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 has been discussed on Twitter by Italian politicians during Covid19 pandemic. Therefore, we have 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 have been downloaded through Twitter API using the function *get_time* of the library *rtweet* of **R**. Before applying the model, the tweets have been preprocess. Firstly, every tweet has been turned in a list of tokens. Then, every list has been filtered so that it would not contain scoring characters, special characters, numbers, articles, conjunctions and pronouns. The names of politicians and parties has also been deleted thereby avoid correlations between them and other words. Moreover, bigrams have been formed starting from the tokens (such as *emergenza_sanitaria*). Finally, the tokens present in less 10 tweets have been discarded thereby not including rare words. We decided to not reduce tokens to their fundamental root since to obtain a better interpretation of the results. We have taken this decision after some tests.

Descriptive analysis

Before applying the model, we did some descriptive analysis. We made some cloud words to display the most frequent words: the bigger the worded, the more frequent it is. The figure 1 shows the wordcloud 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) together with the common word of political and economic Italian debate (lavoro, europa, mes). Lastly, words related to communication are present such as live, diretta, amici.

Figure 1: Wordcloud considering all tweets.

Then we analyzed the wordclouds for some politicians. In Figure 2 there is the worldcloud of Matteo Salvini: the most frequent word are about the social communication and covid19.

In Figure 3 there is the worldcloud of Maurizio Gasparri which is characterized of words of tone polemic such as vergogna, incapace, grillini, demente.

Figure 2: Wordcloud of Matteo Salvini.

Figure 3: Wordcloud of Maurizzio Gasparri.

In Figure 4 there is the worldcloud of Giuseppe Civati which is particular for the absence of the pandemic theme, instead we found words concerning culture and the case of Silvia Romano. Lastly, In Figure 5 there is the worldcloud of Laura Boldrini which is charachtherized by the theme of social rights (donne, lavoro, democrazia).

Figure 4: Wordcloud of Giuseppe Civati.

Figure 5: Wordcloud of Laura Boldrini.

Methodology

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

For our case, we considered some different versions of LDA. This because the tweets are not like common text documents, they consist of few words Moreover we have many documents (tweets) for every author (politicians) and consequently there are a lot of correlation. In such cases, the LDA could be less performing. For this reason, we considered two alternative.

The first is to group the tweets of the same politician in one document and then fit the LDA on this corpus. 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 document is small and they are similar to each other.

The second approach is to apply some extension of LDA. One is called Author Topic Model and take account also of the author of the documents. This model provides a three.step document generations: the author is selected, then given the author the topics are chosen and finally, given a topic, the words of the document are generated. However, this model did not bring to satisfactory results, probably because the model is designed for a corpus of documents where every document has many authors. The second extension of LDA considered is Dirichlet Multinomial Mixture model. It assume that every document has at least one topic. This model seems to fit our data very well since the tweets are short text documents and generally are 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 consider every tweet as single independent document because, in our opinion, it provides the best results. However, we considered only the tweets with more than 200 words in order to have fairly long documents.

For the implementation of LDA we us 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 model for values ranging from 5 to 50. We chose the best model by looking at which one assigned the topic in most appropriate way. We chose the model with 20 topic. I remind you that the output of LDA is a probability distribution of topics for every tweet and a probability distribution of words for every topic. For some topics (1), the probability distribution of words define a specific themes. For others (2) it is more difficult to define a title for the topic from the probability distribution of words.

Then we defined a title to each topic and we assigned to each tweet the topic with the highest estimated probability thereby to highlight the most discussed topics. From 6 it is possible to see that the most common topics reflect the different aspect of the pandemic. The theme of employment is one of the must discussed which is related to the political debate on that has

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

taken place on aid to companies, layoffs and VAT. The other themes highlighted are about the government and its decisions, the health emergency, the economic crisis and European Union.

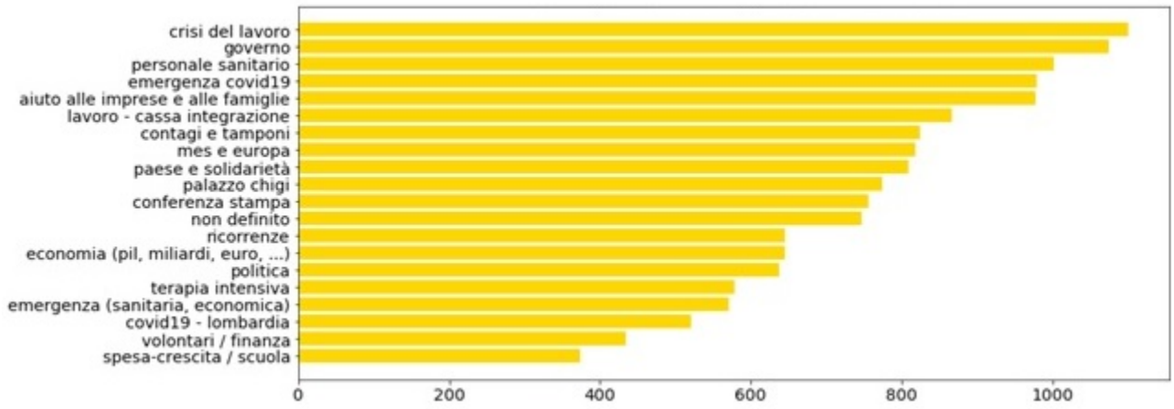
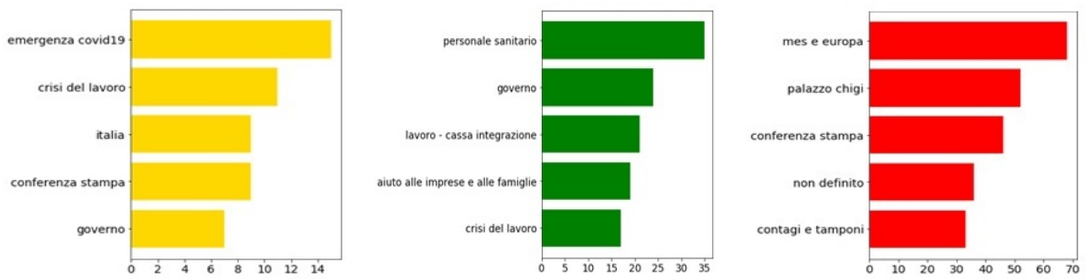


Figure 6: Overall frequency of topics

Finaly, we analyzed the frequence of the topics for some politicians (7). For example, Giuseppe Conte (first minister, Movimento 5 Stelle) discussed more about the covid19 emergency. Nicola Zingaretti (Partito Democratico) dedicated more tweets to healthcare workers while ALberto Bagnai (Lega) presents more tweets about European Union and Mes.



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

Figure 7: Frequency of topics for politician