Statistical Learning Project

2nd Milestone

Group 01 – Stefano Rando 1745758, Negin Amininodoushan 1915791, Marco Muscas 1883544, Leonardo Placidi 1761588, Davide Zingaro 1873954

Research Title

The Pandemic behind the Pandemic: Italian Media response to COVID-19.

Abstract / update

As quarantined citizens of 2020, in the middle of the Covid-19 outbreak, we feel like to be affected by one more spreading desease:the COVID-19 media response!
At the start we will focus on Italian data(and if we will get meaningful results we will extend it to larger pool), since is anyway one of the most affected countries and was the first big outbreak out of China. We will analyze reflected effects of COVID-19 in not only news and journals but also in social media, so we will have an understanding of how different categories of media sources (social media, online newspapers) approached the COVID-19 outbreak day by day, reacted and evolved from the website of Pretezione Civile. We expect to see a development of news similar to the spread of a virus, reaching possible insights on how digital information can mimic a biological pandemic: it's an INFODEMIC!

Main research aim & framework / update

We felt so overwhelmed by the informations about COVID-19, like we can't go online without this name popping out, so we started to think of this info as a desease itself! Our main goal in fact will be to compare the growth and behaviour of COVID-19 and Media news about it, and how news starts to change and oppress internet more and more, becoming a digital pandemic.
We will also study anomalies in Media behaviour, expecially we can analyze texts, expecially comments we parsed from reddit, using natual language processing tools to make sentiment analysis and time-studies about discussed topics. Analyze Coronavirus news day by day: https://opendatadpc.maps.arcgis.com/apps/opsdashboard/index.html#/b0c68bce2cce478eaac82fe38d4138b1 and creating paralles with our analysis. We will try to also trarin prediciton models such as neural network to predict comments sentiment and covid cases, taking features from all the different datasets we create.

Data collection & source(s)

Our pilot database is composed of 4 different datasets:

1. **RedditCommentsData**.
   from the megathreads posts published in the r/italy subreddit at https://www.reddit.com/r/italy. Everyone of the megathread has been posted with the purpose of letting people talk about the Covid situation, so it was our optimal setup. To get this data we used Reddit API and used PRAW: The Python Reddit API Wrapper. The dataset is composed by 2 tables: megathreads_data and comments_data(this last one will be in chunks for checkpoints creation, we will merge the chunks later in a data cleaning process).
   The first one has a row for every megathread (every day a new megathread) with ID of the post, Month, Date (Hour was irregular and we decided to not include that), #Upvotes, #Comments, Upvote Ratio(not for comments because not accesible).
   The comments_data is more important and every row refers to a comment and we collected author(author ID was not so important), Submission ID(ID of the megathread where the comment belongs), text of the comment(main source of interest), hour, minute, second when it has been posted, ID of the comment, Top Level (if the comment is a response or a normal comment), Parent ID (if a response, what was it's parent comment), #Upvotes(Upvote ration not accessible).
   We plan to do sentiment analysis, topic analysis, changing in comments and discussion. As for the others source of data the collection is partial and we can fastly add new days to our analysis.

2. **quotidiani.csv** dataset.
   This dataset was driven from quotidiani.net (an online Italian and international newspaper website). It consists of 2 columns : region and website domain and it shows the domain of websites that was referred to in this website for each region. We used beautiful soup library of python to scrape pages belong to each 20 regions of Italy. Our idea is to see if there is any relation between the number of references to each region in news and the actual spread of the Virus. Also we will use the domain for each region later to get the content of articles for each region.

3. **intopic_it_articles.csv**
   Our news dataset consisted of news collected from many different italian outlets. A website at www.intopic.it provided an index for those articles.
   A web scraping program targeted the pages of intopic.it to extract data: we were able to get:

- The article's title
- A preview of the article's content
- The date of publication of the article
- The authoring news outlet
- Tags (which are not always included) related to the content
- The URL of the specific article

   After that, using another website available at www.quotidiani.net we extracted a series of news outlets websites by region which were put into the quotidiani.csv. Trouble was that such information is not always correct. For example: www.tgcom24.mediaset.it is clearly a national (not specific to a region) therefore requiring some edits to have, in the end, a corrispondence between news outlet's website and region.
   All this data, more than 40000 as of May 11th, will be ideally updated to have the complete content of an article, which will require more specific scrapers.
   The tools used for doing this were Python 3 scripts using libraries such as Pandas for creating and managing Dataframes, BeautifulSoup for parsing content of HTML pages, urllib 3 to manage connections and perform HTTP requests. There were also some native and more "general purpose"

libraries to handle datatypes derived from strings. As stated in the next paragraph we will try to use this dataset to define features absed on sentiment analysis, count of key words and more to create predictive models for covid and sentiment of reddit comments.

4. **ProtezioneCivile_data** data published at https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni. In this dataset we collected all the italian regional observations relative to the covid-19 spreading that we downloaded and worked using python. The dataset is composed of one table: regional_data.
   The main features of interest for us are: the 'data' column, which sorts all the observations for each region relatively to the daily record of our features of interest, it starts from 24/2/20 to the 5/5/20(but will be updated untile probably 30 May). Then we have the 'denominazione regione' column, which lists all the regions, and it'll be very important for finding the actual correlation between media and news attention to the desease relative to the actual spread of the virus in the territory. After some cleaning of useless features, we added features focused mainly on the daily and total number of positive (totale_positivi), new positive (nuovi_positivi), dead (deceduti), tests (tamponi), and tested percentage of the population and hospitalized with symptoms with other new columns if needed. We plan to analyze the features to get insight about the macro and micro diffusion of the virus, regional analysis, plot and visualize the effect of disease in relation to media coverage.
   You will find also some tables, with the same fatures for each region, as we think could be a fast instrument in focused analysis.

Models & Methods - Update

There have not been singificant changes, only updates are that we are thinking on focusing more on sentiment analysis and try to make predictions on the sentiment of new comments. Also we would like to train some neural networks to try and approach in predicting italian data about covid, also we will want to see if creating a model that takes into account italian covid data+analysis on newspaper sources(dataset 2)+sentiment analysys of reddit comments we can train a model in predicting the sentiment of new reddit comments and with the same data a model to predict the covid numbers day by day. We will try also to show the great correlation between the news and the covid outbreak numbers, trying to understanding what we think is a mirroring trend by the news.

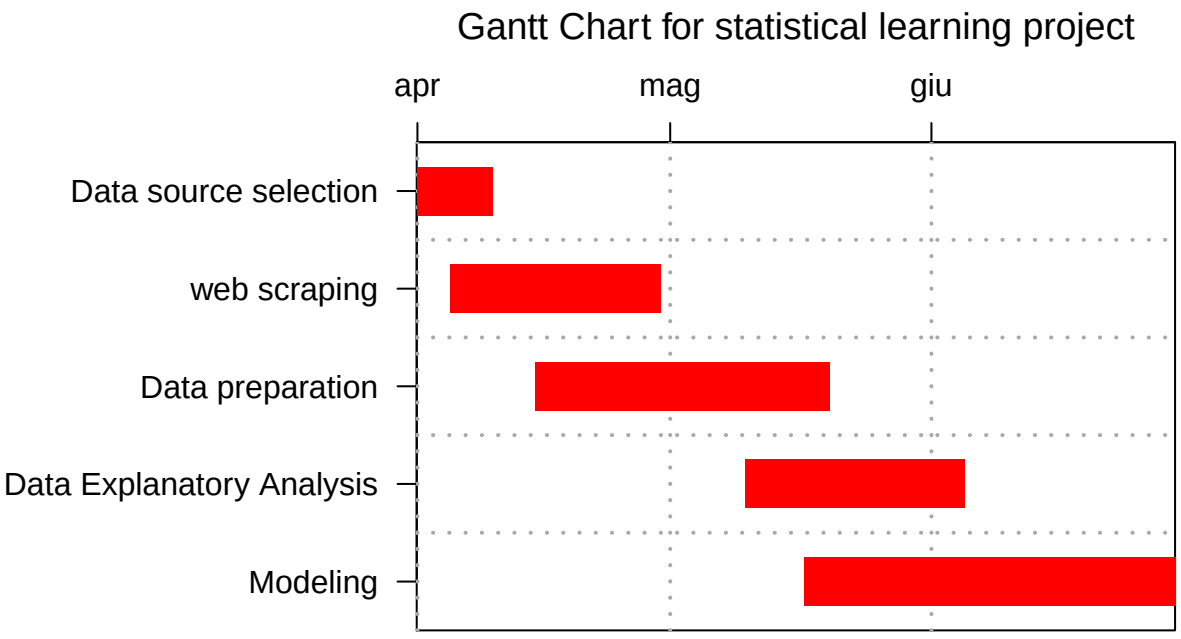
Software/Hardware Toolkit / update

We will also use this tool on italian sentiment analysis https://nicgian.github.io/Sentita/.

Problems so far...

We encountered a problem when deciding from which social network we were going to collect the data. At first we would have liked to parse Twitter tweets, especially because on Twitter we could have found both people from all around Italy as well as celebrities, politicians and such. Unfortunately we eventually had to pass over the idea because of limitations in the Twitter API. We were interested in finding out how the people reacted to the COVID-19 outbreak since it started to manifest as a serious problem. This, as we know, happened around the end of February. We started to collect data by the end of April and the standard Twitter API key could only give us the possibility to scrape Tweets from the last seven days. We thus chose to change the social network and to scrape comments from Reddit instead of Twitter. We are afraid that Reddit could be a pretty biased social platform, but in the end we collected more than 150'000 comments and it's reasonable to expect opinions of every kinds, maybe just with unbalanced presences.
Working on quotidiani.csv and intopic_it_articles.csv we knew that web scraping is not always welcomed and some websites put measures to discourage it. We will maybe have a problem in parsing the content of articles from some news outlets, we had a similar problem with intopic.it. Of course we have always been mindful not to send too many request at once increasing traffic towards a website. Another problem was in determining which region a news outlet belongs to: basically we got half an answer from quotidiani.net . Possibly this will be solved by entering such data manually.

Project Timeline / update



References / update

List any additional (if any) references cited in the previous sections. We did not quote this articles directly, but methods and ideas we will try to materialize come from these papers we read and we would like to imitate some approaches they used in particular for modeling.

- *Modelling and predicting the spatio-temporal spread of Coronavirus disease 2019 (COVID-19) in Italy by Diego Giuliani, Maria Michela Dickson, Giuseppe Espa1, and Flavio Santi 1Department of Economics and Management, University of Trento 2Department of Economics, University of Verona March 23 2020, <https://arxiv.org/pdf/2003.06664.pdf>.*

- The COVID-19 Social Media Infodemic by Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi,Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo2,1, and Antonio Scala, CNR-ISC Roma, Università Ca Foscari di Venezia, Big Data in Health Society Roma, Universit di Brescia, Politecnico di Milano 6CNR-IIT Pisa, <https://arxiv.org/pdf/2003.05004.pdf>.

- *Covid-19 epidemic in Italy: evolution, projections and impact of government measures by Giovanni Sebastiani, Marco Massa and Elio Riboli, 18 April 2020 <https://link.springer.com/content/pdf/10.1007/s10654-020-00631-6.pdf>.*

- ALBERTO: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets by Marco Polignano, Pierpaolo Basile, Marco de Gamnis, Giovanni Semeraro, Valerio Basile, <http://ceur-ws.org/Vol-2481/paper57.pdf>.
