The public opinion on Sanremo Music Festival 2020 expressed on Twitter.

Sanzogni Carlo mat. 935739

June 18, 2020

Contents

List of Figures

List of Tables

1 Introduction

The aim of this project is to evaluate the public response to *Sanremo Music Festival 2020*. The public reaction is represented by the tweets and so the opinion is only circumscribed to Twitter's users but comparing the final ranking of Festival with the results it can be noted that the sample is big enough to represent in good way the complete audience.
The project starts with a brief explanation about the data collection process, then the data are organized in an ordered and structured data-set. The text data are elaborated in order to perform a Sentiment Analysis and then, thanks to a versatile and adaptable expression of the polarity of the emotions the real evaluation of the event could start.
The method consist in recreating a final ranking using different *Indexes* that describe the public's response by single artist who performs at the show. The first of these Indexes will only evaluate the *Presence* of the artists in the the Tweets, then, starting from that, are estimate the *Positive*, *Negative* and *Compound* absolute impacts for each of them. Lastly these three are weighted with the first (the *Presence Index*) and the *Relative Indexes* are created.
Among them the *Relative Compound Index* is chosen as the best one, compared with the final ranking and, as will be seen later, the results are quite satisfying.
At the end the same methodology is applied also for some **Special Guests**, in order to add a management point of view and evaluate which of them has the best response, so who added value at the show and who not, some results came out but not as good as the previous.

2 Data-set description

The data-set used is completely original. Data had been collected by my colleague M. E. Passerini and me accessing the Twitter API thanks to the open source Python library *Tweepy* (for further information about this part look at the code on my GitHub repository).
Our streaming has been made in async mode, downloading data ex-post in several session, so data are, in some part, time inconsistent; this have to be intended that the distribution during the time is missing in some time slots, not because the users did not tweet but for downloading reasons, and this is a problem that did not allow me to complete a timing evaluation to compare to the artists one.
The original data-set was composed by 38 810 tweets all connected with the show by an **Hashtag** (e.g. #sanremo, #Sanremo20, #Sanremo70 etc.), which is stored in the first column of the Dataframe, unluckily this approach bring an other problem, some users use more than one of them so some tweets had been recorded more than one time. To avoid this, the tweets were reduced to 24 793 filtering at unique the **Tweet text** (Third column). Lastly in the second column are contained the **Timestamp**, so tweet's publishing day and hour, in the format yyyy-mm-dd hh-mm-ss, and in the fourth the **User Name** so the User's Twitter Nickname.

3 Methodology

First, to perform a proper Sentiment Analysis, at all the **Tweet text** are applied some clean function in order to remove the non-textual characters, the present URL and some other problematic parts of the tweets.
Starting from the new so called **Tweet puliti** (cleaned tweets) it is now possible to perform a Polarity analysis. In this project is used **Vader**, a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. With **Vader Sentiment Analyzer** every tweet is mapped with four classes of sentiments: **Negative, Positive, Neutral** and **Compound**. The first three are a typical lexicon of sentiment words approach, handling also negations and using a degree modifier to alter sentiment intensity, in fact boosters and dampeners; however the last, **Compound**, needs a special mention because it is either the most performing either the trickiest: it is not simply a computation on the firsts, but it is completely a new normalized evaluation between -1 and 1 weighted with a parameter alpha. Although the evaluation could be similar as before being the positive, negative and neutral scores computed at same time the results are little different but, according to the result of the complete process and a personal random evaluation on the binding tweets/scores performed by me, more precise. Dealing more deeply with this last sentence, due to the fact that **Vader** is built under English language I prefer to check physically the scores extracting random some tweets and the results have been very satisfying.
The real show's evaluation process starts mapping the tweets by the presence or absence of an artist, so 1 if the name is mentioned and 0 if not.

$$Presence\ Index_a = \sum_{i=1}^{N} p_i \tag{1}$$

where

$$p_i = \begin{cases} 1 & a \text{ is present} \\ 0 & a \text{ is absent} \end{cases} \tag{2}$$

Starting from that is possible to build the *Presence Index* (1) which consist in sum of all the times he or she appears and is used to weight the *Absolute Scores* and create the Relatives ones. So, for example, the *Absolute Positivity* (3) will be the sum of all the Positive values associated to an artist and the *Relative Positivity* (4) will be that value divided by the artist's *Presence Index*.

$$Absolute\ Positivity_a = \sum_{i=1}^{N} p_i * pos \tag{3}$$

$$Relative\ Positivity_a = \frac{\sum_{i=1}^{N} p_i * pos}{PI_a} \tag{4}$$

4 Experimental result

The result are very interesting in fact some important information could be noticed: looking at the *Presence Index* (Figure 1) there is one groups of artists that is highly more present than the others and is formed by: Diodato (the winner) and others

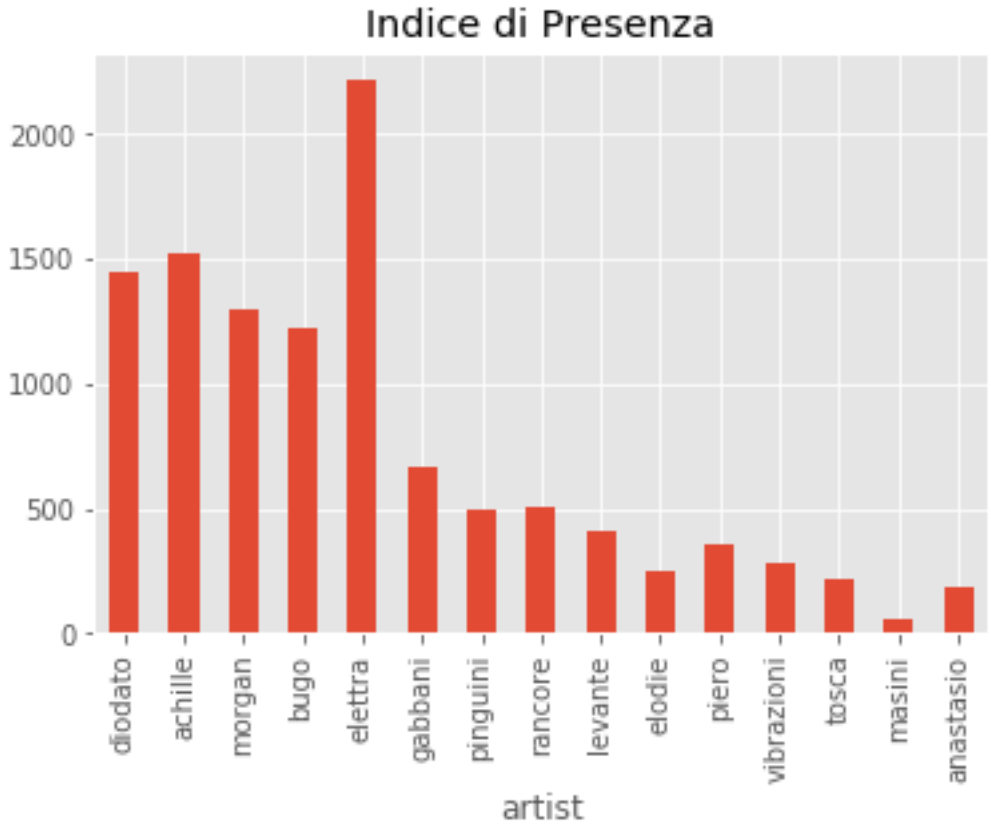

Figure 1: Presence Index

(Achille Lauro, Morgan, Bugo and Elettra Lamborghini) that did not rank so good in the final ranking. These other will be called the **Showmen** in fact they caught lots of audience but in an ambiguous way: checking the *Absolute Scores* (Figure 2) this 'definition' seems lot more meaningful: they have either high positivity either high negativity, and both are correlated with the high presence level. For these reasons fundamental is to look at *Relatives Scores* (Figure 3) that weight the Sentiment Values on the presence and more deeply to take the *Relative Compound*, which, as already said, is a synthesizer of positive and negative values. Ordering by that suddenly three of the four **Showmen** fall at the bottom of the ranking and that is due to the fact that their high scores was caused by the many interactions and not by good impressions, differently from Diodato that confirms his good public's response. In Table 1 there is the comparison between the first ten positions of the final classification of the shows against my ranking built on the *Relative compound score*: the results are quite interesting because nine of the ten artist are present, except for Irene Grandi who in my rank is substituted by Levante, and four of them are also in the right position: Diodato, first, Pinguini Tattici Nucleari, third, Le Vibrazioni, fourth, and Elodie, seventh.

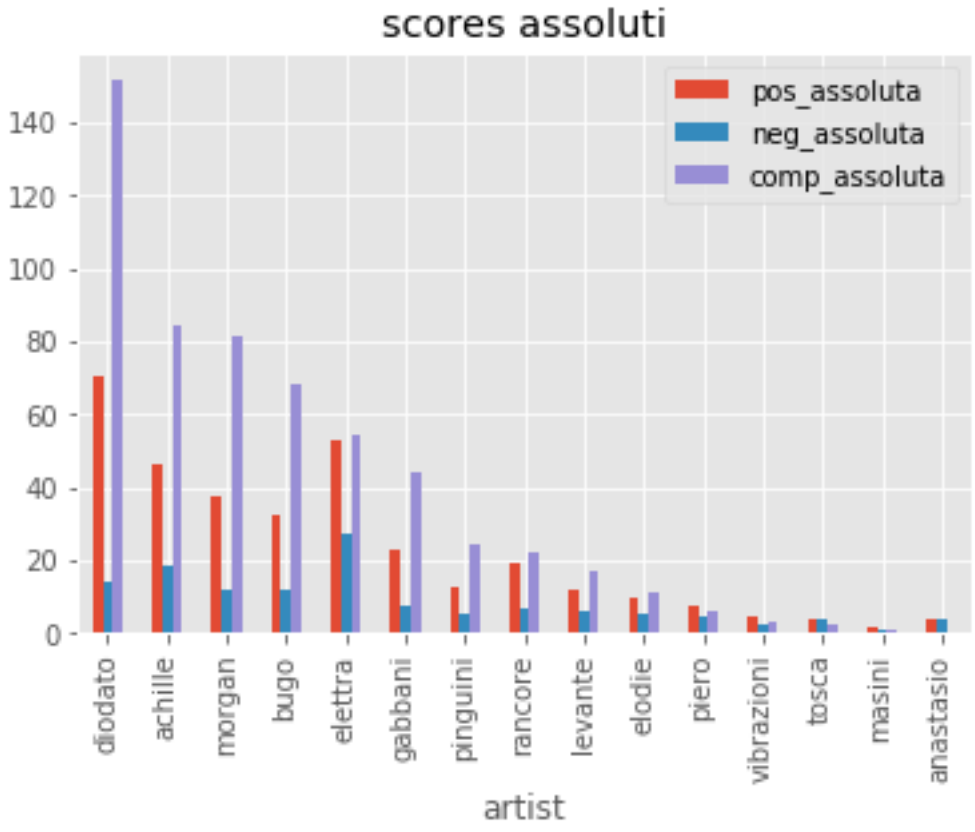

Figure 2: Absolute Scores

5 Conclusion

In conclusion it can be stated that the official final ranking is representative of the Twitter public's opinion and so that the major controversies about the weight of the Press vote are probably baseless. Switching to a more practical point of view if the lead is 'There is no such thing as bad publicity' the **Showmen** are the most important artists even if their music is not so good, they boost the interactions and is good for the TV show point of view.
The method of the *Relative compound score* (Figure 4) is then applied to a basket of **Special guests** that attended the festival. This result have however to be very contextualized in fact the level of interactions of the guests is so much lower than the artist's one (Figure 5). For this reason Coez and Lewis Capaldi became the top two, only because their interaction are a few but positive. The real information capable by this analysis is that Fiorello and Massimo Ranieri have recieved a good response by the public, Diletta Leotta, Roberto Benigni and Tiziano Ferro a discrete one, and, surprisingly, Zucchero and Biagio Antonacci were the worst and the public didn't like them. So, from a management point of view, is a must to call back the first two for the next edition, for the middle three it takes an evaluation and crossing off the last two.

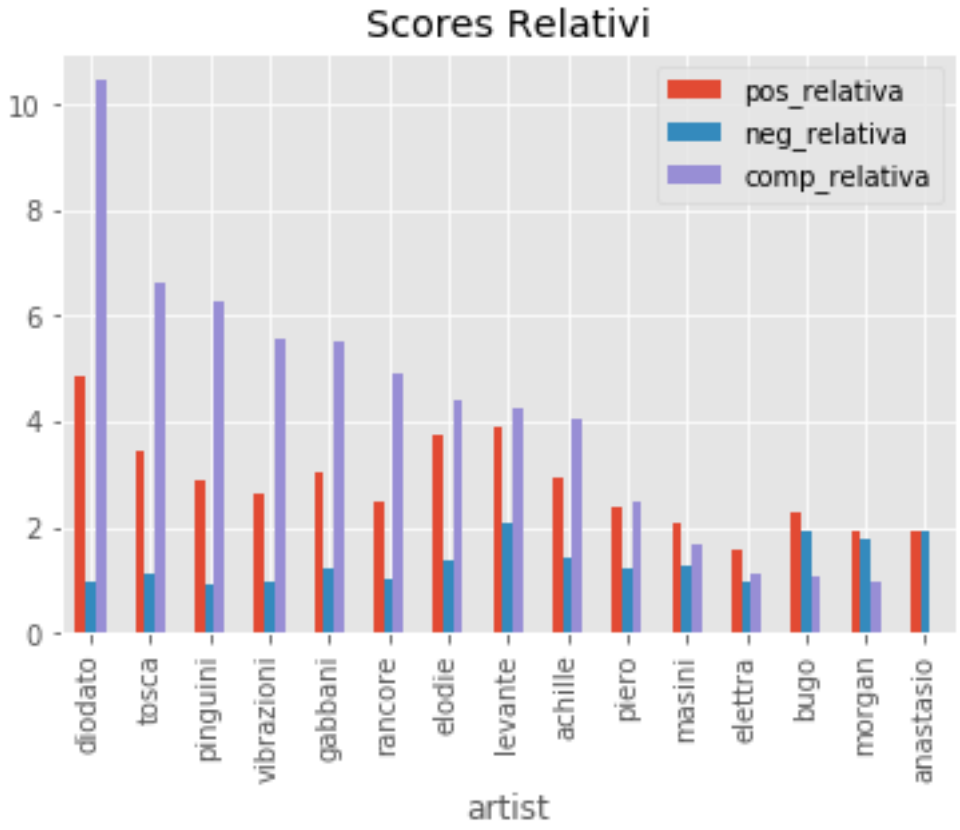

Figure 3: Relative Scores

Further extensions of the project could have be done with a timing homogeneous data-set, very interesting could be studying the fluctuation of the emotions during the phases of the day, during the week of the festival and more deeply during the live TV show and it would be also interesting to analyze the last column of the data-set, **User Name**, in order to split the pure Users by the Magazines and Press Agencies to create two different rankings and discover if is possible to create a more precise prediction by that.

1	Diodato	Diodato
2	Gabbani	Tosca
3	Pinguini TN	Pinguini TN
4	Le Vibrazioni	Le Vibrazioni
5	Piero Pelu	Gabbani
6	Tosca	Rancore
7	Elodie	Elodie
8	Achille Lauro	Levante
9	Irene Grandi	Achille Lauro
10	Rancore	Piero Pelu

Table 1: Final Ranking VS My Ranking

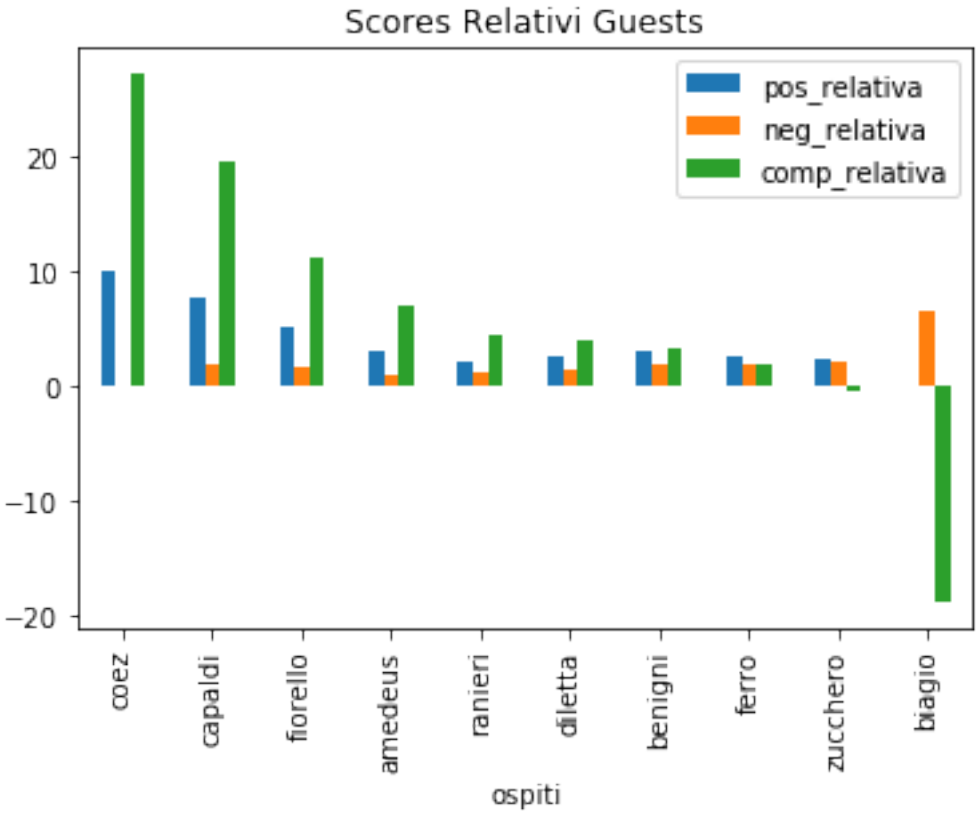

Figure 4: Guest's Relative Scores

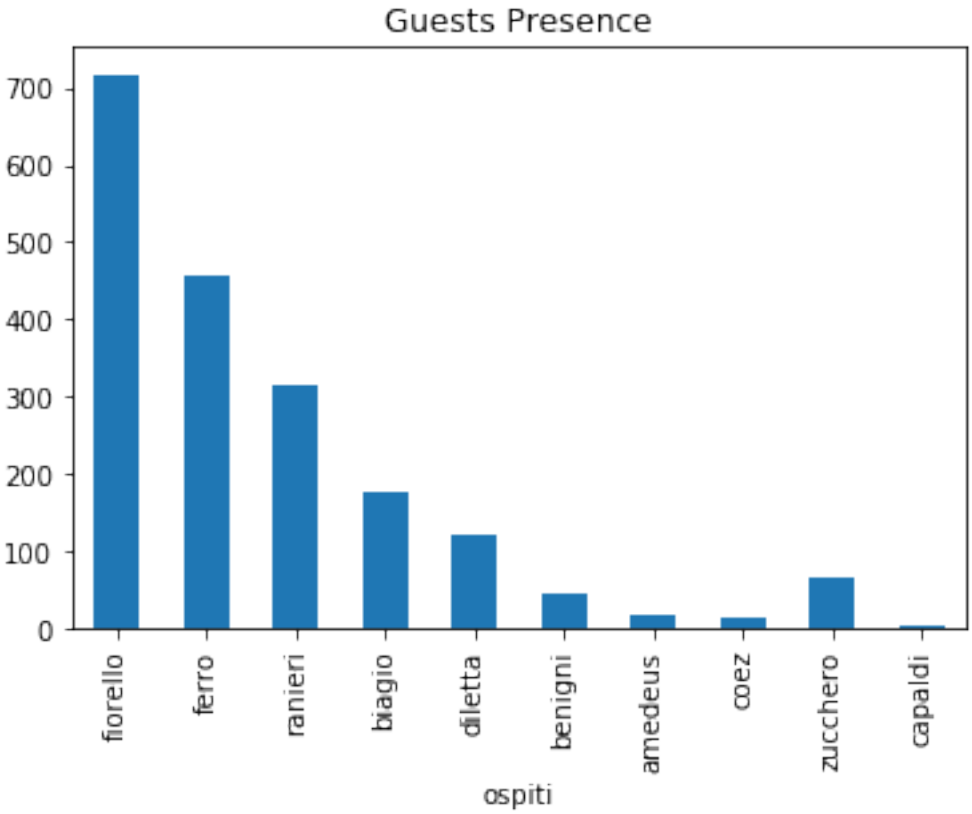

Figure 5: Guest's Presence Index