Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di laurea magistrale in Data Science

ANALISI GARA 6 DELLE FINALS NBA 2021

Progetto a cura di:
Emanuele Artioli
Giacomo De Gobbi
Davide Vercesi

Anno Accademico 2020-2021

Contents

1 Introduzione e Obiettivi

Le finals NBA sono l'atto conclusivo di quello che viene definito il campionato più bello del mondo. Tale evento non viene seguito solamente dai tifosi delle rispettive squadre ma anche da tutti gli appassionati di basket in giro per il mondo. Nonostante le difficoltà di visione legato al fuso orario che rende complicata la fruizione soprattutto in Europa e in Asia, quest'anno gli spettatori medi si sono attestati sui 9 milioni in media, toccando il picco di 16 milioni in gara 6. Quest'ultima gara sarà proprio l'oggetto di analisi in questo progetto, che ha visto confrontarsi i Milwaukee Bucks del due volte MVP Giannis Antetokumbo e i Phoenix Suns del futuro hall of famer Chris Paul.

In particolare, l'obiettivo di questo studio è quello di analizzare come gli utenti Twitter hanno vissuto il match andando a studiare i loro comportamenti nella piattaforma. Come primo task di questo progetto, si è voluto capire qual è l'affluenza di tweet nel corso della serata, partendo dal pre-partita, passando per la partita stessa, per poi concludere con il post-partita. Come secondo task si è condotta sia una sentiment che una emotion analysis curandosi particolarmente del preprocessing del testo per ottenere il massimo risultato da questi algoritmi. Come ultima task si è deciso di studiare attraverso metriche e strumenti della teoria dei grafi, la struttura del network. Particolare attenzione è stata rivolta verso l'individuazione delle community che si sono venute a formare utilizzando come relazione tra i nodi/utenti i concetti di mention e retweet. Infine, utilizzando tali informazioni e integrandole con altri dati provenienti da fonti diverse, si è cercato di raccontare la partita attraverso la creazione di 3 dashboard interattive.

2 Raccolta Dati e Integrazione

La raccolta dati, ovvero la collezione di tweet è stata effettuata tramite libreria *Tweepy*, che gestisce in ambiente Python la API ufficiale di Twitter. Lo storage è avvenuto tramite la collaborazione di Apache Kafka e MongoDB.

La pipeline di questo progetto ha la seguente forma: alle 23:00 del 20 luglio 2021 il programma (un notebook Python) si avvia lanciando 3 funzioni in multi-threading (ovvero in esecuzione parallela) tramite il pacchetto *Threading*:

- Uno StreamListener di *Tweepy* che ascolta tutti i tweet pubblicati in attesa di uno che contenga uno o più hashtag tra quelli indicati, in questo caso: '#NBAFinals', '#FearTheDeer', '#RallyTheValley', '#Suns', '#Bucks'. Se un tweet rispetta questa condizione, esso viene scomposto nei suoi campi di interesse e inviato sotto forma di JSON al producer di Kafka, il quale si assicura che essi non siano duplicati di tweet già inseriti a sistema, e in caso di esito positivo li aggiunge al topic 'sunsbucks0'.

- Un consumer di Kafka che riceve ogni tweet aggiunto al topic dal producer e li invia al database MongoDB. La sua caratteristica di essere non relazionale è fondamentale per questo progetto in quanto i tweet, seppure provenienti dalla stessa fonte e ripuliti di tutti i campi non necessari, non hanno la stessa forma: la maggior parte di essi è una risposta o un retweet, e questa caratteristica non è gestita in modo uniforme: retweet e risposte hanno un campo Retweeted_Status contenente l'intero tweet originale, mentre i tweet principali sono sprovvisti di questo campo. Esso è necessario per gestire correttamente retweet e risposte, dunque va importato da

Tweepy, ma dà errore se cercato nei tweet che ne sono sprovvisti. Tutto ciò la gestione tramite tabelle statiche impossibile.

- Un thread principale che verifica il continuo funzionamento di producer e consumer e li riavvia in caso di crash. Il notebook è stato interrotto alle 9:30 del 21 luglio 2021, e ha collezionato circa 250000 tweet, la maggior parte dei quali durante i minuti effettivi di partita, avvenuta tra le 3:00 e le 5:00 del 21 luglio. La dimensione di questo dataset ha superato di poco i 100MB.

Un'ulteriore fonte di dati, in questo caso asincrona, deriva dallo scraping dati relativi alla telecronaca della partita dalle pagine di ESPN [2] e TheGuardian [10]. La libreria utilizzata per python è *BeautifulSoup*.

Il dataset ricavato da ESPN è una raccolta di eventi di cronaca (azioni rilevanti) di cui è stato collezionato l'istante, la descrizione, e il punteggio in quel momento. Il ruolo della cronaca di TheGuardian è stato permettere l'integrazione tra questi dati e i tweet, infatti il dataset comprende l'informazione sul punteggio per legarlo a ESPN, e quella sul tempo reale (ore e minuti) per legarlo ai tweet. Per poter integrare i tweet, è stato necessario tenere conto dei diversi fusi orari, infatti Tweepy scarica i tweet con fuso orario di Londra, pertanto questi sono stati spostati a +2, mentre TheGuardian calcola dinamicamente l'orario in base alla localizzazione del lettore, infatti non è stato necessario modificare questo campo. Inoltre, TheGuardian include il dato sul quarto di partita in cui ogni azione è avvenuta, che arricchisce ulteriormente l'analisi.

3 Sentiment & Emotion Analysis

E' stata eseguita l'analisi del sentimento sul set di dati utilizzando *VADER* che è uno strumento di analisi del sentimento basato su lessico e regole specificamente adatti allo studio dei testi nei social media. *VADER* fornisce un punteggio composto che assegna un punteggio di polarità al testo complessivo. Questo punteggio varia da -1 a +1. Se il punteggio è compreso tra -0,05 e +0,05, allora è classificato come neutro. Se è maggiore di 0,05, allora è considerato positivo e se è inferiore a -0,05, allora è considerato negativo. Inoltre, si è utilizzato la libreria *NRCLex* per etichettare le parole all'interno di ogni tweet con i corrispondenti affetti emotivi (cioè, la ruota delle emozioni di Plutchik [8] che include rabbia, anticipazione, paura, disgusto, gioia, tristezza, sorpresa e fiducia) basati sul lessico degli affetti del National Research Council Canada (NRC). Prima di poter effettuare tali operazioni il testo di ciascun tweet è stato processato e pulito in modo da ottenere dei risultati più precisi e meno bias. In particolare:

- Conversione dei tweet in minuscolo
- Rimozione Retweet, UserMentions e link
- Rimozione punteggiatura e caratteri speciali
- Rimozione stopwords
- Sostituzione parole 'allungate': una parola allungata è definita come una parola che contiene un carattere ripetuto più di due volte, per esempio, 'Awesoooome'. La sostituzione di queste parole è molto importante poiché il classificatore le tratterà come parole diverse da quelle di partenza abbassandone la frequenza. Tuttavia, ci

sono alcune parole inglesi che contengono caratteri ripetuti, soprattutto consonanti, quindi si è utilizzato il wordnet di *NLTK* per confrontarlo con il lessico inglese.

- Gestione negazioni con antinomie: Uno dei problemi che emergono quando si analizza il sentimento è la gestione della negazione e il suo effetto sulle parole successive. [7] Ad esempio: "I didn't like the movie" e si scartano le stopwords, "I" e "didn't". Così, alla fine, si ottengono i token "like" e "movie", che è il senso opposto a quello del tweet originale. Ci sono diversi modi di gestire la negazione, e ci sono anche molte ricerche in corso su questo; tuttavia, per questo progetto, in particolare, scansioneremo i nostri tweet e li sostituiremo con una antonimia [4] (che otterremo da lemmi in wordnet) del nome, verbo o aggettivo che segue la nostra parola di negazione.

4 Network Analysis

L'ultima analisi del progetto riguarda lo studio del network venutosi a creare dalle interazioni avvenute su Twitter durante la partita. Vengono proposti due tipi di network considerando le due principali relazioni che si possono estrarre dal social: mention e retweet. Le mention servono per taggare altri utenti all'interno di un tweet o di un commento, richiamandone così l'attenzione, mentre nel caso del retweet, un utente B può decidere di retweettare il tweet di un utente A condividendolo all'interno della propria bacheca così come l'utente originale lo ha pubblicato. Per problemi computazionali dati dalle notevoli dimensioni della rete e per evitare di considerare anche i bot, si è deciso di eliminare gli utenti con meno di 100 follower. Vengono perciò calcolate alcune metriche tipiche della teoria dei grafi con particolare attenzione al task della community detection. Le community sono gruppi di nodi che mostrano comportamenti simili all'interno del gruppo e comportamenti dissimili con nodi che appartengono ad altri gruppi. Uno degli approcci più usati per la community detection è l'algoritmo di **Girvan-Newman** [5]

Pseudo Code:

- Step 1. Calcolare edge-betweenness per ogni arco.
- Step 2. Rimuovere l'arco con la betweenness maggiore.
- Step 3. Ricalcolare betweeness.
- Step 4. Ripetere finchè ogni arco è rimosso, o la funzione di modularità è ottimizzata.

Per raggiungere una certa qualità di partizionamento, l'algoritmo di Girvan-Newman utilizza il concetto di modularità [6], che è dato da:

$$Q = \sum_{s=1}^{m} [\frac{l_s}{L} - (\frac{d_s}{2L})^2]$$

Dove la somma è sui m moduli della partizione, l_s è il numero di archi dentro il modulo s, L è il numero totale degli archi dentro la rete e d_s è il grado totale dei nodi nel modulo s. Più il valore della modularità è alto, più strette e simili saranno le connessioni dei nodi all'interno di una comunità rispetto agli altri nodi fuori da tale comunità. Sia la network analysis che le relative visualizzazioni sono state realizzate trami il software open-source **Gephi**.

MENTIONS	RETWEET
Nodes:70370	Nodes: 60931
Edges:92539	Edges: 79631
Average Degree: 1,315	Average Degree: 1,307
Average Weighted Degree: 1,546	Average Weighted Degree: 1,537
Diameter: 15	Diameter: 10
Average Path length: 5.72	Average Path length: 2.12
Modularity: 0,761	Modularity: 0,767
Number of Communities: 3659	Number of Communities: 3275

Table 1:

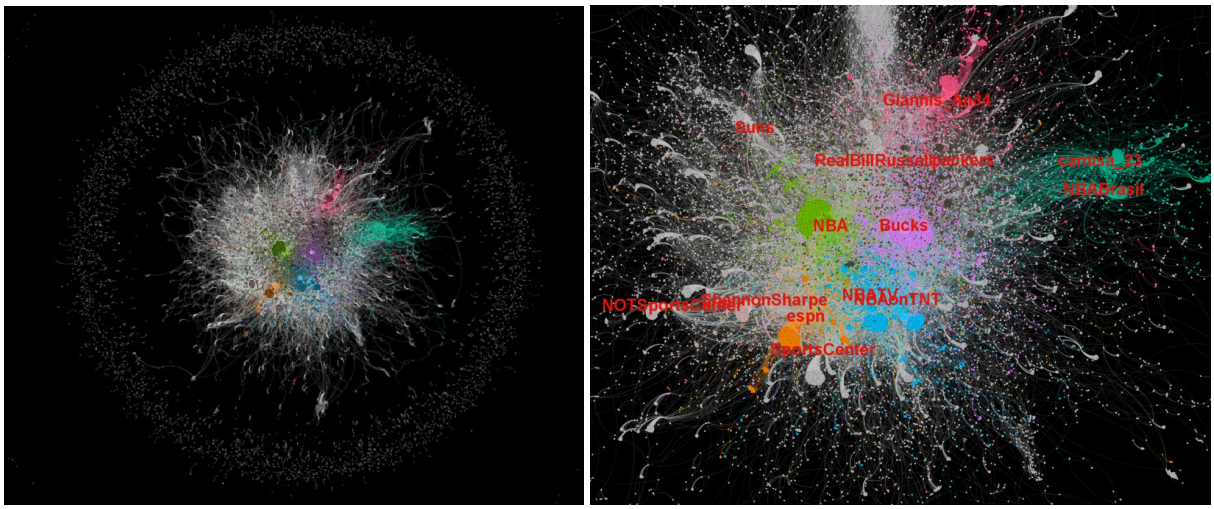


Figure 1: Mentions

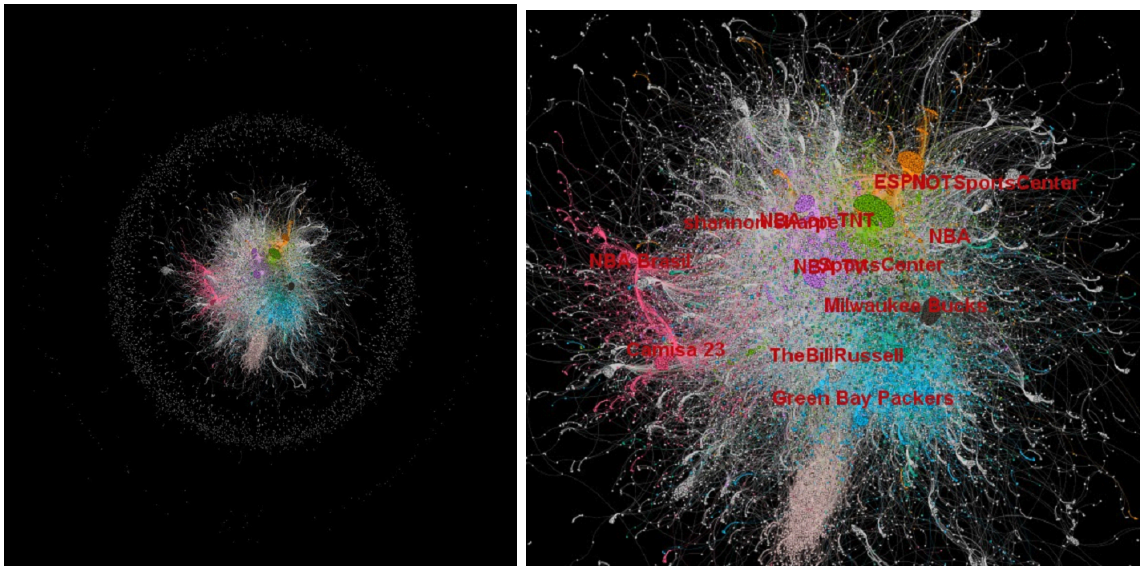
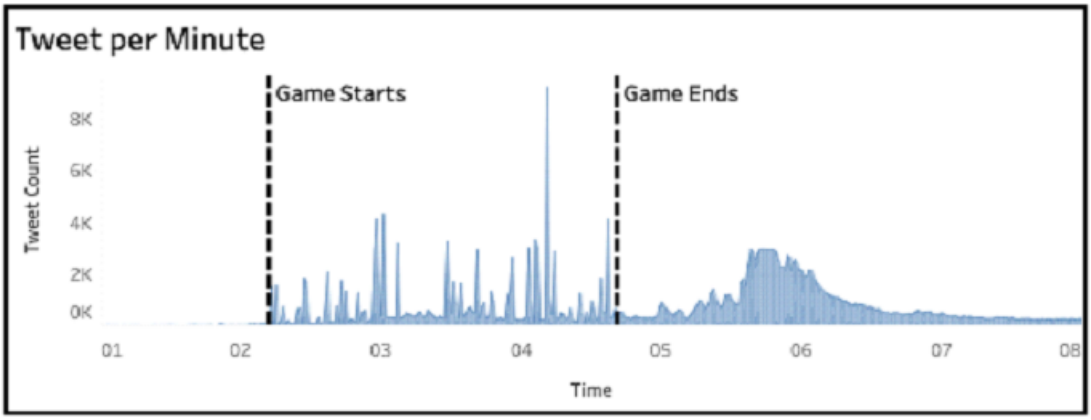


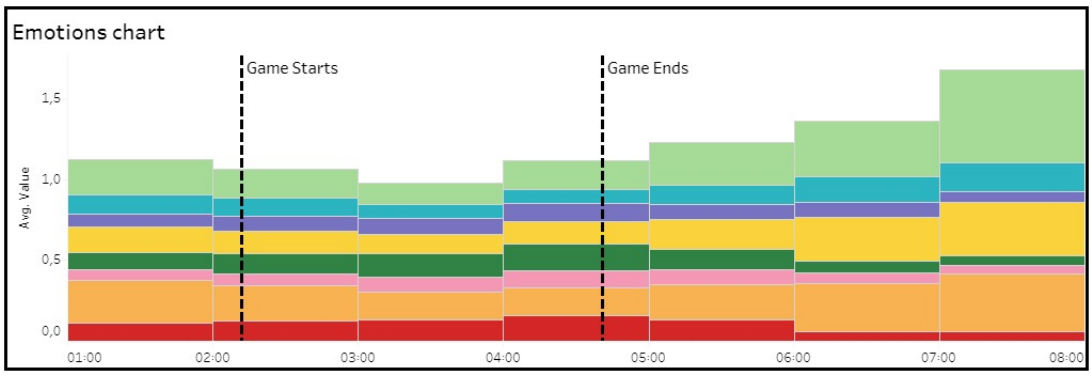
Figure 2: Retweet

5 Visualizzazioni

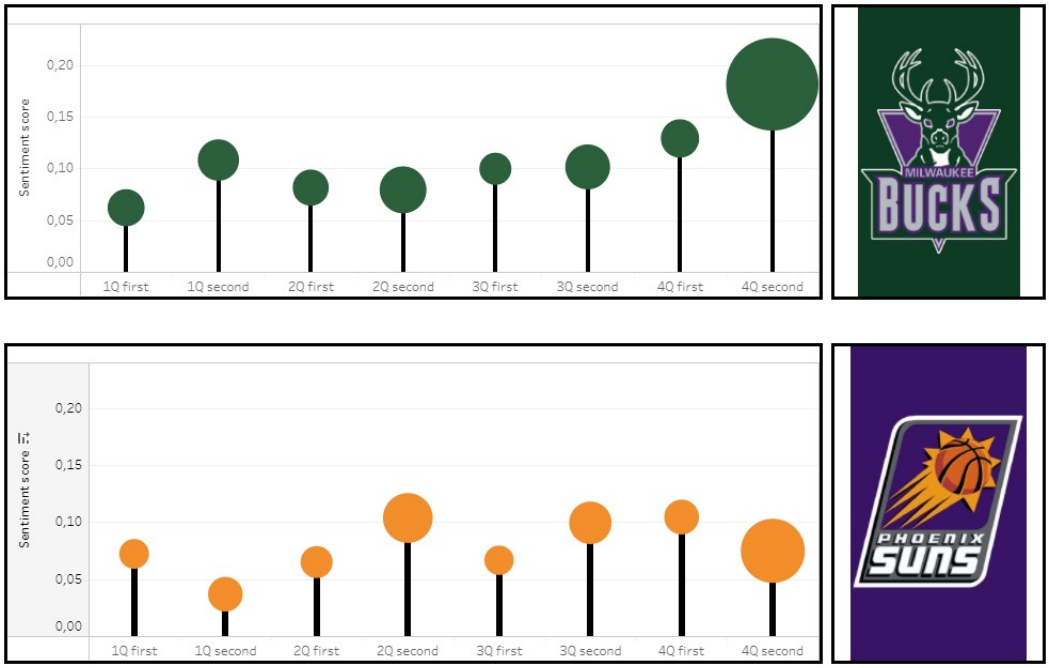
[9]

La prima infografica è composta da due visualizzazioni che vanno lette in maniera speculare. In entrambe nell'asse x viene plottato il tempo (suddiviso in minuti nella prima e in ore nella seconda), mentre l'asse y differisce per le due: nel primo caso si è plottato il numero di tweet raccolti con la possibilità di posizionarsi con il cursore del mouse sopra un qualsiasi punto per vedere il punteggio della partita il quarto e l'azione saliente che è avvenuta in quel momento. Mentre la seconda restituisce i punteggi medi delle otto emozioni calcolate durante la emotion analysis e i colori dei riquadri sono stati scelti in accordo alla rappresentazione della ruota di Plutchik. In particolare: verde chiaro per Trust, celeste per Surprise, blu scuro per Sadness, giallo per Joy, verde scuro per Fear, rosa per Disgust, arancione per Anticipation ed infine rosso per Anger.

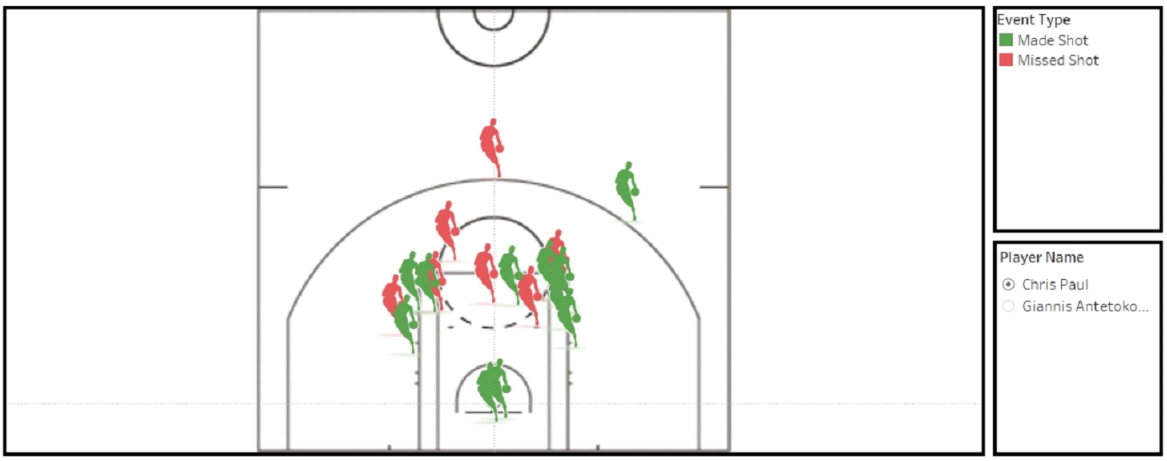




La seconda infografica è composta da due lollipop chart rispettivamente per le due squadre in campo, in cui si è plottato il punteggio derivante dalla sentiment analysis calcolata durante il periodo di gioco. Inoltre, si è deciso di rendere variabile la grandezza dei cerchi rispetto al numero di tweet prodotti nella frazione di tempo in esame (più il cerchio è grande, più si è twittato).



La terza infografica è una rappresentazione interattiva dei tiri tentati dai due giocatori più importanti delle due squadre, Chris Paul [1] e Giannis Antetokoumbo [3].I dati sono stati ricavati attraverso web scraping dal sito NBA Advanced Stats e comprende diverse informazioni tra cui l'asse x e y per poter plottare su un campo da basket la posizione del tiro effettuato, se il tiro è stato segnato oppure no, se il tiro vale 2 o 3 punti e la "Action Type" che fornisce maggiori dettagli su come si è svolta l'azione. Infine, è stato possibile estrarre il link che conduce ad un file mp4 per poter visionare ogni singola azione dei due giocatori.



6 Test Infografiche

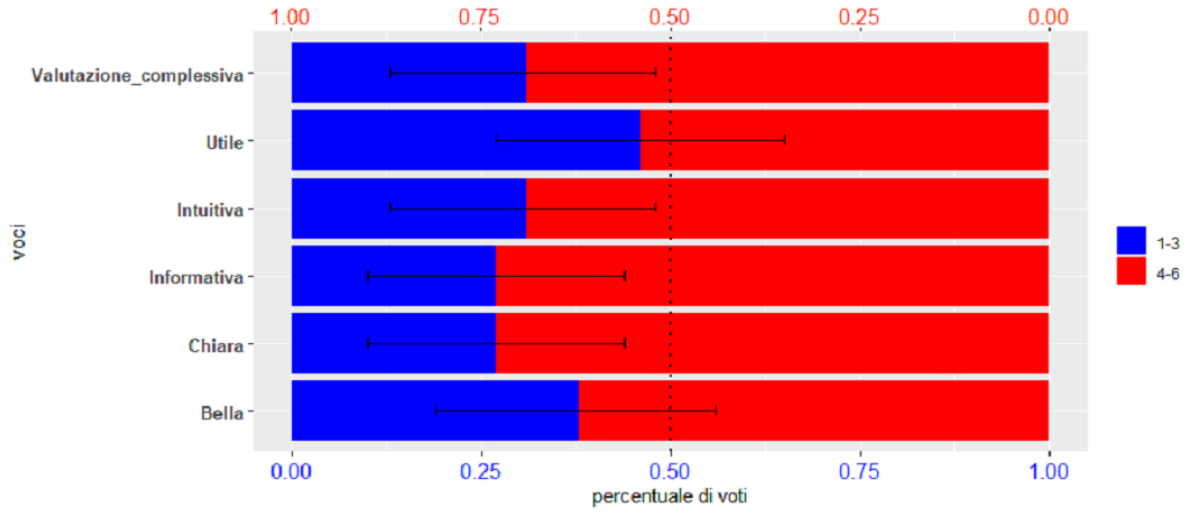
QUESTIONARIO PSICOMETRICO

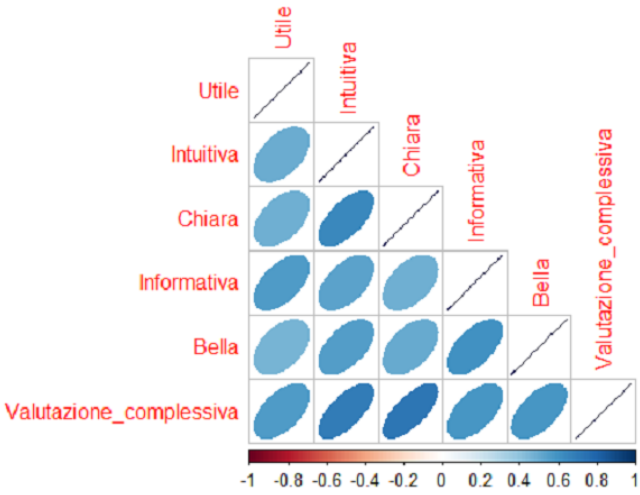
Ad un campione di 24 persone con età media 25 anni, 16 uomini e 8 donne, è stato sottoposto un questionario psicometrico, nel quale per ogni infografica sono state poste le seguenti domande:

- Quanto ritieni l'infografica utile?
- Quanto ritieni l'infografica intuitiva?
- Quanto ritieni l'infografica chiara?
- Quanto ritieni l'infografica informativa?
- Quanto ritieni l'infografica bella?
- Quale valutazione complessiva dai all'infografica?

Per ogni domanda viene chiesto di assegnare un punteggio da 1 a 6, le risposte a queste domande da parte degli utenti sono visualizzate ed analizzate tramite uno stacked barplot e un correlogramma.

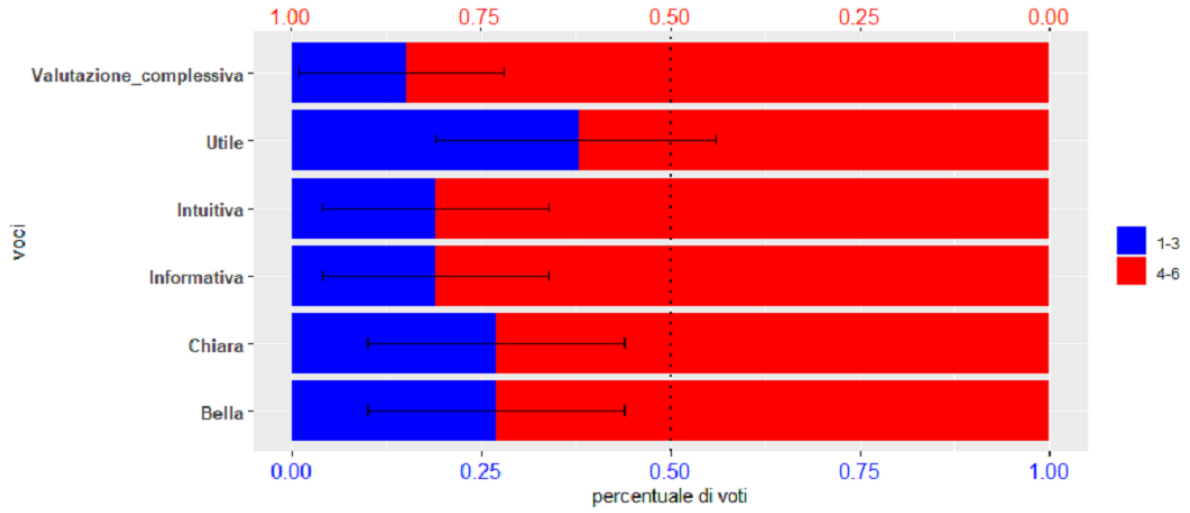
PRIMA INFOGRAFICA

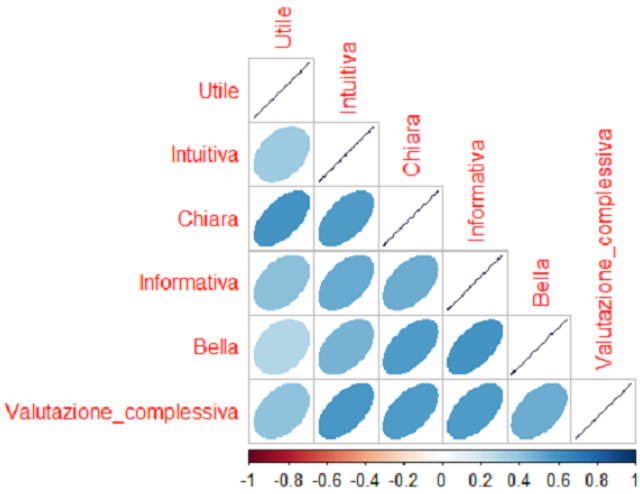




La prima infografica ha riscontrato complessivamente l'apprezzamento del pubblico, poco apprezzata è stata l'utilità dove vi sono parecchie valutazioni negative (nel range 1-3). Si osserva un alto indice di correlazione nelle coppie valutazione complessiva - chiara, valutazione complessiva - intuitiva e intuitiva - chiara, segno che gli utenti hanno apprezzato quest'infografica perchè chiara e intuitiva, e un aspetto influenza l'altro.

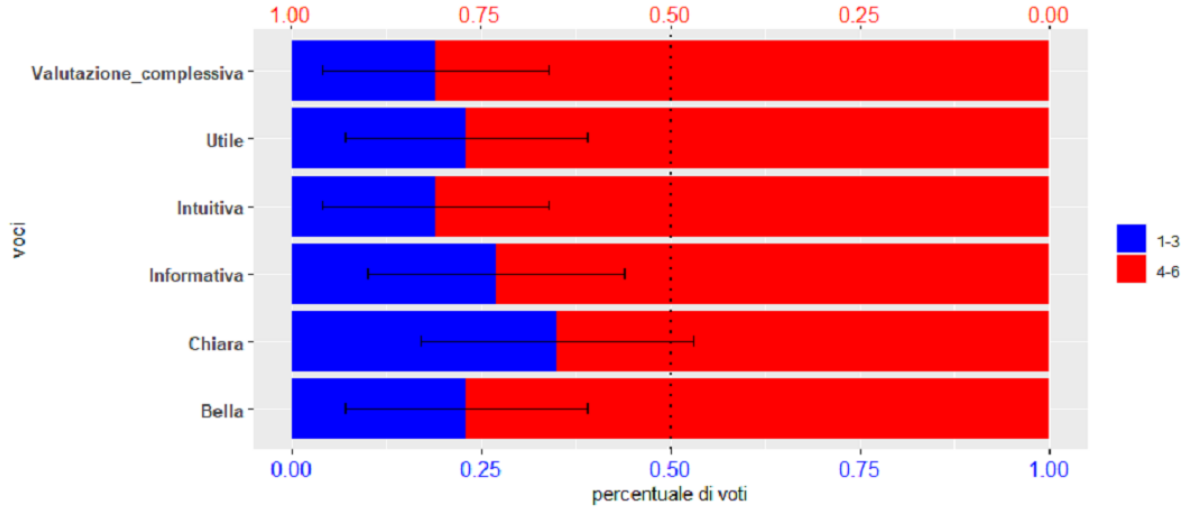
SECONDA INFOGRAFICA

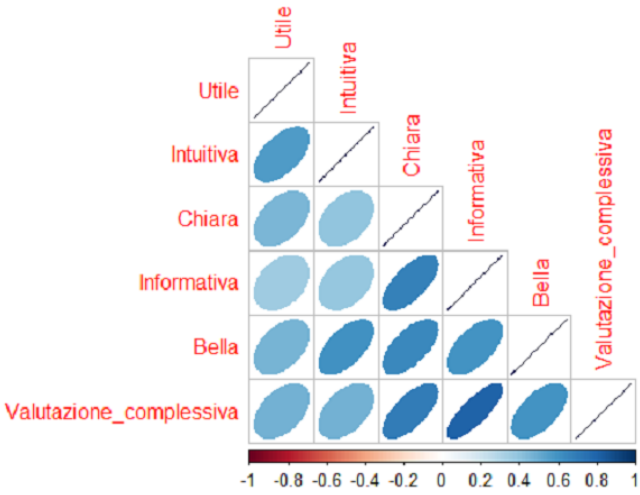




La seconda infografica ha riscontrato l'apprezzamento del pubblico, in quanto vi sono molti giudizi positivi (nel range 4-6). Si osserva correlazione più alta nelle coppie valutazione complessiva - intuitiva, valutazione complessiva - chiara, valutazione complessiva - informativa, informativa - bella e chiara - utile.

TERZA INFOGRAFICA





La terza infografica ha riscontrato più apprezzamento negli utenti, vi sono un alto numero di valutazioni positive (nel range 4-6) e si osserva un alto indice di correlazione nella coppie valutazione complessiva - informativa, valutazione complessiva - chiara e chiara - informativa segno che nel complesso l'infografica è piaciuta perchè esplicativa e ordinata.

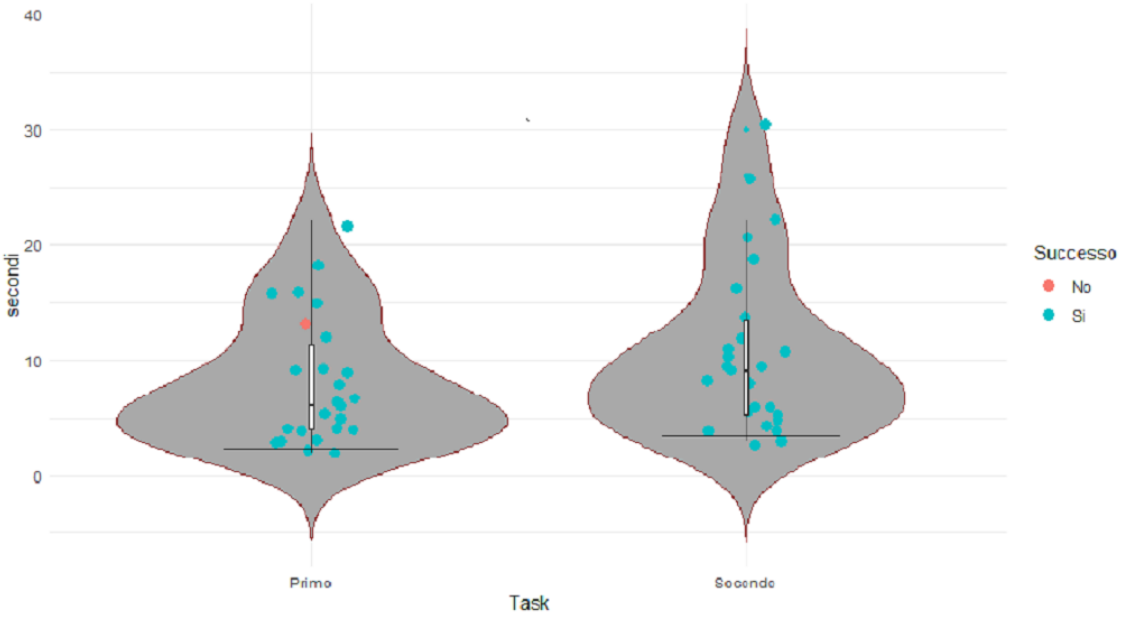
7 User Test

Al medesimo campione, a cui è stato richiesto di compilare il questionario, sono state poste delle domande (due per ciascuna infografica), le cui risposte potevano essere ricavate guardando l'infografica, ed è stato tenuto conto del tempo impiegato a rispondere. Questo è stato fatto al fine di comprendere se le informazioni che si volevano trasmettere tramite l'infografica, fossero percepite anche dagli utenti.

PRIMA INFOGRAFICA

Per la prima visualizzazione è stato chiesto di svolgere i seguenti task:

- In quale periodo è stato maggiore il coinvolgimento emotivo degli utenti?
- In quale periodo si è raggiunto il picco di tweet estratti?

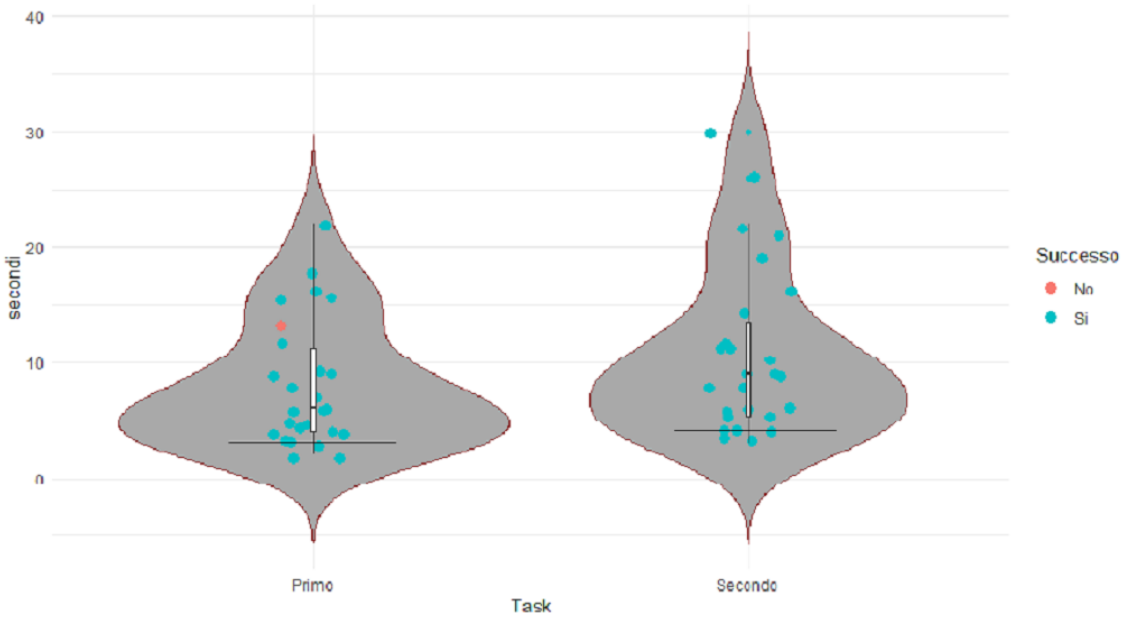


La grande maggioranza degli utenti impiega circa 5-6 secondi per rispondere. Si nota una differenza tra i due task, per il primo i tempi di risposta sono mediamente più rapidi con un massimo di circa 20 secondi, mentre per il secondo il massimo è 30 secondi. Si registra un solo errore nel primo task.

SECONDA INFOGRAFICA

Per la seconda visualizzazione è stato chiesto di svolgere i seguenti task:

- Quale squadra ha avuto in media un punteggio di sentiment maggiore?
- Gli utenti hanno twittato di più i Suns o i Bucks?



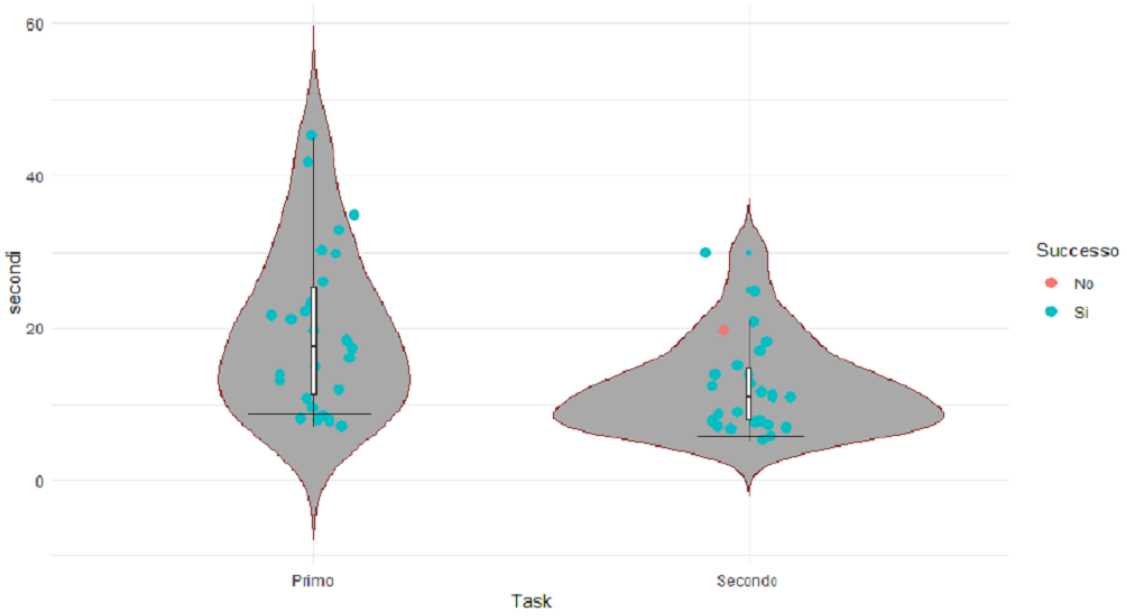
La grande maggioranza degli utenti impiega circa 5-6 secondi per rispondere. Non vi è una differenza significativa nei tempi di risposta tra i due compiti, a parte il secondo che presenta alcuni outliers corrispondenti a tempi di risposta più lenti. Vi è solo una

risposta errata nel primo task.

TERZA INFOGRAFICA

Per la terza visualizzazione è stato chiesto di svolgere i seguenti task:

- Quale giocatore ha preferito tirare di più sotto canestro?
- Quale giocatore ha realizzato più tiri da tre?



I due task risultano essere diversi, nel primo i tempi di risposta sono distribuiti in modo più omogeno in un range che va da 7-8 secondi a circa 30 secondi, con alcuni outliers sui 40 secondi; mentre nel secondo si nota una concentrazione di risposte date in un range che va da 5-6 secondi a meno di 20, con alcuni outliers sui 30 secondi . Quindi la seconda informazione risulta essere più veloce da ricercare per gli utenti, questo potrebbe essere dato dal fatto che gli utenti sono agevolati dal compito precedente in quanto questo permette di prendere già confidenza con l'infografica interattiva, che si traduce in minor tempo di risposta per la seconda domanda. Vi è solo una risposta sbagliata per il secondo task.

ANALISI SUCCESSO/INSUCCESSO COMPLETAMENTO TASK

In conclusione dell'User Test si può vedere come su tutte e tre l'infografiche si è registrato un tasso di errore molto contenuto (pari al 2%). Pertanto, si può ritenere che le tre infografiche riescono a trasmettere in maniera efficace il messaggio che si voleva comunicare agli utenti, essendo inferiore alla soglia di tolleranza fissata arbitrariamente del 10%.

8 Conclusioni

In conclusione possiamo affermare che il progetto è stato un successo e ogni aspetto della ricerca ha fornito informazioni rilevanti. Per quanto riguarda il primo task, possiamo certamente notare dalla prima infografica l'interesse contenuto per la partita. Prima del fischio iniziale i tweet raccolti sono pochissimi e questo conferma varie voci in ambito sportivo che la partita non fosse particolarmente attesa. Infatti, prima dell'inizio dei playoff, le squadre più accreditate per arrivare alle finals erano i Los Angeles Lakers e i Brooklyn Nets. La partita poi presenta vari picchi dimostrando un'attenzione da parte del pubblico non costante con il picco più alto durante l'ultimo quarto di gioco con la partita ancora punto a punto. Sorprendente è ciò che si verifica dalle 5:30 fino alle 6.30 (orario italiano). Infatti, in quest'ora il numero medio di tweet è sempre sopra i 2000, cosa che non accade durante la partita. Una possibile spiegazione è che gli appassionati di basket europei appena svegliati e prima di andare al lavoro, si siano informati subito del risultato della partita ed abbiamo usato il social per guardare highlights, leggere opinioni ed a loro volta scrivere la loro. Per quanto riguarda la sentiment analysis, il risultato non era aspettato dato che il valore durante la partita è quasi sempre più alto per i Bucks che per i Suns, nonostante Giannis Antentokounmpo sia considerato da molti uno dei giocatori più odiati nella lega. Per la network analysis i risultati erano abbastanza prevedibili. Sia per le mention che per i retweet il numero di community è circa 3000 e quelle maggiori si concentrano tra i maggiori canali ufficiali della NBA. I possibili sviluppi futuri possono coinvolgere tutte e tre le task. Per la prima si potrebbe superare il limite di tweet scaricabili per minuto dall'API di Twitter. Per le analisi della sentiment e della emotion sarebbe interessante considerare non solo i tweet di lingua inglese ma anche di altre lingue. Infine, la network analysis potrebbe essere studiata in maniera dinamica collezionando i tweet delle squadre per ogni partita di regular season e playoff e osservando se la rete durante l'anno sportivo evolva notevolmente.

9 Divisione Lavoro

- Raccolta Dati: Emanuele Artioli, Davide Vercesi
- Integraione dati: Emanuele Artioli
- Sentiment & emotion analysis: Giacomo De Gobbi
- Network Analysis: Giacomo De Gobbi, Davide Vercesi
- Prima dashboard: Emanuele Artioli
- Seconda dashboard: Davide Vercesi
- Terza dashboard: Giacomo De Gobbi

Bibliografia e Sitografia

[1] *Chris Paul Stats*. URL: https://www.nba.com/stats/events/?ContextMeasure=FGA&EndPeriod=0&EndRange=28800&GameID=0042000406&PlayerID=101108&RangeType=0&Season=2020-21&SeasonType=Playoffs&StartPeriod=0&StartRange=0&TeamID=1610612756&flag=3&sct=plot&section=game.

[2] *ESPN*. URL: https://www.espn.com/nba/matchup?gameId=401344140.

[3] *Giannis Antetokounmpo Stats*. URL: https://www.nba.com/stats/events/?ContextMeasure=FGA&EndPeriod=0&EndRange=28800&GameID=0042000406&PlayerID=203507&RangeType=0&Season=2020-21&SeasonType=Playoffs&StartPeriod=0&StartRange=0&TeamID=1610612749&flag=3&sct=plot&section=game.

[4] Raffaella Bernardi Laura Aina and Raquel Fernández. "Negated Adjectives and Antonyms in Distributional Semantics: not similar?" In: *Italian Journal of Computational Linguistics, 5-1* (2019).

[5] Girvan M. and Newman M. E. J. "Community structure in social and biological networks". In: *Proc Natl. Acad. Sci. USA 99, 7821-7826* (2002).

[6] M.E.J. Newman. "Modularity and community structure in networks". In: *PNAS June 6, 103 (23) 8577-8582* (2006).

[7] Umar Farooq; Hasan Mansoor; Antoine Nongaillard; Yacine Ouzrout; Muhammad Abdul Qadir. "Negation Handling in Sentiment Analysis at Sentence Level". In: *JCP Vol.12(5): 470-478 ISSN: 1796-203X* (2017).

[8] Henry Kellerman Robert Plutchik. *Theories of Emotion, Academic Press (1980)*.

[9] *tableau*. URL: https://public.tableau.com/app/profile/davide7152/viz/DATAVIZ_DEF/Storia1.

[10] *The Guardian*. URL: https://www.theguardian.com/sport/live/2021/jul/20/nba-finals-2021-game-6-phoenix-suns-v-milwaukee-bucks-live?page=with:block-60f7872f8f08bb57f5f7b268.