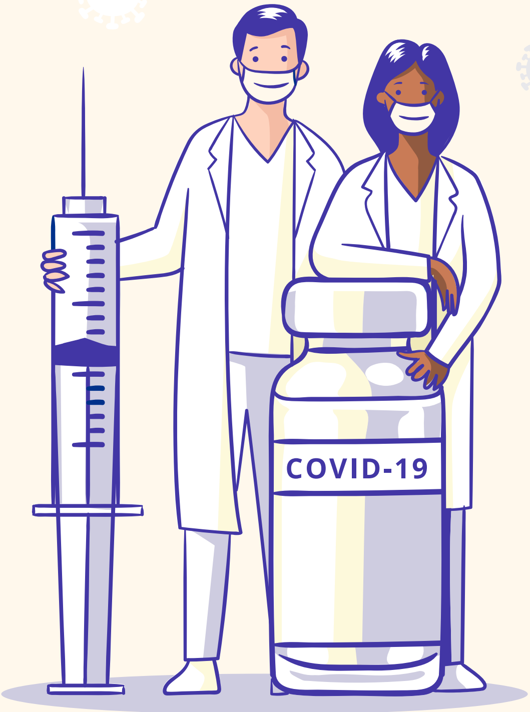
Sentiment analysis regarding tweet about Covid-19 vaccines

Natural Language Processing project

Giovanni Pinna



Goal of the project

The goal is to do the **sentiment analysis** on the tweet of the Italian population has **regarding** the **vaccines** against Covid-19

Vaccines used in the Analysis

Pfizer

Analyze the sentiment of this vaccine that we trust more than the other

Moderna

Analyze the sentiment of this vaccine which is not talked about so much

AstraZeneca

We want analyze the sentiment of this vaccine that has not the trust of the Italian

Sputnik

We consider it to observe the difference in the sentiment with the other vaccines approved by EMA

Dataset

The data that has been used all comes from the Social Network **Twitter**.

The search criteria used are:

data_since = (start date of research) set at **18/04/2021**
data_until = (research end date) set at **28/04/2021**
language = 'it'
search_word = vaccine name

Dataset

Most important labels are:

- tweets
- is_quoted
- lang_user
- source
- user_name
- user_screen_name
- location
- number_of_follower
- is_verified

Hard Pre-processing

Deleted:

- user names
- hashtags
- links
- special characters
- numbers transformed into 0

Processed:

- lower case
- word reduced to its lemma
- POS parsing for take only {'NOUN', 'VERB', 'ADJ', 'ADV', 'PROPN'}.

Soft Pre-processing

Deleted:

- hashtags
- links
- special characters

Processed:

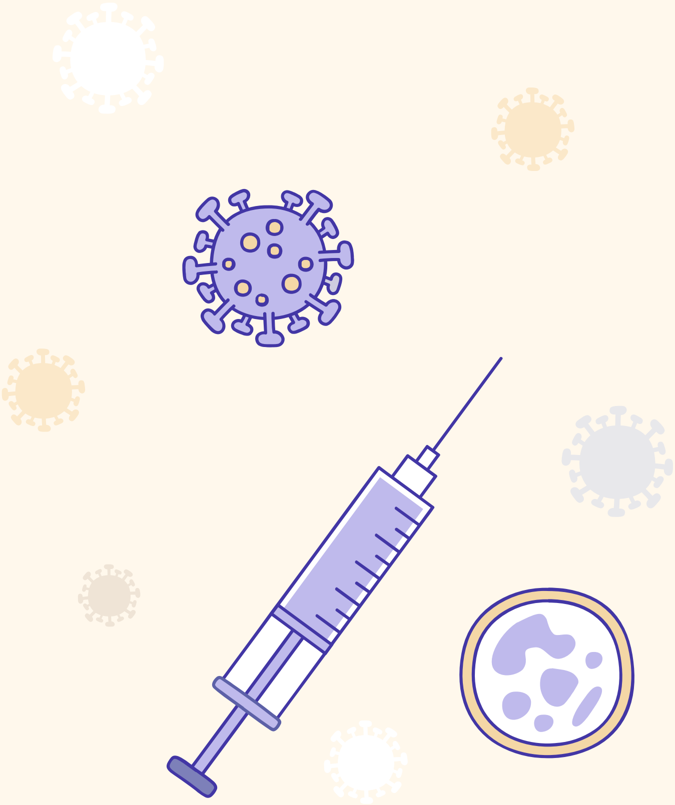
- lower case
- user names replaced with ‘user’

BERT Bocconi University (feel-it).

This model has been used for make the **sentiment** analysis and the **emotion** analysis of the tweet.

For the prediction we have provided BERT with soft pre-processing tweets

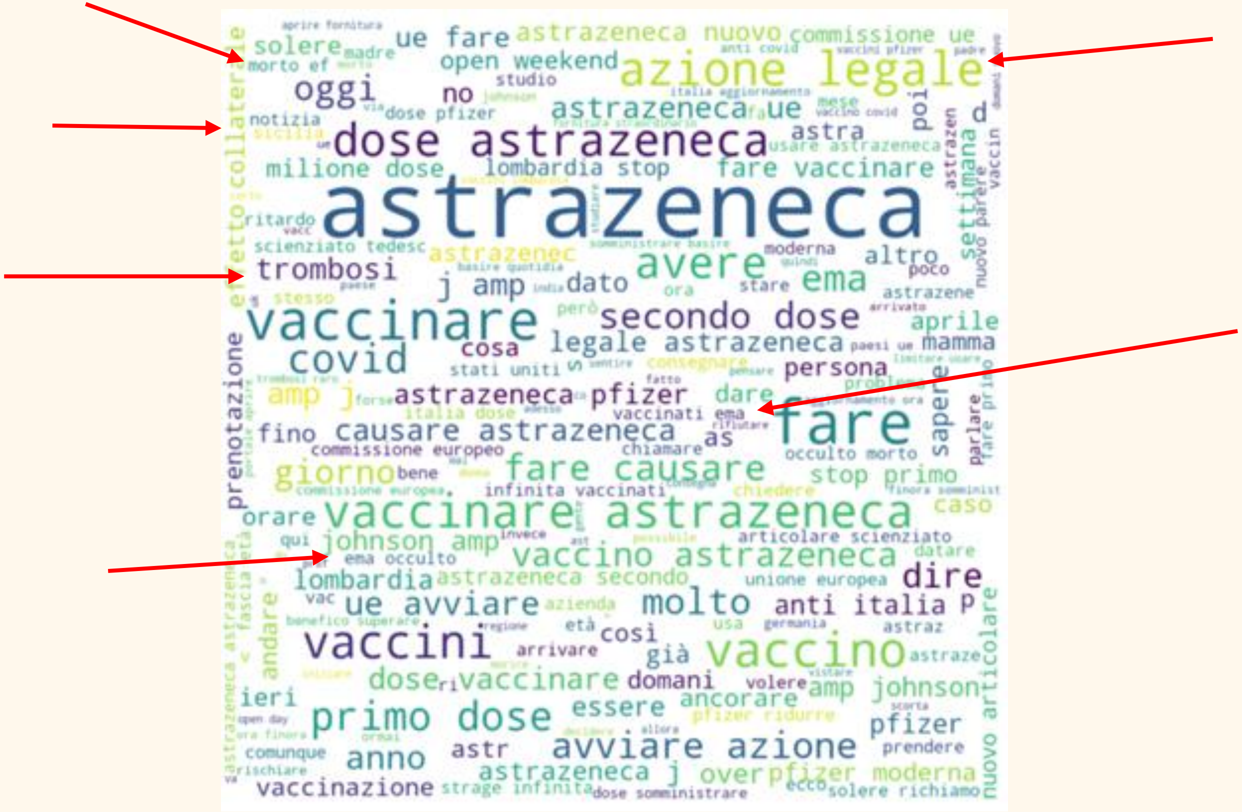
The result of **BERT's predictions is used how truth sentiment** for the other machine learning model.



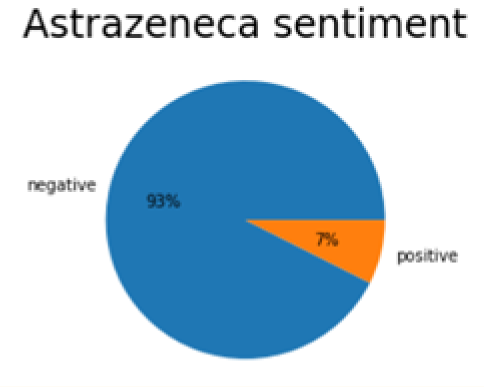
Analysis

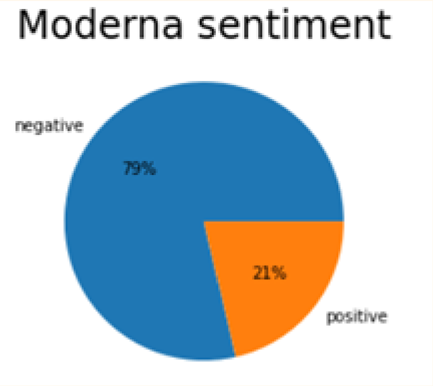
In this part we make analysis of the data and of the distribution of the word

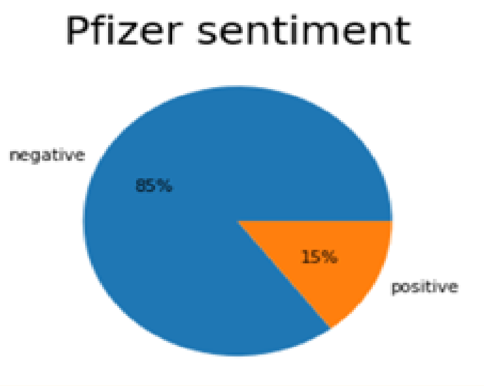
WordCloud Astrazeneca

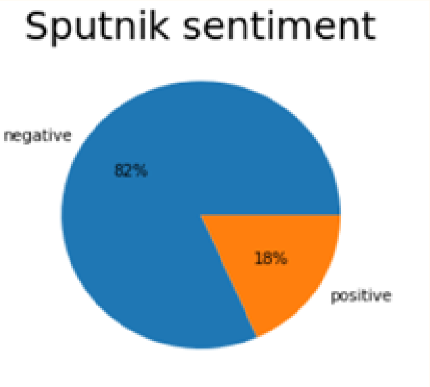


Sentiment analysis

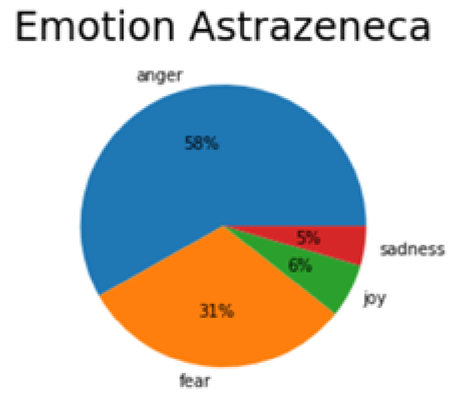

Astrazeneca sentiment

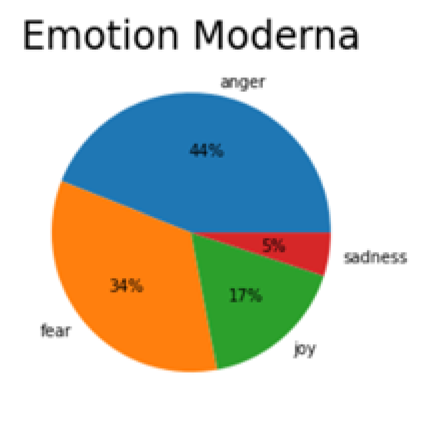

Moderna sentiment

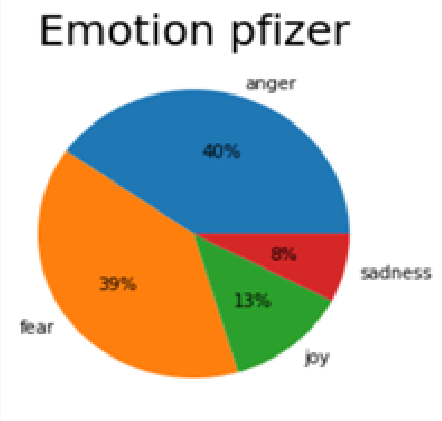

Pfizer sentiment

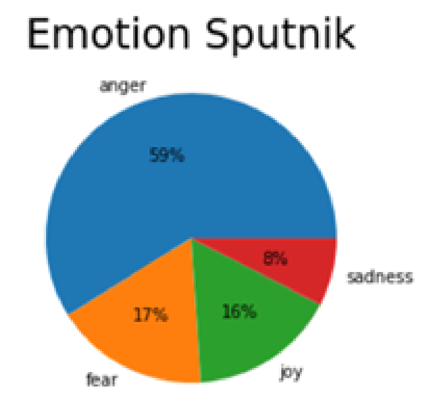

Sputnik sentiment

Emotion analysis


Emotion Astrazeneca


Emotion Moderna


Emotion pfizer


Emotion Sputnik

Other analysis

I tried to figure out if the negative tweets were due to something.

I tried to see if the **verified users** then with more followers could have **influenced others**

I checked if the positive tweets were **contracted in Italian cities most affected by the pandemic**

Other analysis

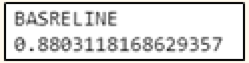
I analyzed the **frequency of unigrams and bi-grams**.

I noticed that some pairs of words that could explain the negativity were "non [nome vaccino]" , "non fare" and “non avere”.

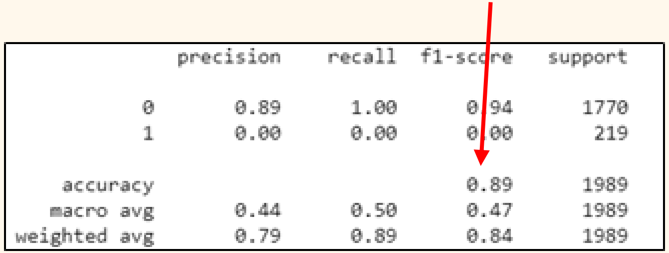
Prediction

Results of the model after applying the selection and regularization.

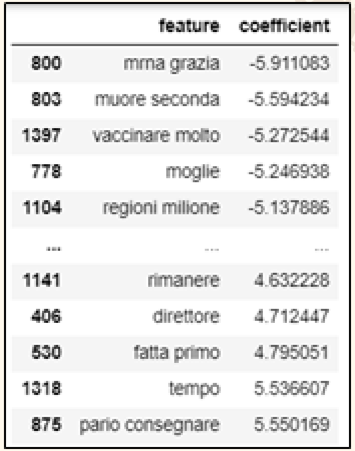
The **baseline and accuracy are particularly close**



```
BASRELINE
0.8803118168629357
```

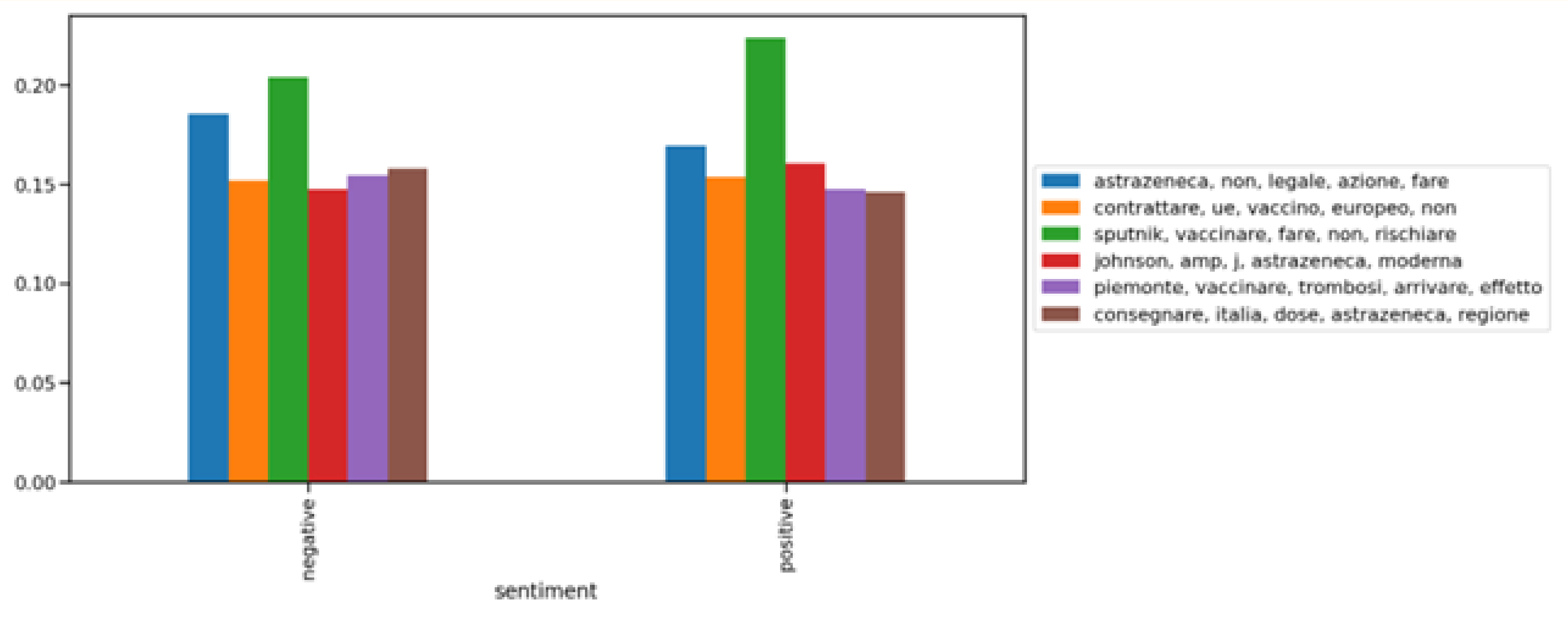


```
              precision    recall  f1-score   support

           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	mma 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

Topic modelling LDA



Future developments

- Try to train model with SVM

- Implement a clustering algorithm and analyze how the observation are grouped together. Maybe we can find other latent information in the data

Conclusion

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. (the people on social are complaining and negative news events have much more influence in public opinion).

The **topics** analysis give us a idea of the argument of the tweets, that are most of which can more easily be **associates to something negative** than something positive.

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