SAPIENZA
UNIVERSITÀ DI ROMA

Final Report -- Group01

Statistical Learning 2019/2020, prof. Pierpaolo Brutti, Università La Sapienza.

Group components:

1. Leonardo Placidi (Team Leader)
2. Stefano Rando
3. Davide Zingaro
4. Negin Amininodoushan
5. Marco Muscas

The pandemic behind The Pandemic

Introduction:

As quarantined citizens of 2020, in the middle of the Covid-19 outbreak, we feel like we are affected by one more spreading disease: the media related to COVID-19!
We focus on Italian data, since it is 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 in according to the covid data from the website of Protezione Civile. In many cases we got insights in which the information spread or even topic or polarity of texts mimic the biological pandemic: it's an INFODEMIC!
Our work has been divided into two parts.

The first part is about sentiment analysis and topic modeling regarding comments coming from r/italy, a community on Reddit. Following these analyses we studied the distributions of the sentiments and topics, relating to the real covid data and obtaining meaningful relations and possible predictive models.
The second partis news analysis starts from a rich collection of news articles that we parsed, then we performed topic modeling to classify each and every article to find the covid related news, in order to achieve some predictions in respect to the national Protezione Civile data. Then, we plotted and analysed the most meaningful results, splitting our observations and insights by regions and zones.
Finally our results have been very interesting and satisfied us as we noticed a real trend between media and covid spread, giving us some pretty accurate predictive models.
We could have done better since our data is a massive set, and we also could have processed them more, but all our work can give light into a new direction of investigation using more sources that seem to mimic an origining phenomenon (covid in our case).

Every method exposed has code implementations that you can find in the GitHub repository:
https://github.com/Gruntrexpewrus/ThepandemicbehindThePandemic.
Data has not been uploaded there since there can be some Intellectual property issues, but ask at placidi.1761588@studenti.uniroma1.it and we will send you everything.
We coded everything using either Google Colab or Jupyter Notebook.
Every knowledge for Neural Networks comes from our personal interests and eventually brief introductions in this first year.

Sentiment and Topic modeling

Data Collection

The analysis is performed on Reddit comments from the CovidMegrathreads that every day was hosted on the subreddit r/italy, we collected data from the 24 february to the 5th of May.
Additional data to perform the analysis have been 540 Amazon reviews(some related to Covid such as masks) and 140 Wikipedia articles (to assure some examples of neutral texts), both to create, together with 500 hand labeled **(by us)** reddit comments, a train set to train/validate our models.

Preprocessing of Data

Every string of text has been processed using stemming, tokenization, stopwords removal and eventually (for the Sentiment analysis) vectorized. The texts were in italian so it took us some time to find and modify targeted italian processors (or even models e.g. BERT for italian text).

Sentiment analysis

Our goal was to predict the sentiment (positive-negative-neutral) of a comment and to do so we trained various models. We used BERT model, developed an SVM, Bayes Classifiers(Gaussian, Binomial, Multinomial).
Every model has been scored based on accuracy on our validation set (we trained our model on 90% of the training data, validated on the remaining 10%).

1) BERT achieved an accuracy of 0.80 and was our best model, from which we did predictions of the sentiments that we used to study our data in the part of the work.

2) Support Vector Machine achieved a 0.67 accuracy, we tried different approaches and the data that we fed was first reduced through a SVD. For SVD we tried with 100, 200, 300 components, the best has been 100. To perform the SVM we first tried a long parameter tuning using the function GridSearchCV().
3) Bayes Classifiers achieved lower scores, for Gaussian 0.42, Bernoulli 0.58 and Multinomial 0.36 (sigh).

Every model we used has been saved and is ready to run.

Topic modeling

We tried to detect different topics (or trends) in the comments collection and thus perform a search over topics distribution. The idea was to try to understand what the main concerns and points of discussion were for the people.

This problem is part of the unsupervised setting of Data Analysis because we didn't have any *a priori* correct labeling. This has been an obstacle which we tried to overcome in two different ways:

1. Collect other examples of documents from another source with available labeling (Wikipedia) and develop a model which we could have eventually applied to our dataset.
2. Embrace the unsupervised nature of our problem and make use of a method which tried to identify not just associations but even the topics themselves.

We opted for the second option for two reasons: first there is no guarantee that a model which works well for Wikipedia articles is also suited for Reddit comments. Instead the large difference between languages in the two setting probably guarantees the inverse.
For second, in this way we reached to detect specific topics and we hadn't to rely on choosing topics in advance.

What we did is apply Latent Dirichlet Allocation to our dataset. A famous algorithm for Topic Modeling which leverages Bayesian statistics (it is, in fact, a hierarchical Bayesian model).

In order to correctly apply this algorithm we needed a way to measure the effectiveness of our results, both for choosing the correct number of topics (which the method isn't able to find out by itself) and for being sure everything was working right.

We searched for the literature and we did find out that the problem of evaluation of a topic model isn't an easy task.. Several different approaches have been

proposed and based on factors like applicability and empirical proved effectiveness we chose two methods.

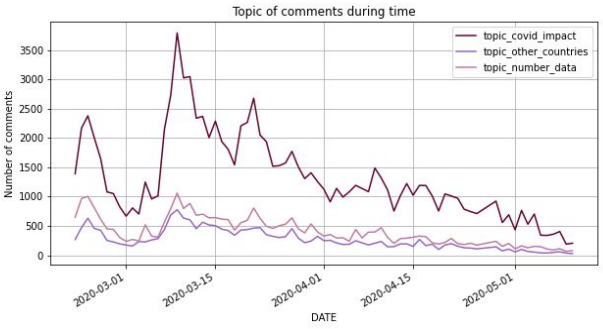
1. Google titles matches. This consists of taking the most important K words per topic (in our case we chose K = 10), then perform a search on Google for every topic over documents containing all those words and counting how many times they appeared on the document title.
2. Word intrusion. We asked some external people to answer a survey where they had to try to detect the word intruder over a set of 6 words.

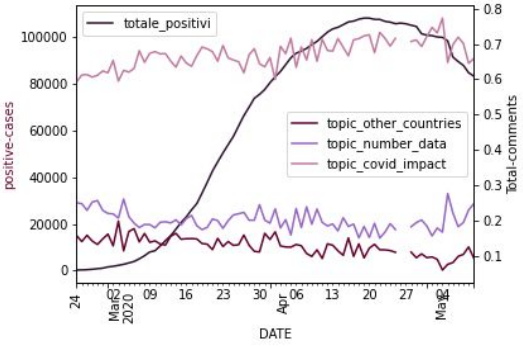
Our evaluation brought us to choosing 3 as the optimal number of topics. We then, of course, took a look directly at our results and we understood that these 3 trends broadly refer to the following three categories:

1. Discussions on how Covid-19 impacted our lives.
2. Discussion on the number of cases and data.
3. English comments or comments talking about the other countries situation.

Our analysis after the predictions based on the best models

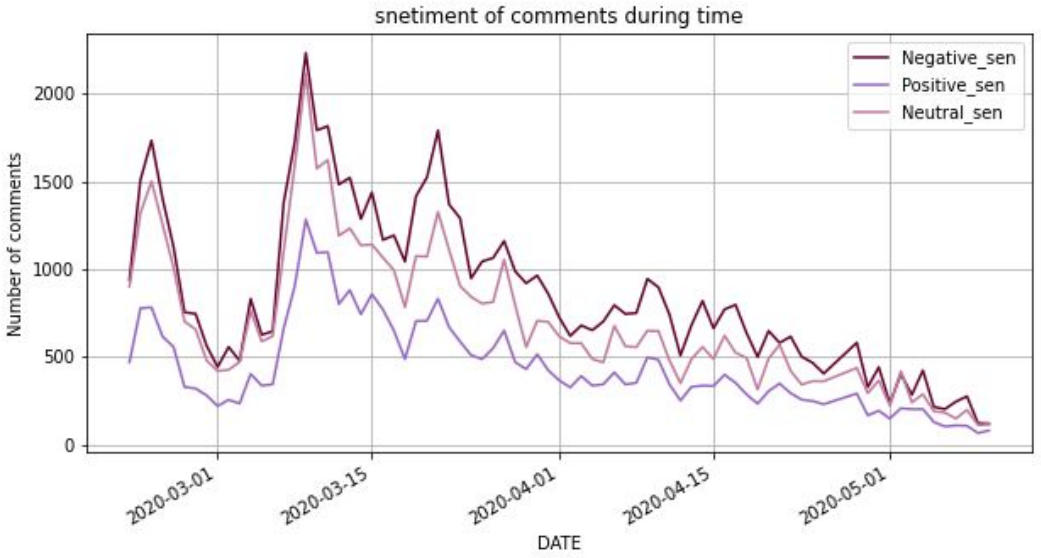
We can see below the plot of distribution of different topics over time. As it was mentioned before, there are 3 different categories: Covid impact, English comments and discussion on numbers and data. As we can see,most of the comments are regarding impacts of covid on life and that is exactly what we expect from social media like reddit, people are willing to talk about their opinion and daily experience rather than numbers and statistics.(On the left , plot of topic distribution over time and on the right, the same plot with normalized values and added the plot of covid positive cases)

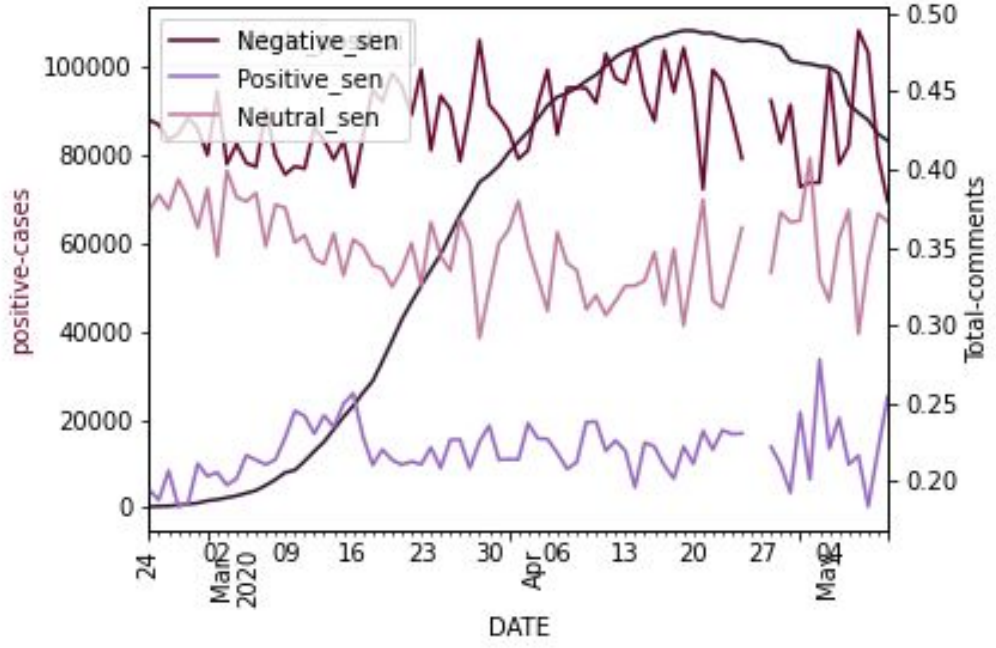




In the plot below, we can observe the result of sentiment analysis on Reddit comments over time. The interesting thing about this plot is that at the first of

the period, the number of Negative and Neutral sentiments are so close to each other but as time passes and the number of positive cases increase, these plots get far from each other. This may be related to the fact that, at first, people did not take the virus so seriously but as the situation gets worse and people are more aware of it; the difference between the number of negative and neutral comments increases: first, the plot of Reddit comments sentiments over time and the second, the same plot with normalized values and added covid positive cases plot.





News Analysis

Data Collection

A first attempt at getting the first bit of data (articles) was done using an indexing website that provided essential information about articles, such as title, author, date and *sometimes* more.
We wrote a first downloading script that iterating over the pages of the website resulted in a pre-dataset.

It did not prove successful as we noticed much of the published data was missing, so our next approach was much more aimed towards having multiple authors from all the regions, that published from February 24th, 2020 to May 31st.
Now we had some of the data, but the content of the articles was missing. Therefore, by means of the articles' URLs and specialized parsers, we were able to retrieve those information also.
A multithreaded approach was preferred as it allowed for getting multiple information at the same time without overloading the single news outlet website. The data from that was ultimately merged and put into the complete articles dataset, which was made of 24584 articles.

Preprocessing of data

The next goal was having a dataset with *only some* information preprocessed for later use. We chose to preprocess the text of each article's title and content.

The raw text was converted to lowercase (easier matching later on), stripped of special characters, stopwords from the italian dictionary defined in the **nltk** library, and was finally stemmed (each word was brought to its root form).

Classification

The main question at this point is: **is this article talking about coronavirus or not?**
We tried three different approaches.

Dummy classifier

We already know some common terms for coronavirus, why not use them?
To perform this kind of classification we checked for term occurrence in the title and content, finding one of the chosen keywords in the text would result in the article being flagged as "talking about coronavirus".

We actually got some interesting results, though deep down we knew that it was a bit too unreliable, in fact adding keywords that some would find relevant would result in matching almost the entire dataset.

Latent Dirichlet Allocation

For this model our data required some transformation. First, we chose the article's content as the relevant field for our classification task.
The text had to be converted to allow for a bag of words model to be assigned to each article. More specifically this bag of words was required to use the entire set of contents.
The LDA classifier had some problems: the covid related keywords were in all the detected topic clusters, therefore ambiguous. Simply put, there was no use for classification.

Non-negative Matrix Factorization

The NMF is another unsupervised learning technique useful when we don't already have the topics assigned to our text data - the articles. It takes the bag of words matrix created in the previous step using the TfIdf vectorizer and reduces its dimensionality. The result will be an easily accessible data structure that allows us to get the prominent topic in the text.
Using NMF this problem did not present itself, but..
The target number of clusters was set to 2 for a simple reason. Having more than that would result in covid-19 related keywords into more than one topic term clusters. Since it would result in an unnecessary complexity for classifying an article in covid related (or not), 2 clusters was maintained as the target.

Region analysis

Here we used the preprocessed dataset to perform another analysis on the various regions of Italy, seeing how the situation evolved for news outlets along the pandemic.
We chose the best performing model from the previous step to perform another topic classification.
We did not have any reference (pre-classified) dataset, so we made a comparison between the dummy model and the NMF. Results showed that the NMF was able to match around 78% of the articles that were flagged by the dummy classifier.

Lombardia

Lombardia was the most affected by the virus, so we thought it might have been a starting point of our analysis - remember, the focus is mainly on how the news behaved.
The graphs showed the new positive cases over time diminished, the news published did not really follow the trend. On the contrary, they seemed to keep the same pace even after the peak - not really in line with the national trend.

Other regions

Using the newly generated data by the NMF we did a visual comparison of the difference between spread of the virus and number of articles - by day and by region.
We noticed that in the northern region, and those more affected the number of news related to covid trended upwards relative to the number of positive cases. Southern regions, which were generally less affected by the virus showed more consistency (correlation) between the news published and number of covid cases.

Zones

when looking for a relation between covid news published by zone we did not find any high correlation, both negative nor positive, between number of news and number of positive cases.

Plotting

Plotting was usually done between two non-directly comparable quantities: the number of articles by day and the number of positive cases by day. That is why an additional operation was required before printing, and that is scaling. More specifically we used the MinMaxScaler.

Predictions

One of the goals of this analysis was performing data prediction from, in this case, news articles published.
Using the same previous techniques we first gave a topic classification to each article, although this time instead of simply having a flag saying "this article talks about covid" or "not", we did hot encoding.

This allowed us to have numerical features representing a given day in the interval of analysis.
The date, which we considered relevant had to be converted to another format since the former *'YYYY-mm-dd'* was not well liked by the model. We chose to use a variant of the UNIX timestamp format, the number of days since January 1st, 1970.
Next step, we grouped up by date by summing the different rows so we would know how many times each term had been used in a day.
This model was then used as training for the regressor: again, all this to try to predict the number of new cases.
As a regressor we used a Random Forest Regressor. After some parameter tuning we found out that having just the parameter bootstrap set to true we got the minimum MSE. Data was scaled before calculating it as it allowed for better interpretation.
Bonus point: we performed the same test but for trying to predict different columns of the dataset from the protezione civile as the scripts allowed for an easy implementation. Surprisingly, different other columns were easily predictable just by using articles.

We would like to thank Professor Pierpa Brutti for all of the support, without his excitement we would never have taken on this project and would never have been able to see it through this shining completion.

Group01 <3

Bibliography

- Berry, Gillis and Glineur "Document Classification Using Nonnegative Matrix Factorization and Underapproximation" 2009
- Okun "Non-negative matrix factorization and classifiers: experimental study" 2008
- Caragea, Cornelia, Jian Wu, Kyle Williams, Das, Khabsa, Teregowda and Giles. "Automatic Identification
- Alghamdi, Alfaqi "A Survey of Topic Modeling in Text Mining" *International Journal of Advanced Computer Science and Application* 2015 (2014).
- Newman, Han Lau, Grieser, Baldwin "Automatic Evaluation of Topic Coherence" *Association for Computer Linguistics* 2009
- Wang "Topic Modeling: A Complete Introductory Guide" *ResearchGate* 2017
- Fundamentals of Sentiment Analysis and Its Applications Mohsen Farhadloo and Erik Rolland, *ResearchGate* 2016
- 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, et al. CNR-ISC Roma, Università Ca Foscari di Venezia, Big Data in Health Society Roma, Università 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, 2020 https://link.springer.com/content/pdf/10.1007/s10654020-00631-6.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
- Yuxin Chen, Jean-Baptiste Bordes, David Filliat, An experimental comparison between NMF and LDA for active cross-situational object-word learning 2016