newspapers. Topic Modelling Section suggest that vaccine has been a crucial topic during the pandemic in Italy. At this regard, not only the above mentioned topic involved the mutation of the virus but also the sense of rejection of the vaccine itself bt the italian population.