

Università degli studi Milano Bicocca

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



Analisi dei sentimenti degli utenti di Twitter durante il mondiale Qatar 2022.

Sabino Giuseppe

Ceccarelli Daniele

Pagani Gabriele

Indice

1. Introduzione
2. Data acquisition
3. Data storage
4. Sentiment Analysis
5. Data Exploration
6. Conclusioni

1. Introduzione

Il calcio è sicuramente tra gli sport più seguiti al mondo, con milioni di tifosi in tutti i continenti e rappresenta un fenomeno culturale di grande rilevanza a livello globale.

Il mondiale in Qatar 2022 rappresenta un segnale di svolta in questo sport.

Per la prima volta infatti un paese del Medio Oriente ha ospitato la competizione calcistica di maggior valore, a simboleggiare la continua crescita che questi paesi stanno vivendo.

Nonostante tutto lo scalpore dovuto alla preparazione di questo avvenimento sportivo, si stima che la finale del mondiale sia stata seguita da circa 1,5 miliardi di persone.

Ciò va in disaccordo con quanto affermato da alcuni esponenti calcistici di valore in questi ultimi anni, secondo i quali le partite stiano diventando meno interessanti, dovuto dallo squilibrio tecnico tra le squadre di prima fascia e le emergenti portando alla visione di partite “noiose”. Si è arrivati al punto di pensare a una rivoluzione di questo sport tramite il progetto della Superlega. Il principale fautore di questa idea è Florentino Perez, presidente del Real Madrid, secondo cui :

[“ i giovani, tra i 14 e i 24 anni, abbandonano il calcio perché li annoia di fronte ad altri divertimenti che preferiscono”](#)

Quali sono le variabili che hanno un impatto sull’opinione delle persone?

Lo scopo del progetto è rispondere a questa domanda confrontando le statistiche principali delle 64 partite disputate al mondiale di Qatar 2022, con l’analisi dei sentimenti dei Tweet riguardanti le partite stesse.

La scelta di Twitter è dovuta dal fatto che a livello internazionale [il 75% degli utenti abbia meno di 30 anni](#) .

Per basarsi sul livello tecnico delle nazionali, per verificare se realmente influisce, è stato preso in considerazione [l’ultimo ranking ufficiale](#) delle squadre corrispondente al 6 ottobre 2022.

2. Data Acquisition

Per la parte di Acquisizione dei dati l'obiettivo è quello di entrare in possesso del ranking FIFA delle singole squadre, delle statistiche delle partite del mondiale 2022, e dei tweet relativi ad esse.

In assenza di un API sul sito FIFA, la libreria Selenium (versione 4.4.3) di Python è stata utilizzata per navigare il sito internet FIFA.com (in lingua inglese) attraverso il browser Microsoft Edge (versione 108 del browser).

L' HTML del sito FIFA.com potrebbe cambiare in futuro. In tal caso, bisognerebbe modificare lo script di Python, aggiornando gli XPATH dei vari elementi del sito internet ufficiale del torneo che sono stati analizzati attraverso la libreria Selenium.

Il primo step consiste nel raccogliere la lista di squadre che hanno partecipato al mondiale in Qatar.

Nella sezione "[TEAMS](#)" del sito FIFA.com sono indicate tutte le 32 nazionali in ordine alfabetico.

I nomi ufficiali delle squadre FIFA sono stati poi utilizzati per l'acquisizione dei tweets pubblicati durante il mondiale.

Quando si accede al sito "TEAMS" di FIFA.com, appare il pop-up dei cookies.

La seguente funzione è stata utilizzata per rifiutare la raccolta dei cookies e chiudere il pop-up:

```
"cookies = WebDriverWait(driver, delay).until(EC.presence_of_element_located((By.ID, 'onetrust-reject-all-handler'))).click()"
```

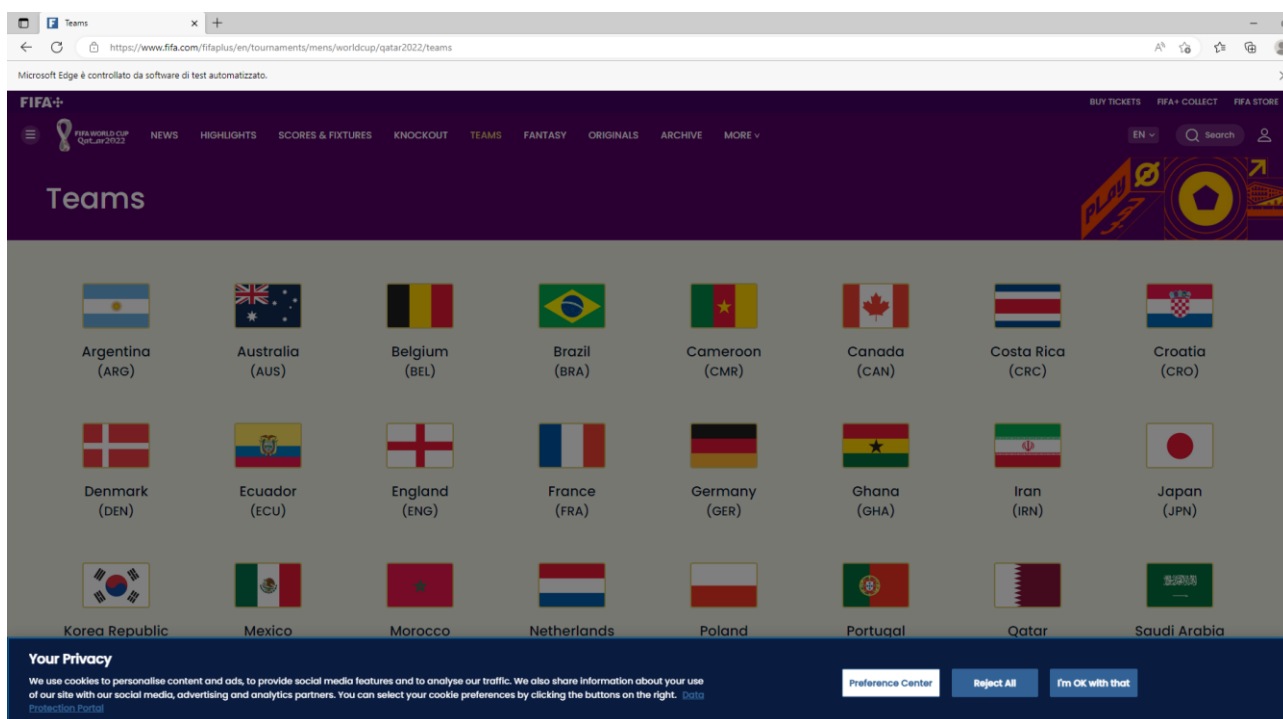


Figura 1 - pop-up cookies sito FIFA.com

Chiudere il pop-up dei cookies permette poi di visualizzare correttamente il codice HTML del sito TEAMS in modo da implementare lo scraper per la raccolta dei nomi delle squadre.

Grazie alla funzione “.text” di Selenium è stato possibile raccogliere il nome di ogni squadra cercando nella pagina web tutti gli elementi del codice HTML con CLASS_NAME uguale a "flag-with-info_flagAbbr__Dn4Ab". Nel sito “TEAMS”, oltre al nome esteso, sono indicate anche le sigle di tre lettere di ogni squadra (e.g. “ARG” per la squadra Argentina) con CLASS_NAME "flag-with-info_flagCountry__Yw8QR".

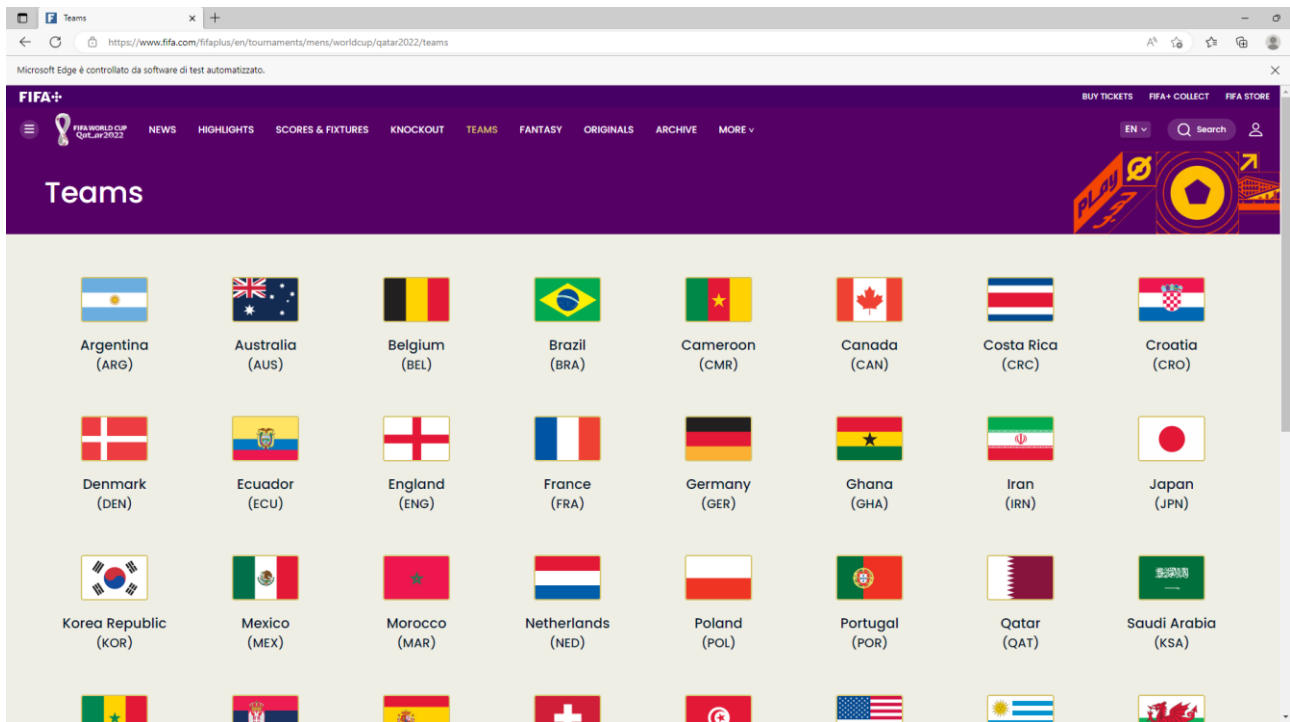


Figura 2 - sezione "TEAMS" del sito FIFA.com

Il secondo step consiste nell'individuare l'URL delle partite del mondiale.

Questo passaggio è fondamentale per poter acquisire i dati relativi alle singole partite (e.g. data e ora del calcio d'inizio, risultato finale, statistiche e nomi dei giocatori che hanno segnato un gol)

Nella sezione "[SCORES & FIXTURES](#)" del sito FIFA.COM, il web address associato ad ogni partita è situato nel codice HTML con un tag "a" ed un href che inizia con "fifaplust/en/match-centre/match/17/255711/".

Le partite presenti nella pagina web "SCORES & FIXTURES" sono elencate in ordine cronologico dal primo match del mondiale "QATAR-ECUADOR" del 20-Novembre-2022 fino alla finale "ARGENTINA-FRANCIA" del 18-Dicembre-2022.

I 64 URL (un URL per partita del mondiale) sono stati poi inseriti in una lista di Python chiamata "games_world_cup_2022_URL_list".

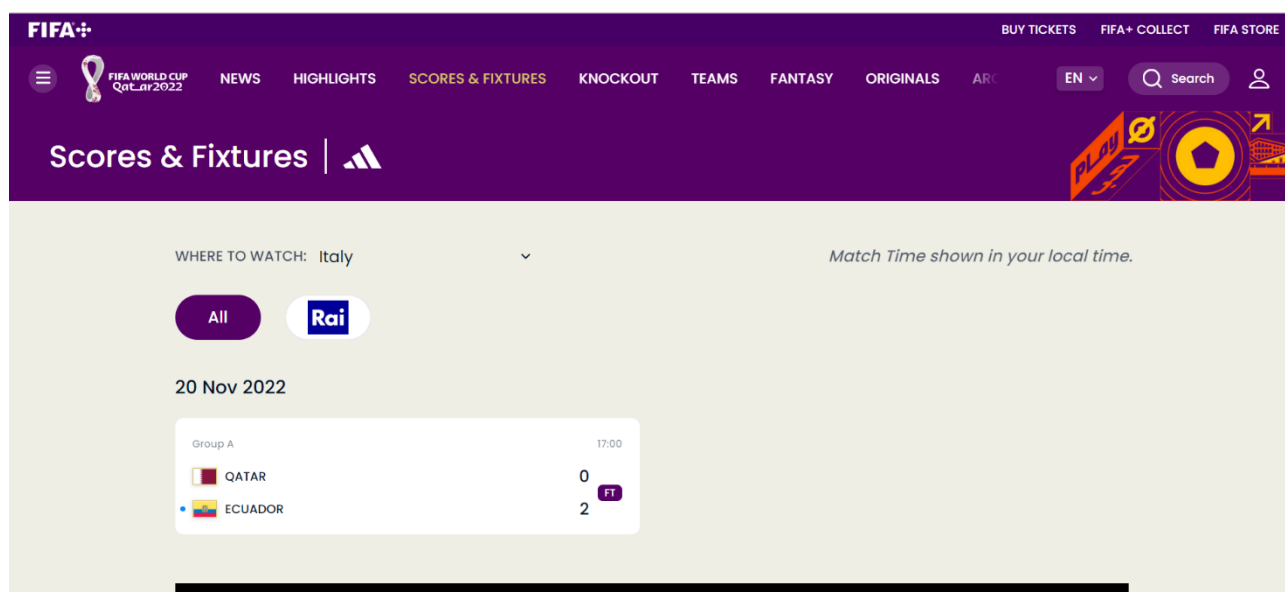


Figura 3 - sito web "Scores & Fixtures"

Di seguito sono elencati i passaggi della procedura automatizzata che è stata implementata per navigare sulle pagine web delle singole partite e raccogliere i dati necessari utilizzando come esempio la prima partita "QATAR-ECUADOR":

1. Una volta aperto il sito internet del match "QATAR-ECUADOR" (vedi Figura 4), è possibile acquisire i seguenti dati:
 - a. data e ora del calcio d'inizio (fuso orario italiano)
 - b. tipo di partita (i.e. gironi iniziali, ottavi, quarti...) → il match "QATAR-ECUADOR" è classificato come "Group A" perché le due squadre appartenevano al girone A di qualificazione
 - c. nome della squadra che giocava in casa (a sinistra) e nome della squadra che giocava in trasferta (a destra)
 - d. risultato finale del match
2. cliccando nella freccia sotto il simbolo della "Rai", possiamo accedere all'elenco completo dei nomi dei calciatori della squadra in casa ed in trasferta che hanno segnato un gol. (vedi Figura 5)
 - a. Per esempio, nel primo match del mondiale, il calciatore Enner Valencia del team "ECUADOR" ha segnato due gol
3. Scorrendo in basso nel sito web, sotto il pulsante "Highlights", è presente il link "STATS". Una volta cliccato sul pulsante STATS, è possibile visualizzare tutte le statistiche del match. (Figura 6)

- a. Nel codice Python, i nomi delle statistiche e XPATH per individuare le statistiche nel sito web, sono indicati rispettivamente nelle liste “list_statistiche_stringhe” e “list_statistiche00”
4. I dati del match “QATAR-ECUADOR” raccolti grazie ai tre passaggi sopra indicati, sono stati “concatenati orizzontalmente” in un dataframe di una singola riga con 44 colonne.
- a. Il numero di colonne dei singoli dataframe è variabile perché ad ogni partita è associato un elenco diverso di giocatori che hanno segnato gol. Per esempio, la partita “SPAIN-COSTA RICA” del 23-Novembre-2022 è terminata 7 a 0 per la squadra in casa ed il dataframe generato dalla procedura automatizzata era composto da 49 colonne.

“User Defined Functions” di Python sono state utilizzate per l’acquisizione dei dati secondo i passaggi 1-3 sopra descritti (“match_header_function”, “gol_scored_function”, “statische_partita_function”)

È stato necessario definire una funzione chiamata “cartellini_rossi_casa_function” per eliminare dall’elenco dei nomi dei giocatori che hanno segnato gol, i nomi di quei calciatori che sono stati espulsi dopo aver ricevuto un cartellino rosso.

Nella pagina web delle singole partite, i cartellini rossi (dove presenti) sono stati inseriti nella sezione dove sono indicati i nomi dei calciatori che hanno segnato un gol.

Quattro calciatori in quattro partite diverse del torneo sono stati espulsi.

I calciatori espulsi appartenevano sempre a squadre che giocavano “in casa”.

I nomi dei calciatori espulsi (“Wayne HENNESSEY”, “Vincent ABOUBAKAR”, “Denzel DUMFRIES”, “Walid CHEDDIRA”) sono sempre situati all’inizio dell’elenco dei nomi di calciatori delle squadre in casa che hanno segnato un gol.

Considerando gli elementi sopra indicati, è stato possibile definire una funzione per assegnare il nome del calciatore espulso ad una nuova colonna chiamata “cartellino_rosso_casa_1”.

I paragrafi precedenti hanno fornito una rapida descrizione della procedura implementata su Python per raccogliere tramite scraper i dati relativi alle 64 partite del mondiale.

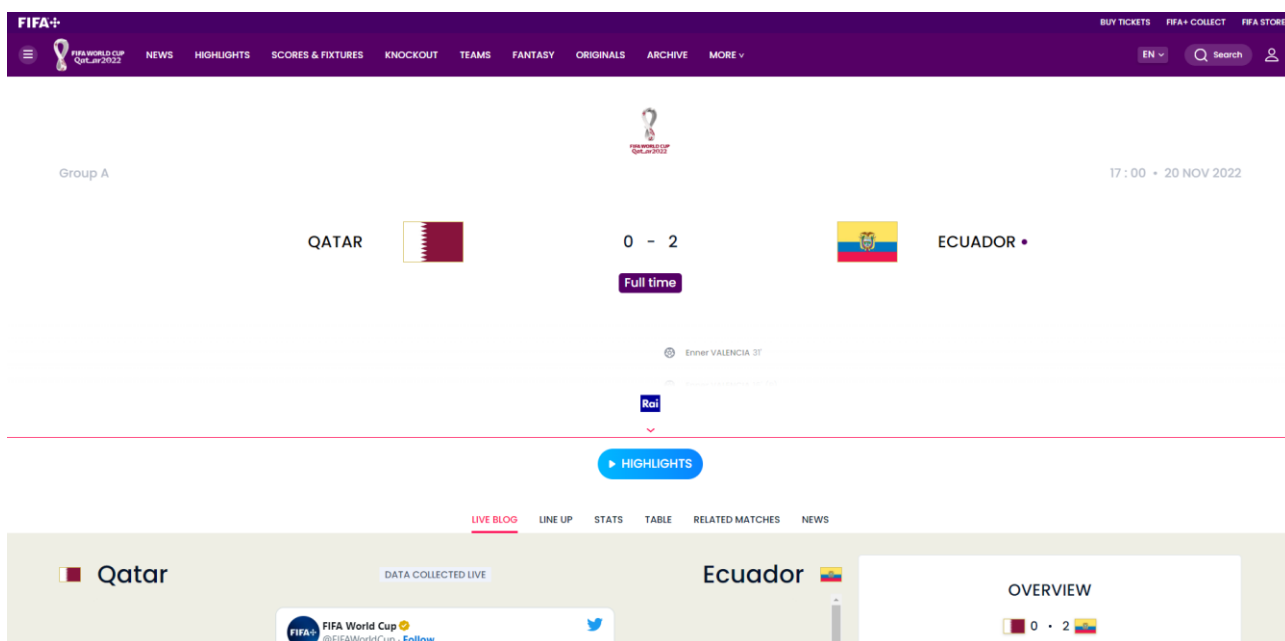


Figura 4 - sito web della partita "QATAR-ECUADOR", prima partita del mondiale 2022

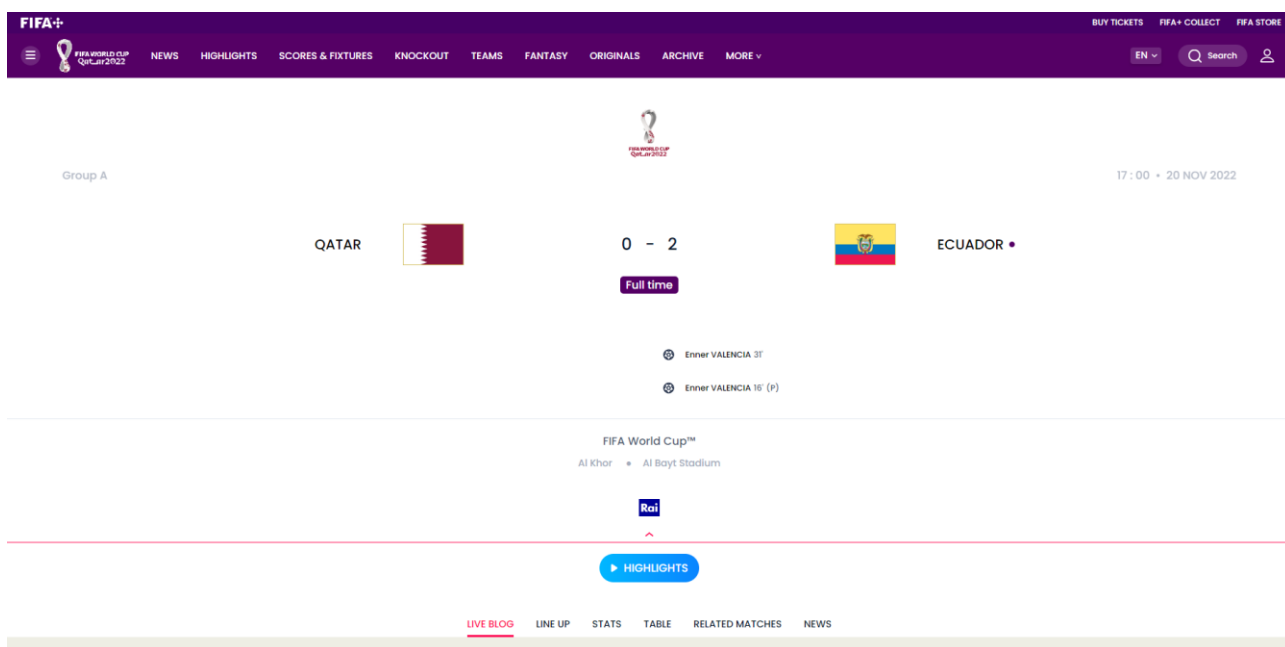


Figura 5 - elenco completo dei giocatori che hanno segnato durante la partita

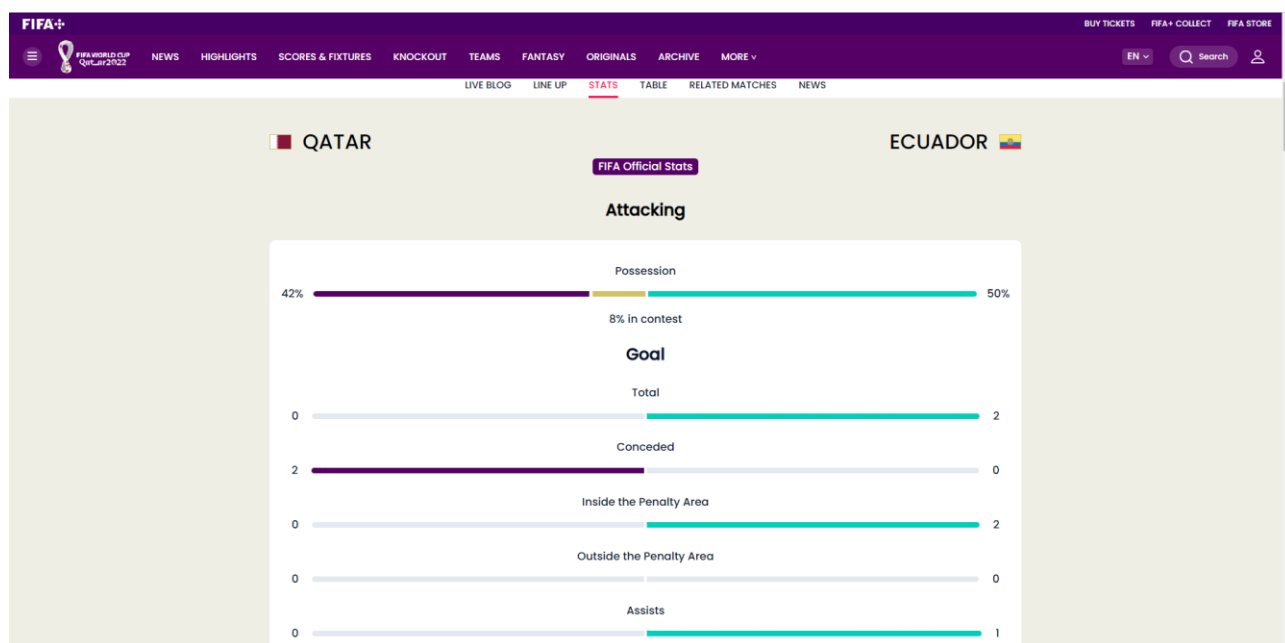


Figura 6 - sezione "STATS" contenente le statistiche del match

Successivamente, è stata definita anche una procedura per acquisire dal sito FIFA i rankings delle squadre nazionali di calcio maschile che hanno partecipato al mondiale in Qatar.

Nel sito "<https://www.fifa.com/fifa-world-ranking/men?dateId=id13869>" è possibile visualizzare l'evoluzione nel tempo dei punteggi del ranking di ogni squadra, utilizzando il "dropdown menu" in alto a destra.

FIFA e Coca-Cola hanno sviluppato un'algoritmo per assegnare ad ogni squadra un punteggio sulla base della performance storica.

FIFA, in data 22 dicembre 2022 (dopo la finale del mondiale di Qatar del 18 dicembre), ha aggiornato i punteggi ed il ranking delle squadre maschili.

Invece, l'ultimo aggiornamento prima dell'inizio del mondiale risale al 6 ottobre 2022.

In questo progetto sono stati utilizzati sia i punteggi ed il ranking aggiornati al 6 ottobre 2022 (prima dell'inizio del torneo) che i punteggi ed il ranking aggiornati al 22 dicembre 2022.

I punteggi relativi al 6 ottobre 2022 sono stati poi utilizzati per testare eventuali relazioni statistiche con il "sentiment" degli utenti di Twitter (le analisi saranno poi descritte in maggiore dettaglio nel capitolo #5)

Le date "06 Oct 2022" e "22 Dec 2022" del dropdown menu situato in alto a destra del sito web, sono state individuate attraverso XPATH nel codice HTML e selezionate grazie alla funzione ".click()" di Selenium.

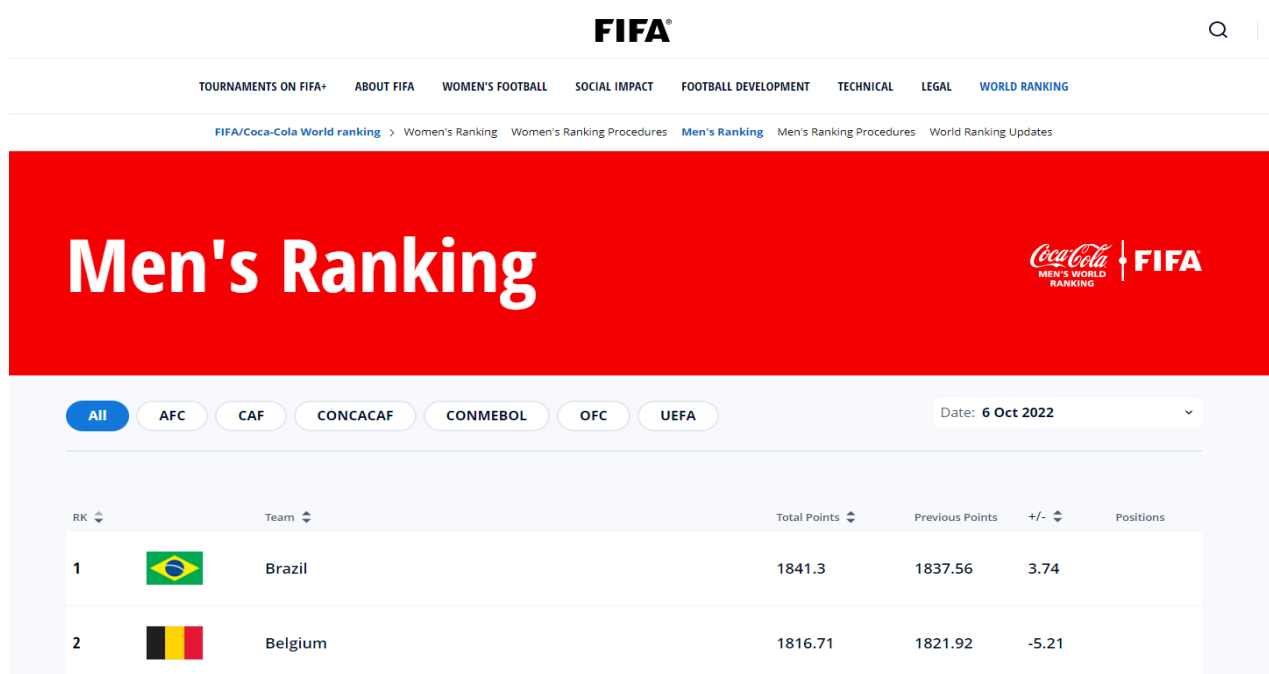
I dati relativi al ranking e al punteggio di ogni squadra sono elencati in formato tabellare nel sito FIFA.

Le singole righe delle tabelle sono state individuate grazie al CLASS_NAME

"row_rankingTableFullRow__Y_A4i ". il testo contenuto all'interno di ogni elemento del codice HTML è stato poi diviso con la funzione “.split('/n’)” e successivamente inserito in un dataframe Pandas.

I nomi delle squadre nazionali elencati nel sito dei rankings FIFA e nel sito web del mondiale 2022 in Qatar sono identici, ad eccezione del nome della squadra statunitense (“USA”) ed iraniana (“IR IRAN”). I nomi di queste due squadre sono state modificate sulla base dell'elenco dei nomi delle nazionali indicati nel sito “[TEAMS](#)” del sito ufficiale del mondiale 2022 (“United States” e “Iran”). Nel sito dei rankings FIFA sono presenti i punteggi di tutte le nazionali associate alla FIFA, quindi è stato necessario filtrare il dataframe con i punteggi e rankings (attraverso un'operazione di LEFT JOIN su pandas tra il dataframe con i nomi delle squadre del mondiale 2022 ed il dataframe con tutti i rankings) per conservare solo le informazioni relative alle nazionali che hanno partecipato al torneo in Qatar.

I dataframes contenenti le statistiche ed i rankings FIFA sono stati convertiti in formato JSON attraverso la funzione della libreria Pandas “.to_json()” ed importati su MongoDB.



The screenshot shows the FIFA Men's World Ranking page as of October 6, 2022. The page features a red header with the 'Men's Ranking' title and the Coca-Cola FIFA logo. Below the header, there are navigation tabs for various football confederations: All, AFC, CAF, CONCACAF, CONMEBOL, OFC, and UEFA. The 'All' tab is selected. A table displays the top two ranked teams:



RK	Team	Total Points	Previous Points	+/-	Positions
1	 Brazil	1841.3	1837.56	3.74	
2	 Belgium	1816.71	1821.92	-5.21	

Figura 7 – esempio di FIFA rankings aggiornati prima del mondiale in Qatar

Per quanto riguarda i tweet invece è stata utilizzata la libreria snsrape.

Innanzitutto è stato ripulito il dataset delle statistiche, poiché per alcune partite sono stati ritrovati pochi riscontri a livello di numero di commenti.

Infatti si presume che poche persone abbiano commentato la partita “Korea Republic-Portugal” ma piuttosto “Korea-Portugal”, ottenendo in definitiva dei riscontri positivi sull’incremento dei tweet.

La ricerca è stata effettuata in lingua inglese essendo più internazionale per avere una visione più ampia sul mondo in generale e non solo in Italia.

Per ogni partita sono stati ricavati 1000 tweet per avere un campione abbastanza ampio per dare valore alle statistiche ricavate; non è stato possibile ricavarne un numero maggiore in quanto per molte partite con questo limite è stato raggiunto l’intero campo selezionabile.

E’ stato necessario limitare l’acquisizione dei tweet ai soli giorni risalenti lo svolgimento della partita, in questo modo si ha una visuale più ristretta con commenti relativi al solo mondo calcistico (una ricerca USA-Iran può portare ad argomenti politici) e anche per risolvere il problema di partite doppie, in questo caso solo Croazia-Marocco.

Il risultato finale comprende 57518 tweet in formato json come il seguente esempio:

```
{'data_partita': '2022-11-20',
  'testo': 'So this is it! Welcome to FIFA World Cup 2022. Official kick-off started tonight. \nThis is just a quick snap for the opening of FIFA World Cup here in Doha, Qatar. Ecuador won against Qatar for tonight’s match. 2-0 amigos. 🥳 #FIFAWorldCup #FIFAWorldCup2022 #QatarWorldCup2022 https://t.co/xfBxSZRCsI',
  'reply_count': 0,
  'like_count': 1,
  'retweet_count': 0,
  'quote_count': 0,
  'hashtags': ['FIFAWorldCup', 'FIFAWorldCup2022', 'QatarWorldCup2022'],
  'partita': 'QATAR-ECUADOR'}
```

Figura 8 - esempio di file JSON con un tweet sulla partita "QATAR-ECUADOR"

3. Data storage

L’acquisizione dati comporta una notevole perdita di tempo se eseguita ad ogni accesso, di conseguenza per migliorare la velocità di esecuzione è stato pensato l’utilizzo di un DB.

Il dataset relativo alle statistiche comprende l’elenco dei marcatori di ogni partita, così come il nome dei calciatori espulsi.

Essendo che non in tutte le partite ci sono state espulsioni ed è ancora più difficile che tutte le partite abbiano avuto lo stesso numero di gol segnati, il dataset ha una modesta quantità di valori nulli.

Per questo motivo la scelta più logica consiste nell’utilizzo di un mongoDB, grazie al quale non sono presenti schemi fissi, di conseguenza nessun valore nullo.

Grazie all'utilizzo della libreria "pymongo" è stato possibile connettere direttamente il DB con il notebook jupyter, dal quale è stato creato prima il DB , all'interno del quale sono presenti 3 collezioni:

- WorldCupStats: che contiene 64 documenti, ognuno dei quali rappresenta le statistiche di ogni partita
- WordlCupTweet: che contiene 57518 documenti dei tweet mostrati in precedenza
- WorldCupRanking: che contiene 32 documenti con il ranking delle nazionali e il valore della squadra

4. Sentiment Analysis

In questa parte dell'elaborato si implementa un modello di textual classification, ovvero, una tecnica di Supervised Machine Learning volta alla predizione della classe di appartenenza dei dati in un set predefinito di classi. In particolare, è stata adottata la tecnica di Sentiment Analysis che è uno sviluppo del Natural Language Processing che elabora sistemi al fine di individuare ed estrarre opinioni da un testo. Considerato che i motivi che spingono le persone a twittare riguardano principalmente reazioni a fatti, trend e notizie, con l'analisi del Sentiment dei tweet è possibile esaminare il loro livello di positività o negatività (polarità). I testi molto lunghi non sempre sono decifrabili in maniera ottimale, quindi, il limitato numero di caratteri imposto da Twitter è sicuramente un elemento che può aiutare la Sentiment Analysis.

Durante la fase di preprocessing, sono state adottate diverse tecniche di pulizia del testo per rendere i dati maggiormente sfruttabili dal modello di Sentiment Analysis.

In particolare:

- Normalization: è un task sempre necessario per le applicazioni di Text Mining e consiste nel ricondurre forme diverse dello stesso elemento ad una forma unica per tutto il corpus. La normalization è uno dei task che lascia generalmente più libertà d'azione poiché è fortemente

dependente dal tipo task di Text Mining che si svolge. Per questo motivo sono state proposte le seguenti tecniche:

- Testo in minuscolo;
- Rimozione degli spazi;
- Rimozione della punteggiatura;
- Rimozione degli URL
- Rimozione dei duplicati;

E' stato deciso di non rimuovere i punti esclamativi perché torneranno utili per la Sentiment Analysis e, per lo stesso motivo, sono state convertite le emoji nelle espressioni "good" e "bad" a seconda del loro significato. Inoltre, sono stati lasciati nel dataset anche gli hashtag per sviluppare la Word Cloud dei 50 hashtag con maggiore frequenza nel dataset.

- Stop-words: consiste nel rimuovere le parole, in questo caso in lingua inglese, che si presentano con ritmo molto frequente nel corpus poiché non sono ritenute utili nel caratterizzare i documenti. È necessario tale smaltimento perché la loro presenza spesso compromette i risultati dei modelli di classificazione.
- Tokenization: step di cruciale importanza che frammenta i testi in singoli elementi, chiamati token, dividendo un flusso di testo in unità dense di significato. Per fare ciò, sono state applicate delle semplici espressioni regolari al testo in esame. Questa è una tecnica di pre-processing fondamentale per applicare al meglio modelli di Text Mining.
- Stemming: riduce le parole alle loro radici, alla loro forma base. A fronte del rischio di perdere il preciso significato delle parole, se sviluppato correttamente incrementa le prestazioni del modello da implementare. E' stato utilizzato l'algoritmo di Porter, uno fra i più utilizzati per lo stemming in inglese.
- Lemmatization: riduce termini flessibili ad una forma base con il vantaggio di ridurre la dimensione del dizionario analizzato. È molto simile allo stemming ma con la differenza che lo stemming opera sulle singole parole indipendentemente, seguendo unicamente regole sintattiche, mentre il lemming considera anche il contesto e la semantica.

Dopo questa fase, E' stata aggiunta al dataset una colonna per ogni tecnica di pre-processing utilizzata in modo da avere una visione globale dei risultati raggiunti. Inoltre, è stata condotta un'analisi esplorativa usando i dati a disposizione per approfondire la struttura del dataset. Per esempio, attraverso la libreria Word Cloud di Python, sono stati identificati i cinquanta hashtag maggiormente rilevati nei tweet.



Figura 9 - World Cup

Come prima impressione, si può notare nella Word Cloud che sono in evidenza le rivelazioni del Mondiale (per esempio il Marocco) e le partite che hanno suscitato più tweets come Marocco vs Croazia, anche perché è l'unica partita del torneo che è stata giocata due volte.

Dopo la fase di pre-processing è stata implementata la Sentiment Analysis utilizzando Textblob.

TextBlob è una libreria Python per il Natural Language Processing (NLP) che offre una semplice interfaccia per accedere a diverse funzionalità, tra cui l'analisi dei sentimenti. Con TextBlob, è possibile analizzare il sentimento di un testo utilizzando un modello addestrato che è stato incorporato nella libreria stessa. Il modello assegna un punteggio di sentimento compreso tra -1 e 1, con -1 indica un sentimento negativo, 0 un sentimento neutrale e 1 un sentimento positivo.

Su un totale di 55023 tweets, sono stati classificati:

- 10197 tweets con sentiment negativo (19%);
- 21036 tweets con sentiment neutro (38%);

- 23790 tweets con sentiment positivo (43%).

5. Data Exploration

Prima di iniziare con la parte relativa all'analisi esplorativa è stato necessario procedere con l'integrazione dei tre dataset,

Innanzitutto, sono state effettuate due query per richiamare i dati da MongoDB sul notebook jupyter.

- `WorldCupCollection.find()`: Per avere a disposizione tutti i documenti, utilizzata per i dataset relativi a statistiche e ranking.
- `WorldCupCollection.aggregate(pipeline)` : con la pipeline inserita di seguito, per aggregare i dati per partita ed avere il numero totale di tweet. L'aggregazione è avvenuta anche sulla colonna "data_partita", poiché Croazia-Marocco è stata giocata due volte in date diverse

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

Figura 10 - pipeline della funzione "aggregate"

Il dataset relativo ai tweet è stato ordinato in ordine di numero di tweet crescenti:

kick_off_day	nome_partita	Negativo	Neutro	Positivo	total_tweets
2022-11-21	UNITED STATES-WALES	7	20	195	222
2022-11-29	IRAN-UNITED STATES	12	19	230	261
2022-11-24	URUGUAY-KOREA	70	121	80	271
2022-12-03	NETHERLANDS-UNITED STATES	4	5	285	294
2022-11-25	ENGLAND-UNITED STATES	7	17	290	314

Figura 11 – esempio di dataframe con i "sentiment" di Twitter per ogni partita

Dei primi 5 record, 4 contengono partite disputate dagli Stati Uniti.

Come spiegato nel Cap. 2 **Data acquisition**, lo stesso procedimento di pulizia riservato alla "Korea" è stato effettuato anche per gli Stati Uniti D'America, non trovando però un valore di dominio di riferimento utile a migliorare la qualità del dato.

E' stato dunque effettuato il merging tra il dataset delle statistiche e quello dei tweet, sulle colonne relative alla partita giocata e alla rispettiva data. Infine si è proceduto con l'integrazione di quest'ultimo al dataset dei ranking, aggiungendo il rank e il punteggio sia della squadra in casa, sia di quella ospite, per ogni partita.

Data analysis

```
query={
  "$and": [
    {"squadra_in_casa": {'$ne': "UNITED STATES"}},
    {"squadra_in_trasferta": {'$ne': "UNITED STATES"}}
  ]
}

query_result = WorldCupCollection.find(query,query_projection)
```

Figura 12 - query da MongoDB escludendo i dati relativi al team U.S.A.

Considerando il numero relativamente basso di tweets associati alle partite giocate dagli Stati Uniti ed anche l'anomala percentuale di tweets classificati come "positivi", i dati relativi alle partite giocate dal team U.S.A. non sono state utilizzate nelle analisi descritte nei successivi paragrafi.

Con la query soprastante, è stato importato il dataset finale ad eccezione delle partite giocate dagli Stati Uniti. "Query Projection" è un dizionario con l'elenco delle colonne selezionate e sono rappresentate di seguito.

0	_id	60 non-null	object
1	tipo_di_partita	60 non-null	object
2	squadra_in_casa	60 non-null	object
3	squadra_in_trasferta	60 non-null	object
4	gol_segnati_casa	60 non-null	int64
5	gol_segnati_trasferta	60 non-null	int64
6	rigori_finali_segnati_casa	60 non-null	int64
7	rigori_finali_segnati_trasferta	60 non-null	int64
8	nome_partita	60 non-null	object
9	tiri_totale_casa	60 non-null	int64
10	tiri_totale_trasferta	60 non-null	int64
11	tiri_in_porta_casa	60 non-null	int64
12	tiri_in_porta_trasferta	60 non-null	int64
13	Negativo	60 non-null	int64
14	Neutro	60 non-null	int64
15	Positivo	60 non-null	int64
16	total_tweets	60 non-null	int64
17	Punti Rank casa	60 non-null	float64
18	Punti Rank ospite	60 non-null	float64

Figura 13 - funzione ".info()" di Pandas

Il dataset è composto da 60 osservazioni, si ha quindi una table completeness del 100%

Questo dataset è stato ulteriormente modificato:

- Essendo il numero di tweet diverso per determinate partite, è stata creata una colonna “percentuale_tweet_positivi”. La percentuale per ogni partita è stata calcolata come rapporto tra il numero di tweets “positivi” ed per il numero totale di tweets.
- Per quanto riguarda Gol, punti Ranking, e tiri in porta, è stata creata una colonna per ciascuna di esse, sommando l’attributo della squadra di casa con la squadra ospite.
- Tipo_partita è stato modificato rimuovendo il girone del gruppo, andando a suddividere in generale le varie fasi del torneo (sostituendo per esempio, il testo “Group A” relativa al girone A di qualificazione con la stringa “Group”)

Ottenendo il dataset con il quale si è proceduto con le analisi:

	tipo_di_partita	nome_partita	gol_totali	tiri_porta_totali	punti_ranking_totali	percentuale_tweets_positivi
55	Quarter-final	ENGLAND-FRANCE	3	11	3502.31	45.9
56	Semi-final	ARGENTINA-CROATIA	3	10	3402.80	37.4
57	Semi-final	FRANCE-MOROCCO	2	3	3323.20	40.0
58	Play-off for third place	CROATIA-MOROCCO	3	6	3190.50	65.6
59	Final	ARGENTINA-FRANCE	12	14	3535.50	62.2

Figura 14 - Dataframe

	gol_totali	tiri_porta_totali	punti_ranking_totali	percentuale_tweets_positivi
count	60.000000	60.000000	60.000000	60.000000
mean	3.183333	8.000000	3230.17300	43.662167
std	2.514309	3.288398	149.59512	7.221910
min	0.000000	0.000000	2905.71000	29.520000
25%	2.000000	6.000000	3135.92250	39.075000
50%	3.000000	8.000000	3226.99500	41.400000
75%	4.250000	10.000000	3322.16500	47.350000
max	12.000000	17.000000	3535.50000	65.600000

Figura 15 - statistiche descrittive

```

partita con numero minore di gol: DENMARK-TUNISIA
partita con numero maggiore di gol: ARGENTINA-FRANCE

partita con numero minore di tiri in porta: URUGUAY-KOREA
partita con numero maggiore di tiri in porta: COSTA RICA-GERMANY

partita con il minore punteggio totale FIFA: QATAR-ECUADOR
partita con la maggiore punteggio totale FIFA: ARGENTINA-FRANCE

partita con la minore percentuale di tweet positivi: URUGUAY-KOREA
partita con la maggiore percentuale di tweet positivi: CROATIA-MOROCCO

```

Figura 16 - funzioni ".idxmin()" e ".idxmax()" di Pandas

	gol_totali	tiri_porta_totali	punti_ranking_totali	percentuale_tweets_positivi
gol_totali	1.000000	0.530939	0.206897	0.069350
tiri_porta_totali	0.530939	1.000000	0.164558	0.142860
punti_ranking_totali	0.206897	0.164558	1.000000	-0.127815
percentuale_tweets_positivi	0.069350	0.142860	-0.127815	1.000000

Figura 17 - correlation matrix

Nella Figura #15 è possibile visualizzare le statistiche descrittive delle variabili quantitative.

Nella Figura #16 sono indicate le partite con i valori minimi e massimi per ogni variabile. Le partite indicate nella Figura #16 sono state individuate attraverso le funzioni “.idxmin()” e “.idxmax()”.

I coefficienti di correlazione lineare delle varie coppie di variabili sono mostrati nella Figura #17

Come prima analisi è stato effettuato un report del dataset, in questo modo si hanno statistiche generali sul dataset come valori null o spazio di memoria occupato.

La sezione degli alert, da suggerimenti sul miglioramento della qualità dei dati, indicando appunto avvertimenti da seguire.

È poi possibile selezionare ciascuna colonna per avere a disposizione statistiche descrittive e istogrammi ad essa relative.

C'è poi una parte di interaction nel quale è possibile selezionare diverse variabili e vedere la correlazione tra loro e l'heatmap con la correlazione tra variabili.

N.B. si consiglia la visione del file HTML “Report Qatar2022” per avere una visione chiara di quanto elencato

Per comprendere la relazione tra le variabili è stato utile prendere in considerazione un pair plot che mette in correlazione le variabili a gruppi di due.

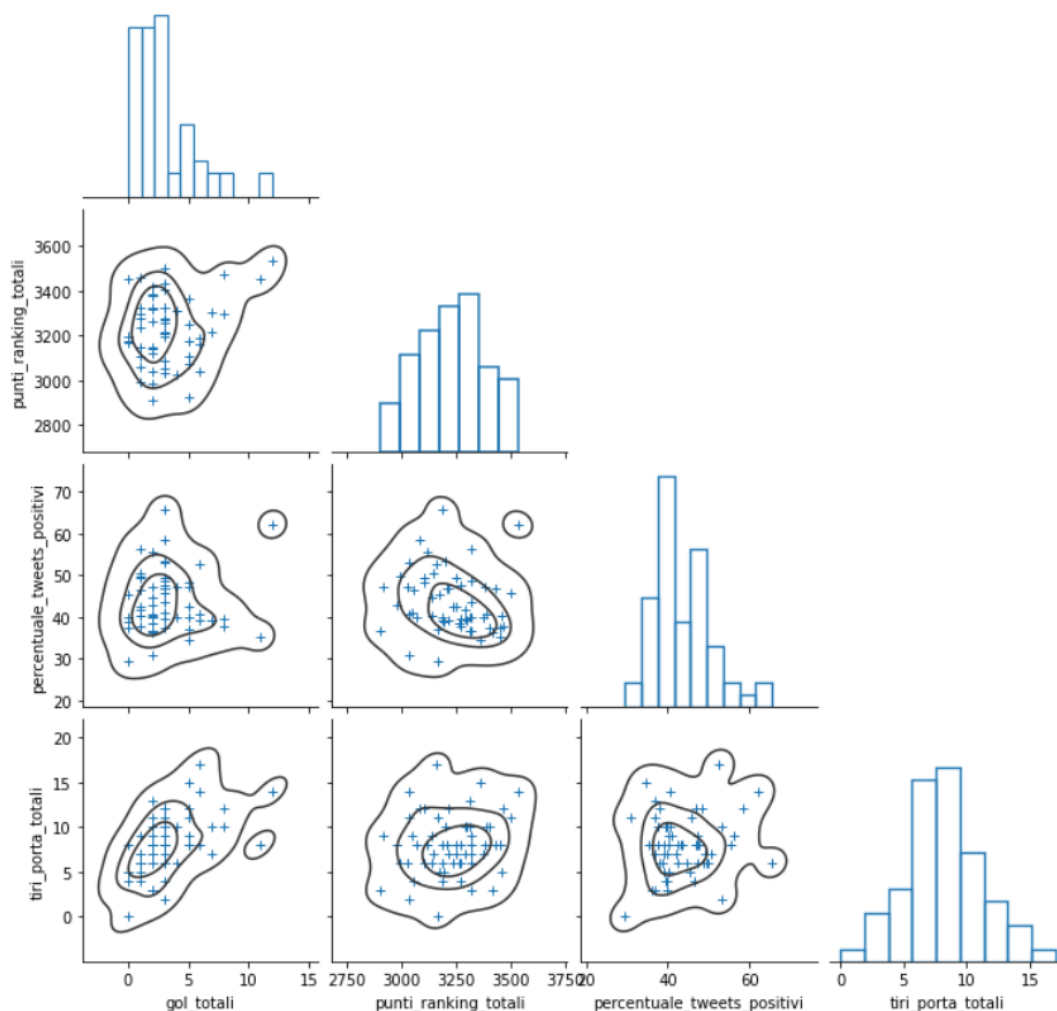


Figura 18 – funzione “`pairplot()`” della libreria *Seaborn*

Ad esclusione della ovvia correlazione tra tiri in porta e gol totali, le altre variabili non hanno una correlazione evidente.

L’obiettivo di queste analisi è rispondere alla domanda iniziale, occorre quindi verificare se i punteggi dei ranking FIFA delle nazionali influisce sulla percentuale dei tweet (**il grafico sulla sinistra**), e allo stesso modo se lo spettacolo della partita, presumibilmente tiri e gol, ha in qualche modo effetto sui tweet (**il grafico sulla destra**).

Per questo motivo è stata tracciata una retta di regressione nel diagramma di dispersione variabile sopra elencate

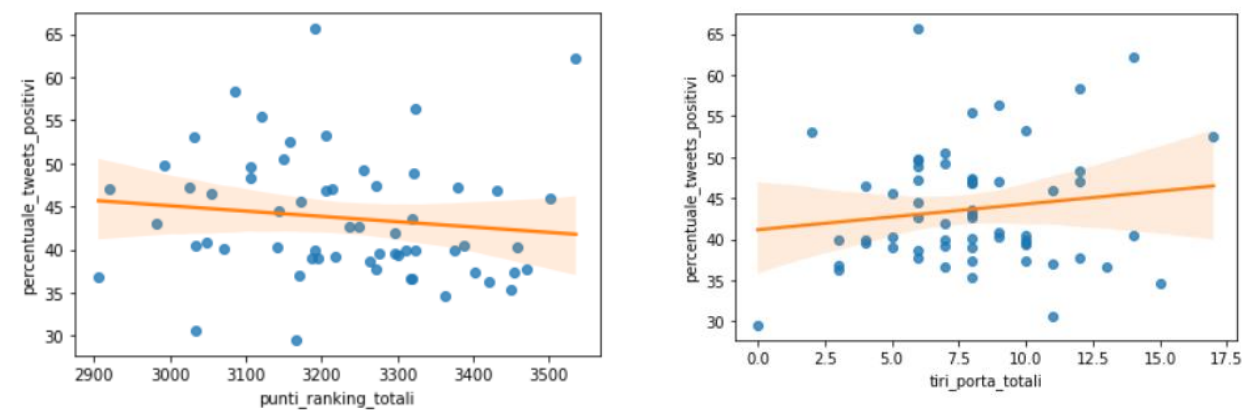


Figura 19 – scatterplots con linee di regressione

OLS Regression Results							
Dep. Variable:	percentuale_tweets_positivi				R-squared:	0.016	
Model:	OLS				Adj. R-squared:	-0.001	
Method:	Least Squares				F-statistic:	0.9633	
Date:	Wed, 11 Jan 2023				Prob (F-statistic):	0.330	
Time:	16:57:45				Log-Likelihood:	-202.77	
No. Observations:	60				AIC:	409.5	
Df Residuals:	58				BIC:	413.7	
Df Model:	1						
Covariance Type:	nonrobust						
	coef	std err	t	P> t	[0.025	0.975]	
const	63.5938	20.330	3.128	0.003	22.900	104.288	
punti_ranking_totali	-0.0062	0.006	-0.981	0.330	-0.019	0.006	
Omnibus:	10.042	Durbin-Watson:		1.932			
Prob(Omnibus):	0.007	Jarque-Bera (JB):		9.810			
Skew:	0.846	Prob(JB):		0.00741			
Kurtosis:	4.031	Cond. No.		7.05e+04			

OLS Regression Results						
Dep. Variable:	percentuale_tweets_positivi				R-squared:	0.020
Model:	OLS				Adj. R-squared:	0.004
Method:	Least Squares				F-statistic:	1.208
Date:	Wed, 11 Jan 2023				Prob (F-statistic):	0.276
Time:	17:15:04				Log-Likelihood:	-202.64
No. Observations:	60				AIC:	409.3
Df Residuals:	58				BIC:	413.5
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	41.1522	2.466	16.690	0.000	36.217	46.088
tiri_porta_totali	0.3137	0.285	1.099	0.276	-0.258	0.885
Omnibus:	7.503	Durbin-Watson:	2.009			
Prob(Omnibus):	0.023	Jarque-Bera (JB):	6.685			
Skew:	0.757	Prob(JB):	0.0353			
Kurtosis:	3.616	Cond. No.	23.2			

Figura 20 - summary delle regressioni lineari con Statsmodels

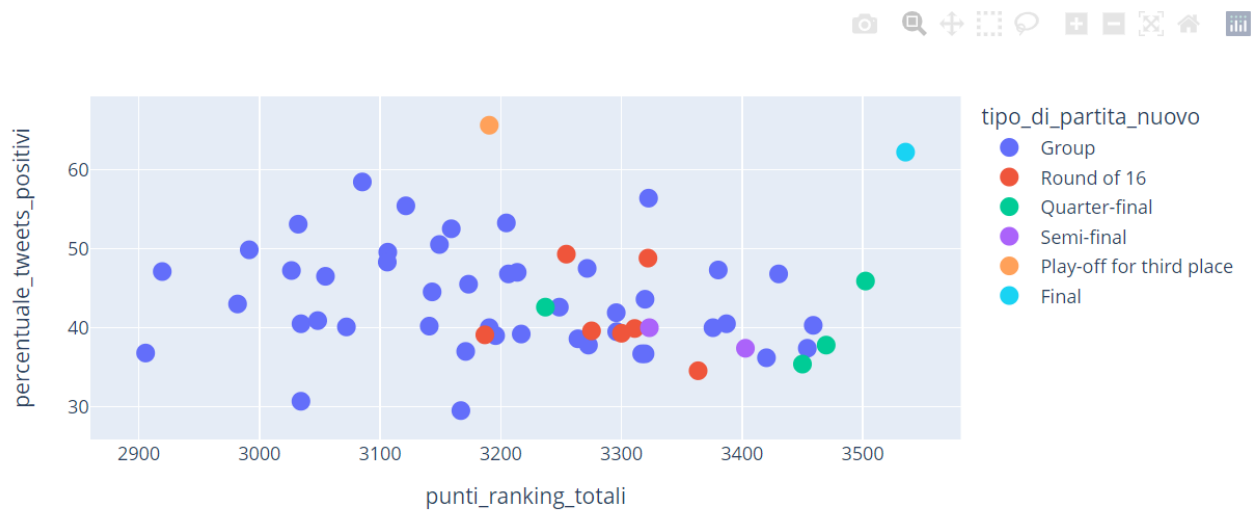


Figura 21 - Plotly - scatterplot

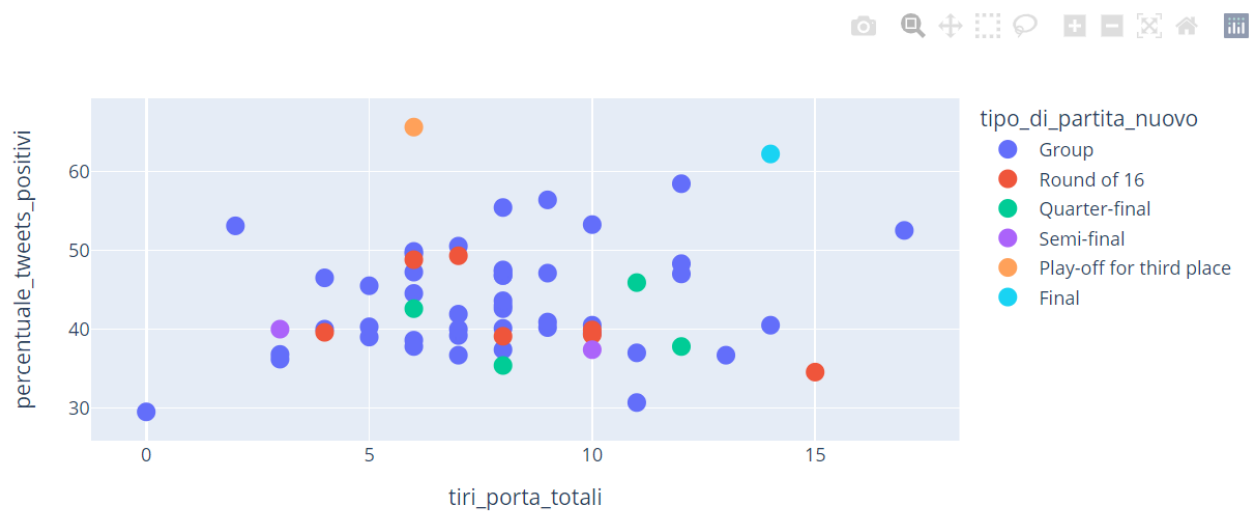


Figura 22 - Plotly - scatterplot

Seppur non in maniera significativa viene dimostrato come la percentuale dei tweet positivi non sia influenzata dal livello tecnico della partita, anzi, si nota che la percentuale di tweet positivi sia maggiore quando si affrontano due squadre con ranking minore.

Si conferma invece il trend secondo il cui per essere piaciuta una partita deve essere divertente e quindi ricca di tiri.

Nella figura #20 è possibile osservare i risultati delle regressioni generati con la funzione “.summary()” della libreria Statsmodels.

Come sopra indicato, i bassi valori di R-squared ed i p-value dei coefficienti di regressione delle variabili indipendenti maggiori del 5%, ci confermano che non esiste una relazione statistica significativa tra le variabili.

Su Python, sono stati anche visualizzati i grafici delle distribuzioni dei residui delle regressioni e gli scatterplots dei residui rispetto ai valori stimati dalle regressioni.

È stata utilizzata anche la libreria Plotly, per creare scatterplots interattivi (i punti del grafico sono stati colorati secondo le classi della variabile “tipo_partita”, ovvero gironi, ottavi, quarti....) (vedi Figura #21 & #22).

Ovviamente il numero di partite analizzate è relativamente piccolo (60 partite circa).

6. Conclusioni

Ma quindi quali sono le variabili che hanno un impatto sull'opinione delle persone?

Considerando i p-value dei coefficienti di regressione delle variabili indipendenti intorno al 30% circa non possiamo affermare che da un punto di vista statistico (i.e., p-value minore di 5%), i coefficienti di regressione siano diversi da zero.

Però, nonostante il dataset relativamente piccolo (60 partite circa), è interessante la correlazione leggermente negativa tra la somma dei punteggi dei rankings FIFA con il “sentiment” degli utenti Twitter (coefficiente di correlazione uguale a -0.1278 come da Figura #17).

Da quanto ricavato da questo studio si può affermare che si ha un trend crescente di divertimento, se la partita è ricca di azioni che portano a un tiro, dimostrandosi noiosa in caso contrario.

Risposta positiva si ha invece per quanto riguarda il livello tecnico delle due squadre. Quanto affermato inizialmente faceva intuire un abbassamento di tweet positivi in caso di partite tra squadre meno blasonate. Si è scoperto invece che mediamente hanno ricevuto più tweet positivi partite con una somma di punti ranking minore.

Un' idea per uno sviluppo futuro è quella di analizzare le statistiche ed i tweets relativi ad un numero maggiore di partite, raccogliendo per esempio i dati dei mondiali del 2014 e 2018.

Si potrebbe prendere in considerazione anche i top 5 campionati europei, che prevedono un divario tecnico notevolmente maggiore in quanto, almeno sulla carta, al mondiale giocano le squadre di maggior importanza. Sarebbe dunque utile analizzare partite dove sono presenti squadre già affermate da tempo e squadre appena promosse, tenendo in considerazione il divario profondo dovuto dalla presenza del calciomercato e della differenza di soldi spesi.

Dopo aver raccolto un numero maggiore di dati sulle partite calcio, si potrebbe valutare la possibilità di implementare altri modelli di machine learning di regressione.