UNIVERSITY OF TRIESTE

Natural Language Processing project

Sentiment analysis regarding tweets about covid-19 vaccines

Giovanni Pinna

Academic Year 2020-2021

Summary

1 Problem

The goal is to do the sentiment analysis on the tweet of the Italian population has regarding the various vaccines on the market against Covid-19. In particular, the vaccines that have been taken into account in this analysis are those approved to date by the EMA (Pfizer, Astrazeneca, Moderna) plus Sputnik.

It was also decided to consider the Sputnik (Russian) vaccine to observe whether the sentiment differs significantly in comparison to the approved vaccines of Italy and the EU.

The data was taken from the social network Twitter some with a developer account allows you to download the tweets and data associated with them quite easily.

To identify the Italian population it was decided to assume that Italian users are all those who have written a tweet in Italian and that have as their place of profile an Italian city or the Italian state.

From this analysis we expect most tweets to be marked with negative sentiment (like 70%). This is because normally people tend more to complain on social networks than to express positive concepts. In addition, complaints are increasingly resonant and affect the user who is therefore more likely to share them.

2 Datasets

The data that has been used all comes from the Social Network Twitter. This is because such a social network allows you to download tweets and other data, which we will see later, easily once you have a developer account.

To download the data it was necessary to get hold of the various tokens and secret keys (accesstoken, accesstokensecret, apikey, apisecretkey) that are made available once the developer account is obtained. When this is done you have used the Python Tweepy library which easily allows you to download the data. In particular, to make the work even easier, we used the function "Cursor " of the library, which automatically downloads tweets according to the search criteria used (the programmer does not have to deal with the tweets pagination).

The search criteria used to download the tweets we are interested in are:

data_since = (start date of research) set at 18/04/2021
data_until = (research end date) set at 28/04/2021
language = set to 'it'
search_word = (keyword) set independently of the vaccine under consideration

As mentioned above we have not only downloaded the text of the tweet, but many other data so as to make, even future analyses, much more complete and consistent. In particular, the lables are:

[tweets , is_quoted, tweet_quoted, name_of_who_I_answered , id user_str, lang_user, creation_date_of_tweet, result_type, number_of_retweet, result_type, number_of_retweet, retweeted, source, user_name, user_screen_name, location, user_description, number_of_follower, number_of_friend, is_verified, creation_date_of_account, like]

(Table A describes them in more detail).

Once all this data was downloaded, we moved on to cleaning it.

3 Pre-processing

The process of pre-processing the text is extremely important in order to make documents less subject to noise and variance.

For this project it was decided to apply two types of pre-processing, one hard and one soft. This is done because not all the algorithms that we are going to use will need a hard pre-processing.

Before applied pre-processing, it was necessary to delete the lines that had the same tweets. This is done because many people simply retweeted without any comment. So in our dataset there were many lines with the same tweet maybe retweeted by two or more different users. This decision was dictated by two factors: the first is that we did not want to assume that: if a person retweets a tweet we assume that he thinks in a equal way to the text that he retweeted. The second motivation concerns the computational cost, in fact working and processing a huge amount of tweets was heavy and extremely time-consuming.

So once you deleted all the rows that had the exact same tweet, you could proceed with the pre-processing bellow.

3.1 Hard preprocessing

For hard pre-processing it was decided to delete all user names, hashtags, links and special characters and all numbers were transformed into 0 using regex. Before deleting it, however, this data was saved in special columns which were used during analysis paragraph.

In addition, the sentence was also processed by placing it in lower case and each word was reduced to its lemma. POS parsing was used to hold words that had the following attributes {'NOUN', 'VERB', 'ADJ', 'ADV', 'PROPN'}.

3.2 Soft pre-processing

As for soft pre-processing, it was decided to delete only the links and special characters. Each username has been replaced with the word "user".

It was decided, in contrast to hard pre-processing, to keep hashtags without the special character "#" assuming that during the training of algorithms this more data could help in a good prediction or a better understanding of the text.

3.3 Other types of pre-processing

For the analyzes that we will see later it was also necessary to clean up the column concerning the location of the various user profiles. This is because there were names that indicated the same place, but written differently (e.g. "Milano","Milan", "Milano, Lombardia", "Milano, Italia" , all these combinations were collapsed in the word "Milano"). In this case, special regular expressions were created for each main city. Creating a regular expression for each city would have been too onerous and even without much sense because very few tweets were collected in non-main cities (as we will see in the analysis part).

Another pre-processing that needed to be done was on the artificially created column used to save the hashtags. In fact, with them were also saved the special characters not useful and not significant for the future analysis. So always through regex these special characters have been deleted.

4 Sentiment analysis with BERT

As the first model I decided to use is the model based on BERT built by Bocconi University (feel-it). This model has the peculiarity of being pre-trained on data in Italian and therefore usable on the tweets I downloaded.

This model had the task of doing the sentiment analysis on the text of the clean tweets with the soft pre-processing procedure. In fact, this neural network was created to understand natural language and therefore there was no need to do hard pre-processing to decrease noise and variance.

The first model used was this not only to get a general idea of sentiment in the various tweets, but also to use its predictions as the true labels for other models. In fact, the predictions provided by the Bocconi model have been used as a labels for all the models that we will see later.

For analysis purposes and to make a check on the predictions of sentiment I also used the analysis of emotion made available always by the same model.

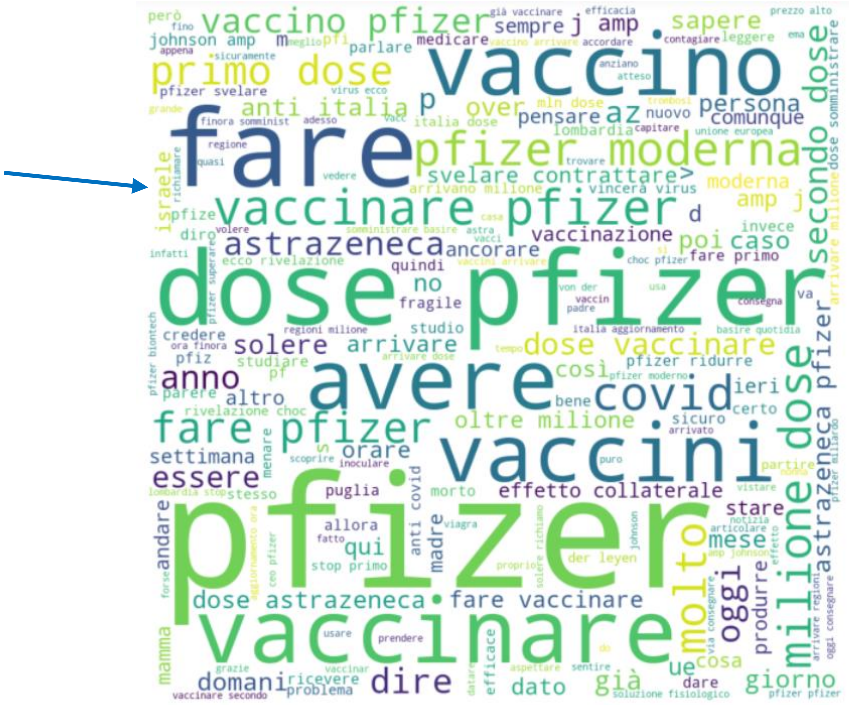
5 Analysis

At this point we have the dataset that has two more columns: one for sentiments and one for emotions.

First we decided to see which words appear the most through a "WordCloud" representation. (In the representation " WordCloud" stopwords have not been taken into account).

5.1 WordCloud

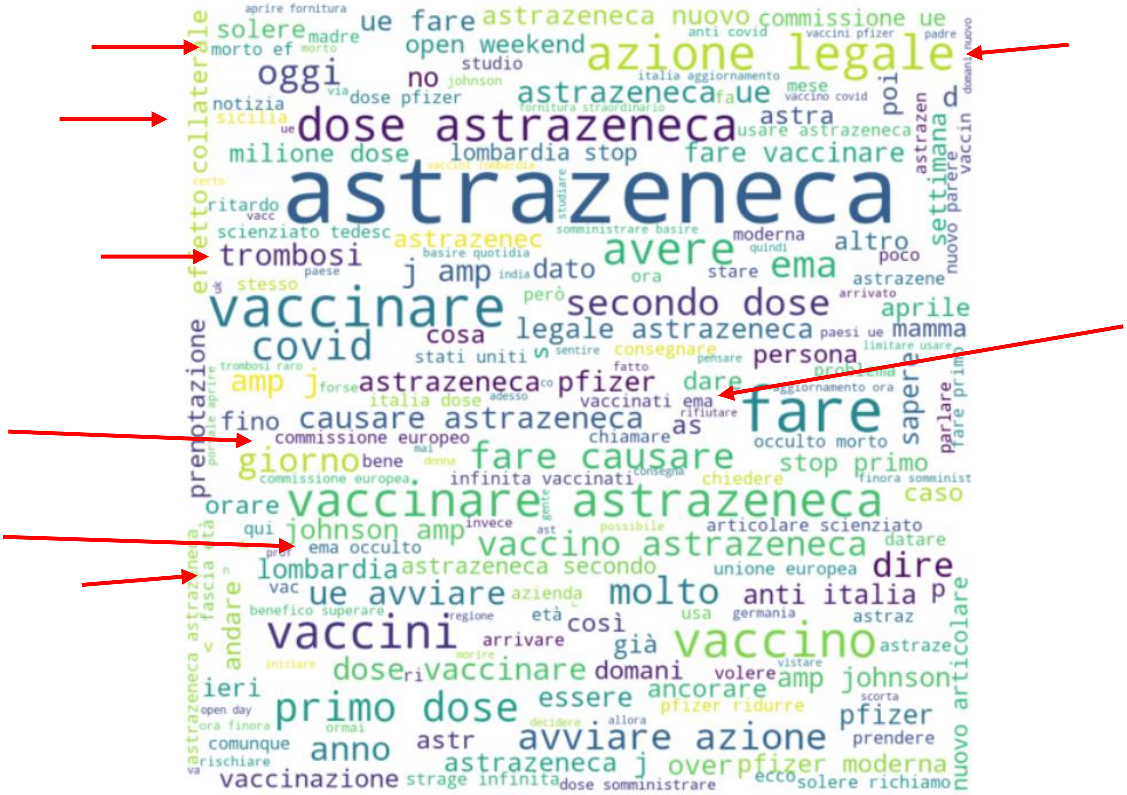
5.1.1 Pfizer



As we can see from the image, you can immediately notice the words that are the biggest correspond to the key words in our search. This is normal because they are all words that will be present in almost every tweet. As far as Pfizer is concerned, we can see the words 'million dose' which makes clear reference to the stocks that should be coming in Italy in this period. You can also notice the names of other vaccines: Astrazeneca, Moderna and j amp j (johnson & johnson). There are no words that stand out with a clearly positive or negative connoinois. It seems that we are talking above: to do the doses, other vaccines and the arrival of a stack of a million doses.

I think it is important to note (blue arrow) how words have also been captured concerning the progress of the pandemic and vaccination situation in other states. First of all Isdraele which to date is the state first in vaccination and which has mainly used Pfizer vaccine.

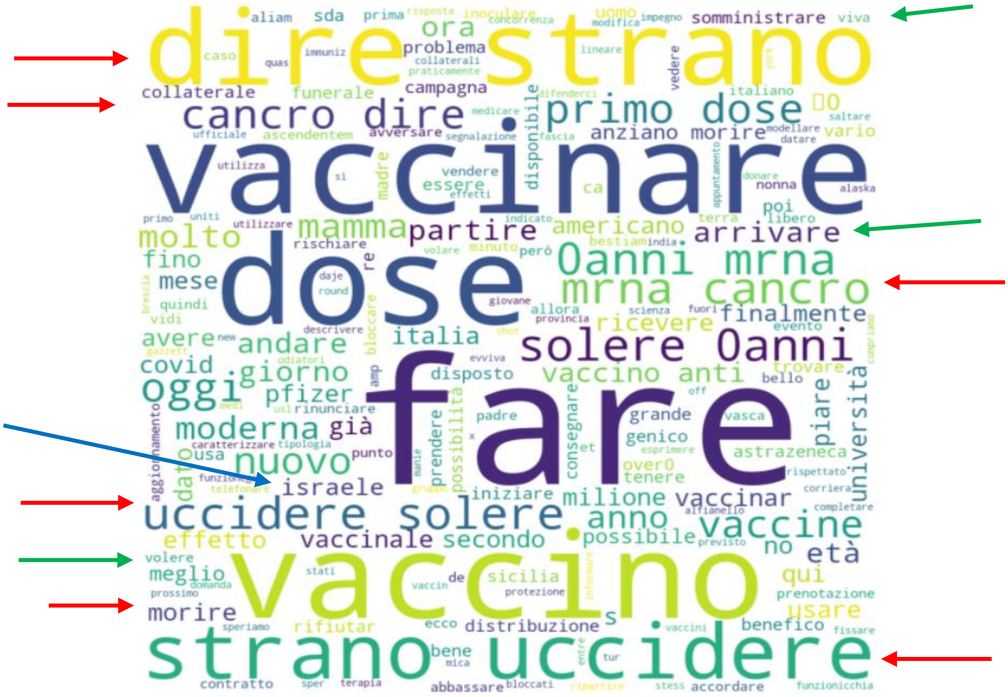
5.1.2 Astrazeneca



As for Astrazeneca's wordcloud in addition to the words obviously present such as "Astrazeneca" and "vaccine" it can be noted that the size , and consequently the importance of words, is sum up all the same. So as far as Astrazeneca is concerned, there are no possible main topics that we can immediately identify.

In any case, a more in-depth analysis can find words that refer to negative topics such as: cases of thrombosis, side effects and legal action against the company. A few words concern the EMA and its approvals.

5.1.3 Moderna



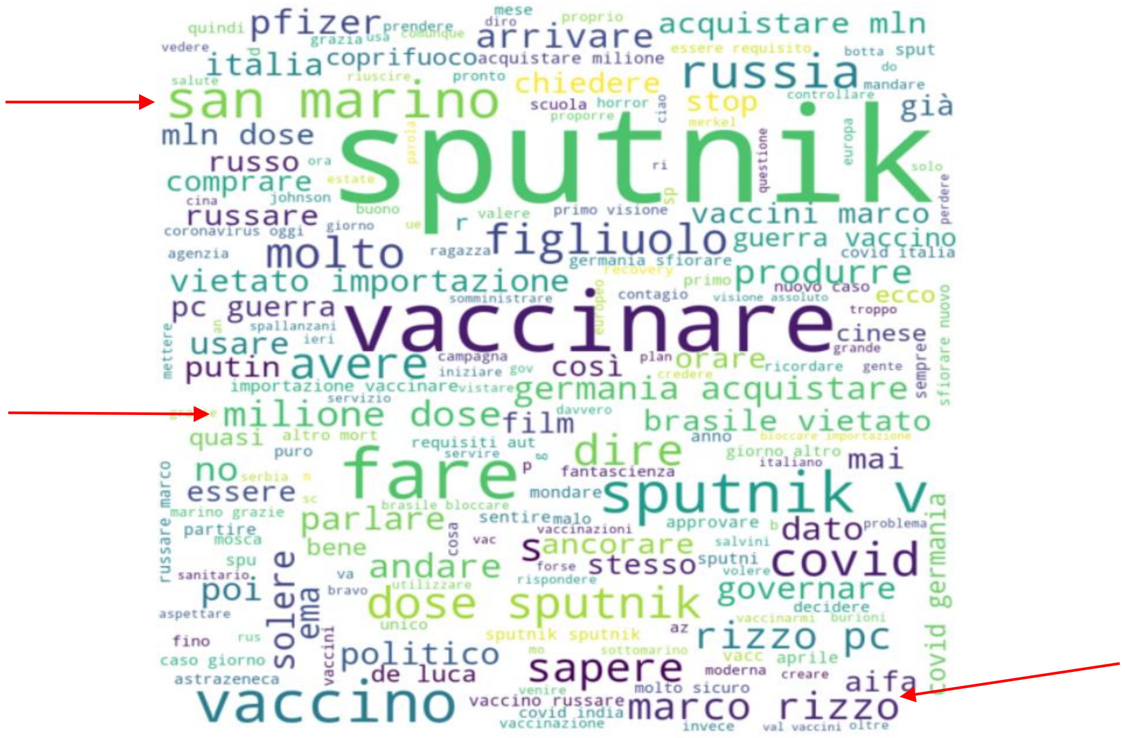
From this image you can already see what are the most important words to which we can intuitively trace back to a positive or negative connoted.

In particular, the words indicated by the red arrows have a negative meaning since they speak of death or diseases. Probably attributable to tweets that want to emphasize the possibility that with vaccination you can have very serious complications.
The green arrows, on the other hand, indicate words that are probably related to tweets of positive feeling.

I think it is important to note (bluearrow), as in the case of Pfizer, how words have also been captured that also concern the progress of the pandemic and vaccination situation in other states. This was also expected because most of the time we talk about Pfizer and Moderna together (as we will see later on the word frequency analysis part).

From the image you can think that there will be more tweets with a negative sentiment having managed to identify more words that lead us to think that they derive from negative tweets.

5.1.4 Sputnik



From this word cloud I expected to see more words that have a negative feeling and mistrust, but surprisingly it didn't happen. In this image you can see the words "russia" and "putin" which are quite obvious given that the vaccine is Russian and the figure of Putin is linked. There are words like "comprare" that can suggest that many of the tweets are about buying this vaccine for to be able to have as many doses as possible.

One of the words that stands out on the others is "san marino". In fact, this nation was the first in the Europe area that vaccinating its population with the Sputnik vaccine.

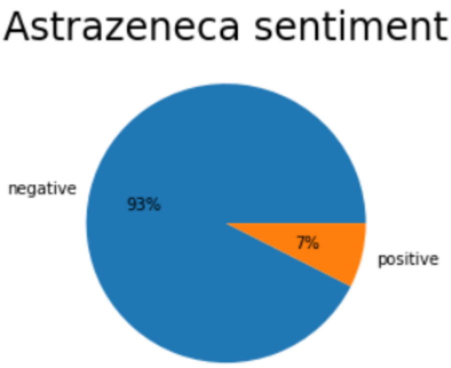
Another pair of words that is visible is "Marco Rizzo" (secretary of the Communist Party) who is an Italian politician who says that the Sputnik vaccine is not possible to use in Europe because: "there is an Atlantic bloc that does not want". He has also expressed a similar opinion on other vaccines such as the Chinese one. His claims must have generated a big response on social networks and that's why we find his name in the image.

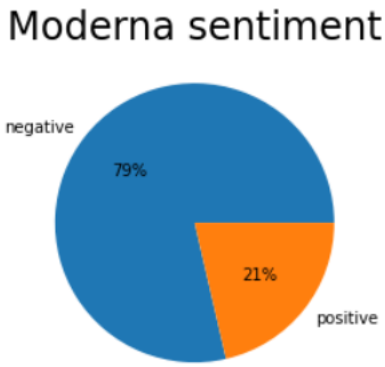
It can also be seen that though all the cleaning in the text and the selection that has been made there are words that do not center like "russare marco".

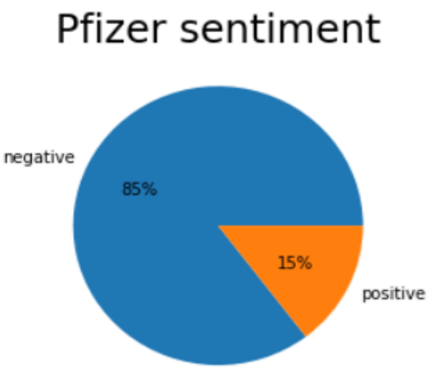
5.2 Sentiment analysis

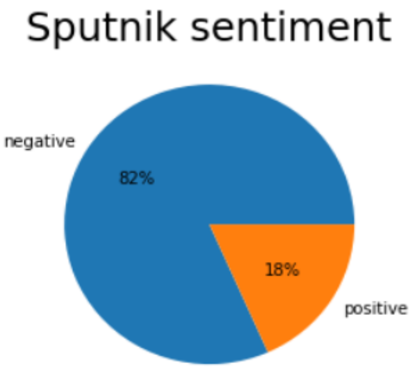
Now let's analyze the sentiment that has been predicted by the BERT-based neural network.

We expect there to be more negatives than positives, this is because the words seen in the word cloud make us suspect about this. In addition, the tendency to complain about people on social networks is to be taken into account.









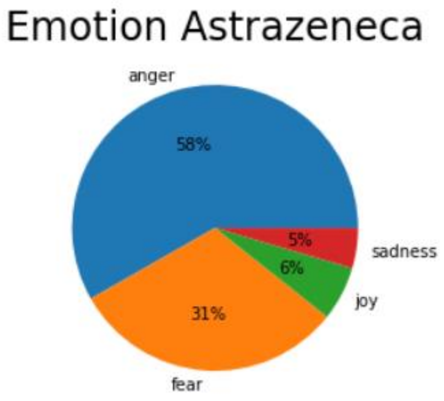
From these charts you can immediately see how much more tweets categorized as negative than positive tweets (far beyond our initial predictions). Astrazeneca was expected to have a predominance of tweets with negative sentiment. This is certainly due to all the scepticism regarding vaccines and reinforced by the fact that the first cases of thrombosis and death occurred precisely with the Astrazeneca vaccine. The fact that the EMA and AIFA have spoken out more about this vaccine have fuelled a feeling of mistrust and fear. The change of name of the company and the vaccine change the bugiardino must also have negatively affected public opinion.

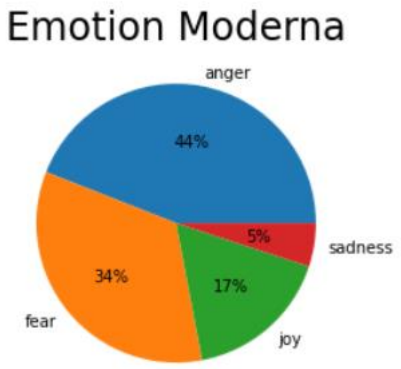
As for Pfizer we expected positive tweets to be in greater numbers than, for example Astrazeneca, but we didn't expect that many negatives. Particularly after the success of Isdraele's vaccination campaign I thought the positive feeling was greater, not much, than the onecurrently found. Perhaps this could be due to public conditioning after the Astrazeneca thrombosis cases. This skepticism could also be due to the fact that most people do not trust vaccines in analysis a priori since they create and distribute in few time. This is because normally for a drug it takes years for approval before it is put on the market.

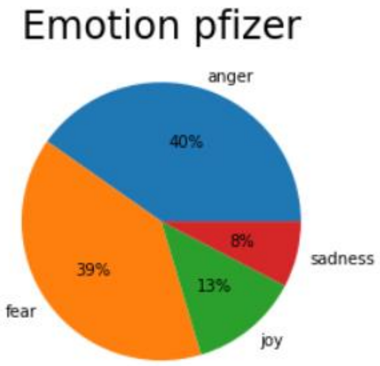
As far as Moderna is concerned, the sentiment of this vaccine is in line with expectations and perhaps also benefits from the fact that not so much is said about it compared to the other two approved by the EMA.

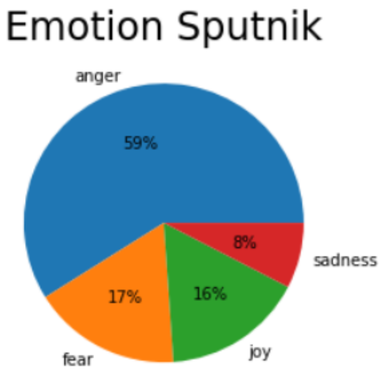
As far as Sputnik is concerned, I expected to find more negatives by not being an une a approved vaccine from Russia, a state that is not transparent in common opinion and does not have much confidence for Europeans. Perhaps the fact of having a sentiment in line with other vaccines could be due to the fact that on the days of downloading tweets there was talk and pushed for it to be approved. In the same days Spallanzani was doing the first tests with Sputnik vaccine. Another event that may have changed people's opinion a little were that some regional governors, popular among citizens, considered buying doses of the Russian vaccine, thus feeding the confidence of the population such a vaccine.

5.3 Emotion Analysis







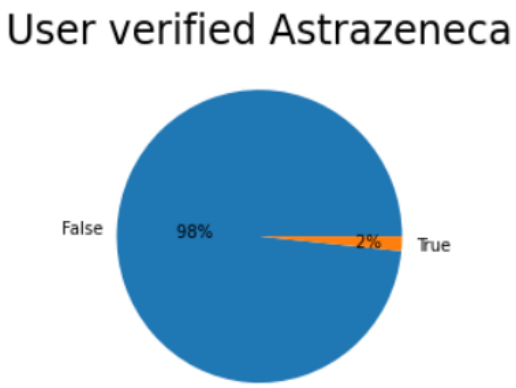


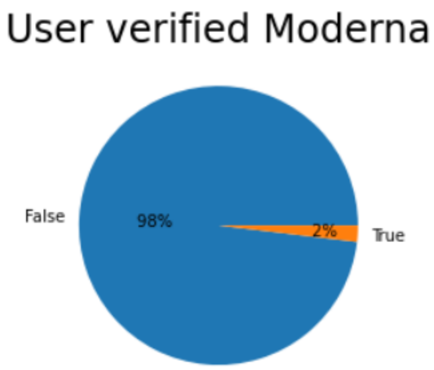
I also used emotion analysis as evidence to really verify that negative tweets were characterized by a "negative" emotion. In fact, this statement can be immediately felt by the charts where the only positive emotion "joy" is in a clear minority compared to the others that have a more negative connotation. In particular, it is always interesting to take into account the case of Astrazeneca in which the predominant emotion are the "anger" and "fear". This is probably due to cases of thrombosis and the fact that the company has not gained the trust of citizens (e.g. change of the bugiardino, various second thoughts of the drug approval bodies and the fact that the EU now wants to proceed legally with the company for not meeting the agreements). As for Pfizer and Moderna the feeling of "fear" is higher than in Astrazeneca this perhaps because the side effects (which probably generate the emotion "anger") following the administration of these vaccines did not have the same media resonance as those of Astrazeneca, but remains the feeling of suspicion about vaccines.

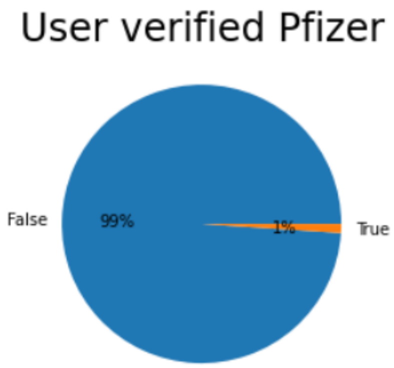
Sputnik has the positive feeling" joy " in greater quantities than other vaccinations, it may be because, as also mentioned in the previous section the vaccine was being considered for possible approval.

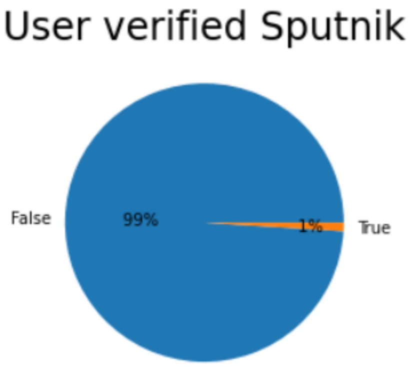
5.4 User verified analysis

To try to better understand where these negative feelings came from and if they were induced by something they tried to make various hypotheses. One of them is that verified users, so those who have more resonance and importance than a normal user of the social network may have conditioned others with their ideas.

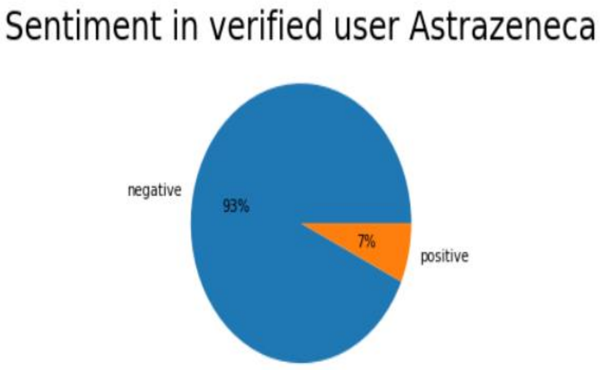


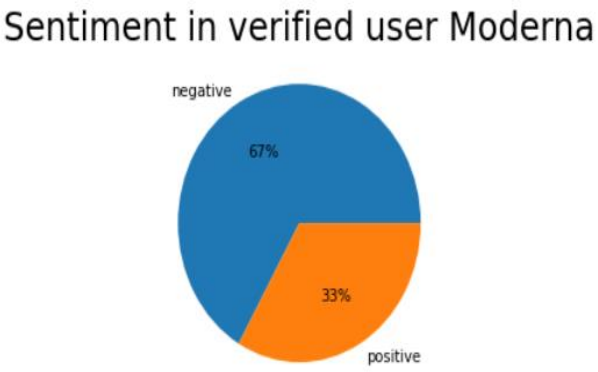


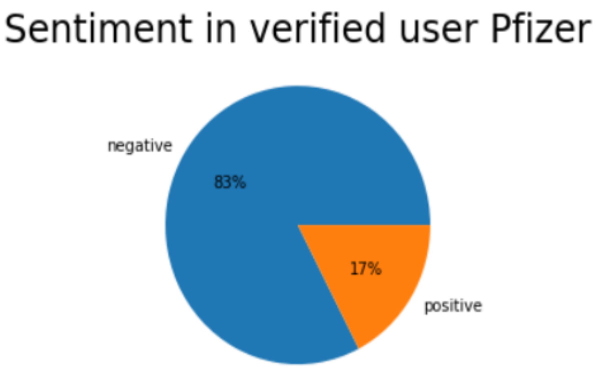


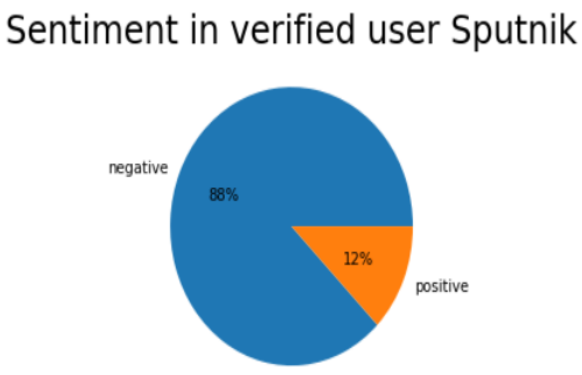


Seeing these charts you can immediately see how the percentage of users verified in our dataset is so small. So that you can immediately reject the hypothesis made previously. Especially since as can be seen from the image below the verified users they sum up the trend of global sentiment.





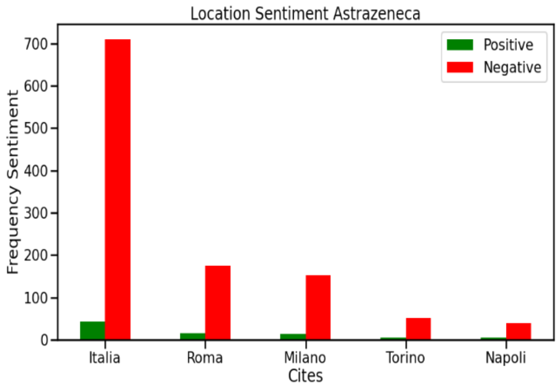


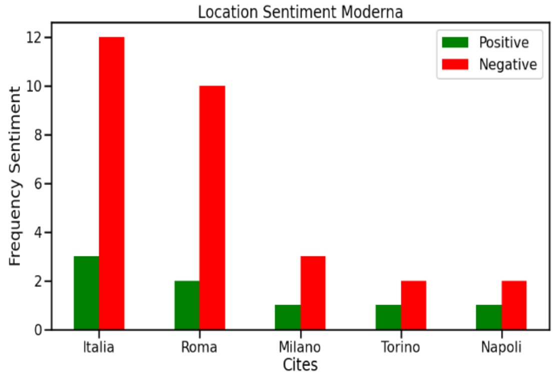


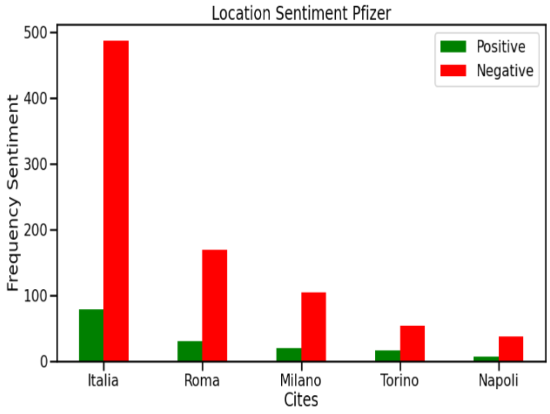
Only for the Moderna vaccine you can see that the percentage of tweets of verified users are a third of the total. It may be for this reason that Moderna has the most present positive feeling compared to other vaccines. (See Appendix A to find out what the most influential users think.)

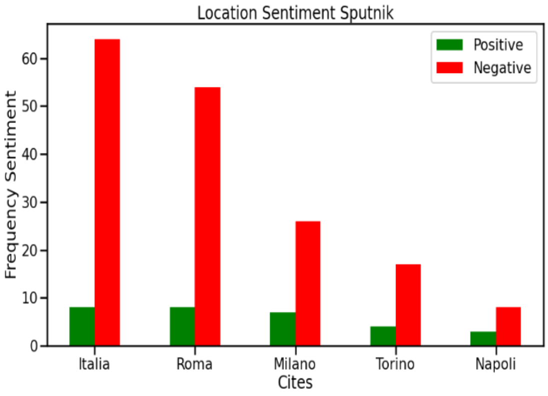
5.5 Location Analysis

In this sub section the intention is to go and see if the distribution of negative and positive tweets changes based on the user's location. In fact, it may happen that cities that have been very affected by the pandemic situation are more likely to get vaccinated and, as a result, there are more positive feelings regarding vaccines.





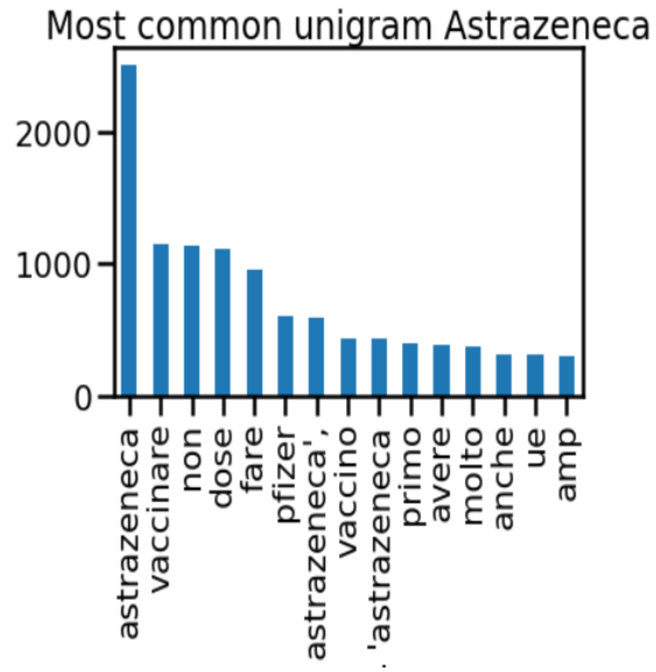


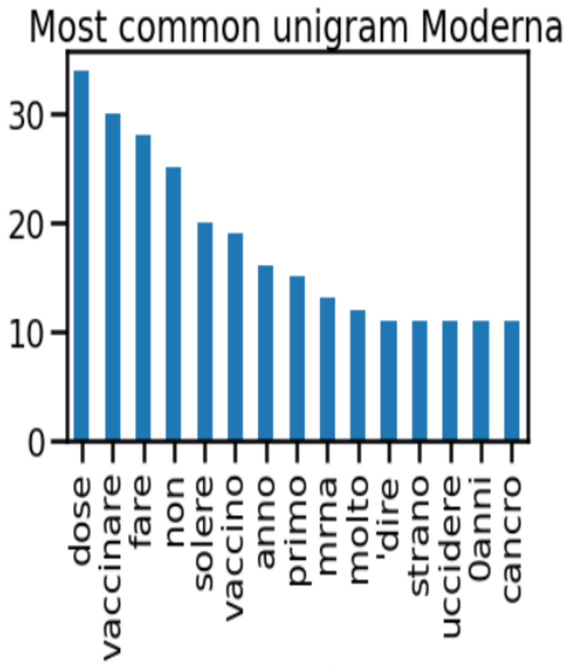


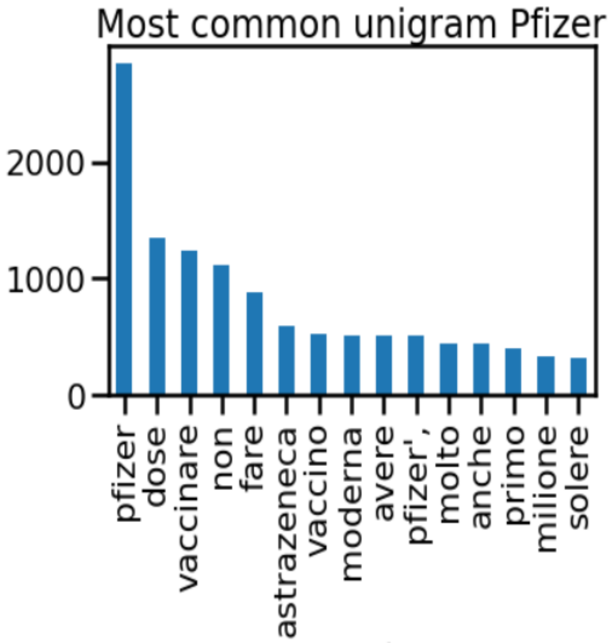
From these histograms we can see how even in the main cities including Milano, which is one of the most affectedcities, the distribution of sentiment remains substantially the same as before. It is therefore not possible to say that according to the user's place it thinks more positively or more negatively than the various vaccines.

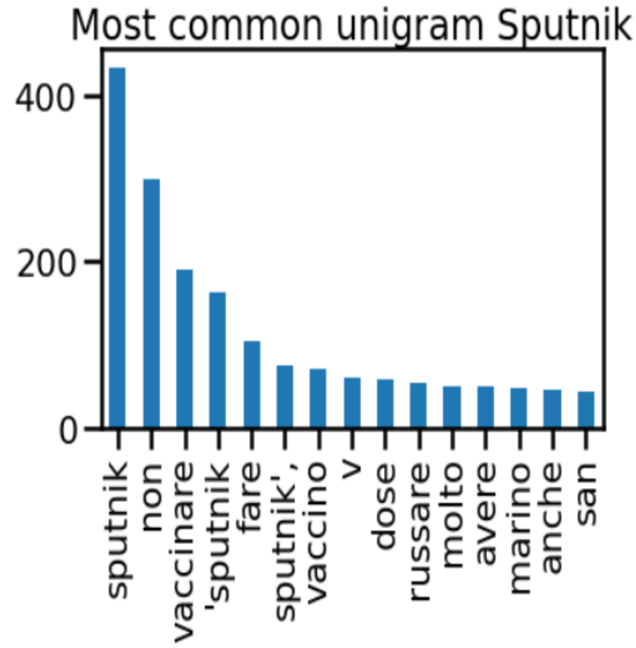
5.6 Most frequent word

In this section we want to analyze the words trying to understand which are the most frequent and what meaning they give to sentences. A bit like what has already been done in the section concerning the word cloud, but now taking into account fewer words, but knowing exactly their frequency.



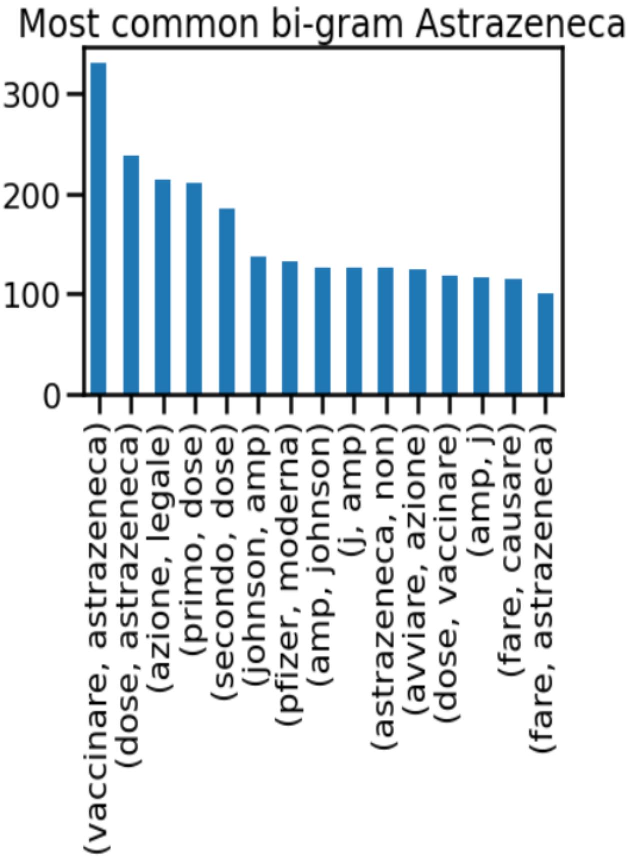


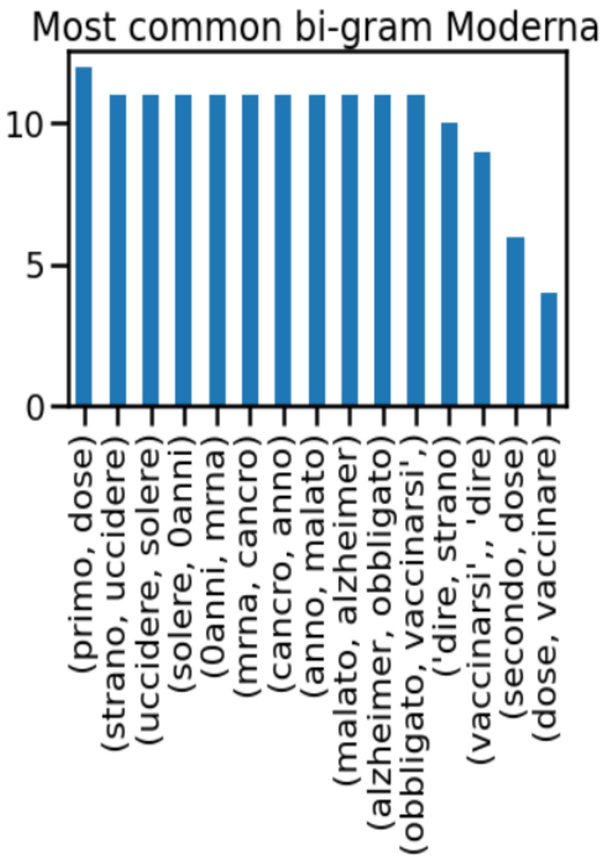


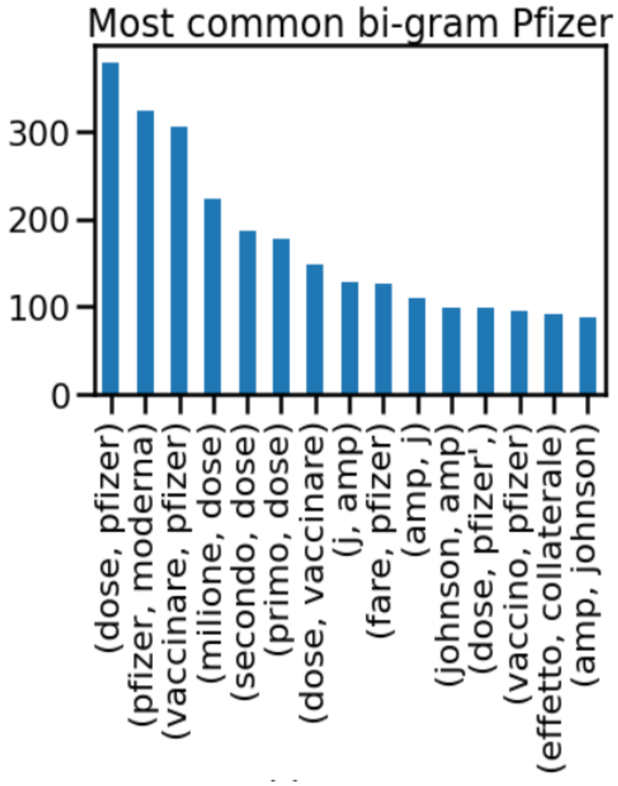


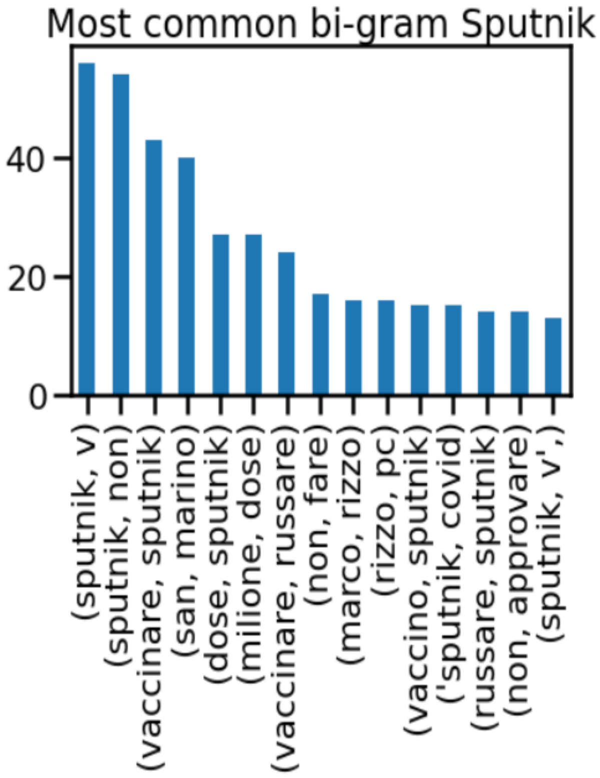
Let's see how in the word cloud that the most frequent words are the names of vaccines themselves and of course the word "vaccino". Very interesting to note, which was not take in the word cloud, how the word "non" appears among the first positions in each chart and with it also the word "fare". It could be that both words appear together thus giving a negative connotation of most tweets (agree with what was analyzed before).

You see how analyzing only single words can't capture all their meaning. In fact, in the last graph we see that the words "san" and "marino" that are analyzed separately when we know very well that it should be together. Consequently, it was also decided to analyze the frequency of bi-grams to try to extract much information as possible from the frequency of words in the corpus.









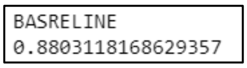
Let's see how the word "non" often occurs near the name of a vaccine, or near the word "fare" which confirms the thesis described above. The only chart where the word "non" doesn't show up is Pfizer's chart. This is because in bi-grams there is noise due for example to combinations with the J&J vaccine that takes several positions. In particular, if we were to take more positions we could look at the combination "pfizer , non" (80 appearances) and "non avere" (80 appearances) respectively in positions 17 and 18 (only the top 15 are shown in the graph).

6 Prediction

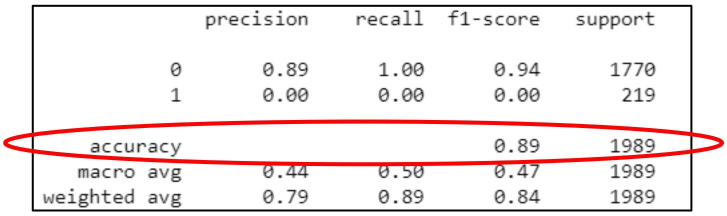
In this part we tried to implement machine learning models based on Logistic Regression and evaluate their performance. In particular, it was noted that these models did not perform much better than their "most_frequence" baseline. Machine learning models perform at a performance comparable, if not the same, of the baseline. This fact could be due to the great imbalance that the dataset presents. In fact, as we recall, more than 80% of the observations have a negative feeling. Having ascertained this, it has been tried to improve the various models by making features selection and regualrization. After applying these two techniques, the model did not show any particular improvements.

As a result, I hypothesized that the problem is caused from the fact that we didn't have enough data to train the model at its best. To try to solve this problem I decided to merge all the datasets creating one with more than 10,000 observations and to fit the Linear regressions on this. Even with this methodology, however, accurancy (a parameter that I took into account to evaluate when the model was good) did not rise much remaining always comparable to the baseline.

Now we will comment on the results deriving from the application of Logistic Regression to the total dataset, but for individual datasets (therefore containing information regarding only one vaccine) can be made similar observation.



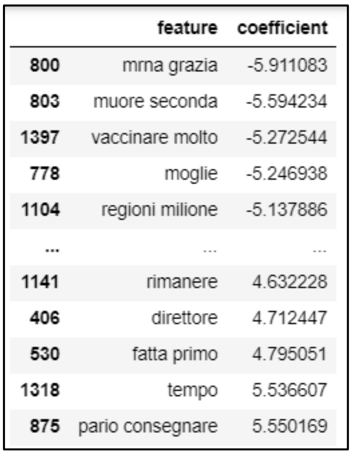
```
BASRELINE
0.8803118168629357
```



```
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      1770
           1       0.00      0.00      0.00       219

    accuracy                           0.89      1989
   macro avg       0.44      0.50      0.47      1989
weighted avg       0.79      0.89      0.84      1989
```

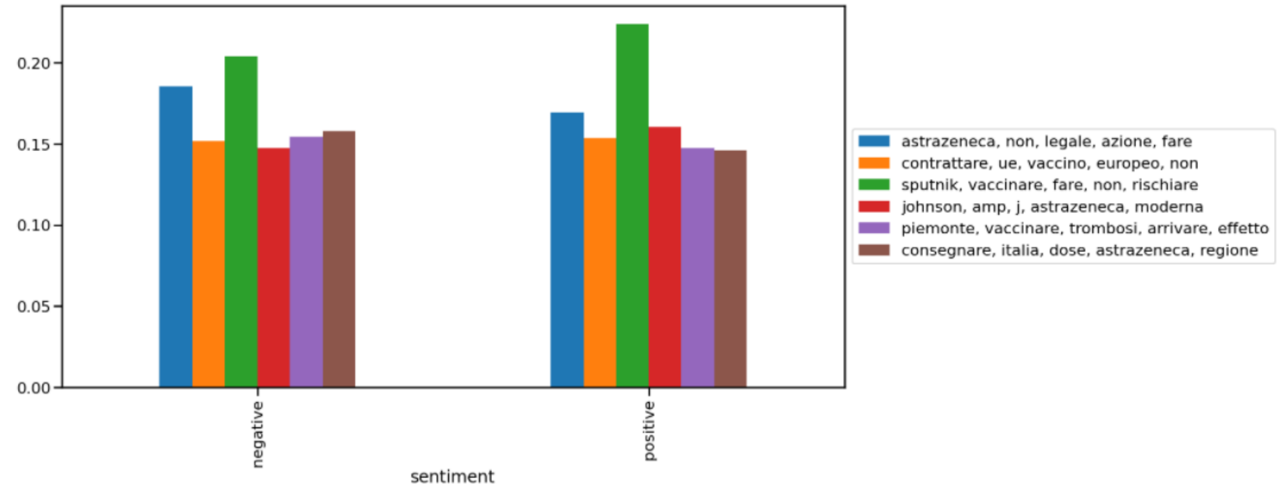


	feature	coefficient
800	mrna grazia	-5.911083
803	muore seconda	-5.594234
1397	vaccinare molto	-5.272544
778	moglie	-5.246938
1104	regioni milione	-5.137886
...	...	...
1141	rimanere	4.632228
406	direttore	4.712447
530	fatta primo	4.795051
1318	tempo	5.536607
875	pario consegnare	5.550169

In the images above we reported the results of the model after applying the selection and regularization of the model parameters. Let's see how the baseline and accurancy are particularly close and therefore comparable performance. We also wanted to bring back the coefficients that have features in our model. We see that the feature "vaccinare molto" and "mrna grazia" have very negative coefficients, while "tempo" and "pario consegnare" have a very positive coefficients. We cannot deduce much more from this simple table. Consequently, it was considered necessary to deepen with specific techniques the topics of the various documents and to understand if our results make sense.

7 Topic modelling

To analyze the topics of the various tweets we used the LDA technique. We applied it first only on negative texts and then only on positives. So you can understand for each feeling what the latent topics in our data. Finally, we applied the technique to the entire dataset to get an overall view of the topics. To choose the best number of topics we used the coerence scores techniques (UMass and CV) with which we found that the best number of topics is in between 5 and 8 (we chose 6). After that we extracted our descriptros. To see if these topics were more related to positives or negatives we did the aggregation based on the sentiment as you can see from the graph below.



This figure suggests some interesting observations.

In particular, we see how the green topic is more present in positive observations. The words that make up this topic suggest that we talk about the Russian vaccine and how maybe it is not a good idea to vaccinate with this vaccine. It seems correct that this topic is slightly more present among the positives. The tweets will probably concern the fact that people think that the Russian vaccine is risky and that it is more appropriate to get vaccinated with another vaccine.

The topic identified by the blue color is most present among negatives. Reading the words describing the topic one can say that we are talking almost unequivocally about Astrazeneca and the legal action that the EU intends to take against the company.

The purple topic is also slightly more present among the negatives. The words that compose it have a connoted that on average is negative, especially the word "trombosi". Probably this topic refers to the side effects of vaccines.
For the word "piemonte" it is not possible to find a great connection with the other words of the topic.

Finally, the last topic that seems interesting to comment is the brown one. In fact, this topic is more present among the negatives and it seems that it speaks of the fact that not enough vaccines arrive to Italy by pharmaceutical companies. This most likely generates angry and indignant tweets and consequently categorized as negative.

8 Conclusions

From the analyses made it can be concluded that most tweets express a negative feeling. This is due to multiple factors, both psychological and news facts. In fact, among the psychological/behavioral factors we can find the fact that people tend more to complain on social media thus generating more tweets with negative sentiment. In addition to this, in recent years they have been thinking worse and worse than vaccines. Therefore we are more suspicious on this topic especially for vaccines concerning Covid-19. In fact, people are very suspicious mainly about the fact that such vaccines were created and sold in a very short time compared to other drugs (although the news teaches us that in reality this is due to bureaucratic speeding, and investments both in terms of money and out-of-the-ordinary minds). People often express negativity on these topics also because side effects, lawsuits and other news events generate a lot of media importance and a huge stream, in our case, of negative tweets. In fact, negative news events have much more influence in public opinion than positive ones.

Following all these deductions we can say that it is normal to have a huge percentage of negative tweets.

As far as the prediction part is concerned, needless to say, we have not been able to have good predictions and that, at least with the logistic regression model, we cannot get far beyond the baseline. This leads us to think about implementing other, more complex models.

The topic analysis it gives us a good insight into the topics related to our tweets, most of which can more easily be traced back to something negative than something positive.

8.1 Possible future developments

Future developments could be to try to train other types of predictive algorithms such as SVMs and try to use perhaps even the data coming out form emotion analysis. In this way, more complex algorithms could be trained, but they probably have better accuracy.

Another possible implementation might be to implement a clustering algorithm and see how observations are grouped together. This could perhaps uncover other latent information in the data.

9 References

Bocconi BERT model for sentiment:
https://huggingface.co/MilaNLProc/feel-it-italian-sentiment

Bocconi BERT model for emotion:
https://huggingface.co/MilaNLProc/feel-it-italian-emotion

Twitter developer account and data dictionary:
https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

BERT original paper:
https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

LDA :
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Appendix A

Following the analysis on the verified users we tried to see also the user/users who have multiple followers, which does not necessarily mean that they are verified. It's fair to analyze them too as their tweets could get to a lot of people.

	Pfizer	**Astrazeneca Municipality**	**Sputnik**	**Modern**
user_name	La Stampa, 10	Tgcom24	iodonna	ilGiornale
user_screen_name	LaStampa, 1st	MediasetTgcom24	IOdonna River	the news
Tweet	'EU-Pfizer agreement for 1.8 billion doses, Von der Layen: 'By July 70% of European citizens will be vaccinated'	"AstraZeneca vaccine, Ema reiterates: no limit use for age groups #astrazeneca "	Science fiction lovers, honestly, how can you miss a movie with such a Cold War title, so vintage? https://t.co/JROypbrYe2	"There are no changes to the plan, commitments fulfilled." #Fontana explains the decision to allocate #Pfizer #Moderna to the first doses and reassures those who have to make the call with #AstraZeneca https://t.co/Aqip8kqM0X
number_of_follower	1195115	1155901	117191	537196
is_verified	True	True	True	False property
Sentiment	negative	Positive, 20	negative	Negative, 20
Emotion	Fear	Fear	sadness	Anger

From this table you can immediately see that most "users" with more followers are newspapers and institutions already verified, so we can stick to the category of before. Unfortunately, the table also shows how the tweet about sputnik doesn't actually focus on vaccines, but it's about a Cold War movie. Unfortunately, it is inevitable that we will not be able to skimme this information a priori and therefore our data will always be in the presence of noise or "outlayers". This is probably just a random error since the words analyzed in the section dedicated to wordcloud are consistent with the scope of vaccines.

Table A

Label	type property	signified
Tweet	String	text of the tweet or comment to a tweet
is_quoted	Boolean	This field only surfaces when the Tweet is a quote Tweet. This field contains the integer value Tweet ID of the quoted Tweet.
tweet_quoted	String	This field only surfaces when the Tweet is a quote Tweet. This attribute contains the Tweet object of the original Tweet that was quoted.
name_of_who_I_answered	String	*Nullable*. If the represented Tweet is a reply, this field will contain the screen name of the original Tweet's author.
id user_str	String	The string representation of the unique identifier for this Tweet. Implementations should use this rather than the large integer in id
lang_user	String	*Nullable.* When present, indicates a BCP 47 language identifier corresponding to the machine-detected language of the Tweet text, or und if no language could be detected.
creation_date_of_tweet	String	UTC time when this Tweet was created.
result_type	String	Metadata.resulte_type
number_of_retweet	Int	Number of times this Tweet has been retweeted.
retweeted	Boolean	Indicates whether this Tweet has been Retweeted by the authenticating user.
Source	String	Utility used to post the Tweet, as an HTML-formatted string. Tweets from the Twitter website have a source value of web.
user_name	String	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the name of user
user_screen_name	String	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the twitter name of the user
location property	String	The user who posted this Tweet. See User data dictionary for complete list of attributes. We selected the location of the user if he/she put it in his/her twitter profile
user_description	String	The user who posted this Tweet. See User data dictionary for complete list of attributes. In particular we extraxt the user bio
number_of_follower	Int	Number of followers of the user
number_of_friend	Int	Number of follow
is_verified	Boolean	If the user is verified by twitter
creation_date_of_account		UTC time when the user profile was created.
like property	Int	Number of likes to the tweet