Mike Tyson vs Roy Jones

ANALISI DEI RISULTATI UFFICIALI E DELLE OPINIONI ESPRESSE DAGLI SPETTATORI

Paparella Matteo 812561 Msc Data Science
Zottola Gabriele 812363 Mcs Data Science

February 7, 2021

Contents

1 Introduction

Il giorno 29 novembre 2020 viene organizzato il match evento di pugilato "Tyson vs Jones". Con 1.2 milioni di pay-per-view venduti, il galà, detiene il record di evento pugilistico con più incassi nel 2020. Mike "iron" Tyson e Roy Jones Jr., entrambi ex campioni dei pesi massimi, si sfidano sopra il ring per un totale di 8 round da 2 minuti ciascuno. Peculiarità del match è la giura, che assegna il voto dei cartellini a distanza senza trovarsi a bordo ring come di consueto.
L'esito ufficiale dell'incontro termina con un pareggio, divenuto oggetto di dibattito tra gli esperti e i fans, causato anche dalla modalità di assegnazione dei punteggi a distanza, voluto per il rispetto delle norme anti-covid.

1.1 Report Goal

Obiettivo del progetto è affrontare velocità e varietà dei big data.
Il fine dell'analisi prevede di analizzare le opinioni espresse dagli utenti Twitter relative agli atleti Mike Tyson e Roy Jones Jr., tramite utilizzo di algoritmi di sentiment analysis e confrontare i risultati ottenuti con le votazioni ufficiali e le statistiche del match.

1.2 Metodologia ed architettura software

Per il monitoraggio in tempo reale dell'evento si sfruttano le API fornite da Twitter, l'accesso ai dati è permesso tramite iscrizione con un account developer, dove, una volta eseguito il log in, vengono fornite le credenziali per l'accesso al servizio. Per il download e l'immagazzinamento dei dati si utilizzano, rispettivamente, la piattaforma di stream processing Kafka e il database non relazione MongoDB.
Si effettua la creazione di un database e di una collezione su MongoDB e si procede alla creazione di un topic su Apache Kafka, di un producer e di un consumer. I dati, presi in streaming vengono immagazzinati in un formato Json binario all'interno della collezione creata.
I Tweets, scaricati in lingua inglese, vengono presi in tre differenti momenti:
- durante la serata della cerimonia del peso, il 27 novembre utilizzando gli hashtags: MikeTyson, Tyson, IronMike, TysonvsJones, tysonvsjones, TysonJones2020, RoyJones, Jones, RoyJonesJr
- il giorno 29 novembre 2020 in un arco temporale che va dalle 00:00 alle 11:00 del mattino circa , utilizzando i seguenti hashtags: MikeTyson, Tyson, IronMike, TysonvsJones, tysonvsjones, TysonJones2020, RoyJones, Jones, RoyJonesJr.
- nel pomeriggio del 29 novembre, dopo aver riscontrato la presenza di varie polemiche legate all'incontro
I dati ufficiali relativi all'incontro vengono presi, in formato CSV, dai siti web: www.cbssports.com e http://beta.compuboxdata.com/fighter tramite utilizzo dell'estensione Chrome "Web Scraper".

Le fasi di data exploration, data pre processing e data integration vengono effettuate utilizzando JupyterNotebook con linguaggio di programmazione Python. La visualizzazione dei dati è permessa sfruttando il software Tableau.

2 Data Extraction

I Tweets vengono caricati sul Notebook tramite collegamento al database MongoDB, così facendo risultano essere immediatamente disponibili per l'analisi.
I dati relativi alle statistiche ufficiali del match vengono caricati su Notebook tramite comando "read csv".
Per aumentarne la maneggevolezza e poter integrare i dati di diverso formato, si convertono in Data frame sia il dataset relativo ai tweets scaricati sia i dataset relativi alle statistiche ufficiali e ai voti dei giudici.

3 Data Exploration

I tweets scaricati risultano essere in totale 767.473 e vengono scaricati con i seguenti attributi:
- Id
- Screen-Name
- Position
- Text
- Date-time
- Name
- Followers
- Number-Retweet

Le statistiche scaricate calcolano, per ogni atleta e per ogni round:
- Totale dei pugni andati a segno / totale dei pugni tirati
- Totale dei jab (diretto sinistro) andati a segno / totale dei jab tirati
- totale diretto destro andati a segno / totale diretti destri tirati

I dati relativi al punteggio mostrano il giudizio dato dai 3 giudici per ogni atleta e per ogni round. Viene assegnato un punteggio di 9 all'atleta considerato perdente nel round, 10 all'atleta considerato vincente e 8 se il pugile è stato atterrato durante i minuti di ripresa del combattimento. Viene dichiarato vincitore dell'intero match il pugile a cui vengono assegnati più round vincenti.

4 Data Pre-Processing

Nella fase di data pre processing si mira a pulire i dati in modo tale da aumentarne la qualità. Oggetto di questa fase sono i tweets scaricati, le statistiche e

i punteggi ufficiali che, necessitano tutti di essere trattati per poter adoperare una integrazione dati e renderli adatti all'analisi preposta.

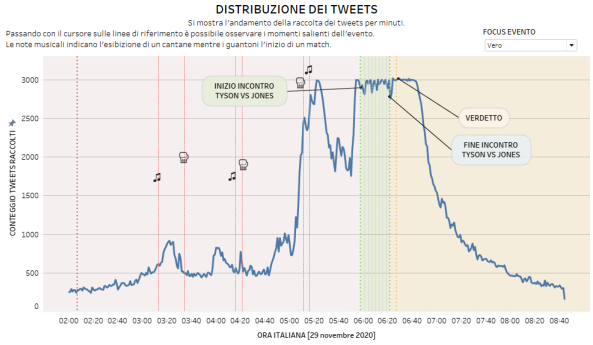
4.1 Individuazione BOT

Al fine di incrementare la qualità del dataset si adotta una strategia per l'individuazione di possibili BOT.
Si conta il numero di volte che uno 'Screen name' twetta nel corso della serata e si osservano gli utenti con un numero di messaggi elevato. Si considerano bot le pagine che presentano un numero di followers basso, contenuto del tweet ripetitivo e condivisione di contenuto in una breve distanza di tempo.
Si è proceduto all'eliminazione degli utenti che presentano un comportamento anomalo basato sulle considerazione sopra fattel.

4.2 Date time

I tweets vengono raccolti su virtual machine che riporta l'orario di pubblicazione un'ora indietro rispetto al fuso orario italiano. Primo passaggio fondamentale è la conversione dell'orario in formato "Date time" e l'aggiunta di un'ora all'intera colonna.
Si considera a questo punto la necessità di visualizzare la distribuzione dei tweets che sono stati raccolti, evidenziando i momenti salienti della nottata.

url https://public.tableau.com/profile/matteo4985#!/vizhome/VisualizzazioeConteggioTweetsRVolumetweets



4.3 Feature selection

Dall'osservazione degli attributi nella fase di data exploration, si decide di eliminare le seguenti colonne ritenute ridondanti per l'analisi: 'Screen-name', 'Name', 'Followers', 'Number-retweet', 'Position'. Il dataset relativo ai punteggi, scaricati tramite web scraping, viene trattato eliminando colonne che presentavano unicamente valori nulli e non davano alcuna informazione aggiuntiva. Si rinominano inoltre delle colonne per rendere più comprensibile il contenuto di

riferimento.
Alla fine del processo si ottengono due Data frame con soli attributi utili per l'analisi.

4.4 Feauture extraction

In questa si fase si utilizzano gli attributi esistenti per poter estrarre nuove proprietà utili allo studio preposto.
In particolare si sfrutta il testo dei tweets per estrarre gli hashtags presenti, il risultato viene quindi inserito in una nuova colonna.
Si utilizzano inoltre i dataframe relativi a statistiche e punteggi scaricati per estrarre i seguenti attributi:
- jab (diretto sinistro) tirati tyson: tot dei jab tirati da Tyson
- jab tirati jones: tot dei jab tirati da Jones
- jab (diretto sinistro) atterrati tyson: tot dei jab andati a segno da Tyson
- jab atterrati jones: tot dei jab andati a segno da Jones
- diretti destro atterrati da Tyson: tot dei diretti destri atterrati da Tyson
- diretti destri atterrati da Jones: tot dei diretti destri atterrati da Jones
- Tot punch tirati tyson: totale dei pugni tirati da Tyson
- Tot punch tirati jones: totale dei pugni tirati da Jones
- CONSUM pugni tirati tyson: somma cumulata dei pugni atterrati da Tyson
- CONSUM pugni tirati jones: somma cumulata dei pugni atterrati da Jones
- CONSUM pugni landed tyson: somma cumulata dei pugni tirati da Tyson
- CONSUM pugni landed jones: somma cumulata dei pungi tirati da Jones

4.5 Text cleaning

Per ottimizzare il processo di estrapolazione delle opinioni ed assegnare ad ognuna di essa un valore numerico, si rielabora il testo in modo tale da performare al meglio i modelli di sentiment analysis. A tal proposito, si adottano le seguenti fasi, elencate in ordine di applicazione, per il trattamento del testo dei tweets scaricati:
- Rimozione di emoticons (occhi, naso, bocca, "smiles", simboli, pictographs, transport e map symbols, bandiere etc.)
- Normalizzazione del testo: trasformazione delle lettere da maiuscolo a minuscolo
- Rimozione di HTML tags
- Rimozione di mentions (@)
- Rimozione di numeri
- Rimozione di URLs
- Tokenization
- Rimozione di stop words
- Rimozione di punteggiatura
- rimozione "manuale" di espressioni (es: 'r', 'nd', 'nhttps', 'nstay' ecc.)
- Lemmatizzation

Per poter controllare al meglio il processo di preparazione del testo si sono visualizzate, dopo la rimozione di articoli e punteggiatura, le parole più frequenti nel testo in modo tale da poter monitorare e migliorare la qualità del processo.

4.6 Split

La fase del progetto chiamata “splitting” rappresenta la necessità di identificare, per ogni tweet scaricato, l’atleta ed il round di riferimento.
Si ottengono due dataframe distinti utilizzando una funzione in grado di identificare all’interno della colonna degli hasthtags una serie di parole chiave, opportunamente scelte, identificative del pugile: “Tyson”, “MikeTyson”, “Ironmike”, “tyson”, “Jones”, “Roy”, “RoyJones”, “roy”. Il fine della funzione è assegnare ad ogni tweets, in una colonna denominta ”Match parola”, la nomenclatura ”Tyson” o ”Jones” a seconda del fatto che la parola chiave identificata sia relativa ad uno dei due pugili.
Così facendo è possibile suddividere il dataset di partenza in due dataset distinti per combattente.
Si procede all’assegnazione di ogni tweet il round di riferimento. Si costruisce quindi una colonna “ROUND” per entrambi i data frame dei pugili, si individuano le fasce orarie relative al round di appartenenza e si aggiunge un minuto di scarto rappresentante il tempo di recupero in cui i due atleti riposano.
In particolare si riporta:
- 05:58:00 – 06:02:00 : 1 round
- 06:02:01 – 06:05:00 : 2 round
- 06:05:01 – 06:08:00 : 3 round
- 06:08:01 – 06:11:00 : 4 round
- 06:11:01 – 06:14:00 : 5 round
- 06:14:01 – 06:17:00 : 6 round
- 06:17:01 – 06:20:00 : 7 round
- 06:20:01 – 06:23:00 :8 round

Ad ogni tweet che appartiene ad una delle fasce orarie sopra indicate viene assegnato il valore del round di riferimento.
Per tutti i tweets che sono stati condivisi prima o dopo lo svolgimento del match, viene aggiunto valore ”null” alla colonna ROUND.

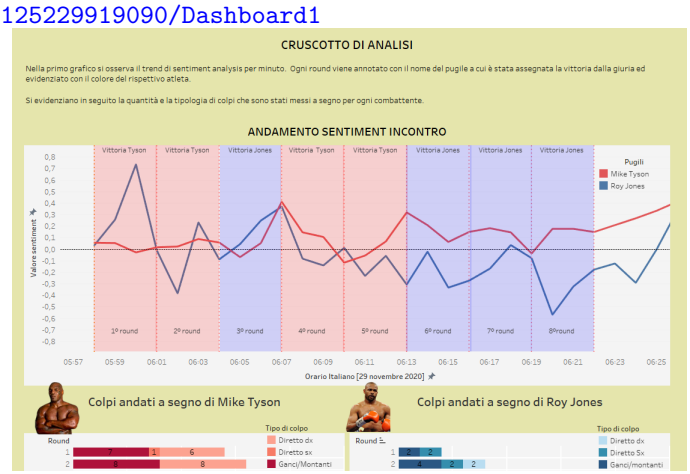
5 Data enrichment

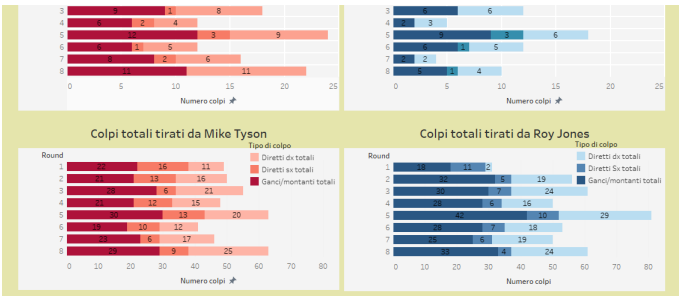
Nella fase di data enrichment si sfruttano le colonne relative al numero del round, costruite sia per il dataframe relativo a Tyson che per il dataframe relativo a Roy Jones. Si considera questo come elemento in comune, in modo tale che, tramite un LEFT OUTER JOIN, si possa andare ad arricchire ad ogni tweet il voto dei giudici e le performance (statistiche) del rispettivo round.
Si ottengono così due dataframe unici, con i dati dei tweet saricati affiancati ai valori statistici ed il risultato dei giudici per round e per atleta.

6 Sentiment Analysis

Per l'assegnazione di un valore al testo del tweets si utilizza il modello di sentiment analysis "afinn".
Afinn utilizza un approccio "Lexicon based approach", in cui ogni parola all'interno della lista dei vocaboli ha un valore che va da -5 a +5: ad ogni corpus di testo viene assegnato un punteggio relativo alla somma dei valori delle parole presente all'interno di esso.Il modello scelto, che si basa su un lessico di 2477 parole in lingua inglese già polarizzate, ha la peculiarità di modificare automaticamente il valore di una frase nel momento in cui trova costrutti lessicali come la presenza di negazione ('not') o presenza di avversativi (es: 'but'). Un valore complessivo maggiore di zero indica un tweets positivo, minore di zero un tweets negativo ed uguale a zero un tweets neutro.
Sulla base di quanto detto si procede all'assegnazione di un voto numerico al testo precedentemente processato.
Per potere effettuare delle analisi vengono calcolati per ogni round e per ogni atleta il valor medio di sentiment e la rispettiva deviazione standard.
Si costruisce infine un cruscotto di analisi, tramite utilizzo di Tableau, contenente una visualizzazione relativa ai valori medi di sentiment e le varie statistiche.

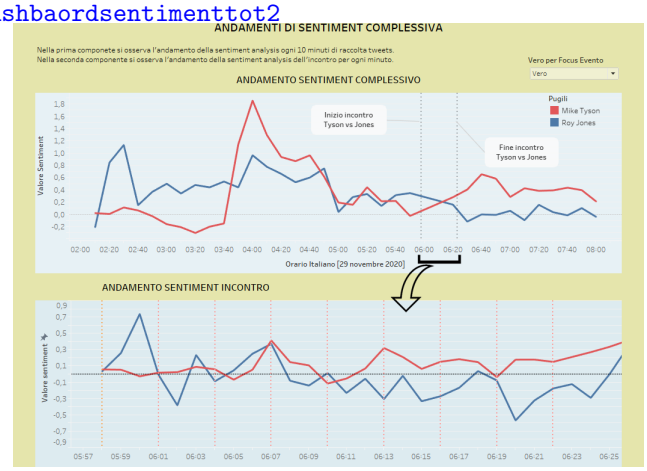
https://public.tableau.com/profile/matteo4985#!/vizhome/Analisidelmatch_16125229919090/Dashboard1





Si osservano inoltre i valori medi di sentiment analysis durante tutto la nottata mettendo a confronto i risultati dell'analisi, calcolati ogni 10 minuti, con i risultati ottenuti durante il match.

https://public.tableau.com/profile/matteo4985#!/vizhome/AndamentoSentimentComplessivo/dashbaordsentimenttot2



7 Conclusione

Dall'osservazione dei valori ottenuti tramite sentiment analysis è possibile osservare un tifo del pubblico altalenante per i primi cinque round. Nelle ultime riprese è invece possibile osservare una polarità prevalentemente positiva verso Mike Tyson, a discapito del suo avversario, il cui valor medio, in alcuni round, cade sotto la soglia della neutralità. Si osservano delle discrepanze tra l'assegnazione della vittoria e il voto dei giudici, in particolar modo nel primo round e negli utlimi tre.
Focalizzandosi sugli ultimi minuti di incontro si rilevano i seguenti risultati:
- il volume dei colpi tende ad essere maggiore per il pugile Roy Jones nonostante quelli andati a segno siano minori di quelli del suo avversario
- Osservando il diretto destro e il totale degli altri colpi, Tyson va a segno più volte rispetto al suo avversario
- negli ultimi tre round l'attività del jab (diretto sinistro) è quasi assente, in

totale Jones va a segno con solo due colpi mentre Tyson con tre

8 Criticità

Le criticità maggiori sono state:
- ottenere una visualizzazione intuitiva ed idonea allo scopo preposto
- scelta dell'algoritmo di sentiment
- gestione del fattore tempo

9 Sviluppi Futuri

Sviluppi futuri:
- infografica più interattiva
- estendere l'analisi osservando parti del corpo più colpite (dati non sempre reperibili)
- analisi delle aree del ring più coperte durante il match (es: corde, angoli, parte centrale ecc.)