Università degli studi Milano Bicocca

Corso di laurea magistrale data science, a.a 2022/2023

World Cup 2022 Twitter Sentiment Analysis

Sabino Giuseppe

Ceccarelli Daniele

Pagani Gabriele

Contents

1.Introduction

In this paper, we delve into the 2022 FIFA World Cup in Qatar, one of the biggest and most watched sporting events in the world. The tournament brings together some of the best national teams from around the globe to compete for the coveted trophy. In order to gain a deeper understanding of the tournament and its impact, we have conducted a data-driven analysis that combines statistics from the FIFA website with a Sentiment Analysis of tweets related to each match. Our goal was to provide a comprehensive view of the tournament by incorporating both quantitative and qualitative data.

Data visualization plays a crucial role in this analysis, as it allows us to visually represent the events in an intuitive format. Through the use of various charts techniques, we aim to shed light on the performance of the teams, the mood of the fans, and the overall sentiment towards the event. By exploring the relationships between the various data points, we hope to uncover patterns and insights that would otherwise remain hidden.

The results of our analysis provide a perspective on the World Cup and offer new insights into the tournament and its impact with the purpose of answering the following questions: how did people react to one of the most watched events in the world? Is there a relationship between match statistics and user sentiments?

2. Dataset explanation

2.1 Data Acquisition

For the Data Acquisition part, the goal is to obtain the FIFA ranking of individual teams, the statistics of the 2022 World Cup matches, and tweets related to them. In the absence of an API on the FIFA website, the Python Selenium library was used to navigate the FIFA.com website through the Microsoft Edge browser.

The first step is to gather the list of teams that participated in the World Cup in Qatar. In the "TEAMS" section of FIFA.com, all 32 national teams are listed in alphabetical order. The official FIFA team names were then used to acquire the tweets posted during the World Cup.

The second step is to identify the URL of the World Cup matches. This step is crucial in order to acquire data related to individual matches (e.g. starting time and date, final result, statistics).

The following are the steps of the automated procedure that was implemented to navigate on the pages of individual matches and collect the necessary data:

1. Once the match website is opened , the following data can be acquired: date and time of the kick-off (Italian time zone), type of match (i.e. initial groups, eighth, quarterfinals...)
2. Scrolling down on the website, under the "Highlights" button, there is a "STATS" link. Once the "STATS" button is clicked, all the match statistics can be viewed.

Subsequently, a procedure was also defined to acquire the FIFA rankings of the male national soccer teams that participated in the World Cup in Qatar from the FIFA website. On the website "https://www.fifa.com/fifa-world-ranking/men?dateId=id13869", you can view the evolution over time of the ranking score of each team, using the dropdown menu at the top right. FIFA and Coca-Cola developed an algorithm to assign each team a score based on historical performance. FIFA updated the scores and rankings of male teams on December 22, 2022 (after the World Cup final in Qatar on December 18). The last update before the start of the World Cup was on October 6, 2022.

Regarding the tweets, the snscrape library was used. Firstly, the statistics dataset was cleaned because for some matches there were few comments found. This is because it is presumed that few people commented on the match "Korea Republic-Portugal" but rather "Korea-Portugal", ultimately resulting in positive findings on the increase in tweets. The research was conducted in English as it is more international, providing a wider vision of the world in general and not just in Italy. 1000 tweets were obtained for each match to have a large enough sample to give value to the statistics obtained; it was not possible to obtain a larger number as for many matches this limit reached the entire selectable field. It was necessary to limit the acquisition of tweets to only the days the match was played.

This provides a more narrow view with comments related only to the football world (a USA-Iran search may lead to political topics) and also solves the problem of double matches, in this case only Croatia-Morocco.

2.2 Data Storage

Regarding the data storage, the most logical choice is to use a MongoDB, thanks to which there are no fixed schemes, therefore no null values.

Thanks to the use of the "pymongo" library, it was possible to directly connect the database with the Jupyter notebook, from which the database was created. The database has three collections:

- WorldCupStats: containing 64 documents, each of which represents the statistics of each match
- WorldCupTweet: containing 57518 documents of the tweets shown earlier
- WorldCupRanking: containing 32 documents with the ranking of the national teams and the team value.

2.3 Sentiment Analysis

We implemented a Sentiment Analysis model, which is a development of Natural Language Processing that develops systems to identify and extract opinions from a text. Given that the reasons behind people tweeting are mainly reactions to events, trends, and news, by analyzing the Sentiment of tweets it is possible to examine their level of positivity or negativity (polarity). Long texts are not always decipherable in an optimal way, so the limited number of characters imposed by Twitter is certainly an element that can help Sentiment Analysis. During the pre-processing phase, various text cleaning techniques were adopted to make the data more usable by the Sentiment Analysis model. In particular:

- Normalization: is a task always necessary for Text Mining applications and involves reducing different forms of the same element to a unique form for the entire corpus. Normalization is one of the tasks that generally leaves more room for action as it is strongly dependent on the type of Text Mining task being carried out. For this reason, the following techniques were proposed: o Lowercase text; o Removing spaces; o Removing punctuation; o Removing URLs o Removing duplicates; Exclamation marks were not removed because they will be useful for Sentiment Analysis and, for the same reason, emojis were converted to the expressions "good" and "bad" depending on their meaning. In addition, hashtags were also left in the dataset to develop the Word Cloud of the 50 most frequent hashtags in the dataset.
- Stop-words: involves removing words, in this case in English, that occur very frequently in the corpus because they are not considered useful in characterizing the documents. Such removal is necessary because their presence often undermines the results of classification models.

- Tokenization: a crucial step that fragments texts into single elements, called tokens, by dividing a text stream into dense units of meaning. To do this, simple regular expressions were applied to the text in question. This is a fundamental pre-processing technique for applying Text Mining models.
- Stemming: reduces words to their roots, to their basic form. Despite the risk of losing the precise meaning of the words, if developed correctly it increases the performance of the model to be implemented. The Porter algorithm, one of the most used for stemming in English, was used.
- Lemmatization: reduces flexible terms to a base form with the advantage of reducing the size of the analyzed dictionary. It is very similar to stemming but with the difference that stemming operates on individual words independently, following only syntactic rules, while lemmatization also considers context and semantics.

After the pre-processing phase, the Sentiment Analysis was implemented using Textblob. TextBlob is a Python library for Natural Language Processing (NLP) that provides a simple interface to access various features, including sentiment analysis. With TextBlob, it is possible to analyze the sentiment of a text using a trained model that has been incorporated into the library itself. The model assigns a sentiment score between -1 and 1, with -1 indicating a negative sentiment, 0 a neutral sentiment, and 1 a positive sentiment. Out of a total of 55023 tweets, the following were classified:

- 10197 tweets with a negative sentiment (19%);
- 21036 tweets with a neutral sentiment (38%);
- 23790 tweets with a positive sentiment (43%).

2.4 Data Integration

Before starting the data visualization part, it was necessary to proceed with the integration of the three datasets. First, a query was made to aggregate the tweets. In this way for each match is available the number of Positive, negative and neutral tweet.

- WorldCupCollection.aggregate(pipeline): with the following pipeline inserted, to aggregate the data by match and have the total number of tweets. The aggregation was also done on the "data_partita" column, as Croatia-Morocco was played twice on different dates.

```
pipeline=[{ "$group": {"_id":["$partita","$sentiment","$data_partita"], "sentiment_count": { "$count": {} }}}]
```

Directly from the tableau data source it was possible to integrate the resources in such a way as to have statistics, the number of positive tweets and the pre and post-world cup rankings of both teams for each match.

3. Data Visualization

Data visualization is the process of translating data into graphs, it is also a huge part of the data analysis process and an efficient way to communicate information universally, quickly and effectively for everyone.

Data arrives with such an overwhelming speed, volume and variety that one is not able to understand it without applying a layer of abstraction such as the visual one.

This chapter try to explain what, how and why Data Visualization was designed and developed.

The entire presentation is present at the following link in order to have an interactive view of the whole:

https://public.tableau.com/app/profile/giuseppe4109/viz/WorldCup2022SentimentAnalysis/WorldCup2022TwitterSentimentAnalysis?publish=yes

3.1 Which country received the greatest number of positive tweets?

This dashboard represents an interactivity map with all the countries that participate in the tournament except England and Wales which are shown as "United Kingdom" on the map. Each continent was divided into its own confederation, because in the history of the world cup, only teams belonging to UEFA and CONMEBOL have won the tournament.

So, the question at hand is:

Could this be a factor influencing user sentiment, because people follow the advantaged team, or do users prefer to follow the "underdog" team in the tournament?

As can be seen from the image below, the countries are represented in the form of bubble graphs, within which the Average number of positive tweets that each team received per game is represented. This was necessary because not all the teams played the same number of games, in fact Argentina, reaching the final, played 7, while the teams eliminated in the group stage only 3 games.

In this way the number of tweets received is not affected by the number of games played.

It has been noted that the country that has received the most tweets on average is Morocco, the revelation of the world championship, the first African team to ever reach the semi-finals of the world cup.

In third place, together with the finalist France, is Japan, another revelation of the world cup that eliminated Germany and won against Spain

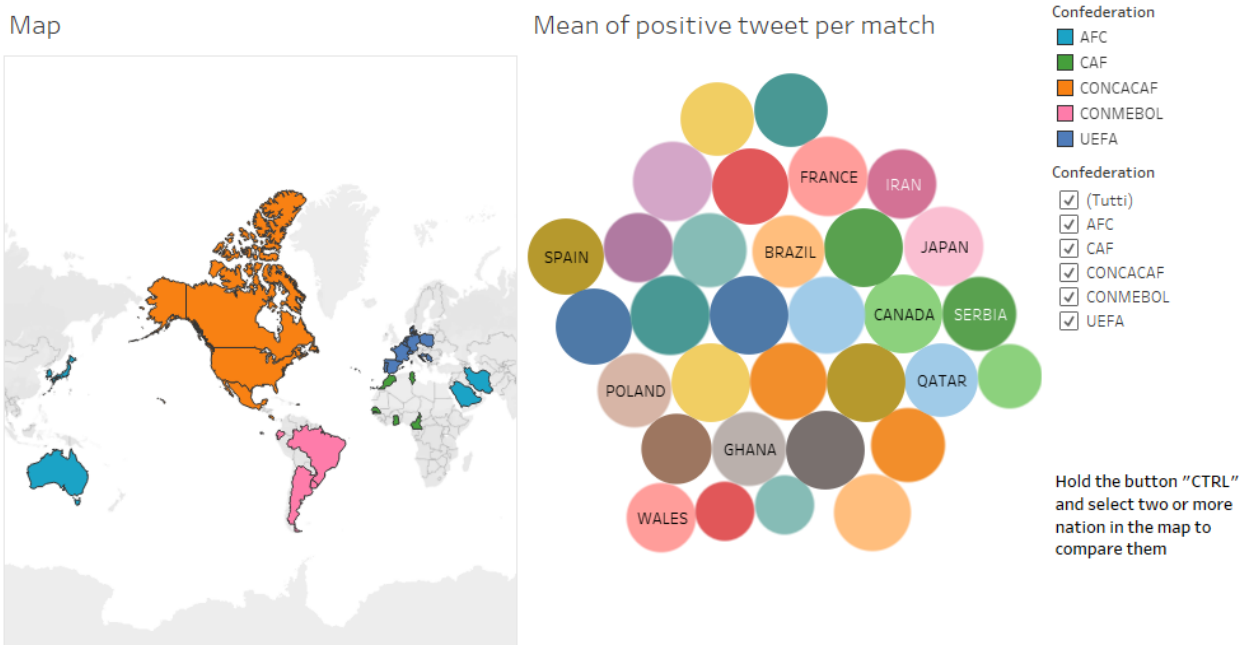


Figure 1

3.2 Which stats had the greatest effect on viewers' sentiment?

In general, the statistics that can most influence a viewer's state of mind are the number of goals scored, a symbol of entertainment and fun. On the other hand, the variable "ball possession" can have the opposite effect on viewers' sentiment. When one team dominates the entire match and controls the ball most of the time, viewers will perceive the game as boring. The main objective of the following visualization is to identify whether there is a correlation between the number of goal scored and Number of positive tweet (the two scatterplots at the top of Figure 2), and a correlation between total passes and number of positive tweet (the two scatterplot at the bottom of Figure 2).

There is the possibility to compare the global correlation of all the match played, with a team that is possible to choose through the filter.

In the example below, it is compared the Argentina's match, with all the other match, in order to identify a correlation. It is easy to note that the goal scored affects in a positive way the general sentiment of a user, but it is more difficult to highlight this correlation even in the graph representing ball possession.

Obviously, this does not apply to all teams, for example for Spain that gets the same number of positive tweets both in games in which it scored 0 goals, and in those in which 8 goals were scored. For this reason, it was given the opportunity to select the team to highlight.

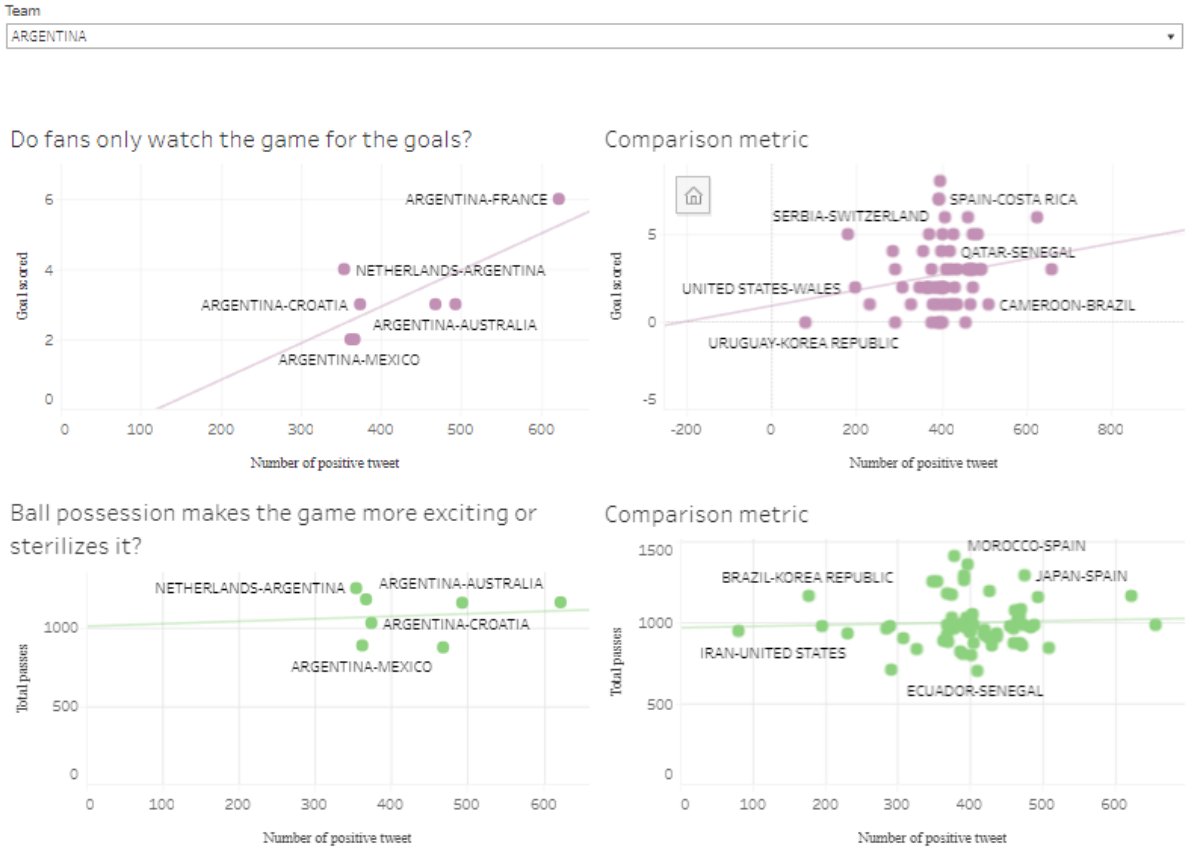


Figure 2

3.3 How does ball possession affect viewers' sentiment of a World Cup match?

To better deal with the previous speech, it has been analyzed in more detail how much a "static" game affects the viewers' sentiment. Since Pep Guardiola started coaching the Barcelona football team, "tiki-taka" has become an increasingly effective way of playing the game. But do the spectators like this style of play or does it slow down too much the game and make it boring to watch?

On the x axis there are the names of the games played, while on the y axis the number of total passes. Each game takes on a color based on the percentage of positive tweets received. the red color indicates that the match had a low percentage of positive comments, the green color indicates the opposite.

It is highlighted the mean of the total passes of all the match, which is around 1000 passes. Then this number is used like a comparison metric to determine if the user is more interested in a match with less or more than this number. In general, there are more matches with at least 50% positive tweets when less than 1000 passes are made in the match. This is very evident by removing the last matches of the World Cup with the appropriate filters (final and playoff for third place) because they are the ones with the highest number of positive tweets.

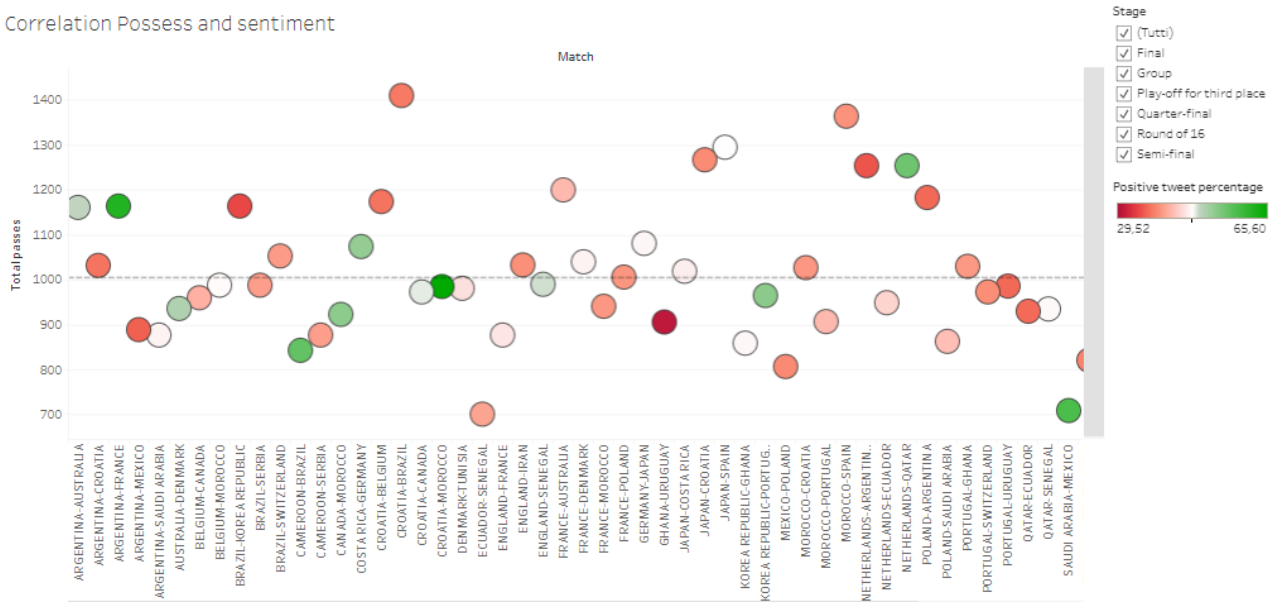


Figure 3

3.4 Does ranking affect viewers' sentiment about a match?

The president of Real Madrid, Florentino Perez, says that young people are moving away from the world of football, and one of the reasons could be the difference in the technical level of the two teams. Do available data from Twitter validate Florentino Perez's statement?

As previously mentioned, FIFA developed its own algorithm to assign a score to each national team based on their historical performance. The scores are then used to rank all the national teams within the FIFA organization.

The chart below shows the difference in rankings for each team before and after the World Cup in Qatar.

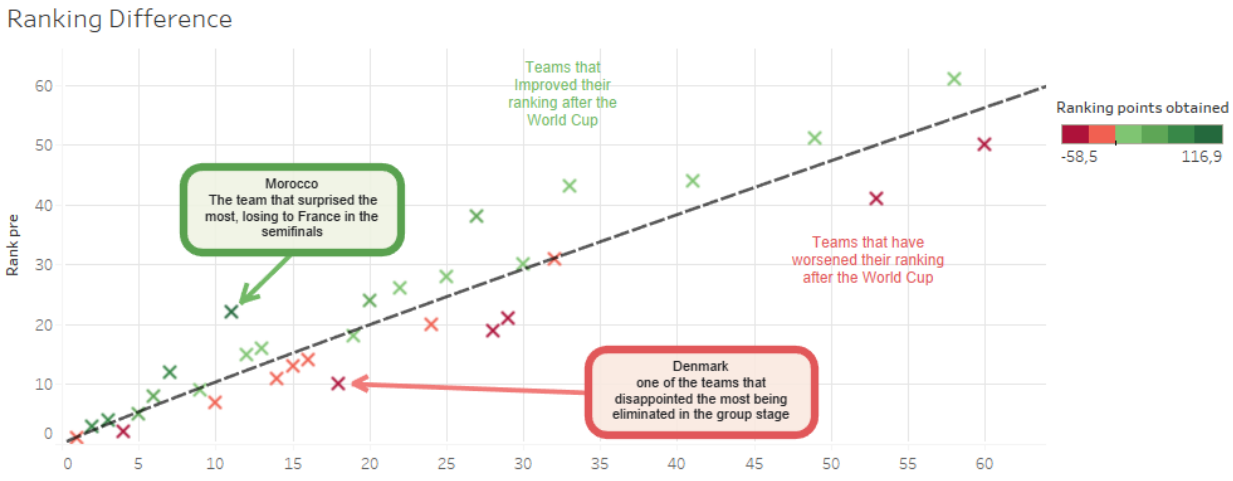
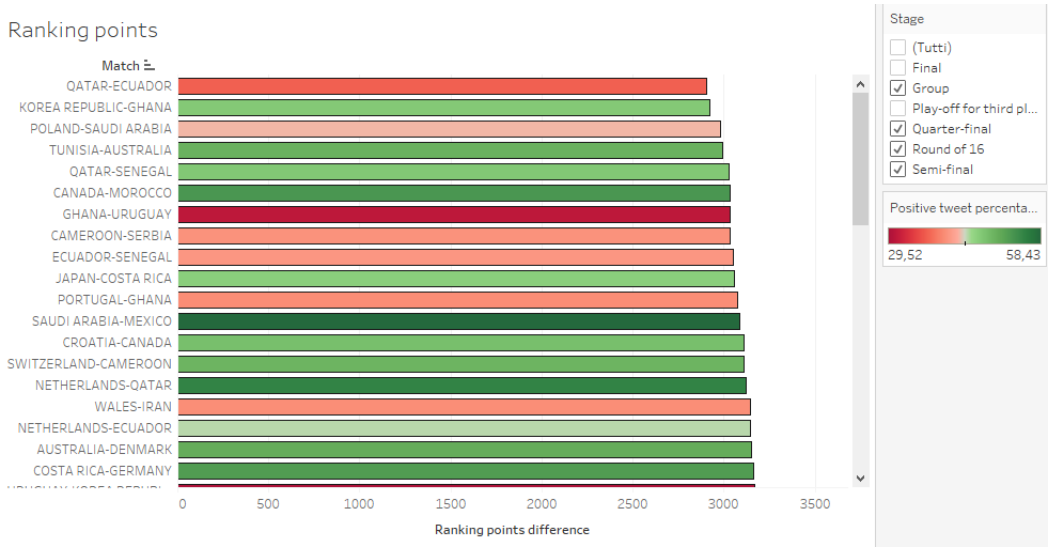


Figure 4

In fact, you can see how Morocco, that surprised analysts and viewers by reaching the semifinals, had one of the biggest increases in points, unlike Denmark, who did not meet expectations by being eliminated in the group stage.

It is now possible to see the effect that team rankings (according to FIFA's own algorithm) has on the viewers' sentiment about each game. On the x-axis, the sum of the ranking scores of the two teams is shown. The higher the score, the higher the two teams are placed.



By
filtering the matches to obtain games up to the semi-final, it almost seems that there is a greater involvement in the matches with a lower ranking score, in contrast with what the president Florentino Perez said.

Figure 5

4. Evaluation

This chapter explains how previous dashboards have been evaluated.

Three quality assessments were used:

1. Heuristic evaluation:3 user interact with the infoviz, asking them to «think aloud», and detect usability problems;
2. User Test: 7 user perform a task requiring to interact with the data viz;
3. Psychometric questionnaire: 18 users evaluate some quality dimensions of the interaction with the data viz, using the Cabitza-Locoro scale.

4.1 Heuristic evaluation

The main problems were found in the first dashboard, as the interactive use of the map does not seem to be clear, for this reason a small paragraph has been inserted to explain how to use it. Initially there was a second legend used to identify nations by color, but it was deemed redundant as the name is present within each bubble.

In the second slide it was possible to use a comparison to understand the general progress of the tournament, so as to be able to compare the selected nation.

In dashboards 3 and 4 the main problem was due to the use of colours, in fact, since most of the values are close to each other, it was difficult to be able to distinguish them quickly.

was solved using a red-green scale, where red points to a negative value and green to a positive value.

4.2 User Test

Seven individuals took part in the user test and were asked to perform some tasks on the interactive dashboards.

Three different tests were designed for the users. Data related to success rate and execution time were recorded for each exercise. In the following paragraphs the results of each test will be described in greater detail.

The user test consisted in performing the following tasks:

1. In Dashboard #1, please complete the following tasks:
    a. select the semi-finalist of World Cup 2022 (Argentina, Croatia, Francia and Morocco). Which semi-finalist was mentioned the most on Twitter?
    b. Remove from the filter the teams belonging to UEFA or CONMEBOL. What team was mentioned the most on Twitter?
2. Looking at Dashboard #2, is there a statistical relationship between number of scored goals and positive tweets for Argentina, France and Spain? For the same three team, please analyse whether the number of total passes is correlated to the percentage of positive tweets.
3. Please remove the final game, the semi-finals and the playoff for third place from the filter of Dashboard #3. Which type of game did Twitter users enjoy the most? The ones with more than a 1000 passes or games with less than 1000 passes.

The table below provides a summary of median time (expressed in seconds) for each task and the corresponding success rate.

Based on median values, among users who completed their objectives, Task #1 took the longest time to complete.

Only one user failed to complete Task #2 and every single user was able to successfully finish Task #3.

Task	Success	Number of Users	Median Time (Seconds)
1	No	2	106
1	Yes	5	83
2	No	1	74
2	Yes	6	55
3	No	0	n.a.
3	Yes	7	28

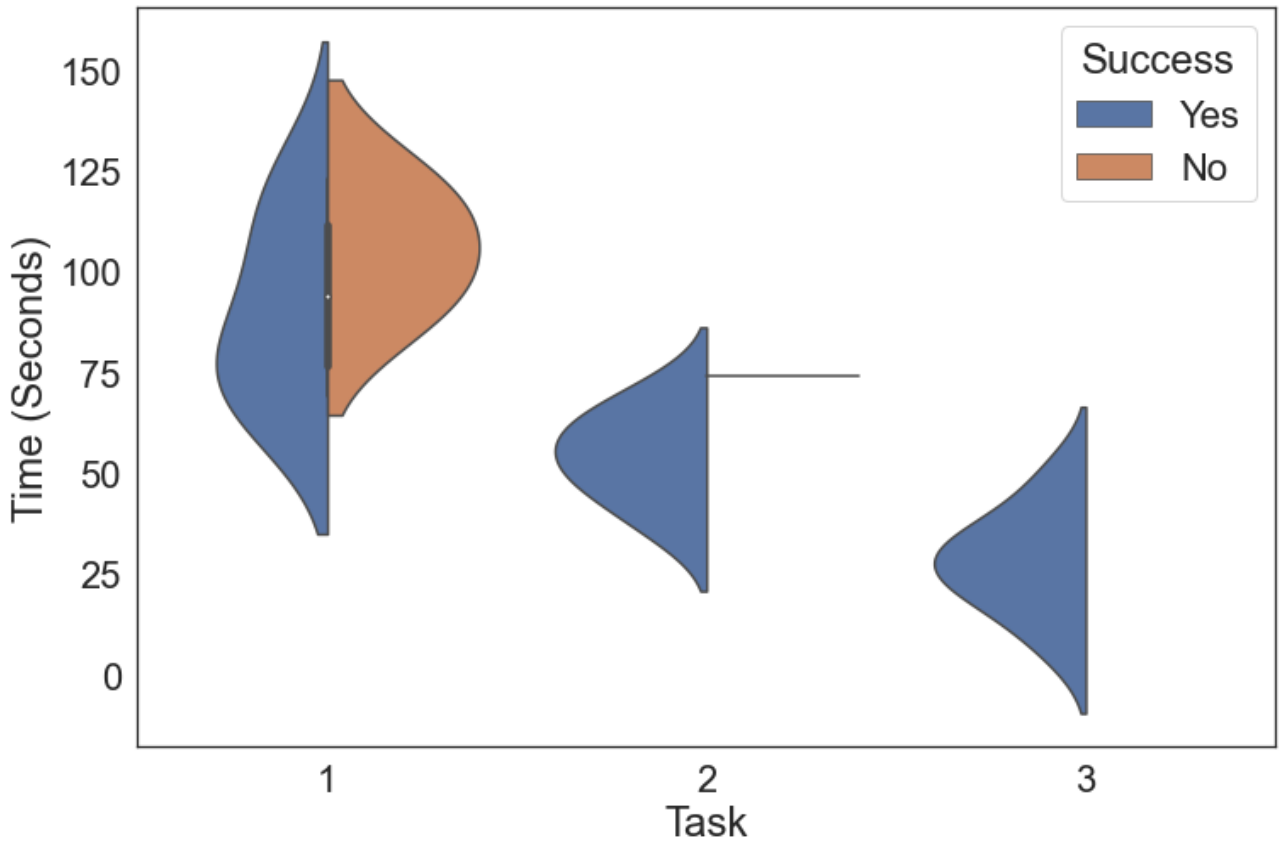


Figure 6

The violin plots shown above provide greater insight on the distribution of the time spent by users in performing each task. “Hue” parameter of the Seaborn library was used to separate the distribution of execution time of user that successfully completed each task from users who failed to achieve the objective.

4.3 Psychometric questionnaire

In the following step of the Evaluation phase of the data visualization project, a psychometric questionnaire was sent to 18 individuals. The main objective of this questionnaire was to assess specific quality dimensions of the dashboards.

The Cabitza-Locoro scale was used to design on Google Forms the questionnaire. Users had to fill out the questionnaire and rate the interactive dashboards on a scale from 1 (lowest grade) to 6 (maximum grade) the utility, intuitivity, beauty and informativity of each dashboard.

Once the results of the questionnaire were collected from Google Forms, median values and distribution functions were analysed.

Distributions of the answers submitted by users were visualized using violin plots.

Looking at the table below, it appears the four different dashboards received a median grade between 3 and 4.

Dashboard #3 was rated a "4" (median value) across all the different quality dimensions. On the other hand, Dashboard #2 received a median grade of "3" for both "Beauty" and "Intuitivity" dimensions.

We could therefore assume that the individuals who filled out the form had a neutral opinion of the four different dashboards.

Across all four dashboard and based on median values, the quality dimensions that would best describe the visualization project are "Utility" and "Informative". Whereas the quality dimension that received the lowest grade (based on median values) is "Intuitivity" (only Dashboard #3 received a median grade above 3 for "Intuitivity").

Hence, additional work will be required to make the dashboards more intuitive and easier to use.

Median Values	Dashboard 1	Dashboard 2	Dashboard 3	Dashboard 4
Utility	4	4	4	4
Intuitivity	3	3	4	3
Informative	4	4	4	4
Beauty	4	3	4	4

Standard Deviation	Dashboard 1	Dashboard 2	Dashboard 3	Dashboard 4
Utility	0.99	1.06	0.82	0.96
Intuitivity	1.23	1.18	1.02	1.45
Informative	0.8	1.09	1.21	1.19
Beauty	1.25	1.05	1.2	1.54

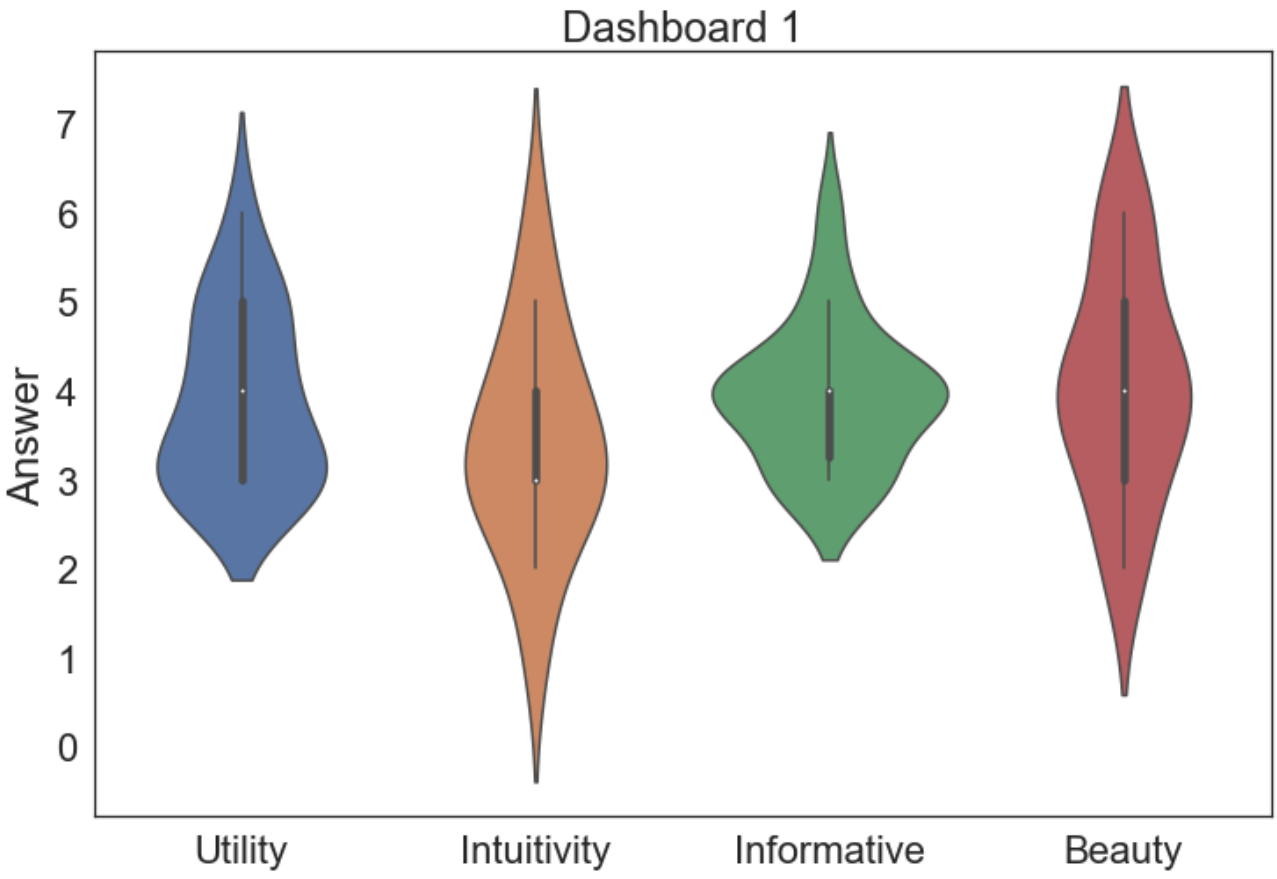


Figure 7

Analysing the distribution functions of the grades for Dashboard #1, it appears that “Intuitivity” and “Beauty” have larger standard deviations compared to “Utility and “Informative” quality dimensions.

“Intuitivity” and “Beauty” quality dimensions appear to have more symmetric distributions, whereas “Utility” and “Beauty” are positively skewed.

All four dimensions received at least one grade equal to “6”. One user rated the “Intuitivity” dimension a “1” (i.e., minimum grade)

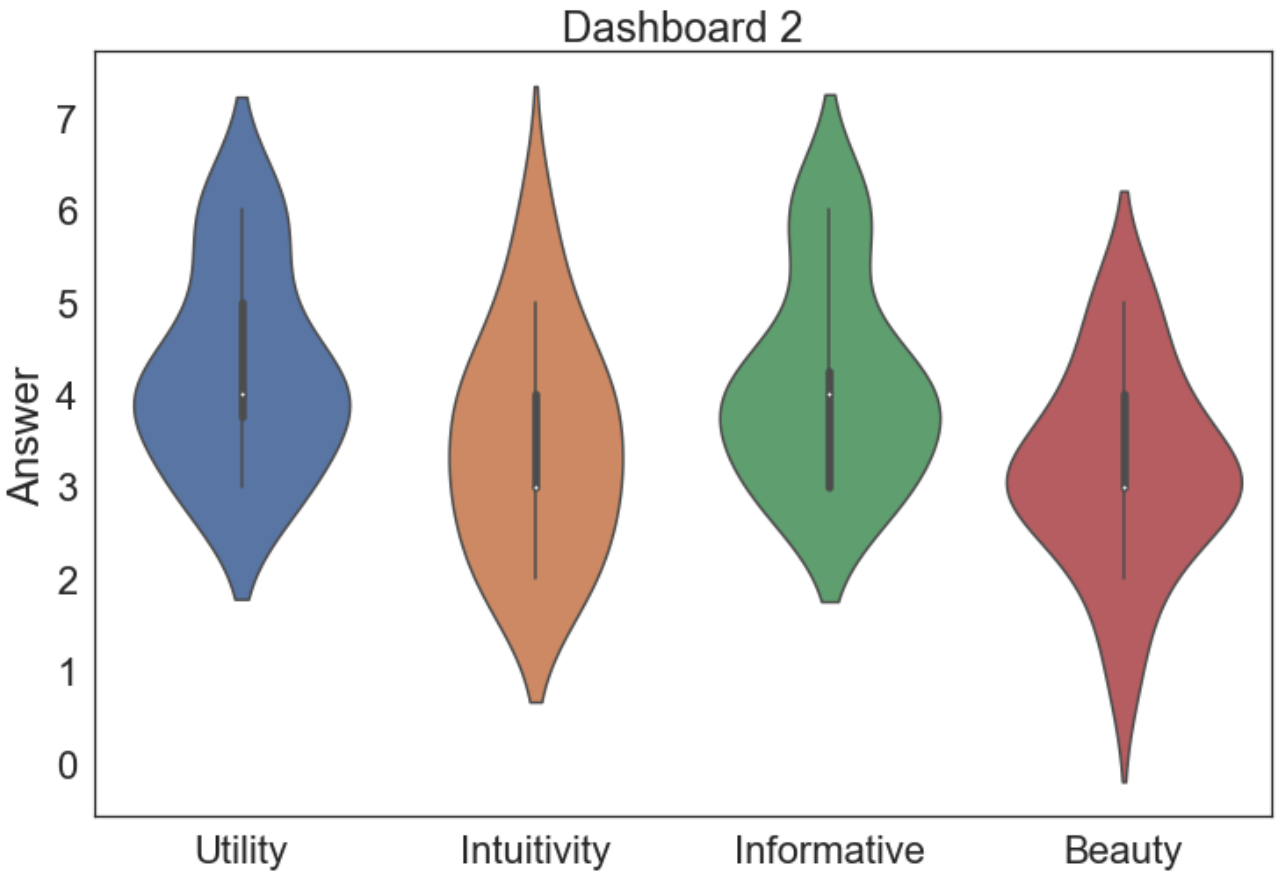


Figure 8

Probability density functions of the quality dimensions of Dashboard #2 have similar values of variability. Standard deviations range between 1.05 for “Beauty” and 1.18 for “Intuitivity”.

In this case, no user rated “Beauty” a 6. One user also assigned the minimum grade of 1 to “Beauty” dimension.

Based on users’ feedbacks, work needs to be done to improve the “quality” dimension of this dashboard.

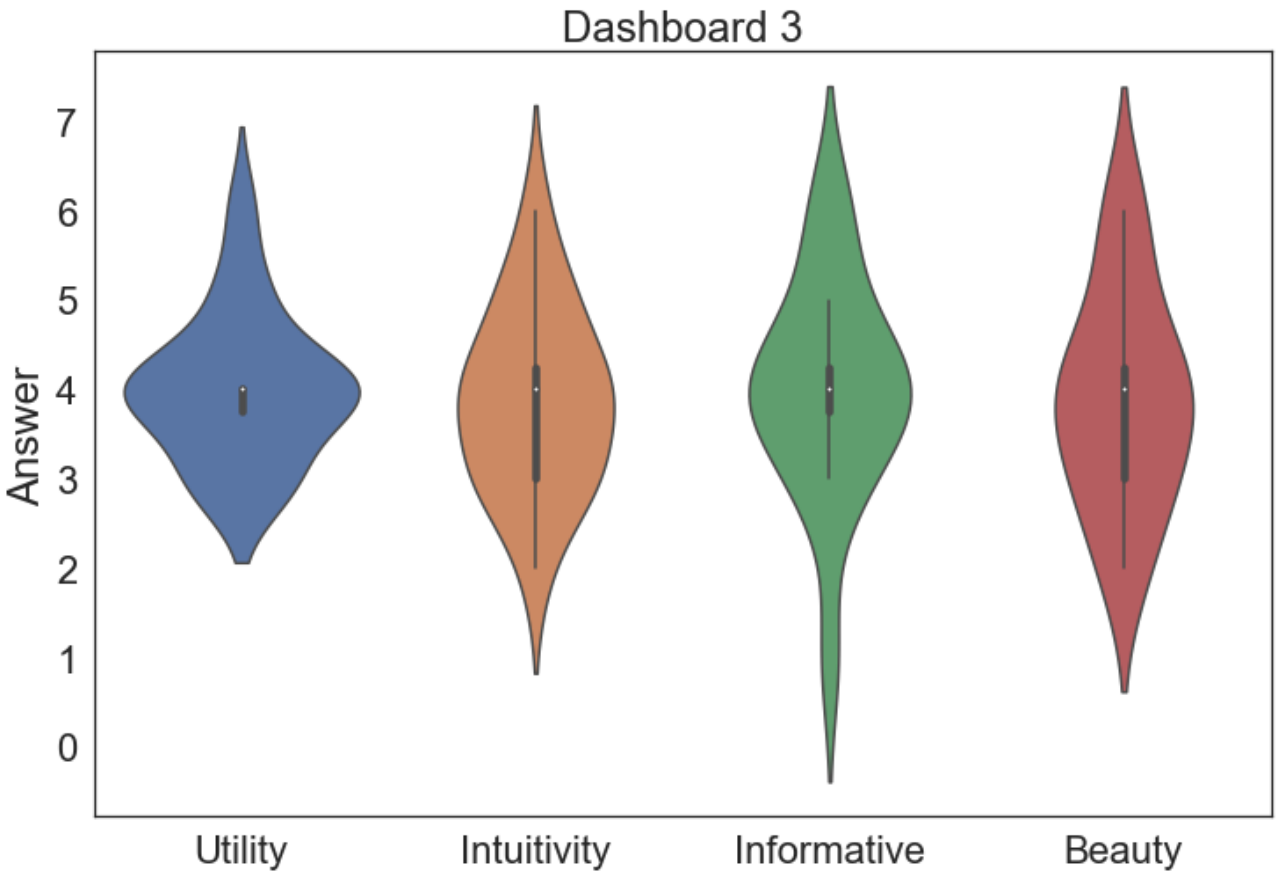


Figure 9

As stated above, the median grade of all four quality dimensions of Dashboard #3 is equal to 4.

“Utility” dimension had the lowest standard deviation in grades (0.82).

The minimum grade assigned by users to the “Utility” dimension is 3.

“Informative” dimension received both the minimum and maximum grades.

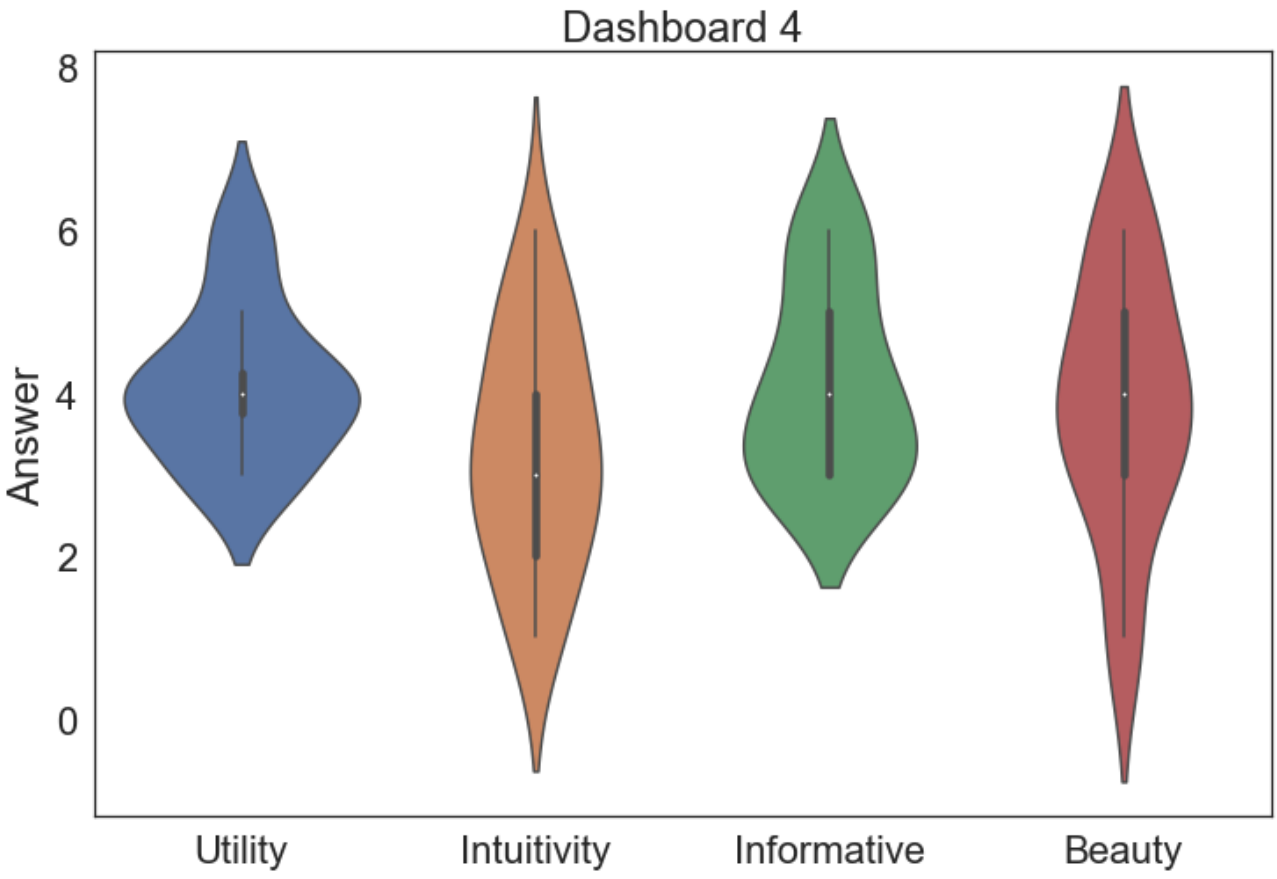


Figure 10

Apart from “Intuitivity” (median score equal to 3), every other dimension received a median grade equal to 4.

Out of the four different quality dimensions of Dashboard #4, “Beauty” has a standard deviation of 1.54, approximately 60% more than the standard deviation of “Utility” (0.96), suggesting greater variability in users’ response regarding the “Beauty” dimension.

“Utility” and “Informative” did not receive a single mark below 3.

All four dimensions received at least one mark equal to 6.

5. Conclusion

The goal was to find an answer to the question "which statistic influences the sentiment about a football match the most?".

In part, the answer can be given because a positive correlation has been noted between the number of goals scored (therefore of an event that creates a show) and the satisfaction of the match, and a slight negative correlation between positive tweets and number of passes. So users are more interested in seeing

games with many verticalizations and less tactics. Furthermore, the affirmation of the president Florentino Perez, creator of the Superlega to challenge higher level teams, was found to be untruthful.

As a final observation, it has been noted that users tend to follow teams that surprise and exceed expectations.
it is a "partial" success as the search is limited to the world tournament. A possible future development to give greater accuracy to the search is to have a greater number of matches available, even of events of less depth to have a more generic perspective.