FINALE DI CHAMPIONS LEAGUE 2019/20: COME GLI SPETTATORI UTILIZZANO TWITTER PRIMA, DURANTE E DOPO LA PARTITA

Alessandro Fossati, Giorgio Nardi, Luca Pretini

Università degli studi Milano Bicocca, cdLM Data Science

1 Abstract

La finale di Champions League è uno degli eventi sportivi più seguiti al mondo, è quella partita che non solo i tifosi delle squadre in campo seguono, ma é un evento che coinvolge con passione tutti gli amanti del gioco del calcio. In riferimento a questa partita, la stima degli spettatori in tutto il mondo si aggira intorno a 5 milioni e 800 mila.
Quest'anno le squadre in campo rispondono al nome di Bayern Monaco e Paris Saint Germain, rispettivamente squadre vincitrici di coppa e campionato nazionale tedesco e francese per la stagione 2019/2020, entrambe dunque con la possibilità di conquistare la tripletta di trofei in caso di vittoria della partita.
All'interno di questo progetto si andrà ad analizzare come gli utenti Twitter hanno espresso la loro opinione riguardo a determinati episodi della partita, ma non solo, in quanto si andrà ad analizzare anche la frequenza di tweet degli utenti stessi durante il pre-partita ed il post-partita, in generale ed in riferimento ad alcuni giocatori in particolare.

2 Introduzione

Valutare come gli spettatori interagiscono durante la partita o come reagiscono a determinati eventi, può aiutare a capire come si sta svolgendo il match in corso. Come primo obiettivo di questo progetto, si vuole capire quale 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, analizzando anche la lingua in cui i tweet sono stati pubblicati.
Dopo aver fornito uno sguardo generale sull'andamento dei tweet e sulle statistiche generali, si cerca di capire come gli utenti reagiscono a determinati eventi, come ad esempio l'errore sotto porta di Mbappè allo scadere del primo tempo,

al gol di Coman o al palo colpito da Lewandowski, il grande atteso della partita. Si cerca inoltre di capire come gli spettatori hanno reagito alla prestazione non all'altezza di Neymar, altro grande atteso del match, in relazione alle aspettative che c'erano verso di lui prima dell'incontro. Infine, si vuole mostrare in che modo gli utenti italiani hanno utilizzato Twitter per dire la loro riguardo alla partita.
La relazione si suddivide in più sezioni relative alle varie fasi di lavoro del progetto. La prima sezione concerne le modalità di raccolta e storage dei dati da Twitter, dei dati inerenti alle statistiche generali della partita e dei dati inerenti alla cronaca del match. In secondo luogo, vi sarà una sezione dedicata al management, con l'obiettivo di intraprendere una corretta trattazione dei dati ottenuti, al fine di renderli utilizzabili per produrre le infografiche. Successivamente verrà presentata una sezione dedicata alle visualizzazioni dei dati, prodotte al fine di rispondere alle domande di ricerca precedentemente presentate. All'interno di questa sezione verrà introdotto anche il lavoro di Sentiment Analysis e di Word Cloud effettuato sui testi dei tweet.

3 Raccolta e storage dati

Per la fase di raccolta e storage dei dati ci si è affidati ad Apache Kafka, un'architettura producer-consumer che ha permesso di creare una coda condivisa tra la libreria usata per la raccolta dei tweet, ovvero Tweepy, e PyMongo, driver Python per immagazzinare dati in MongoDB. Il nostro producer, Tweepy permette l'interazione con le API Twitter e immette nella coda i tweet raccolti in tempo reale. Da essa PyMongo, ovvero il consumer, raccoglie i tweet che vengono processati e inseriti nel database MongoDB.

Ogni tweet, prima di essere immagazzinato, viene filtrato all'interno del producer, ed alcuni campi superflui vengono scartati giá in fase di cattura. La raccolta dei tweet inizia alle 19:15 italiane circa e termina alle 00:30 italiane, permettendo di raccogliere non solo dati inerenti ai 90 minuti della partita, bensì anche dati relativi al pre-partita, al post-partita e ovviamente all'intervallo del match. Si raccolgono in tutto circa 759000 tweet, che verranno immagazzinati in formato BSON per un totale di circa 200MB. La ricerca dei tweet di interesse avviene per hashtag, e in particolare si cercano tutti i tweet che all'interno del loro testo presentino un hashtag inerente all'evento, in particolare:
#PSGBayern, #UCLFinal, #UCL, #ChampionsLeague, #PSGFCB.
Una seconda fase di acquisizione dati viene effettuata mediante lo scraping delle statistiche inerenti alla partita, come ad esempio la percentuale del possesso palla delle due squadre, oppure il numero di tiri in porta, o il numero di cartellini gialli per squadra, ottenute selezionando il codice HTML della tabella d'interesse dal sito di Football Critic.
Queste statistiche verranno poi inserite in un apposito Dataframe ed utilizzate a scopo esplorativo nella fase di visualizzazione dei dati. Inoltre, viene effettuato scraping dei dati inerenti alla cronaca della partita minuto per minuto, dal sito

di Repubblica. Partendo dal testo della cronaca, utilizzando Pandas si arriva ad ottenere un set di dati in formato CSV non solo contenente la cronaca minuto per minuto, ma anche alcuni campi all'interno dei quali figura chi è il soggetto dell'azione e che tipo di azione è stata svolta. Per effettuare web scraping si utilizza la libreria Python "BeautifulSoup", mentre per creare i file CSV e lavorare sui Dataframe si è utilizzata la libreria "Pandas" di Python.
L'uso di scraping per la raccolta di un testo non cosí corposo come la cronaca di una singola partita puó sembrare esagerato rispetto al task da svolgere, tuttavia si è voluta sviluppare questa parte del progetto pensando anche ad una futura estendibilitá, implementando uno scraping che fosse per lo piú autonomo: il codice riconosce infatti automaticamente le azioni dal contesto (es. palo) e chi le ha svolte, ovvero i giocatori. L'unica azione manuale richiesta è stata l'immissione della lista dei giocatori, permettendo quindi un'ampia riusabilitá del codice.

4 Data Management

Una volta ottenuti ed immagazzinati i dati, inizia un lavoro mirato ad ottenere dei Dataframe utilizzabili per rispondere alle domande di ricerca attraverso le infografiche. Sfruttando PyMongo si ha facilmente accesso alla collezione MongoDB creata durante la raccolta in streaming dei dati.
Sfruttando blocchi di codice Python si decide di sottoporre la collezione a una prima lavorazione, eliminando eventuali duplicati di tweet, inoltre, viene creato un indice sul campo relativo alla data. Questa operazione di indexing permette di effettuare in maniera più rapida ed efficace ricerche sul campo "timestamp", al fine di ottenere un subset della collezione originale per ogni momento della serata: pre-partita, primo tempo, intervallo, secondo tempo e post-partita.

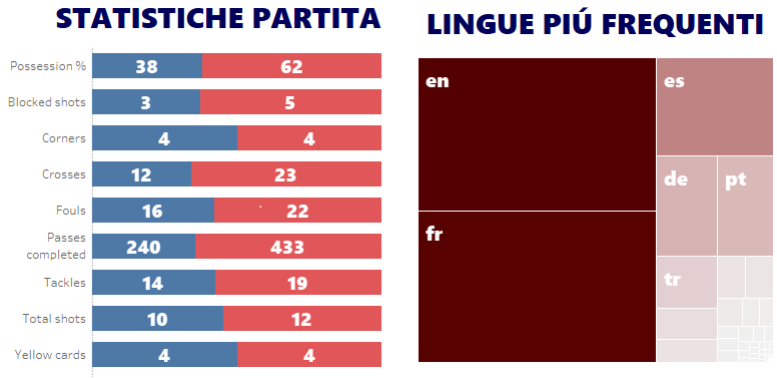
Questa suddivisione risulta utile dal momento che le visualizzazioni riportate successivamente saranno relative a differenti momenti della serata, come anticipato. Viene inoltre aggiunto il campo "minuto_partita" alle collezioni inerenti ai novanta minuti del match, ottenuto tramite manipolazione della data, per facilitare le operazioni di arricchimento che verranno successivamente presentate. Per completezza il campo viene inserito anche nelle altre collezioni, con valore "None". In seguito, mediante l'utilizzo della libreria Pandas di Python, si ottengono dei Dataframe a partire dalle 5 collezioni precedentemente citate, e per ognuna di esse si effettua un'operazione di data cleaning, che permette di rendere più comprensibile e leggero ognuno dei 5 Dataframe. L'operazione di cleaning avviene alleggerendo i campi "Hashtags", "User_mentions" e "Urls" di ogni tweet, eliminando i sottocampi non utili o ridonadanti e passando da una struttura a dizionario ad una a lista di valori dove possibile.
Una volta predisposte all'utilizzo le collezioni create, è necessario effettuare un'operazione di cleaning anche per il Dataset relativo alla cronaca della partita, poiché in alcuni minuti potrebbe essere accaduto più di un evento. Questa caratteristica potrebbe destare problemi nel momento in cui si volesse andare ad

arricchire il Dataframe relativo ai tweet nei 90 minuti di partita con il dataset contenente la cronaca del match, basandosi sul campo comune “minuto_partita”. Senza un’operazione di cleaning si rischierebbe di ottenere duplicati indesiderati nel dataset finale. Si decide dunque di ottenere righe non duplicate in relazione al campo “minuto_partita” del dataset riguardante la cronaca del match, unendo i campi relativi alle righe duplicate in liste Python, garantendo di fatto una semplice accessibilità, e ricomponendo il dataset. A questo punto è possibile effettuare l’enrichment del Dataframe relativo ai 90 minuti di partita con quello inerente alla cronaca. Nel dataset risultante ogni tweet ha nel campo ”testo” la cronaca riferita al minuto di partita in cui lo stesso tweet è stato pubblicato. Per i tweet relativi a minuti in cui sono avvenuti più episodi di cronaca vengono riportati nella variabile ”testo” tutti gli eventi in una singola lista.
Questo set di dati sarà poi concatenato al Dataframe relativo all’intervallo, per poter sviluppare in maniera completa il Linechart che verrà presentato nella sezione dedicata alle visualizzazioni. Ovviamente, le variabili acquisite dal dataset relativo ai tweet della partita in seguito all’enrichment con il file di cronaca (minuto_effettivo, testo, soggetti, eventi) vengono inseriti anche nelle collezioni non riguardanti i 90 minuti di partita, logicamente con valore ‘None’. In seguito la stessa operazione verrá effettuata anche per il dataset completo, relativo a tutti i momenti della serata, il quale verrà utilizzato per valutare l’andamento dei tweet nel tempo totale raccolto.

5 Visualizzazioni

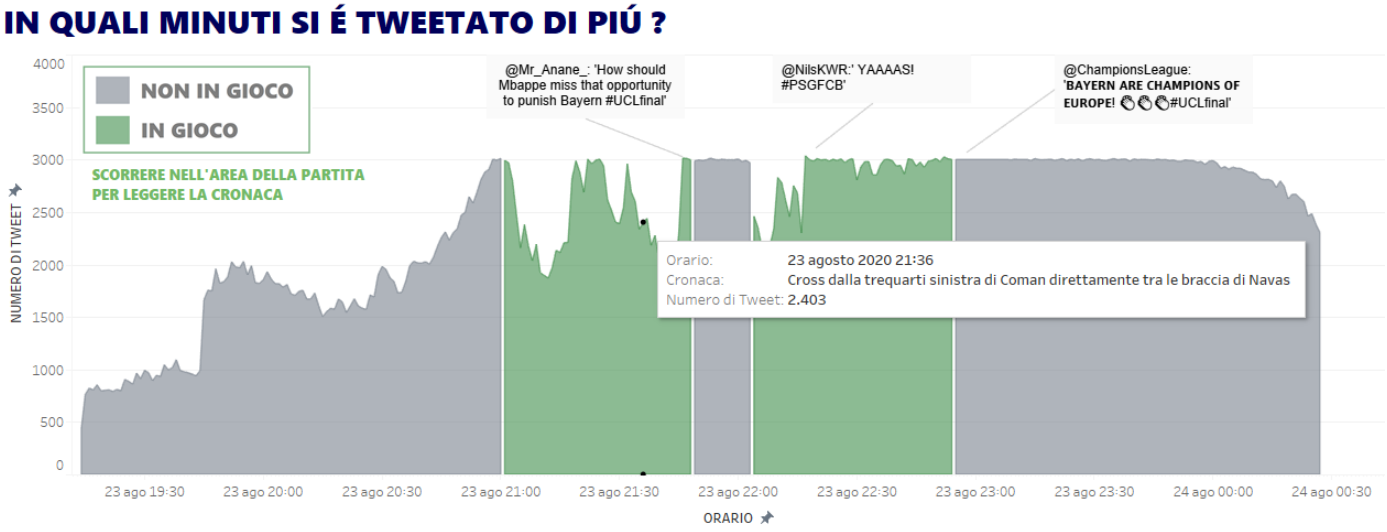
5.1 Infografiche Introduttive

Per uno sguardo introduttivo sulla finale di Champions League, vengono presentate le statistiche della partita, ottenute precedentemente tramite scraping, ed un tree map composto dalle diverse lingue presenti nei tweet.



Per rispondere alla prima domanda di ricerca, ovvero, quale è stata l’affluenza di tweet durante tutta la serata, viene presentata una visualizzazione che mette

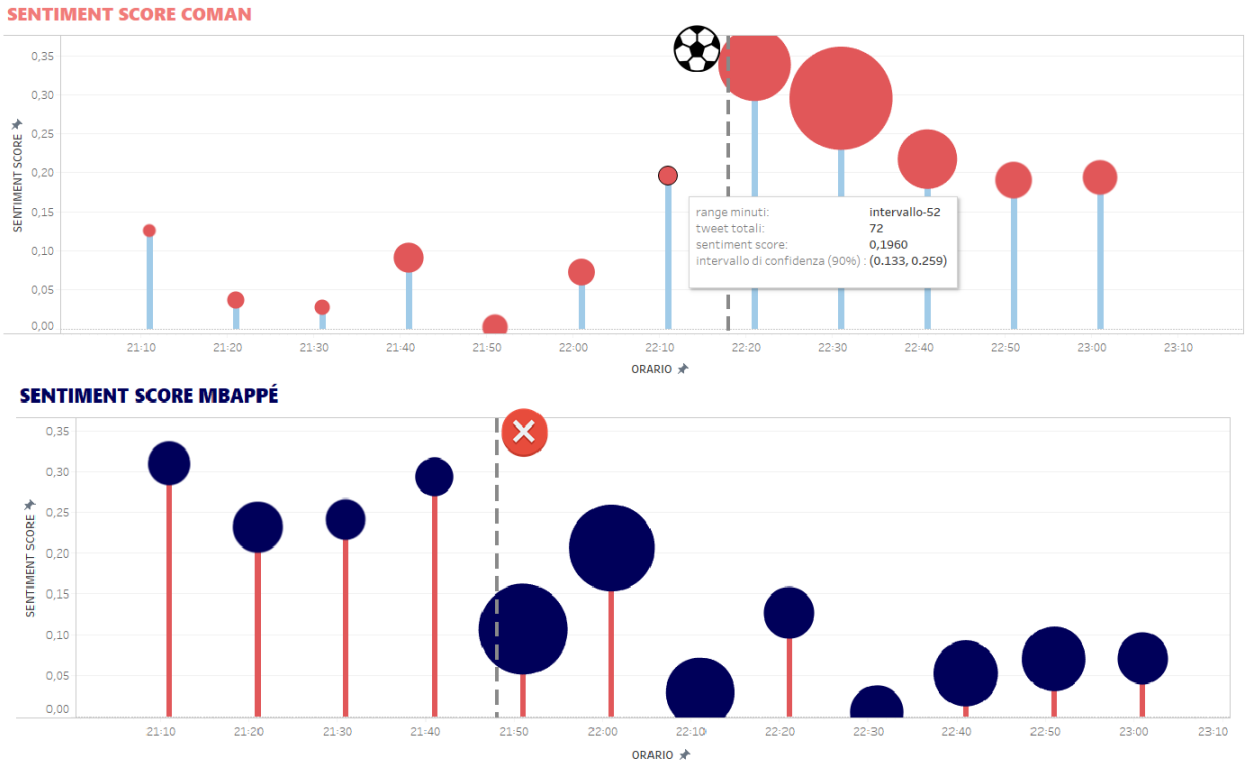
in chiaro come il numero di tweet postati abbia avuto un andamento crescente fino a stabilizzarsi sulla quota massima raggiungibile dall'API di Twitter di 3000 tweet al minuto circa nel post-partita, momento in cui si discute in maniera più animata sul match appena concluso. Posizionandosi sulla parte verde del grafico l'etichetta riporta la cronaca relativa a quel minuto.



5.2 Sentiment Analyis e Visualizzazioni

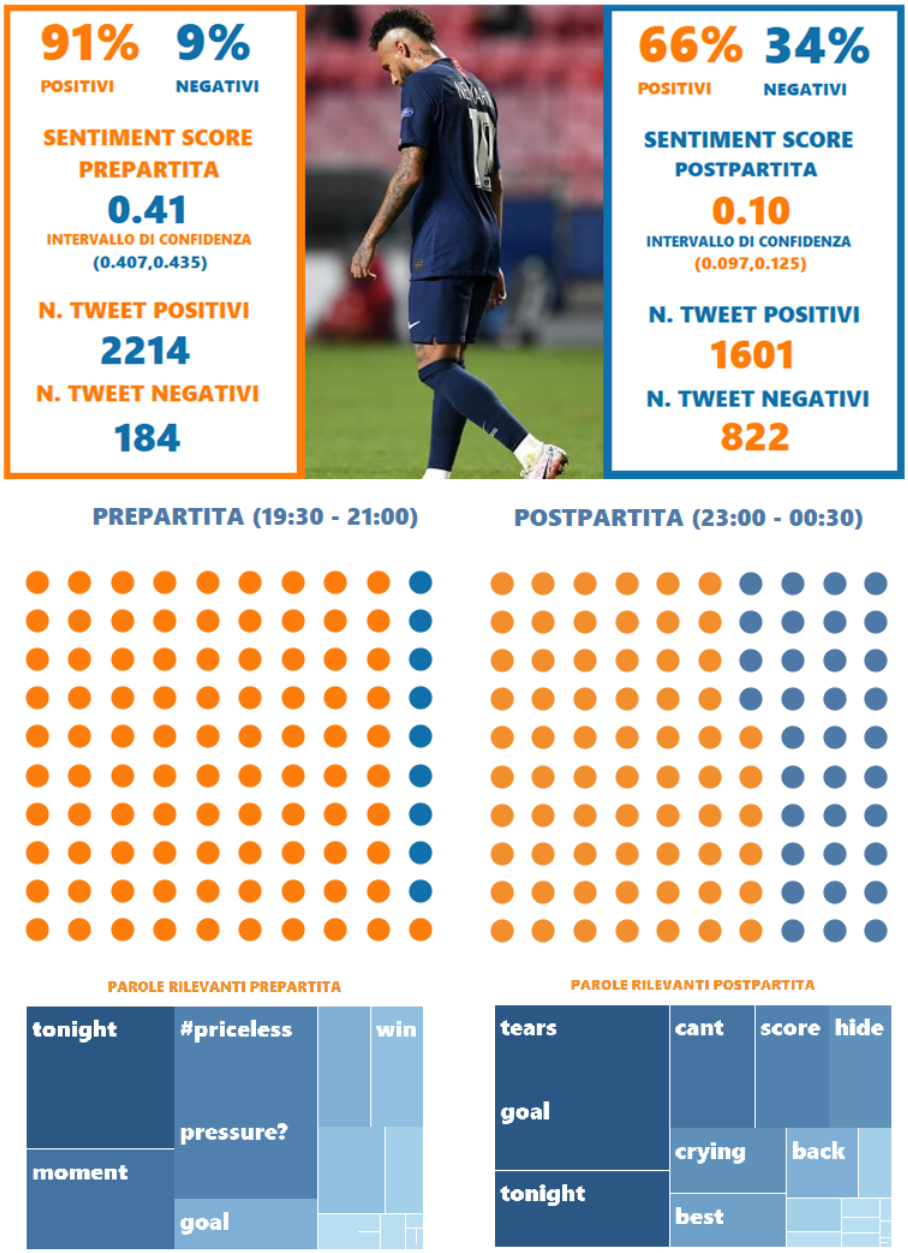
Dopo aver mostrato le visualizzazioni introduttive e dopo aver risposto alla prima domanda di ricerca, si presenta il lavoro svolto inerente alla Sentiment Analysis del testo dei tweet che sono stati raccolti. Per effettuare questo tipo di lavoro, si sfrutta la libreria TextBlob di Python. Mediante questa libreria si possono analizzare tutti i testi dei tweet presi in considerazione, ed in base ad un dizionario di parole calcolarne la polarità, assegnando al testo di ogni singolo tweet un punteggio, denominato *Sentiment Score*, compreso tra -1 e 1. Dove il punteggio assegnato sarà maggiore di zero, il testo del tweet risulterà avere polarità positiva, dove il punteggio relativo al testo del tweet sarà inferiore a zero, quest'ultimo risulterà avere polarità negativa, e dove il punteggio equivarrà a zero, il tweet sarà considerato come neutro. Prima di calcolare il punteggio relativo ad ogni tweet, si effettua una pulizia del testo eliminando caratteri speciali non di interesse, attraverso la regex implementata nella funzione "clean_tweet" di Python, al fine di ottenere una valutazione più veritiera possibile. Una volta presentato il modo in cui si assegnano i punteggi ai testi, si entra nello specifico delle visualizzazioni prodotte.

La prima visualizzazione ha come obiettivo quello di analizzare il cambiamento di polaritá dei tweet in lingua inglese in seguito a due eventi importanti avvenuti nella partita: il gol di Coman al quarto d'ora del secondo tempo e l'errore sotto porta di Mbappè allo scadere del primo tempo. Si sceglie la lingua inglese in quanto essa è la lingua universale ed è la lingua in cui si è tweetato di più, come si evince dal tree map presentato nella precedente sezione della relazione.
È in realtá possibile trovare online delle librerie aggiuntive per TextBlob che implementano dizionari sia in lingua francese che tedesca. Queste sono state testate ma, essendo ritenute sperimentali e avendo restituito risultati non conformi alla libreria ufficiale in inglese, si è scelto di limitare l'analisi alla sola lingua inglese.
Viene dunque suddiviso il tempo relativo al match (comprensivo di intervallo) in sezioni di 10 minuti a partire dal fischio di inizio fino ai primi minuti successivi al fischio finale. Dopodichè vengono selezionati i tweet riguardanti i due giocatori considerati, mediante la ricerca del loro nome all'interno del testo di ogni tweet della collezione, per ognuno degli intervalli di 10 minuti sopra citati.
Si va a mostrare come nel caso di Coman il *Sentiment Score* calcolato nei periodi di tempo successivi a quello in cui si registra il gol sia superiore in media rispetto ai periodi precendenti, mentre, nel caso di Mbappè, si può notare che il *Sentiment Score* calcolato negli intervalli successivi all'errore sia inferiore in media.



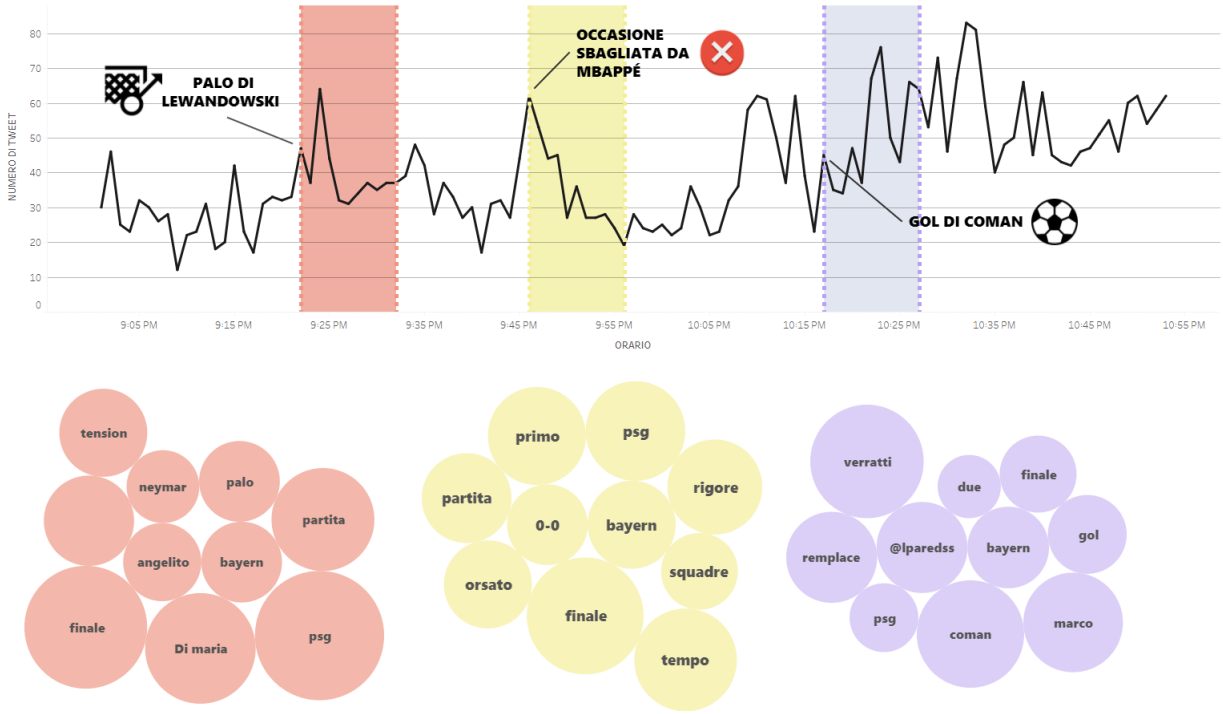
Nel grafico l'area del cerchio del lollipop aumenta al crescere del numero di tweet raccolti nei dieci minuti di riferimento. Si noti come nell'etichetta in figura sia presente l'intervallo di confidenza al 90% per il *Sentiment Score*.

La terza infografica prodotta ha come obiettivo quello di mettere in luce la differenza tra la polaritá dei tweet degli utenti che ruota attorno a Neymar nel pre-partita e quella nel post-partita, differenza dovuta ad una prestazione non all'altezza della situazione da parte del campione brasiliano del PSG. In questo caso si vanno a creare due "Dot Matrix", una per il pre-partita e una per il post-partita, che mirano a rappresentare in maniera chiara la percentuale di tweet positivi, sulla somma di positivi e negativi, inerenti a Neymar, ricercati anche in questo caso matchando tutti i testi dei tweet del pre e post-partita che contenessero il nome del giocatore.
Nella tabella al fianco delle matrici vengono riportati anche i punteggi di *Sentiment Score* del pre-partita e del post-partita, calcolati nella medesima maniera della visualizzazione precedente.
Oltre ad una dimunizione della quota percentuale di tweet classificati come positivi dall'algoritmo di Sentiment Analysis e ad una diminuzione del *Sentiment Score*, si nota una netta differenza nei termini piú frequenti tra i due word cloud, a documentare il fatto che le aspettative nei confronti di Neymar prima del match non sono state rispettate. Anche in questo caso le parole sono estratte dai tweet in lingua inglese, per i motivi precedentemente elencati:



L'ultima infografica mira a mostrare come gli italiani hanno utilizzato Twitter durante i 90 minuti del match. All'interno del Linechart, ottenuto tenendo conto esclusivamente dei tweet in lingua italiana in questo caso, vengono indicati con delle bande colorate i 10 minuti successivi alle tre principali occasioni

del match. Sotto al grafico vengono indicate le 10 parole più utilizzate dagli utenti nei 10 minuti successivi all'evento considerato. Percorrendo il Linechart all'interno di ognuna delle tre bande colorate, il word cloud sottostante si modifica al variare del minuto di riferimento, fornendo un conteggio di ognuna delle 10 parole chiave in relazione allo specifico minuto selezionato.

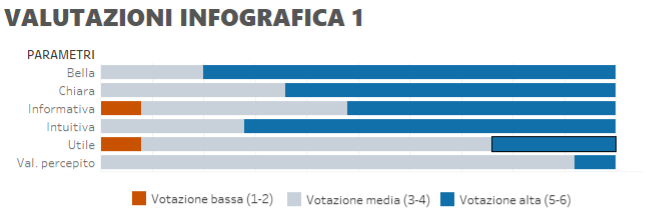


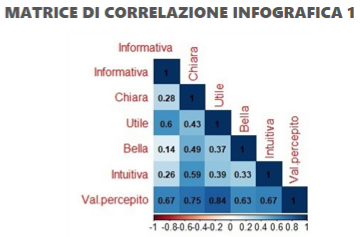
6 Test Infografiche

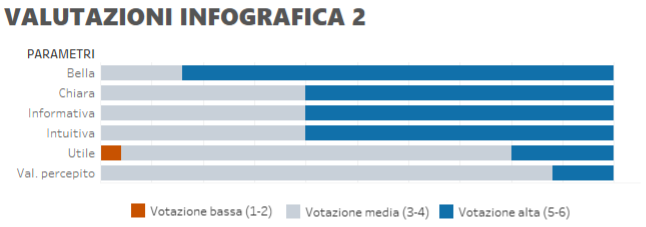
In questa sezione vengono riportati i risultati delle 3 tipologie di test effettuati sulle visualizzazioni dei dati:

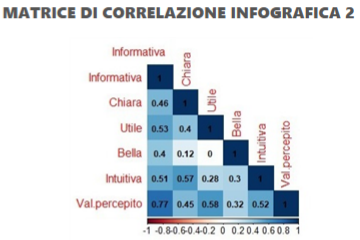
6.1 Test Psicometrico

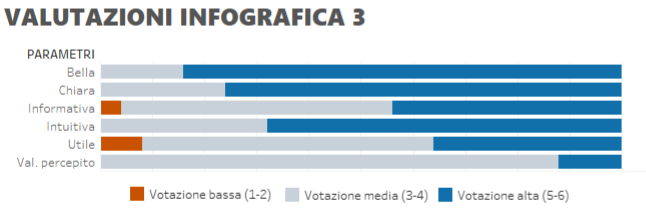
Per questa tipologia di test, vengono coinvolti 25 utenti di età media pari a 30 anni, 15 maschi e 10 femmine. Ad ognuno viene richiesto di assegnare un punteggio da 1 a 6 in quanto a bellezza, chiarezza, informatività, intuitività e utilità ad ognuna delle 4 infografiche. Sono anche state prodotte matrici di correlazione relative ai diversi paramentri di giudizio.

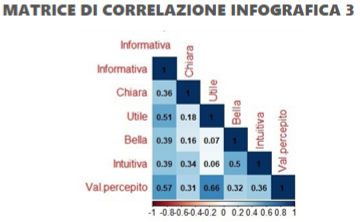


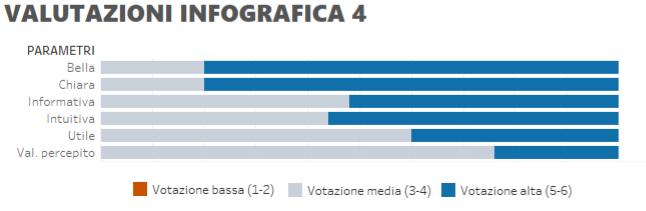


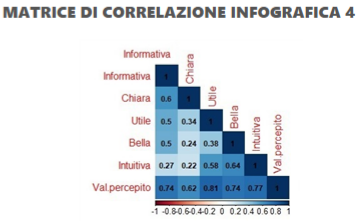






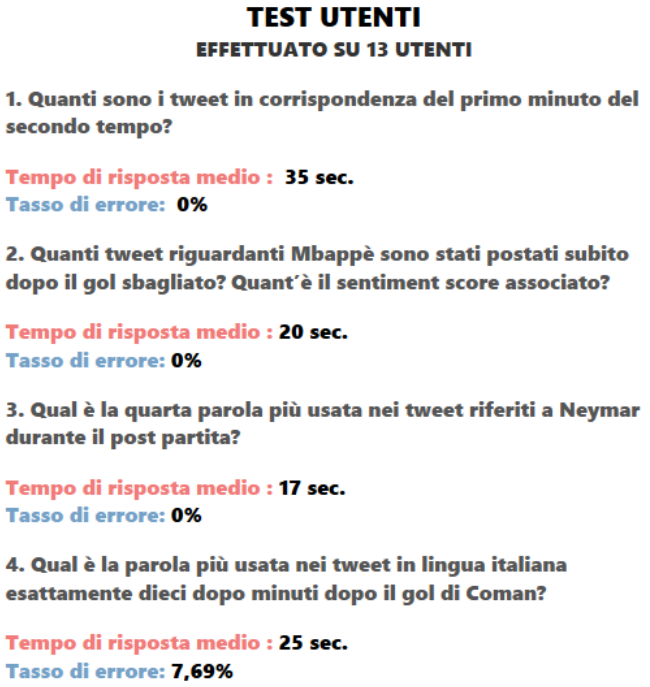






6.2 User Test

Per questa seconda tipologia di test, vengono coinvolti 13 utenti di età media pari a 28 anni, 8 maschi e 5 femmine. Ad ognuno di essi viene richiesto di rispondere ad una domanda per ciascuna infografica. Si riportano il tempo medio di risposta ed il tasso di errore per ogni domanda:



6.3 Test Euristico

Per questa tipologia di test vengono coinvolti 7 utenti. Dal test emerge una differenza nella intuitivitá delle infografiche da parte degli utenti di sesso maschile e femminile, dato il contesto calcistico.
La prima infografica é stata in parte criticata per l'abbondanza di informazioni, nonostante gli utenti abbiano sempre evidenziato semplicitá nel comprenderne i singoli componenti. Il concetto di *Sentiment Score* é risultato stimolante per gli utenti, ma non sempre si é rivelato di immediata comprensione. Anche per la quarta infografica é stata apprezzata l'interattivitá a discapito di una maggior lentezza nella comprensione del suo funzionamento.
Alcuni utenti hanno tentennato nel rispondere alla terza domanda dell'user test, in quanto hanno riscontrato difficoltá nel distinguere la differenza tra le aree dei quadrati relativi a frequenze simili di parole.

7 Conclusioni e Sviluppi Futuri

Si può concludere che il progetto è riuscito nel tentativo di dare risposte soddisfacenti alle domande di ricerca prefissate. Come da aspettativa, in seguito ad un episodio importante nel corso della partita il numero di tweet postati aumenta. In aggiunta, si può constatare come nel caso dei singoli giocatori presi in analisi, la polarità dei tweet si muova di conseguenza rispetto a un gol sbagliato o a un gol segnato. Un'altra conclusione importante si può trarre dalla quarta infografica, infatti, si evince come gli utenti italiani sembrino essere più interessati all'entrata in campo dell'italiano Verratti piuttosto che al gol di Coman, dimostrando come spettatori di diversa nazionalitá possano avere interessi differenti.
Analizzando il risultato dei test effettuati, le infografiche sono parse molto chiare e intuitive agli utenti, anche se meno utili ed informative. Essi sono riusciti ad interagire con le visualizzazioni in maniera corretta in poco tempo, dimostrando di averne compreso il senso. Dalle matrici di correlazioni riguardanti il test psicometrico si nota generalmente una bassa correlazione tra utilitá e bellezza.
Un possibile sviluppo futuro per questo progetto potrebbe essere quello di riuscire a superare il limite di 3000 tweets raccolti al minuto imposto dalle API di Twitter, al fine di riuscire ad effettuare analisi ancora più approfondite e convincenti, soprattutto nei momenti in cui si riscontrano episodi importanti per il match, nei quali non si ha la reale conoscenza di quanti possono essere realmente stati i tweet al minuto. Inoltre, in futuro potrebbe essere utile effettuare scraping anche delle statistiche relative ai singoli giocatori, per confrontare i *Sentiment Score* ottenuti con dati oggettivi che descrivono la performance del calciatore. Un altro possibile sviluppo potrebbe essere quello di estendere la sentiment analysis a piú lingue.

8 Sitografia

Cronaca Repubblica : *https://www.repubblica.it/sport/live/calcio/europa/champions-league/2019/diretta/psg-bayern%20m.*

Scraping Statistiche: *https://www.footballcritic.com/uefa-champions-league-paris-saint-germain-fc-fc-bayern-munchen/match-stats/2136879*
Tweepy: *http://docs.tweepy.org/en/v3.4.0/streaming_how_to.html*

Streaming dati Twitter: *https://gustavosaidler.com/data%20analysis/twitter-stream/*

Kafka-Mongo: *https://towardsdatascience.com/kafka-python-explained-in-10-lines-of-code-800e3e07dad1*

Word Cloud: *https://www.youtube.com/watch?v=_p6XdVw7fdk*

Link Infografiche: *https://public.tableau.com/profile/alessandro1080!/*
https://public.tableau.com/profile/luca.pretini!/

9 Divisione Lavoro

Alessandro Fossati (mat.819499), Data Science:

- Raccolta dati in streaming ed immagazzinamento
- Data management
- Viz Neymar
- Viz Mbappè-Coman
- Word cloud Neymar

Giorgio Nardi (mat.819961), Data Science:

- Raccolta dati in streaming ed immagazzinamento
- Data management
- Scraping statistiche
- Viz Linechart
- Viz Statistiche partita
- Viz Introduttiva

Luca Pretini (mat.864014), Data Science:

- Raccolta dati in streaming ed immagazzinamento
- Sentiment analysis
- Scraping cronaca
- Viz Neymar
- Viz Mbappè-Coman
- Viz Introduttiva