Covid Italian's tweet
Natural Language Process project
University of Trieste - Master Degree's in DSSC

Alessia Paoletti

1 Introduction

Every day on Twitter millions of tweets talking about the latest news are posted. Covid-19 is an hot topic, so there is a huge quantity of tweets concering it. The aim of this project is to analyze the **italian tweets about Coronavirus**.
In this project 4 different periods have been considered, each one representative of a particular moment of the italian emergency [1]:

- Sunday 23 February - Sunday 1 March, the emergency starts
- Sunday 15 March - Sunday 22 March, the global lockdown has just started
- Sunday 19 April - Sunday 26 April, italian citizens ask when the lockdown will end, more than 1 month elapsed since it started
- Sunday 17 May - Sunday 24 May, the Phase 2 starts, life begins to return to normal

In Figure 1 some of the most imporant events of the Covid's italian emergency have been reported, as the creation of the "red zone", the start of the lockdown and the begin of Phase 1 and Phase 2.

Data have been retrieved for each period and variuous tasks have been performed:

- Text generation

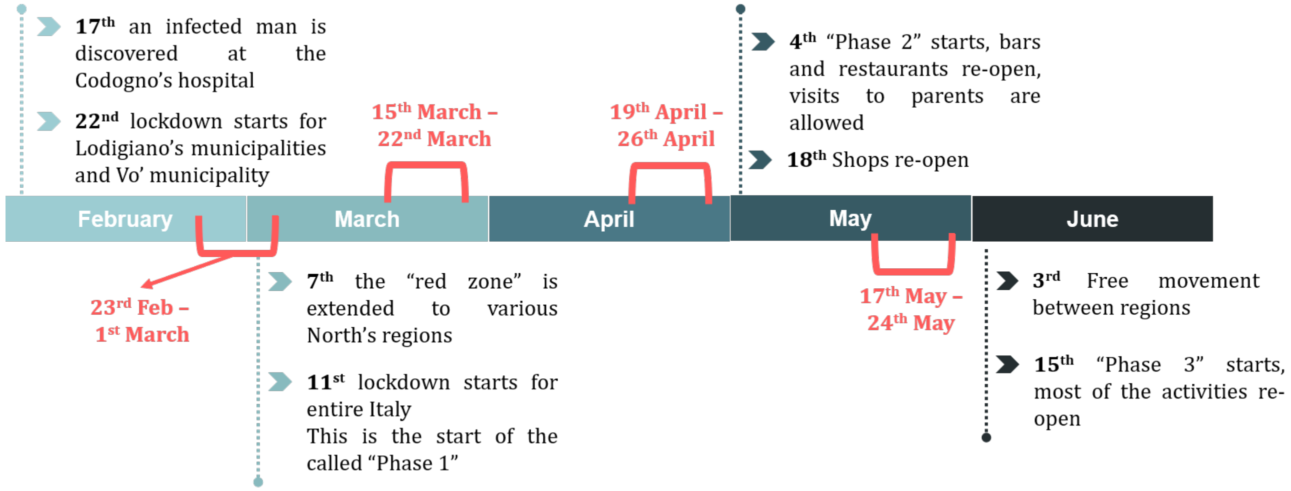

17th an infected man is discovered at the Codogno's hospital
22nd lockdown starts for Lodigiano's municipalities and Vo' municipality
15th March – 22nd March
19th April – 26th April
4th "Phase 2" starts, bars and restaurants re-open, visits to parents are allowed
18th Shops re-open
February
March
April
May
June
23rd Feb – 1st March
7th the "red zone" is extended to various North's regions
11st lockdown starts for entire Italy This is the start of the called "Phase 1"
17th May – 24th May
3rd Free movement between regions
15th "Phase 3" starts, most of the activities re-open

Figure 1: Italian Covid's timeline

- Topic modelling
- Sentiment analysis

2 The data

Italian tweets about the Coronavirus emergency are the protagonists of this project. 40wita [2] is part of the effort by the Italian Association of Computational Linguistics to collect and track resources to alleviate the national and global crisis following the COVID-19 outbreak in 2020.

The data have been collected daily from February 1st, 2020, by filtering TWITA (a collection of italian tweets) [3] with some significant words, like: *covid, covid19, covid-19, corona virus, coronavirus, quarantena, autoisolamento, auto-isolamento, iorestoacasa, stateacasa, COVID19Italia, redditodicittadinaza, eurobond, coronabond, restiamoacasa, preghiamoinsieme, NoMes, milanononsiferma, bergamononsiferma, abbraccciauncinese, iononsonounvirus, iononmifermo, aperisera, quarantena covidunstria, italiazonarossa, bergamoisrunning, , COVID19Pandemic*

For each day the italian tweets filtered using the significant words previously mentioned have been retrieved. For each tweet different information are reported:

- Id
- Text
- Language (italian for all tweets)
- Screen_name: twitter username
- Date, timestamp, year, month, day and hour
- Latitudine and longitudine (and location j_son)
- Source
- Urls
- Description
- Statuse_count
- Followers_count
- Friend_count
- Media

For this task only the text is taking into consideration. More sophisticated analysis could be done taking into account also timing and location, but it is out of the scope of this project.

In Figure 2 the tweets number for each week is reported.

2.1 Italian Twitter users

One quick little observation abouts italian Twitter's users needs to be made. According to [4][5] only 3,7 millions of italian citizens are active users of Twitter, representing about 6% of the overall population. Going deeper, 38,7% are women and 61,3% are men and the age range is 35-54 years old. Thus, all the analysis made on tweets are the thoughts of a very small (and not representative) sample, so the results cannot be generalized to the entire

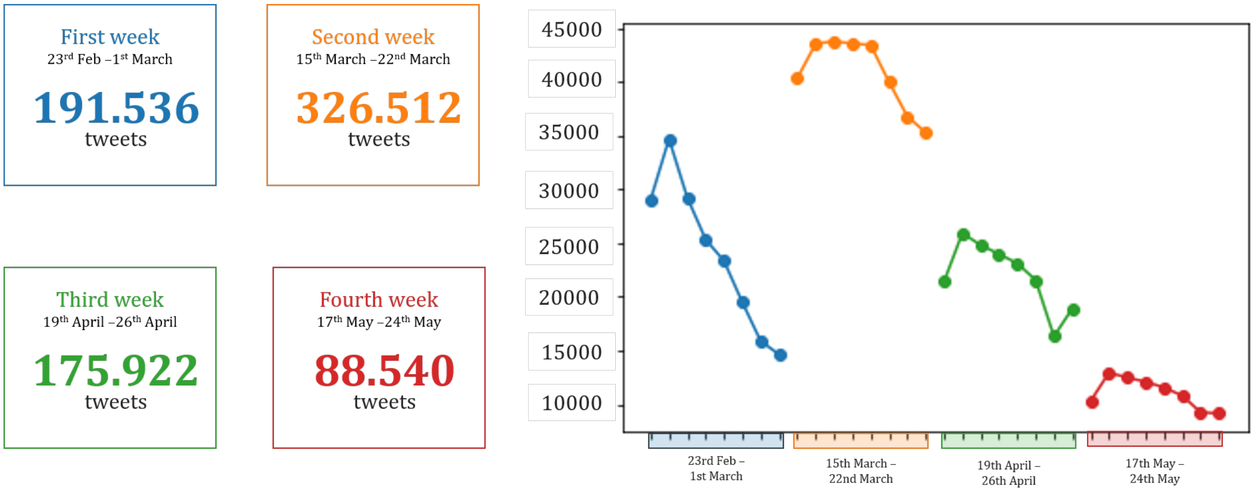


Figure 2: Number of tweets for week

Italian citizens. In the following, expressions like ”italian sentiments” or ”italian thoughts” are going to be used but only for simplicity, and not preteding to make general assumptions.

3 Preprocessing

Preprocessing is an important and critical step since it is used for extracting interesting and non-trivial knowledge from unstructured text data. It is essentially a matter of deciding which parts of text and/or documents should be retrieved to carry out a specific task.

So, the first step of this project has been the preprocessing phase, divided in:

1. Clean the tweet
2. Clean the text

3.1 Clean tweet

In tweets a lot of other extra information over text is present: urls, emojis, hashtags and mentions. Using regular expressions:

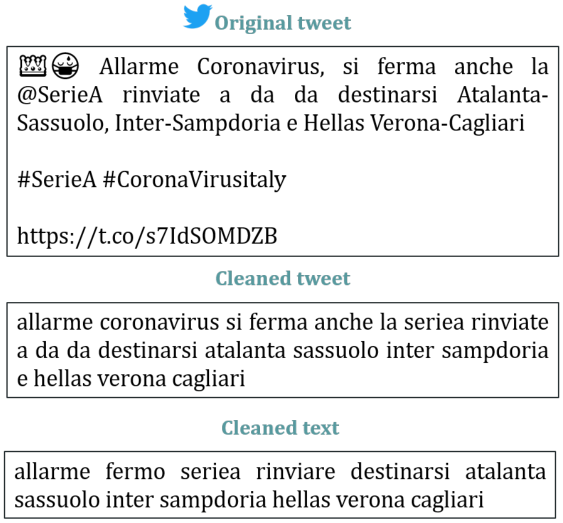
- All the puntuaction has been removed
- Emojis have been removed
- Hashtags have been removed from the tweets, but have been stored separately
- Urls have been removed
- Numbers have been removed
- For the mentions only the mention symbol @ has been removed, but the user mentions has been kept (cause if removed tweets can completly lose the sense)

3.2 Clean text

At this time in the tweet only text information is present and we can work on text using *Spacy* [6] library:

- Stopwords have been removed. In addition to the canonical stopwords some Coronavirus specific words have been added to this list, like *coronavirus, virus, corona, covid* and similar ones.
- Only *'ADJ', 'ADV', 'NOUN', 'NUMERAL', 'NUM', 'PROPN','VERB'* have been kept
- The lemmatize version of each word is kept

In Figure 3 an example of tweet preprocessing is reported.


Original tweet

🏟😷 Allarme Coronavirus, si ferma anche la @SerieA rinviate a da da destinarsi Atalanta-Sassuolo, Inter-Sampdoria e Hellas Verona-Cagliari

#SerieA #CoronaVirusitaly

https://t.co/s7IdSOMDZB

Cleaned tweet

allarme coronavirus si ferma anche la seriea rinviate a da da destinarsi atalanta sassuolo inter sampdoria e hellas verona cagliari

Cleaned text

allarme fermo seriea rinviare destinarsi atalanta sassuolo inter sampdoria hellas verona cagliari

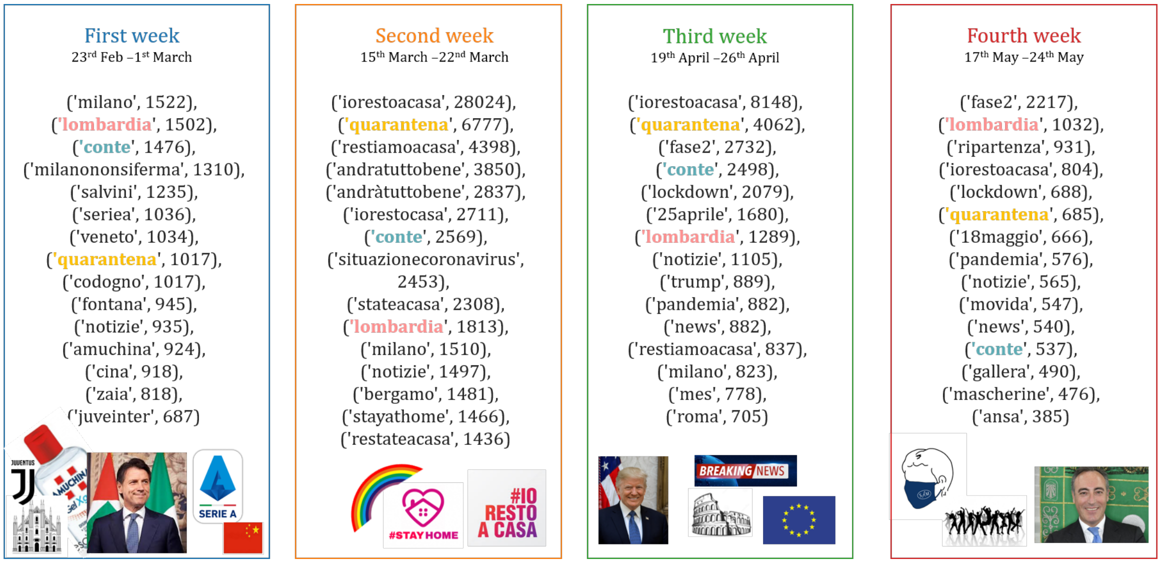
Figure 3: Tweets preprocessing

4 Popular hashtags

During the preprocessing phase for each tweet its hashtags have been removed from the text but have been saved separately. Exploring the most popular hashtags for each period some interesting highlights can be grasped (Figure 4). Some hashtags that can be thought as emblematic words of the emergency, are present in each week, like *#Conte, #Lombardia, #quarantena*. In the first week hashtags regarding the start of the emergency like *#Codogno, #Cina, #amuchina* can be found, as well as hashtags about the football championship. The second week is characterized by hashtags regarding the lockdown as *#stayhome, #stateacasa*. In the third week *#Trump* appears, since the Coronavirus landed in the USA. Finally in the fourth week life start to come back to normal and the *#movida* (party) with *#mascherine* (protections) begins.

5 Protagonists

A very simply task that has been carried out was to find the "protagonists" of each week. Given the cleaned version of the tweet and using *Spacy* library [6], only the proper name (*PROPN*) has been kept. To crown the top 10



First week
23rd Feb –1st March

('milano', 1522),
('lombardia', 1502),
('conte', 1476),
('milanononsiferma', 1310),
('salvini', 1235),
('seriea', 1036),
('veneto', 1034),
('quarantena', 1017),
('codogno', 1017),
('fontana', 945),
('notizie', 935),
('amuchina', 924),
('cina', 918),
('zaia', 818),
('juveinter', 687)

Second week
15th March –22nd March

('iorestoacasa', 28024),
('quarantena', 6777),
('restiamoacasa', 4398),
('andratuttobene', 3850),
('andràtuttobene', 2837),
('iorestocasa', 2711),
('conte', 2569),
('situazionecoronavirus', 2453),
('stateacasa', 2308),
('lombardia', 1813),
('milano', 1510),
('notizie', 1497),
('bergamo', 1481),
('stayathome', 1466),
('restateacasa', 1436)

Third week
19th April –26th April

('iorestoacasa', 8148),
('quarantena', 4062),
('fase2', 2732),
('conte', 2498),
('lockdown', 2079),
('25aprile', 1680),
('lombardia', 1289),
('notizie', 1105),
('trump', 889),
('pandemia', 882),
('news', 882),
('restiamoacasa', 837),
('milano', 823),
('mes', 778),
('roma', 705)

Fourth week
17th May –24th May

('fase2', 2217),
('lombardia', 1032),
('ripartenza', 931),
('iorestoacasa', 804),
('lockdown', 688),
('quarantena', 685),
('18maggio', 666),
('pandemia', 576),
('notizie', 565),
('movida', 547),
('news', 540),
('conte', 537),
('gallera', 490),
('mascherine', 476),
('ansa', 385)

Figure 4: Popular hashtags

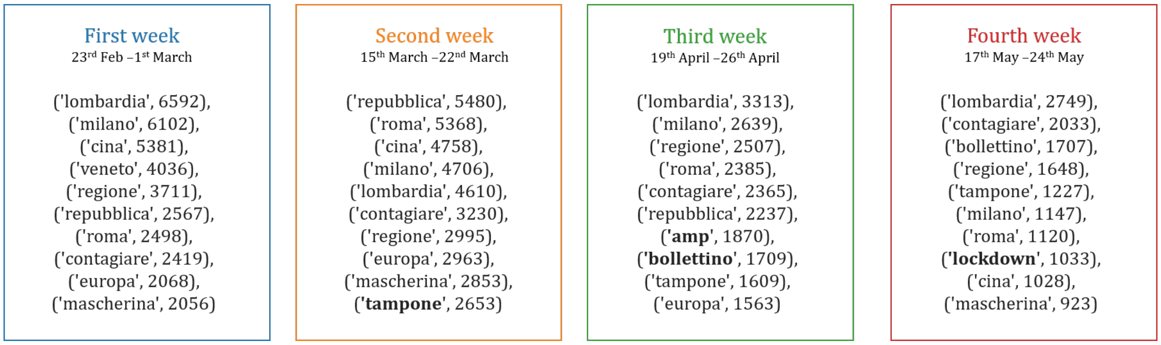
celebreties for each period, occurences have been retrieved and sorted. In Figure 5 results are reported. Unfortunately taking into account only the top 10 protagonists we do not have a lot of diversification (as could be expected). Maybe some differences could be grasped taking into account the top 50 protagonists for each period or excluding the celebrities that appears in each week, but it is out of the scope of this ”toy” task.

Another consideration that can be done is that some terms are identified as proper names, even if they are verbs or nouns. This highlight the well known ”lacks” for languages dfferent for English.

6 Text generation

Tweet generation is a simple and funny task. Starting from a language model based on tweet’s text of each week some new tweets have been generated. Tweets can be generated giving some initial words and in this case two starting points have been choosen:

- *Coronovirus è* (Coronavirus is)



First week
23rd Feb –1st March

('lombardia', 6592),
('milano', 6102),
('cina', 5381),
('veneto', 4036),
('regione', 3711),
('repubblica', 2567),
('roma', 2498),
('contagiare', 2419),
('europa', 2068),
('mascherina', 2056)

Second week
15th March –22nd March

('repubblica', 5480),
('roma', 5368),
('cina', 4758),
('milano', 4706),
('lombardia', 4610),
('contagiare', 3230),
('regione', 2995),
('europa', 2963),
('mascherina', 2853),
(**'tampone'**, 2653)

Third week
19th April –26th April

('lombardia', 3313),
('milano', 2639),
('regione', 2507),
('roma', 2385),
('contagiare', 2365),
('repubblica', 2237),
(**'amp'**, 1870),
(**'bollettino'**, 1709),
('tampone', 1609),
('europa', 1563)

Fourth week
17th May –24th May

('lombardia', 2749),
('contagiare', 2033),
('bollettino', 1707),
('regione', 1648),
('tampone', 1227),
('milano', 1147),
('roma', 1120),
(**'lockdown'**, 1033),
('cina', 1028),
('mascherina', 923)

Figure 5: Protagonists of each week

- *Conte* (the Italian prime minister who has become ”famous” during the emergency)

The language model used for generating the tweets is an 4-grams model with a smoothing factor of 0.001. Increasing the n-gram size leads to more sense made tweets but decrease the ”generation space”, since only few tweets can be generated (and very similar so the original ones).

Some generated tweets are the following:

- coronavirus è una follia uccide di più l influenza il sole ore (first week)
- coronavirus è vicino il blocco cominciato il marzo darà per forza di cose i suoi frutti a partire dal decreto (second week)
- coronavirus è il nostro integratore naturale (third week)
- coronavirus è giallo spariti morti governo cambia i dati della lombardia di oggi (fourth week)
- conte dice che un ospedale non ha seguito i protocolli è colpa del viva fastweb (first week)
- conte chiudo l italia è zona rossa (second week)
- conte frena sulla fase dovremo imparare a convivere col coronavirus (third week)
- conte e il consenso popolare il premier sopravvivrà al coronavirus (fourth week)

7 Clustering & Topic model

Topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. It is an unsupevised task in which a priori k representing the number of topics/clusters needs to be determined.

7.1 Number of clusters and topics

To find the number of topics for LDA the **coherence score** has been used. Looking at the graphs an initial guessing of the number of topics was done.

To find the initial guessing of the number of clusters k the **silhoutte score** for each week was computed. Due to computationaly limitation, it was calculated on a random sample of 8000 instances (so a very small subset) and taking into account the 500 features (trials with higher values were done but the Google Colab session crashed every time without providing any useful results). So in pratice the initial guessing of k was quite random, but taking into consideration that, as shown in the silhoutte graphs, usually we have few number of cluster (expect for the fourth week). Some ideas about the number of clusters were also taken looking at the coherence scores.

In the file *coherence_silhoutte.pdf* both silhoutte and coherence graphs are reported for each week.

7.2 *k*-means ++

k-means clustering is a method that aims to partition observations (in our case tweet's text) into k clusters. k-means++ is a variation of the classical algorithm in which a procedure to initialize the cluster centers before proceeding with the standard k-means optimization iterations is performed.

k-means++ has been applied to each week. For each one some specific topic cluster related to the period's news have been found, in addition to general clusters with words about the emergency like *positivo, decesso, morto, contagio*. For a detailed report of the clusters obtained for each week see *kmeans_cluster.pdf* in this folder.

7.3 Latent Dirichlet Allocation

LDA is one of the most popular topic modeling methods. Each document (in our case tweet) is made up of various words, and each topic also has several words belonging to it. The aim of LDA is to find topics a document belongs to, based on the words in it.

As k-means, LDA has been applied to each week and topic related to latest news can be identified. Unfortunately topics are less "defined" respect to the k-means clusters.

For a more detailed report of the topics obtained for each week see *lda.pdf* in this folder.

8 Sentiment analysis

Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Applying it to the Covid's italian tweets can reveal the italians' emotions during the emergency.

8.1 Train datasets

Two datasets of tweets sport (about football teams and players) [7] have been used to train the models. These datasets contains in total more than 400k tweets of 4 sentiment: positive, negative, neutral and mixed. The mixed sentiment has been deleted since it was in very few tweets (less than 0.5% of the total tweets).

Another problem to solve was the very unbalanced number of tweets for each category:

- 11% of positive tweets
- 4% of negative tweets
- 85% of neutral tweets

Two different strategies have been used to address the unbalanced categories problem:

- **Brutal under-sampling**, taking the less numerous category and consider the same number of tweets for the other 2 categories.
- **Over-sampling with SMOTE** [8]

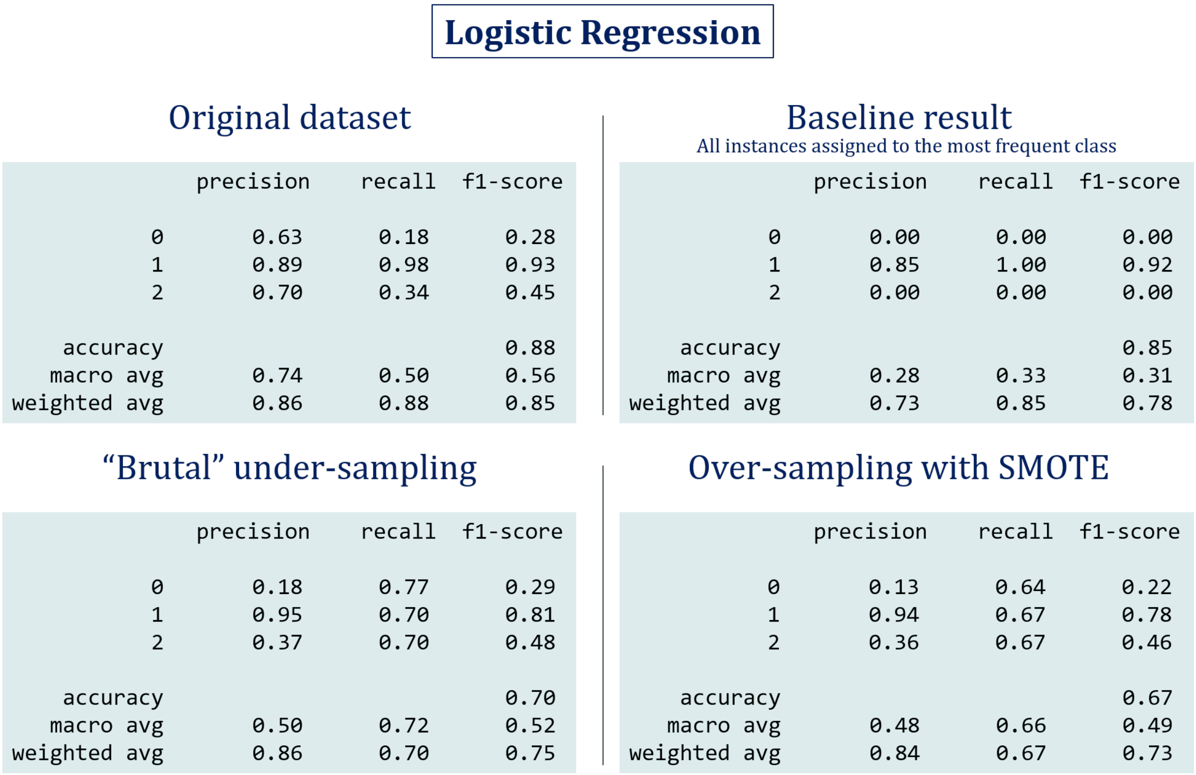
8.2 Model and results

8.2.1 Logistic regression

Simple logistic regression from *scikit-learn* package [9] is the first model that was used. To train the model a vectorizer with *ngram_range=(1,4)*, *min_df=0.001* and *max_df=0.75* has been used.
In Figure 6 the results of logistic regression are reported. For each model the dataset used during the training phase is indicated.

The model trained on the original dataset gives an accuracy of 88% but it is a misleading positive result. This is confirmed by the fact that assigning each instance to the most frequent class will lead to an accuracy of 85%. The same conclusion can be reached taking a look to the recall values, that appears to be near to 1 for the neutral sentiment class.

The models trained with the under-sampling and over-sampling reach accuries of 70% and 67% respectively. These models present more ”balanced” recall values, but of course it comes with cost for precision values.



Logistic Regression

Original dataset

```
              precision    recall  f1-score

           0       0.63      0.18      0.28
           1       0.89      0.98      0.93
           2       0.70      0.34      0.45

    accuracy                           0.88
   macro avg       0.74      0.50      0.56
weighted avg       0.86      0.88      0.85
```

Baseline result

All instances assigned to the most frequent class

```
              precision    recall  f1-score

           0       0.00      0.00      0.00
           1       0.85      1.00      0.92
           2       0.00      0.00      0.00

    accuracy                           0.85
   macro avg       0.28      0.33      0.31
weighted avg       0.73      0.85      0.78
```

“Brutal” under-sampling

```
              precision    recall  f1-score

           0       0.18      0.77      0.29
           1       0.95      0.70      0.81
           2       0.37      0.70      0.48

    accuracy                           0.70
   macro avg       0.50      0.72      0.52
weighted avg       0.86      0.70      0.75
```

Over-sampling with SMOTE

```
              precision    recall  f1-score

           0       0.13      0.64      0.22
           1       0.94      0.67      0.78
           2       0.36      0.67      0.46

    accuracy                           0.67
   macro avg       0.48      0.66      0.49
weighted avg       0.84      0.67      0.73
```

Figure 6: Logistic Regression results

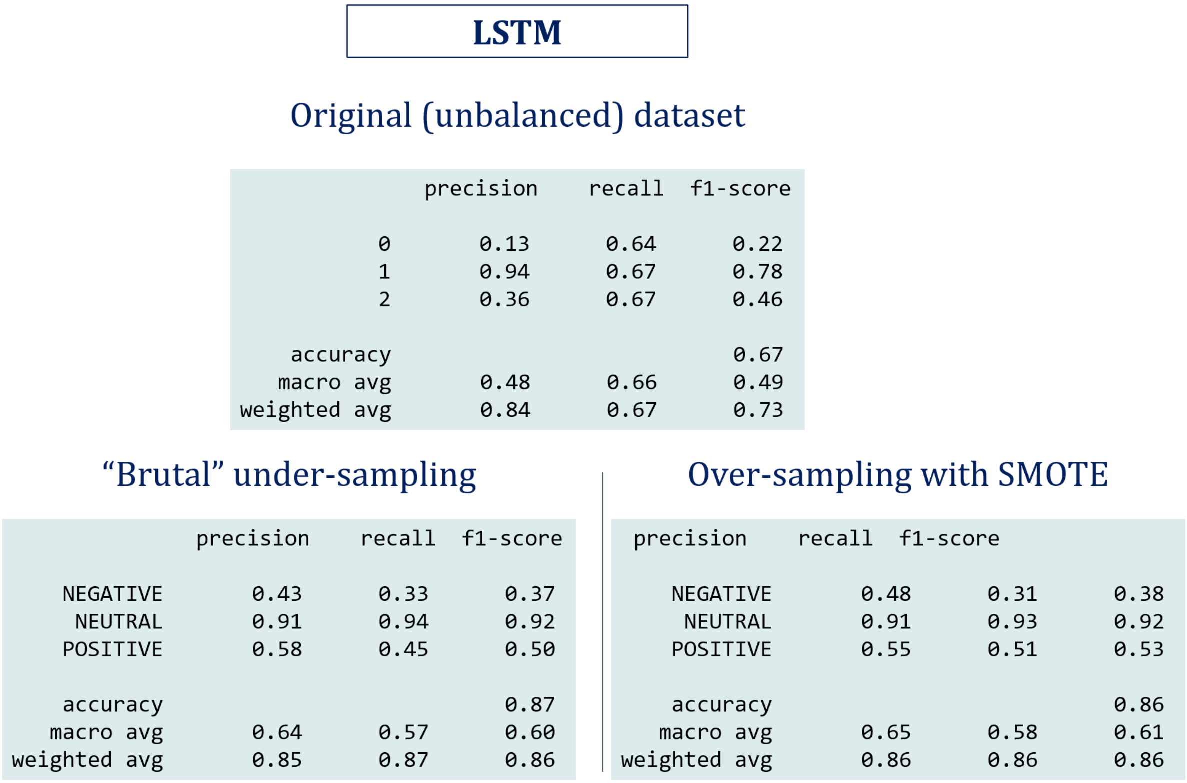
8.2.2 LSTM

A RNN is the other model trained on the sport data and later used to make predictions. The neural network is the following:

```
model = Sequential()
model.add(Embedding(output_dim=64, input_dim=vocab_size, input_length=None))
model.add(LSTM(lstm_dims))
model.add(Dense(hidden_dims))
model.add(Dropout(rate=0.2))
model.add(Activation(’relu’))
model.add(Dense(num_classes))
model.add(Activation(’softmax’))
```

The first layer is of course an embedding layer in order to represent tweets as dense vectors. Then we have the LSTM layer that represents the recursion core. Finally dense layers, combined with dropuout and activation layers, are used.

In Figure 7 the result of the RNN trained on different datasets are reported. With the model trained on the original dataset an accuracy of 67% is reached but, as can be noticed looking to the F1-score, the model works well mainly for the neutral sentiment class (the most frequent). The model trained using under-sampling and over-sampling datasets have been reported, obtaining higher accurancies .



LSTM

Original (unbalanced) dataset

```
              precision    recall  f1-score

           0       0.13      0.64      0.22
           1       0.94      0.67      0.78
           2       0.36      0.67      0.46

    accuracy                           0.67
   macro avg       0.48      0.66      0.49
weighted avg       0.84      0.67      0.73
```

“Brutal” under-sampling

```
              precision    recall  f1-score

    NEGATIVE       0.43      0.33      0.37
     NEUTRAL       0.91      0.94      0.92
    POSITIVE       0.58      0.45      0.50

    accuracy                           0.87
   macro avg       0.64      0.57      0.60
weighted avg       0.85      0.87      0.86
```

Over-sampling with SMOTE

```
 precision    recall  f1-score

    NEGATIVE       0.48      0.31      0.38
     NEUTRAL       0.91      0.93      0.92
    POSITIVE       0.55      0.51      0.53

    accuracy                           0.86
   macro avg       0.65      0.58      0.61
weighted avg       0.86      0.86      0.86
```

Figure 7: LSTM results

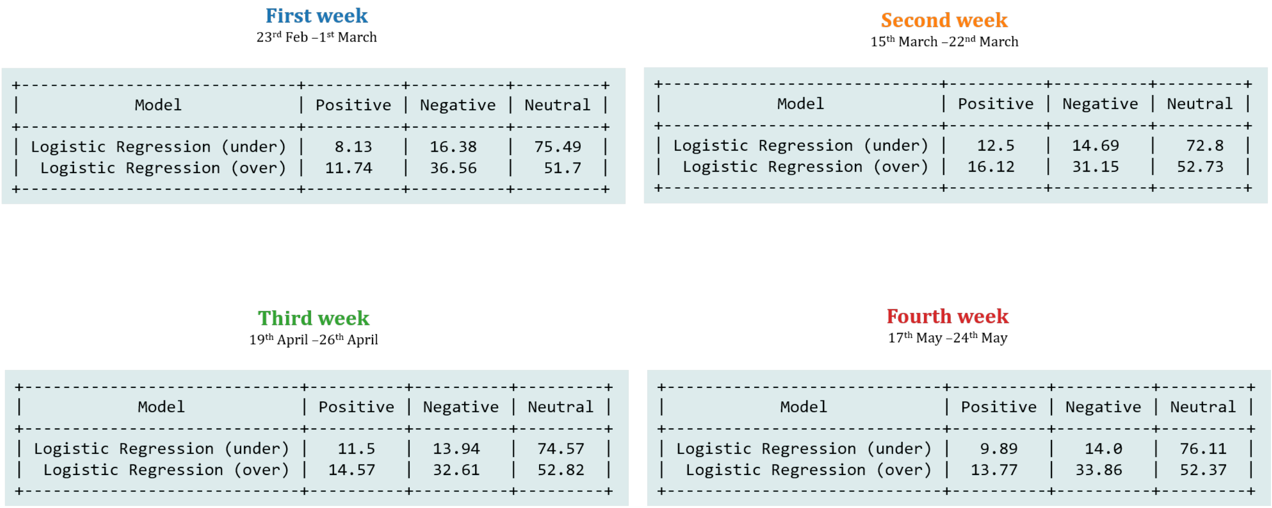
8.3 Prediction on Covid's tweets

The models trained on the sport tweets have been used to predict the sentiment of the Covid's italian tweets.

8.3.1 Logistic regression

The results using the logistic regression models are reported in Figure 8. As one can notice the most tweets are classified as having neutral sentiment. This can be justified since a lot of tweets are about update on Coronavirus numbers (victimis, infected, recovered) or about new laws and restrictions. So most of them are tweets about objective current news.

The main difference between over-sampling and under-sampling models predictions lies in tweets classified as negative rather than neutral.



First week
23rd Feb –1st March

```
+-----------------------------+----------+----------+---------+
|            Model            | Positive | Negative | Neutral |
+-----------------------------+----------+----------+---------+
| Logistic Regression (under) |   8.13   |  16.38   |  75.49  |
|  Logistic Regression (over) |  11.74   |  36.56   |   51.7  |
+-----------------------------+----------+----------+---------+
```

Second week
15th March –22nd March

```
+-----------------------------+----------+----------+---------+
|            Model            | Positive | Negative | Neutral |
+-----------------------------+----------+----------+---------+
| Logistic Regression (under) |   12.5   |  14.69   |   72.8  |
|  Logistic Regression (over) |  16.12   |  31.15   |  52.73  |
+-----------------------------+----------+----------+---------+
```

Third week
19th April –26th April

```
+-----------------------------+----------+----------+---------+
|            Model            | Positive | Negative | Neutral |
+-----------------------------+----------+----------+---------+
| Logistic Regression (under) |   11.5   |  13.94   |  74.57  |
|  Logistic Regression (over) |  14.57   |  32.61   |  52.82  |
+-----------------------------+----------+----------+---------+
```

Fourth week
17th May –24th May

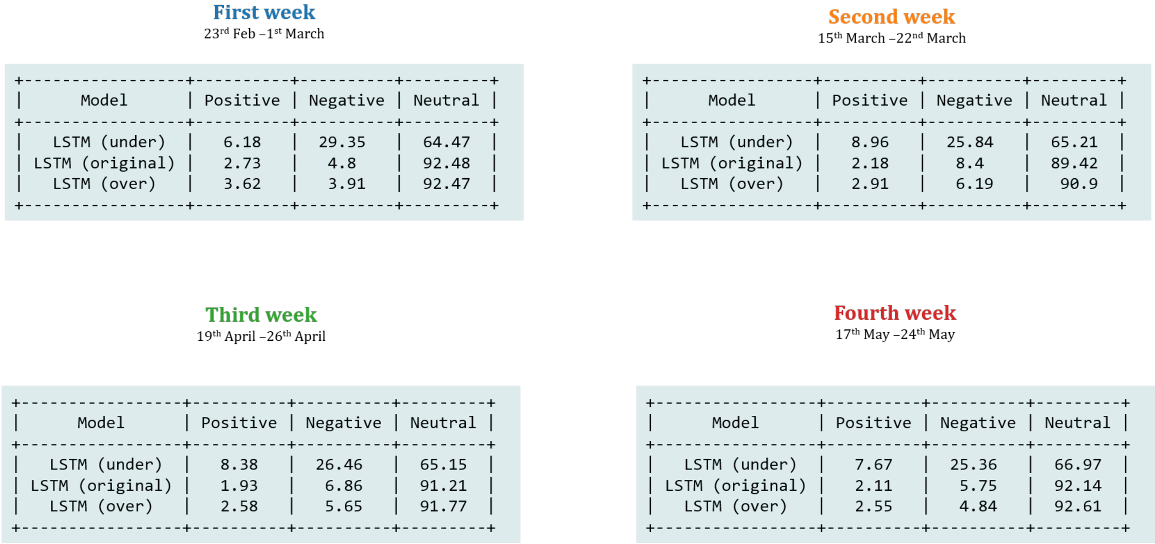
```
+-----------------------------+----------+----------+---------+
|            Model            | Positive | Negative | Neutral |
+-----------------------------+----------+----------+---------+
| Logistic Regression (under) |   9.89   |   14.0   |  76.11  |
|  Logistic Regression (over) |  13.77   |  33.86   |  52.37  |
+-----------------------------+----------+----------+---------+
```

Figure 8: Logistic Regression on Covid's tweets (%)

8.3.2 LSTM

The results using the LTSM models are reported in Figure 9. Using both the models trained with the original and the over-sampling datasets most of

tweets are classified as neutral. As before this can be justified by the fact that a lot of tweets are about objective current news.
The model trained with the under-sampling dataset classifies more tweets as negative rather than neutral ones.



First week
23rd Feb –1st March

```
+-----------------+----------+----------+---------+
|      Model      | Positive | Negative | Neutral |
+-----------------+----------+----------+---------+
|   LSTM (under)  |   6.18   |  29.35   |  64.47  |
| LSTM (original) |   2.73   |   4.8    |  92.48  |
|   LSTM (over)   |   3.62   |   3.91   |  92.47  |
+-----------------+----------+----------+---------+
```

Second week
15th March –22nd March

```
+-----------------+----------+----------+---------+
|      Model      | Positive | Negative | Neutral |
+-----------------+----------+----------+---------+
|   LSTM (under)  |   8.96   |  25.84   |  65.21  |
| LSTM (original) |   2.18   |   8.4    |  89.42  |
|   LSTM (over)   |   2.91   |   6.19   |   90.9  |
+-----------------+----------+----------+---------+
```

Third week
19th April –26th April

```
+-----------------+----------+----------+---------+
|      Model      | Positive | Negative | Neutral |
+-----------------+----------+----------+---------+
|   LSTM (under)  |   8.38   |  26.46   |  65.15  |
| LSTM (original) |   1.93   |   6.86   |  91.21  |
|   LSTM (over)   |   2.58   |   5.65   |  91.77  |
+-----------------+----------+----------+---------+
```

Fourth week
17th May –24th May

```
+-----------------+----------+----------+---------+
|      Model      | Positive | Negative | Neutral |
+-----------------+----------+----------+---------+
|   LSTM (under)  |   7.67   |  25.36   |  66.97  |
| LSTM (original) |   2.11   |   5.75   |  92.14  |
|   LSTM (over)   |   2.55   |   4.84   |  92.61  |
+-----------------+----------+----------+---------+
```

Figure 9: LSTM on Covid's tweets (%)

8.3.3 Overll sentiment

Looking to the results in the four different periods of the emergency we can notice that there were not a significant changing. Each classifier predict almost the same sentiments for each periods, without any meaningful increments or decreases. So in general we can states the overall sentiment of italians Twitter's active users was neutral or, better says, that everyone became a reporter during the emergency, constaty (re-)posting about latest news.

8.3.4 Considerations

Sentiment analysis is a diffused but at the same time challenging task. In this case sports tweets where used as labeled example to discover the Covid tweet's sentiment. Sport and Coronavirus are for sure very different worlds

and this may lead to some inaccurate results. A very stupid example is the italian word *positivo*: in a normal cirmustance this word is associatied to somenthing good, a very peaceful situation or event, but during the Coronavirus emergency it was used to refer to the number of infected people. So a tweet like *un nuovo caso positivo* (a new positive test) is of course a negative event, but without any contest it can be intepreted as a postive thing.
The best solution to this problem would be to have a labeled dataset about Covid tweets sentiment (that unfortunately is not still available). Another way to attenuate this problem can be to provide list of words for each sentiment and, for our particular example, insert *positivo* in the negative (and/or neutral) list. But this type of task cannot be done by everyone, linguistics experts are needed (to obtain good results).

Another possible improvement could be to use emojis and/or hashtags to collect clues about tweets sentiment. For example, if in a tweet appears an hasthags with some offensive words, we can imagine that is a negative one. But, as before, understand which hashtags and emoji are associated to each sentiment is a very delicated task and can lead to really biased results if not done properly. Another limitation to this task is that thousands of millions of hashtags exist, so classify them is a very hard task.

Finally humorism and dialectal expressions definitly do not help during the sentiment analyis task.

9 Conclusions

Twitter is a social network on which each second are posted 2.200 new tweets (from all over the world). During a national lockdown that forces people to stay home, social networks can become the way to express opinions, so interesting highlights can be collected. Thus to "understand" Italian thoughts, italian Covid's tweets were analyzed.

In this project topic modelling and sentiment analysis tasks have been carried out.

Topic modelling on different periods of the emergency was able to grasp the most important news about that period, reflecting which topics worried the italian citizens most.

The sentiment analysis task lead to the conclusion that most of tweets during the emergency was (re-)posting of currents news or daily updates about Coronavirus numbers. Only a small part of users used Twitter to express their actual opinions.

References

[1] *Pandemia di Covid-19 del 2020 in Italia*, `https://it.wikipedia.org/wiki/Pandemia\_di\_COVID-19\_del\_2020\_in\_Italia`

[2] *40twita 1.0: An collection of Italian Tweets during the COVID-19 Pandemic*, Basile Valerio and Tommaso Caselli, `http://twita.di.unito.it/dataset/40wita`

[3] *TWITA - Italian tweets* , Valerio Basile and Malvina Nissim, `http://valeriobasile.github.io/twita/about.html`

[4] *Digital 2020 Italia* `https://wearesocial.com/it/digital-2020-italia`

[5] *Gli utenti dei social media in Italia* `https://blog.ofg.it/gli-utenti-dei-social-media-in-italia#:~:text=Twitter\%3A\%20le\%20donne\%20\%E2\%80\%9Ccinguettano\%E2\%80\%9D,visitatori\%20attivi\%20nel\%20nostro\%20paese.`

[6] *SpaCy Python Library, Industrial-strenght Natural Language Processing* `https://spacy.io/`

[7] *Open datasets for sentiment analysis*, `https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis`

[8] *imblearn.over_sampling.SMOTE* `https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html`

[9] *Logistic Regression model using Scikit-learn Library* https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html