INTER - MILAN
IL DERBY DELLA MADONNINA

Tommaso Cattaneo[1], Lorenzo Sasso[1]

[1]CdLM Data Science, Università degli Studi di Milano – Bicocca

Il derby di Milano, è la stracittadina calcistica che mette di fronte le due principali squadre di Milano, i nerazzurri dell'Inter e i rossoneri del Milan. Conosciuto anche come derby della madonnina, è una delle partite più attese e affascinanti della seria A, seguita in tutta Italia e nel resto del mondo. Storicamente il tifo per l'Inter era emanazione della borghesia cittadina, a differenza di quello del Milan, supportato a maggioranza dalle classi popolari.

Le tifoserie nella storia si sono sempre sfidate a colpi di cori e coreografie; inoltre, negli ultimi anni è scoppiata una vera e propria ”guerra” sui social network a colpi di tweets.

Quanti sono effettivamente questi tweets? Com'è il loro andamento durante la gara? In quanti e in quali stati viene seguito il derby di Milano? E chi sono i giocatori più twittati?

Questo progetto nasce proprio con l'intenzione di rispondere a queste domande tramite l'utilizzo di piattaforme e metodi per l'acquisizione e manipolazione dei dati.

ACQUISIZIONE DEI DATI

La prima parte di raccolta dei dati è avvenuta in tempo reale, scaricando tutti i tweets relativi al match e ad alcuni giocatori, pubblicati prima, durante e dopo la partita. Per estrarre soltanto i tweets di interesse, sono stati presi in considerazione gli hashtags che nel corso dei derby passati sono risultati essere quelli più utilizzati per menzionare la partita (#INTMIL, #derbydellamadonnina, #derbyMilano,#INTERMILAN) e quelli per menzionare alcuni giocatori delle due squadre. I giocatori presi in considerazione sono stati Padelli, Brozovic e Lukaku per l'inter, mentre per il Milan, Donnarumma, Rebic e Ibrahimovic. Successivamente, sono state raccolte le azioni salienti della partita[1](gol, occasioni da gol, parate) ed è stato creato un file contenente quest'ultime con il rispettivo minuto di gioco.

Infine, per effettuare un ulteriore confronto tra i calciatori, sono stati scaricati i dati[2] riguardanti le rispettive prestazioni in campo .

DESCRIZIONE DATASET

Il dataset relativo ai tweets del match (*match.csv*) si riferisce ad una fascia oraria che inizia alle ore 20.30 e termina intorno alle 22.50 circa; per ogni tweet è stato registrato il nome utente, il contenuto del post, l'orario, la localizzazione geografica e gli hashtags utilizzati.

[1] www.diretta.it

[2] https://wyscout.com/it/football-platform/

Gli stessi campi si ritrovano anche nei tweets relativi ai giocatori del Milan (*tweetMilan.csv*) e dell'Inter(*tweet_inter.csv*).

Il file relativo alla cronaca della partita (*CronacaPartita.csv*) contiene la descrizione dell'azione di gioco, l'orario (UTC+1 time), il minuto della partita e l'indicazione del tempo in cui c'è stata l'azione (1° o 2°).

Infine i file contenenti le statistiche dei calciatori indicano per ognuno di essi il numero di palle recuperate, di palle perse, i tiri effettuati, i dribbling, passaggi e gol fatti.

PROCESSING

La prima fase riguardante l'acquisizione in tempo reale dei tweets è stata sviluppata grazie all'utilizzo delle API di Twitter e della libreria *tweepy*, disponibile in Python, attraverso la quale viene reso possibile ottenere i tweets in tempo reale in base agli hashtags scelti.

Successivamente, sempre grazie ad alcune librerie disponibili in Python (KafkaProducer, KafkaConsumer), è stato possibile ricorrere all'utilizzo di Apache Kafka, una piattaforma open source di stream processing. Questa tecnologia utilizza delle applicazioni chiamate producer/consumer, le quali rispettivamente scrivono e leggono flussi di record su una particolare struttura chiamata topic.

Nel nostro caso abbiamo definito 3 producer differenti, all'interno di tre file in Python: uno per la partita (*acquisizioneTweetMatch.ipynb*), uno per i giocatori dell'Inter e uno per i giocatori del Milan. Ciascun producer ha scritto sul rispettivo topic i tweets acquisiti (*match_topic, milan_topic, inter_topic*). I file differiscono quindi solo per il nome del topic e gli hashtags utilizzati.

Il consumer di Kafka, invece, è stato implementato all'interno di un'altra piattaforma: Apache Nifi, un software progettato per automatizzare il flusso di dati tra i sistemi software. In questa struttura sono stati utilizzati 3 nodi *Consume Tweets,* utili per leggere i tweets scritti in precedenza sui differenti topics dai tre producers; i tweets ottenuti dal consumer, in formato JSON, sono stati processati attraverso il nodo *PutMongoRecords,* il quale fa uso della tecnologia MongoDB per la memorizzazione dei dati. MongoDB infatti è un DBMS non relazionale orientato ai documenti, il quale si allontana dalla struttura tradizionale basata su tabelle dei database relazionali, in favore di documenti in stile JSON. La tabelle che troviamo nei database relazionali sono sostituite dalle collezioni, mentre i records vengono definiti documenti. Ogni tweet viene quindi memorizzato in una delle tre collezioni differenti (match, milan, inter), a loro volta contenute all'interno del database *Twitter.*

Successivamente, grazie all'utilizzo della libreria *pymongo* di Python, è stato possibile accedere direttamente al database e alle collezioni in MongoDB. Queste sono state convertite in file CSV per permettere una manipolazione e un'integrazione dei dati più immediata (*DaMongoaCSV.ipynb*). I tre file CSV creati a seguito di questa fase sono rispettivamente *match.csv, tweetMilan.csv e tweet_inter.csv*.

Una volta creati i file CSV, è iniziata la fase di ispezione e pulizia del dataset. Le prime operazioni sono state convertire in formato *datetime* data e ora dei tweets e, successivamente, modificarne l'orario, in quanto questo faceva riferimento ad un fuso orario (UTC time) diverso da quello italiano (UTC+1 time).

L'analisi dei dati è iniziata dal file contenente i tweets del match (match.csv), dove è stato effettuato un raggruppamento in base all'orario e

conteggiato il numero di tweets al minuto relativi agli hashtags della partita.

Dopodichè, è stata eseguita un'integrazione con il file *CronacaPartita.csv,* unendo le due tabelle rispetto alla colonna relativa all'orario; nei minuti in cui non era presente alcuna azione saliente, il file è stato arricchito inserendo il minuto e il tempo di gioco, in modo da avere per tutto il periodo in cui sono stati raccolti tweets le stesse informazioni.

La nuova tabella a seguito di questa integrazione (*tweet_partita.csv*) contiene quindi l'orario, il minuto di gioco, il numero di tweets per minuto, la descrizione di un eventuale azione saliente e il tempo di gioco.

Questo file è stato utilizzato per fornire la prima rappresentazione grafica riguardo l'andamento dei tweets durante la partita, con l'obiettivo di rispondere alle prime due domande iniziali.

Sempre a partire dal file *match.csv,* è stato effettuato un altro raggruppamento, questa volta in base alla locazione geografica; è stato creato quindi un nuovo file (*LocationMatch.xlsx*) contenente la città e il rispettivo numero di tweets. Tutti i comuni sono stati raggruppati per provincia, mentre per le locazioni estere che facevano riferimento ad una nazione è stata presa in considerazione la capitale. Dal file sono state escluse tutte le locazioni inesistenti. Il risultato ottenuto è stato utilizzato per la seconda infografica utile per rispondere alla domanda: in quanti e quali stati è seguito il derby di Milano?

Infine, per rispondere alla domanda riguardo quali fossero i giocatori più menzionati, sono stati estratti dal file tweetMilan.csv e tweet_inter.csv i tweets dei giocatori di ogni squadra attraverso il codice Python progettoIntegrazioneInter2.ipynb e progettoIntegrazioneMilan2.ipynb.

A causa di un numero ridotto di tweets per i giocatori Padelli, Brozovic, Donnarumma e Rebic, sono stati presi in considerazione solo Lukaku e Ibrahimovic rispettivamente con gli hashtags [#Lukaku, #Romelu] e [#Ibrahimovic, #Zlatan, #izback].

I tweets di entrambi i giocatori sono stati raggruppati in base all'orario ed è stato conteggiato il numero di tweets al minuto. I due file contenenti queste informazioni sono rispettivamente *tweetIbraMinuto.csv* e *tweetLuka.csv* .

Dal momento che non in tutti i minuti è presente almeno un tweet, i due file sono stati arricchiti inserendo i minuti mancanti con il corrispondente numero di tweets uguale a zero.

LE INFOGRAFICHE

Le tre infografiche[3] utilizzate per rispondere alle domande poste inizialmente sono state sviluppate attraverso l'utilizzo di Tableau, un software specializzato nella data visualization.

- La prima infografica è servita per visualizzare la quantità e l'andamento dei tweets con riferimento al derby, registrati prima, durante e dopo la partita. Il numero totale di tweets raccolti è di circa 18mila. Come si può notare dalla visualizzazione, il loro andamento è fortemente condizionato dagli eventi che si verificano durante la partita; in particolare vi è una crescita sostanziale in corrispondenza dei gol segnati. Tuttavia, il picco più alto di tweets si è registrato al termine del match, quando, presumibilmente, sono iniziati i classici

[3] https://public.tableau.com/profile/lorenzo.sasso#!/

sfottò tra le tifoserie e sono stati pubblicati tweets relativi al risultato finale della partita.

- Per visualizzare dove e in quante nazioni è stato seguito il derby di Milano, si può far riferimento alla seconda infografica. I tweets registrati provengono da quasi 80 nazioni, collocate in 5 diversi continenti. Escludendo l'Italia (5468 tweets), il numero maggiore di tweets è stato rilevato in Indonesia. Questo risultato, in un paese apparentemente distante dal calcio italiano, si potrebbe spiegare dal fatto che nel 2013 il 70% delle quote della società Inter sono state acquistate dall'indonesiano Erick Thohir, rivendute successivamente nel 2016. Questo fatto ha contribuito ad incrementare la passione e l'interesse dei cittadini indonesiani nei confronti della società nerazzurra e del calcio italiano. Inoltre, nello stesso Paese si registra anche una forte presenza di tifosi rossoneri, circa 30mila[4].

  Un numero importante di tweets è stato rilevato anche in Arabia Saudita, segno che le ultime due edizioni della supercoppa italiana, disputate rispettivamente a Gedda e Riad (2018 e 2019), sono riuscite a promuovere il calcio italiano nello stato asiatico.

  Oltre a ciò, negli ultimi anni entrambe le società hanno fatto importanti investimenti per diffondere i propri colori nel mondo. Si contano scuole calcio di Milan e Inter rispettivamente in 25[5] e 20[6] paesi in tutto il mondo.

- La terza e ultima infografica realizzata mette a confronto i giocatori simbolo delle due squadre e i più menzionati sui social: Ibrahimovic e Lukaku. Per fare ciò, sono state registrate le statistiche più significative riguardo i due attaccanti come gol, palle perse e recuperate, tiri effettuati (*ibra.xlsx, Luka.xlsx*) e si è fornita una rappresentazione grafica dell'andamento dei tweets con riferimento agli hashtag utilizzati per i due giocatori. Il numero di tweets rilevati presentano una netta differenza in termini numerici nonostante entrambi abbiano segnato un gol nella partita. Ibrahimovic (845 tweets) rispetto a Lukaku (253 tweets) gode di maggiore fama sui social con circa sei milioni di followers su Twitter, contro i quasi due milioni di Lukaku.

Successivamente, le tre infografiche sono state sottoposte ad una valutazione euristica. A cinque utenti è stato chiesto di commentare ad alta voce le infografiche. L'output di questa valutazione è una lista di problemi che forniscono una serie di spunti validi per ottenere un miglioramento al lavoro svolto (*Valutazione euristica.docx*).

In seguito si è passato agli user tests. Sei utenti hanno risposto a tre domande (una per infoviz) attraverso l'interazione con le infografiche. Sono stati registrati i tempi di esecuzione di ogni test e il tasso di errore nelle risposte[7].

Infine, è stato somministrato un questionario psicometrico a 28 persone che, attraverso la scala

[4] milanreporter.it/milanisti-indonesia/
[5] acmilan.com/en/academy
[6] inter.it/it/interacademy
[7] public.tableau.com/profile/lorenzo.sasso#!/vizhome/usertests/Dashboard1

Cabitza-Locoro, hanno effettuato una valutazione qualitativa delle infografiche[8].

CONCLUSIONI

L'analisi svolta ha evidenziato l'importanza del calcio italiano a livello nazionale e internazionale, vista la grande quantità di tweets ricavati e i molteplici Stati in cui questi sono stati registrati. Questo risultato è segno che gli investimenti effettuati negli ultimi anni dalle varie società e dalla Lega Calcio, hanno portato ad un incremento sostanziale dell'interesse mondiale verso il campionato italiano.

Viceversa, le menzioni dei calciatori sui social network non sono state così rilevanti come quelle riguardanti il match, nonostante Ibrahimovic e Lukaku fossero i giocatori più seguiti delle due squadre su Twitter.

Infine, in base alla provenienza dei tweets raccolti le società potrebbero programmare investimenti in aree come Africa centrale, paesi dell'Europa orientale, dove è stato riscontrato poco seguito del derby sui social e quindi si suppone poca conoscenza del calcio italiano.

8 public.tableau.com/profile/lorenzo.sasso#!/vizhome/test_15834082587520/Dashboard1