Universita degli Studi di Padova

Department of Statistical Science

Analysis of the topics discussed on Twitter during Covid19 pandemic by Italian politicians

Project for Statistics Method for Big Data

Authors

Boscarini Alessandro, Cifelli Lorenzo, Zampieri Daniele

2020/2021

Introduction

The target of our project is to identify which topics were covered on Twitter by Italian politicians during the Covid19 pandemic. To do so, we analyzed the tweets from 1st February 2020 to 4th May 2020 from 40 politicians by applying the Latent Dirichlet Allocation which assigns to each tweet a distribution of topics.

Preprocessing

The tweets were downloaded by the Twitter API using the function *get_time* of the library *rtweet* of **R**. Before applying the model, the tweets were pre-processed according to the following steps:

- tokenization,
- filtering of scoring characters, special characters, numbers, articles, conjunctions and pronouns,
- filtering of politicians' and parties' names to avoid strong correlations,
- creation of bigrams (e.g. *emergenza_sanitaria*),

After some tests, we decided to not reduce tokens to their fundamental root to obtain a better interpretation of the results. Finally, tokens present in less than 10 tweets were discarded so as not to include rare words.

Descriptive analysis

First of all, we started the analysis with some plots. The following figures show word clouds where the most frequent words are displayed and the size of the word is proportional to its frequency.

Figure 1 shows the word cloud that considers all tweets. In addition to general words such *Italia, governo, paese*, the most frequent words are about pandemic (*coronavirus, covid19, emergenza, casa, decreto*) along with those related to the Italian political and economic debate (*lavoro, europa, mes*).

Focusing on the tweets of individual politicians, Figure 2 shows the word cloud of Matteo Salvini: the most frequent words are about social communication and Covid19. In Figure 3 there is the word cloud of Maurizio Gasparri which is characterized by words with a controversial tone such as *vergogna, incapace, grillini, demente.*

Figure 4 shows the word cloud of Giuseppe Civati which is characterized by the absence of the theme of the pandemic. Conversely, there are words concerning culture and the case of Silvia Romano. Lastly, in Figure 5 there is the word cloud of Laura Boldrini which is characterized by the theme of social rights (*donne, lavoro, democrazia*).

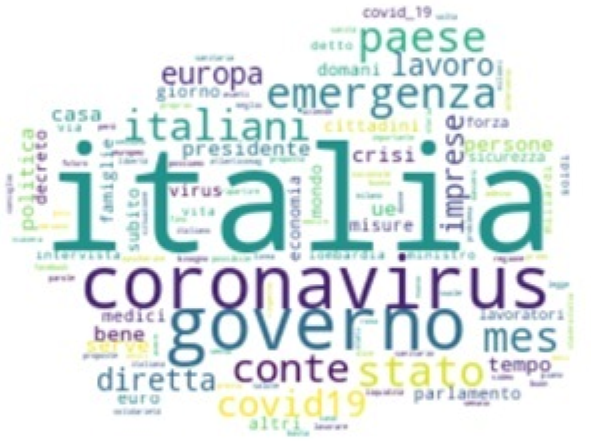


Figure 1: Wordcloud considering all tweets.

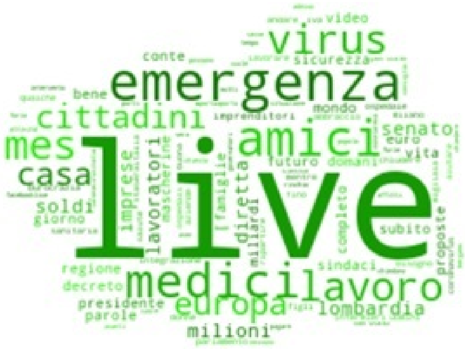


Figure 2: Wordcloud of Matteo Salvini.

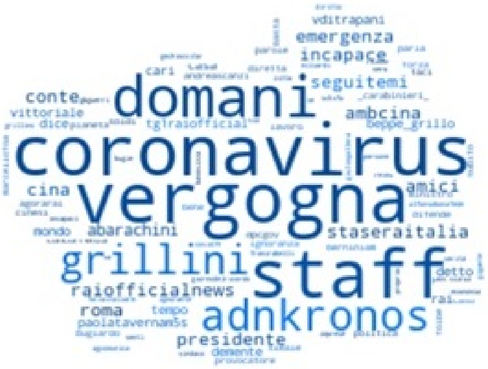


Figure 3: Wordcloud of Maurizzio Gasparri.

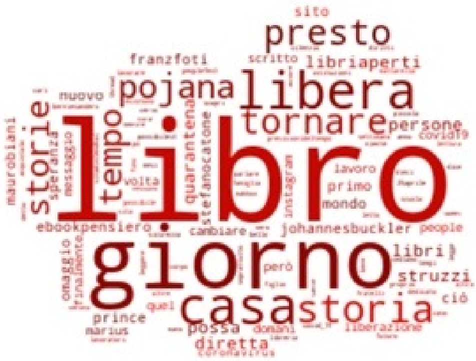


Figure 4: Wordcloud of Giuseppe Civati.

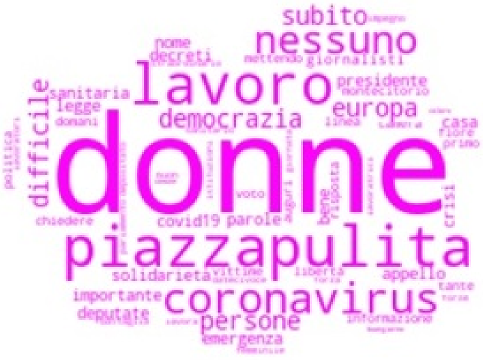


Figure 5: Wordcloud of Laura Boldrini.

Methodology

The project aims is to assign to each tweet a topic in such a way as to identify which themes were more discussed during the pandemic. Hence, we applied the Latent Dirichlet Allocation (LDA). The idea behind this method is that each text document can be described as a mixture of topics and in turn, the topics can be described as a mixture of words. More specifically, each text document is described as a probability distribution on latent topics and the probability of a topic in a document is described by a prior Dirichlet distribution. In turn, each latent topic is described as a probability distribution of words and the probability of a word in a topic is described by a prior Dirichlet distribution. In our case, the text documents are tweets and the authors of the document are the politicians.

In our case, we also considered some versions of LDA. That's because the tweets are not like ordinary documents but are characterized by a few words. Moreover, there are many documents (tweets) for each author (politicians) and consequently, a strong correlation structure is present. In such cases, the LDA could be less performing, and for this reason, we considered two alternative methods.

The first one is to put the tweets of the same politician in one document and then fit the LDA on this tweet collection. In this way, we obtained longer documents and eliminated the correlation between the tweets of the same politician. However, this approach has not been effective because the corpus of documents is small (few politicians) and they are quite similar to each other.

The second approach is to apply some extensions of LDA.

One is called Author Topic Model and takes the author of the documents into account. This model presents a three-step document generation: the author is selected, then given the author topics are chosen and finally, given a topic, the words of the document are generated. However, this model did not bring satisfactory results, probably because the model is designed for a corpus of documents where each document has many authors.

The second extension of LDA considered is the Dirichlet Multinomial Mixture model. It assumes that each document has at least one topic. this method seems to be able to fit our data very well since the tweets are short text documents and are generally about one topic. Unfortunately, we didn't find an implementation in Python.

In conclusion, we decided to present the results of the initial approach that considers each tweet as a single independent document because, in our opinion, it provides the best results. However, we considered only tweets with more than 200 words to have fairly long documents.

For the implementation of LDA, we used the library *Gensim* in Python. The estimate of the model is based on *Variational Bayes.*

Result

The model needs to set the total number of topics. We fit different models for values ranging from 5 to 50. We chose the best model by looking at which one assigned the topics most

Topic 4 aid to enterprises	**Topic 6 covid19 emergency**	**Topic 7 economy**	**Topic 14 hospital**	**Topic 16 Europe and Mes**
imprese	coronavirus	euro	medici	mes
liquidità	emergenza	miliardi	infermieri	ue
famiglie	governo	stato	lavoro	europa
aziende	misure	risorse	personale	bce
lavoratori	covid19	pil	operatori	eurogruppo

Table 1: Consistent topics

Topic 1 intensive care unit	**Topic 10 economic crisis**	**Topic 11 infections and swaps**	**Topic 18 economy school**	**Topic 8 undefinable**
storia	sicurezza	contagi	crescita	*staseraitalia*
terapia	p.iva	tv	crescita	vita
intensiva	Draghi	social	investimenti	Bergamo
parlamento	crisi	tamponi	scuole	persone
leggere	misure	video	contagi	sotto

Table 2: Inconsistent topics

properly. The best-resulting model was the one with 20 topics. The output of LDA is a probability distribution of topics for each tweet and a probability distribution of words for every topic.

For some topics (table 1), the probability distribution of words defines specific themes. For others (table 2) it is more difficult to define a title for the topic from the probability distribution of words.

Then we defined a title for each topic and we assigned to each tweet the topic with the highest estimated probability thereby highlighting the most discussed topics. Table 6 shows that the most common topics reflect the different aspects of the pandemic. The theme of employment is one of the most discussed which is related to the political debate that has taken place on aid to companies, layoffs and VAT. The other themes highlighted are the government and its decisions, the health emergency, the economic crisis and the European Union.

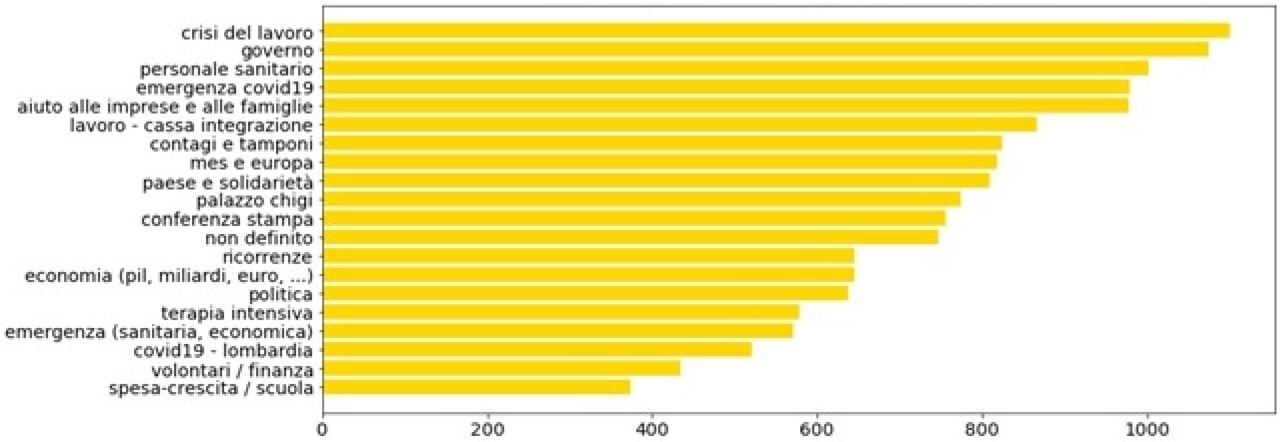
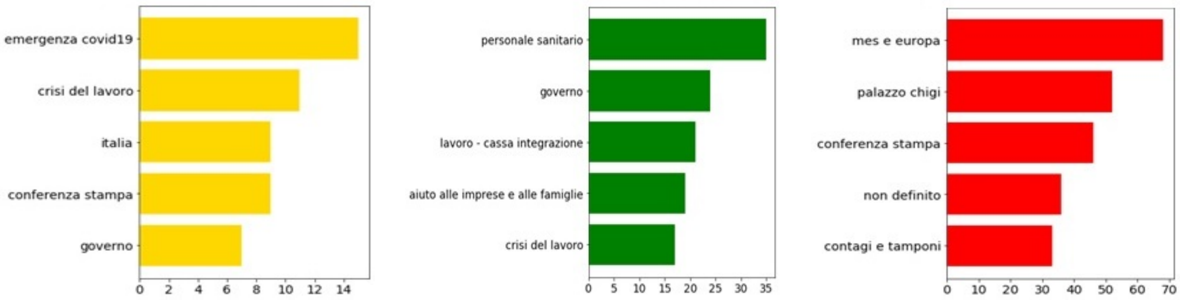


Figure 6: Overall frequency of topics



(a) *Giuseppe Conte* (b) *Nicola Zingaretti* (c) *Alberto Bagnai*

Figure 7: Frequency of topics for politician

Finally, we analyzed the frequency of the topics for some politicians presented in the table (7). Giuseppe Conte (first minister, Movimento 5 Stelle) discussed the Covid19 emergency, Nicola Zingaretti (Partito Democratico) dedicated more tweets to healthcare workers while Alberto Bagnai (Lega) presents more tweets about the European Union and Mes.