CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17824

The use of non-official data source for the analysis of public events: evidences from the Eurovision Song Contest 2022

Alessia Forciniti[1], Andrea Marletta[2], Magda Moretti[2]

[1]Department of Humanities, IULM University, Milan, Italy, [2]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy.

How to cite: Forciniti, A.; Marletta, A.; Moretti, M. 2024. Paper The use of non-official data source for the analysis of public events: evidences from the Eurovision Song Contest 2022. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17824

Abstract

The use of non-official data sources as Twitter has been implemented for the monitoring of social and public events in many different fields during last years. Following this issue, this work proposes to analyse a very well-known musical event, the Eurovision Song Contest (ESC) 2022 using tweets pooled by the official hashtag of the competition. From a methodological point of view, text mining techniques have been applied to detect the most influencing terms and topics tweeted by users during the show and to compare the official results of the contest with a ranking only based on the appreciation of the Twitter users on posts relative to the participant countries.

Keywords: *Twitter Data; Eurovision Song Contest; Text Mining.*

1. Introduction

The statistical analysis of public events has always been topic of interest of quantitative stakeholders from many points of view. If for these events a competition is considered, the study of the factors influencing the final result covered many contributions. For example in sport statistics, the best predictions are achieved taking into account a very huge quantity of statistics from official sources.

On the other hand, during last years, the social media have been recognized as one of the most powerful data source able to add a substantial contribution in forecasting procedures. The decision to place a like or to share a content on social media can be interpreted not only as a form of appreciation but also like the expression of a behaviour. Starting from this, the stakeholders can use this information in the decision making process. It is important to note that social media data are not necessarily quantitative, as number of likes, shares and views but they can also involve qualitative aspects as comments, opinions and answer to open questions making the single word a vector of information.

The contact point between the analysis of public events and the rising communicative power of the social media, lies in the possibility to arrange a link between the event of interest and the social media users using a common hashtag. In this way, all the people are following the public event live or online could comment and/or express their appreciation or dislike in real time. For some events as Tv show or musical events in which a competition is provided, this new form of interaction is translated into a voting system that could influence the final result of the competition (Demergis, 2019, Kumpulainen et al., 2020).

The aim of this study is to provide an analysis based on Twitter data of the Eurovision Song Contest (ESC) 2022 using the tweets collected by the official hashtag of the competition in order to detect the most influencing topic and terms used by Twitter users during the show. The ESC is a very well-known musical event organized yearly by the European Broadcasting Union (EBU), it is a competition among countries in which each country is represented by an artist and the winner country is declared after a voting system based on a combined ranking using quality juries and phone-in-vote. Another objective of this study is to compare the final result of the contest with an alternative ranking only based on the tweets in which the participant country is mentioned. The final goal is to understand if a correspondence exists between the real and a Twitter-based ranking in order to measure the different perception of the Twitter population respect to the real world for this event.

The paper is structured as follows: after the introduction, a second section is dedicated to the methodologies used to answer the research objectives. A third section will show the description of the dataset and some preliminary results. Finally, some conclusions will follow.

2. Text Mining techniques

The authomatic ananlysis of textual data, also known as text mining, is a set of tools allowing the traslation from word to information. This technique is based on the availability of textual and non-structured data, that after some operations of data cleaning can be elaborated to be transformed in structured data, able to be necessarily analysed (Tuzzi, 2003).

A complete definition of text mining can be retrieved in Feldman and Sanger (2007): ‘Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval and knowledge management’. In particular, when the text mining objects are referred to social networks, this technique is also known as web mining or social media mining.

The starting point of the authomatic ananlysis of texts is the corpus, that is, the set of textual units coherent to be analysed (Bolasco and De Mauro, 2013). It is therefore a homogeneous collection for treated topic, structure of the sentence and length able to be observed without systematic distortions. On the basis of the extent and of the composition of the corpus, two typologies of corpus are identified: collection of texts and collection of fragmented texts. A second classification of the corpora detected a single document or a set of records in which each row is differently generated. For social media text mining the corpus is a collection fo fragmented text with a set of rows, that are posts, tweets, reviews generated by social media users.

Before starting the textual analysis, a pre-procssing procedure is needed using some enhancements useful for this category of data. Firstly, a process of micro-fragmentation (tokenization) of the corpus is applied linking a numerical code to each word. Secondly, all the uppercase characters are detected and transformed into lowercase, this step simplifies the analysis but it could generate a loss of information. Thirdly, all the numerical string and the separators (comma, dot, etc..) has to be deleted. Fourthly, some entire words could be deleted from the corpus, they are empty enitities and their elimination did not create loss of meaning. This list usually contains articles, conjunctions, prepositions prononuns, possessive adjectives, common verbs. Finally, the Term-Document Matrix (TDM) is composed by all the words that are still in the corpus after the application of the previous steps. In the TDM, each row is a word, each column is a document of the corpus and each cell is the occurrency of the single word in the single document.

After the pre-processing phase, in order to extract information from textual data, the existing techniques can be classified into three groups: the frequency analysis, the cluster analysis and the sentiment analysis. The frequency analysis consists in the counting of the words contained in the corpus in order to create a dictionary of most important terms. This kind of analysis is descriptive and not sufficient to catch the total information from the corpus. Beyond the removed stopwords, the terms with the highest frequency are often generic and expected because related to the interest topic. For this reason, in this study the attention is focused only on some categories of words as countries and artists. The cluster analysis regards the possibility to create groups of terms that recur together using hierarchical and non-hierarchical models already used for quantitative data. The Sentitment Analysis (SA) represents a discipline based on Natural Language Processing (NLP) and aimed at identifying opinions, emotional, behavioral and attitudinal dimensions expressed in natural language texts (Pang and Lee, 2008). The main objective is to determine the semantic orientation of texts written in natural language by classifying documents based on their polarity; where polarity refers to the linguistic distinction between affirmative and negative terms. There are three different levels of semantic ordering characterized by different granularity: a) subjectivity/objectivity (SO orientation): aimed at determining factual nature or subjective judgment; b) positivity/negativity (PN orientation) (Liu, Hu & Cheng, 2005): with the purpose of determining whether it expresses a positive or negative opinion, and often neutral; c) strength of positivity/negativity (PN strength):

aimed at indicating different levels of intensity of positivity and negativity; from the most negative to the most positive (Nielsen, 2011).

The three different levels of semantic analysis can be performed on the individual document, at the sentence level, and at the aspect level. In the first case, the positive or negative polarity is returned at the general document level, while in the second case, the document is segmented into sentences and the semantic orientation is detected for each sentence. Polarity attribution can occur through lexicon-based approaches (generated manually or automatically, semi-automatically), machine learning, deep learning, hybrid strategies, or those recurring to artificial intelligence. The choice of the lexicon to adopt represents an ongoing debate, and the application context is also an area of continuous research, dividing lexicons into generalists and domain-based. Further discussions in the literature focus on language, as most lexicons are developed for English, not covering other languages or providing limited resources (Zavarrone & Forciniti, 2023). Another possibility is represented by using multilingual language models, that classify and cluster tweets into homogeneous groups irrespective of their language, thereby facilitating a more comprehensive summarization of opinions expressed across various languages.

In this study, a strategy of lexicon-based classification at the document level is adopted, where each tweet is a different document. Specifically, the NRC (National Research Council of Canada) lexicon developed by Mohammad and Turney (2010; 2013) for sentiment analysis of Twitter, and available for various languages including Italian, was used. The NRC lexicon allowed the detection of PN orientation: positive, negative, and neutral opinions were detected as -1, 0, +1 to indicate respectively negative, neutral, and positive terms. Additionally, we utilized the word–emotion association developed by Mohammad and Turney through Amazon's Mechanical Turk. The NRC is based on annotations of the eight primary emotions suggested by Plutchik (1962, 1994): joy, sadness, anger, fear, disgust, surprise, trust, and anticipation.

3. Data collection and results

Data have been collected through the official Twitter API using all the tweets containing the official Italian hashtag of the competition #ESCita. The event took place in Turin, Italy from 10th to 14th May 2022 in three shows. The contest was won by Ukraine Since the event has a duration of more days with two semi-finals show and a final show, multiple extractions of tweets have been achieved starting from the day of the first semi-final until to the day after the final. For this study, only tweets about the day of the final show have been considered for a total of n=61,464 tweets. The variable of interests for each single tweet are: the text in Italian language, the exact time of publication, the username of the author, the number of likes and the number of retweets.

As reported in the previous section, a huge procedure of data cleaning has been applied on the text variable before obtaining the TDM, after the transformation into lowercase and the removal of numerical string and the separators (comma, dot, etc..), a list of words has been added to the usual stopwords list, because marked as empty terms. For example all the terms directly related to the hashtag #ESCita like "eurovision, esc, eurovsiontv, eurovisionsongcontest, esc, eurovisionrai" have been removed as the stopwords. After the construction of the TDM, a dictionary of the entire corpus composed by 33,107 words has been obtained. The word with the maximum frequency in the dictionary is "canzone" (4,403 times) the Italian term standing for "song", this could be expected since the ESC is primarly a music contest. In the top-10 of the most present words, there are also three countries (Spain, Ukraine and Italy), two Italian presenters (Malgioglio and Laura) and two Italian singers (Blanco and Mahmood). As expected, the dictionary is very focused on the Italian side of the competition, this is a limitation of the study but since the show take place in Italy, this choice could be justified.

A good way to synthetize the entire dictionary is the use of a wordcloud, a graph in which all the words over a fixed threshold are represented with a different colour and size, the terms with a bigger type are those with a bigger frequency. In figure 1, all the terms with at least 600 presences in the entire corpus have been represents using a wordcloud. Looking at this figure, it is possible to note that most of the terms indicate name of the participant countries or artists involved in the show as presenters, singers and guests.

Figure 1. Wordcloud of the terms with at least 600 presences in the ESC textual corpus. Source: Twitter Eurovision data (2022).

This issue advices to focus the attention on some specific categories of words, for example the participant countries, in order to search for a link between the final ranking of the contest and the appreciation that these countries received from Twitter users that used the #ESCita hashtag.

The final ranking of ESC has been obtained by the sum of points of televoters and juries from all the participant countries, with the exception that each country can not to vote for itself, therefore the Italian voters ranking did not consider Italy. A simplified version of the obtained dictionary only considering the mentioned countries could be considered as a proxy of the appreciation ranking by Twitter users. The produced ranking has many distortions because it only considers tweets using an Italian hashtag and above all because a mention in a post do not correspond necessarily to an appreciation. Nevertheless all these presuppositions, a simple correlation coefficient between the official ranking and the mentions ranking has been computed and it is equal to 0.52 (p-value = 0.007) denoting a positive association between the two rankings. In Figure 2, a simple scatterplot between the two rankings denotes some potential clusters of country, for example France and Germany are in bottom right part of the scatterplot because nevertheless a very bad position in the official ranking, they have a high number of mentions in Italian tweets, probably because they are in the big 5 group (a group of countries directly admitted into the final show) and a lot of users commented their disappointing result. Other correlations between rankings have been computed, for example comparing the mention ranking with the televoters ranking in Italy (available on the official website of ESC: https://eurovision.tv/), to avoid the bias due to the fact that the mention ranking is only based on Italian tweets, but the correlation coefficient is still about 0.50 (p-value = 0.012).

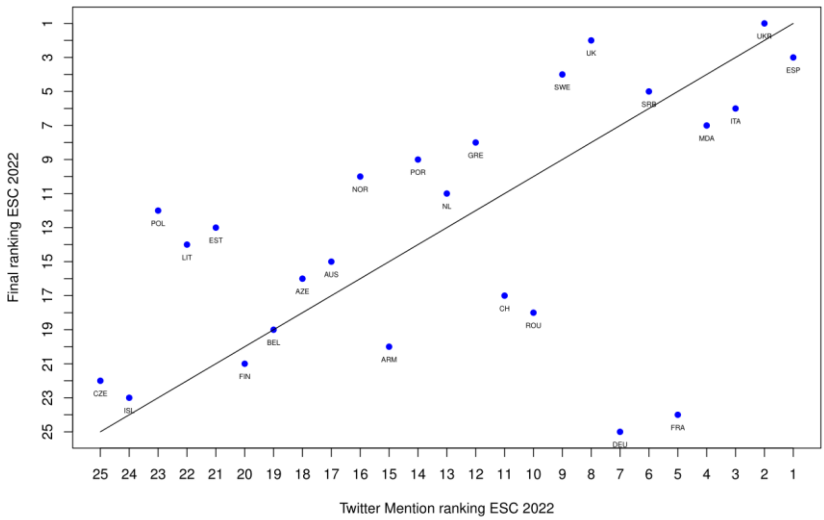

Figure 2. Comparison between official ranking ESC and a ranking based on Twitter mentions. Source: Twitter Eurovision data (2022).

Finally, a sentiment analysis on the entire corpus of the final show of ESC has been realized using the NRC lexicon, according to this classification each tweet could be categorized as positive, negative or neutral. The 91.3% of the tweets are classified as neutral tweets, the 5.1% as positive tweets and the remaining 3.6% as negative tweets.

In figure 3 a representation of the sentiment analysis is about the eight primary emotions described in the previous section. The anticipation is most present emotion, it is present in 20.1% of the tweets, the trust is following with the 18.9%, followed by sadness and joy with respectively 12.9% and 12.5%. A limitation of this study is represented by the fact that sentiment analysis doesn't consider any irony or sarcasm detection: irony is really important when managing comments about TV contests.

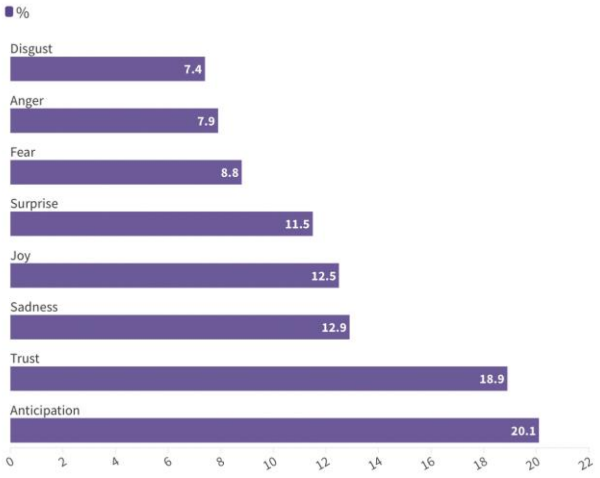

Figure 3. Eight primary emotions for tweets about the final show of ESC. Source: Twitter Eurovision data (2022).

4. Conclusions

The aim of this work was to analyse a very well-known musical event as the ESC 2022 using non-structured data from social media. In particular, a text mining analysis has been conducted on 61,464 tweets containing the official hashtag for Italy, #ESCita and posted during the last day of the event, detecting some preliminary evidences. From a methodological point of view, after an important pre-processing procedure of data cleaning, a dictionary has been created identifying the participant countries as relevant category of words. On the basis of the mentions received by the countries, a proxy of an appreciation ranking has been compared with the official ranking denoting a good correlation. Finally a sentiment analysis on the entire corpus revealed the anticipation as main emotion in the tweets. The study is still in progress and many limitations are present, both for the nature of data and because a direct connection between mentioning and appreciation is missing. Following this issue, some future works could regard the enhancement of the mentions ranking taking into account like and retweets received by mentioned posts and introducing the publication hour as control variable.

References

Bolasco, S., & De Mauro, T. (2013). *L'analisi automatica dei testi: fare ricerca con il text mining*. Carocci editore.

Demergis, D. (2019, October). Predicting Eurovision song contest results by interpreting the tweets of Eurovision fans. In 2019 Sixth international conference on social networks analysis, management and security (SNAMS) (pp. 521-528). IEEE.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Kumpulainen, I., Praks, E., Korhonen, T., Ni, A., Rissanen, V., & Vankka, J. (2020, September). Predicting Eurovision Song Contest Results Using Sentiment Analysis. In Conference on Artificial Intelligence and Natural Language (pp. 87-108). Cham: Springer International Publishing.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*.

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, *29*(3), 436-465.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, *2*(1–2), 1-135.

Plutchik, R. (1962). *The Emotions*. New York: Random House.

Plutchik, R. (1994). *The psychology and biology of emotion*. HarperCollins College Publishers.

Tuzzi, A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Carocci editore.

Zavarrone, E., & Forciniti, A. (2023, July). CSR & Sentiment Analysis: A New Customized Dictionary. In *International Conference on Deep Learning Theory and Applications* (pp. 466-479). Cham: Springer Nature Switzerland.