

# Analisi dei tweet durante le NBA Finals 2019 e valutazione sentimento dei tifosi

Matteo Gaverini, Federico Manenti, Paolo Mariani

## Sommario

I social nel XXI secolo sono ormai diventati parte integrante della vita di una persona: Twitter, Facebook, Instagram sono app che vengono aperte molte volte durante la giornata. Tra tutti Twitter è quella che sta attraversando un cambiamento radicale per quanto concerne il suo uso globale. Oggi l'app viene sempre di più utilizzata da figure di rilievo come politici, capi di stato per commentare non solo articoli o notizie ma anche per comunicare decisioni importanti quali accordi e trattati internazionali (es. Donald Trump). Oltre a questo Twitter sta cambiando la user experience degli eventi sportivi live; oggi i tifosi commentano sempre di più le prestazioni in tempo reale di un giocatore piuttosto che di una squadra e quindi i tweet possono diventare dei veri e propri indicatori del sentimento (umore) delle tifoserie. Il progetto si focalizza sull'analisi dei tweet registrati in real-time di uno degli eventi sportivi più seguiti in America subito dopo il Super Bowl NFL: le NBA Finals (2019). Questi tweet, una volta memorizzati in un database, sono stati utilizzati per valutare il sentimento dei tifosi durante la serie e integrati con le statistiche delle squadre in real time per individuare o meno una qualche relazione o dipendenza tra sentimento e le prestazioni della franchigia. Per ultimo sono state confrontate le statistiche dei 6 giocatori più rappresentativi per le due squadre nelle Finals 2019 e nella Regular Season 2018-19.

## Indice

<b>Introduzione</b>	
<b>1 Architettura</b>	
1.1 Raccolta dati . . . . .	2
1.2 Archiviazione . . . . .	3
<b>2 Dati aggiuntivi</b>	
2.1 Statistiche giocatori . . . . .	4
2.2 Posizioni geografiche . . . . .	5
<b>3 Operazioni effettuate</b>	
3.1 Sentiment Analysis . . . . .	5
3.2 Merging . . . . .	6
3.3 Data Integration . . . . .	6
3.4 Conteggio Hashtag . . . . .	8
3.5 Conteggio Retweet . . . . .	8
<b>4 Risultati</b>	
4.1 Risultati generali . . . . .	9
4.2 Risultati specifici . . . . .	11
4.3 Confronto stats giocatori . . . . .	13
<b>5 Conclusioni</b>	15

Questi possono essere etichettati con l'uso di uno o più *hashtag*, parole o combinazioni di esse concatenate precedute dal simbolo # attraverso il quale l'utente crea un collegamento ipertestuale a tutti i messaggi che citano lo stesso hashtag [1]. Twitter negli ultimi anni sta attraversando un periodo di declino a causa dell'aumento di abbandoni del social: l'azienda stessa ha dichiarato che negli ultimi mesi del 2018 ha perso 5 milioni di utenti [2]. Questo fenomeno però ha portato ad un cambiamento radicale del servizio: grazie alla modifica della struttura dei tweet e all'aggiornamento del design dell'esperienza utente il social sta guadagnando sempre più popolarità tra le squadre dei principali eventi sportivi [3] (basket, calcio, football americano, baseball). Squadre di basket blasonate come i Los Angeles Lakers o i Golden State Warriors possiedono infatti un proprio account Twitter, ed ciascuna di esse possiede degli hashtag personalizzati da utilizzare durante le partite per riportare fatti salienti (live-tweeting) [3]. Di conseguenza sempre più utenti commentano e giudicano le prestazioni dei loro beniamini durante gli eventi sportivi. Il progetto si incentra quindi sulla raccolta ed analisi dei tweet per valutare il sentimento dei tifosi in relazione alle statistiche delle squadre durante le NBA Finals 2019.

Le NBA Finals[4] rappresentano la fase conclusiva del campionato di basket americano della NBA (National Basketball Association) in cui al meglio delle sette par-

## Introduzione

Twitter è un servizio di notizie e microblogging fornito dalla società Twitter, Inc. con il quale gli utenti postano e interagiscono con messaggi chiamati *tweet*.

tite viene consegnato il famoso trofeo *Larry O'Brien Championship Trophy* (si veda Figura 1) alla squadra vincente. Le squadre che si sfidano in questa serie sono le vincitrici della Western Conference e della Eastern Conference. Le due franchigie protagoniste del 2019 sono state Golden State Warriors (West) e Toronto Raptors (East) che, al contrario dei pronostici, ha conquistato il titolo. Per concludere sono state confrontate le statistiche dei 6 giocatori più rappresentativi delle due franchigie durante le Finals 2019 e la Regular Season 2018-19 per scoprire se effettivamente gli atleti migliorano le loro prestazioni nella fase conclusiva della stagione.



**Figura 1.** I Toronto Raptors festeggiano la loro prima vittoria alzando il trofeo al cielo

## 1. Architettura

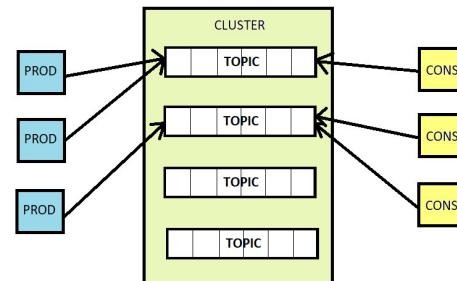
### 1.1 Raccolta dati

L'architettura implementata per raccogliere i dati prevede l'uso di tre strumenti fondamentali: un sistema di messaggistica (*Kafka*), due API (*Twitter Api* [5], *All Sports Api* [6]) e due librerie di Python (*kafka*, *tweepy*) per integrare le componenti.

#### Kafka

Kafka è una piattaforma streaming distribuita che permette di archiviare ed elaborare flussi di record in tempo reale. All'interno di Kafka agiscono principalmente 3 attori: *producer*, *consumer* e *il cluster*[7]. Il primo si occupa di pubblicare e scrivere messaggi sui *topic*<sup>1</sup>, il secondo è incaricato di leggere i messaggi da un *topic* a cui si è iscritto, mentre il terzo utilizza tecniche di *message queuing* per organizzare i messaggi nei vari *topic*.

<sup>1</sup>topic: stream di messaggi di un certo tipo



**Figura 2.** Architettura kafka

#### API

Le API utilizzate per la raccolta dei dati sono Twitter API[5] e AllSports API[6]. Il primo servizio è stato utilizzato per ottenere tutti i tweet durante l'orario temporale definito (1 ora prima dell'inizio della partita fino a 1 dopo la fine), ed è stato stabilito di filtrare i tweet in base a due parametri: la lingua e gli hashtag. Per la lingua si è deciso di ottenere solo i tweet in lingua inglese mentre per gli hashtag si è deciso di filtrare i tweet in base agli hashtag ufficiali delle due squadre oltre a quelli della NBA (si veda Tabella 1). La seconda API invece è servita per ottenere le live stats delle due squadre durante le partite.

	Hashtag
<b>Golden State Warriors</b>	DubNation, WarriorsGround, StrengthInNumbers, GoldenStateWarriors
<b>Toronto Raptors</b>	WeTheNorth, Raptors, TorontoRaptors, RTZ
<b>NBA</b>	NBAFinals, NBAPlayoffs, GswvsTor, TorvsGsw

**Tabella 1.** Hashtag utilizzati per la ricerca

#### Librerie Python

Le principali librerie utilizzate per la raccolta dei dati sono *kafka* [8] e *tweepy* [9]. La libreria *kafka* è servita per implementare la struttura kafka in Python mentre la seconda per accedere alle API di Twitter.

L'architettura globale si incentra sulla realizzazione di tre script Python per gestire lo stream dei dati. Il primo, chiamato *richiesta\_stats*, effettua una richiesta ad AllSport API ogni 30 secondi, filtra le statistiche rilevanti e genera una struttura-dizionario JSON (poi assegnata ad una variabile nel database di IPython). In fase di creazione del dizionario si verificano due casi particolari: nell'ora precedente all'inizio del match saranno contenuti dei valori nulli (non ancora generati in quanto la partita non ha ancora prodotto statistiche), mentre nell'ora successiva conterrà le ultime statistiche registrate al termine della partita. I campi salvati nell'oggetto stats sono:

1. *event\_date*: data della partita
2. *event\_key*: id univoco dell'evento
3. *time*: momento in cui viene effettuata la richiesta
4. *event\_status*: minuto della partita
5. *final\_score*: punteggio aggiornato live
6. *scores*: punteggi dei vari quarti
7. *statistics*: statistiche relative alla partita e ai quarti

Il secondo script, chiamato *Final\_producer*, ottiene i tweet in formato JSON, filtra i campi rilevanti, aggiunge ad ogni tweet il risultato ottenuto in *richiesta\_stats* e infine scrive sul topic l'oggetto finale (tweet + statistiche); nel caso l'oggetto restituito da Twitter sia un retweet si salva, nel campo *text*, il testo originale (quando l'utente non scrive nessun commento durante la fase di retweet), altrimenti viene memorizzato il commento stesso. I campi mantenuti nell'oggetto tweet sono i seguenti:

1. *\_id*: id univoco tweet generato da Twitter
2. *n\_tweet*: numero tweet
3. *time\_tweet*: ora tweet
4. *date\_tweet*: data tweet
5. *source*: device utilizzato per tweettare
6. *retweet\_count*: numero retweet del tweet
7. *placeTweet*: coordinate provenienza tweet
8. *placeUser*: luogo in cui si è iscritto a Twitter l'utente
9. *text*: testo del tweet
10. *hashtag*: hashtag contenuti nel tweet

Il campo *n\_tweet* è stato utilizzato solo come controllo dell'ordine di salvataggio degli oggetti, successivamente verrà eliminato. Dopo la raccolta dati si è notato che il campo *place\_tweet* risulta quasi sempre vuoto, per cui anch'esso è stato rimosso. Infine se il tweet è un retweet viene aggiunto il campo *retweetId* corrispondente all'id del tweet originale.

Il terzo script, chiamato *Final\_consumer*, si occupa di leggere le informazioni contenute nel topic su cui sta scrivendo il producer e in seguito di dividere i tweet provenienti dalle diverse tifoserie.

Per decretare quale fosse la squadra sostenuta dall'utente si è considerato che i tifosi solitamente utilizzano solo gli hashtag della propria franchigia mentre chi commenta la partita in modo più generale usa hashtag della NBA oppure inserisce quelli delle due squadre indifferentemente (si veda Tabella 1). Per questo motivo si è deciso di dividere i tifosi in questa maniera:

- **Tifoso Golden State Warriors** se il tweet contiene esclusivamente hashtag associati ai Golden State Warriors o contemporaneamente hashtag relativi alla squadra e alla lega NBA
- **Tifoso Toronto Raptors** se il tweet contiene esclusivamente hashtag relativi a Toronto Raptors o contemporaneamente hashtag della squadra e della lega NBA
- **Tifoso Neutro** in tutti gli altri casi

## 1.2 Archiviazione

Per archiviare gli oggetti JSON (tweet + stats) letti dal consumer si è deciso di utilizzare *MongoDB* [10], un database non relazionale (NoSQL) di tipo documentale. Il motivo di tale scelta è legato principalmente a due fattori: il primo è la flessibilità in quanto la struttura dei tweet e delle statistiche non è fissa; infatti se l'oggetto tweet fosse un retweet verrebbe aggiunto un ulteriore campo rispetto a quelli già definiti. Il secondo fattore è che sia i tweet che le stats sono già in formato JSON e quindi è molto semplice memorizzare gli oggetti senza modificare la loro struttura nelle collection del db.

Il database realizzato si chiama *test* ed è costituito da 3 collection:

1. *GoldenState*
2. *Toronto*
3. *Neutro*

Il database occupa 1.13 GB su disco, ma la dimensione totale dei documenti supera i 6 GB. La prima collection contiene i tweet di tutte le gare dei tifosi di GoldenState, la seconda i tweet dei tifosi dei Toronto Raptors e infine la terza i tweet dei tifosi considerati neutrali. Oltre a questa struttura si avrebbe anche potuto utilizzare una collection per gara. Tra le due però si è preferito usare la prima. L'archiviazione viene effettuata sfruttando la libreria Python *PyMongo* [11]. Essa permette in maniera molto semplice e intuitiva di interfacciarsi con MongoDB. L'operazione vera e propria di archiviazione viene effettuata dal consumer: una volta che ha letto l'oggetto dal topic e ha capito la

squadra per cui simpatizza l'utente, accede al db `test` e attraverso un'operazione di `insert_one` salva l'oggetto nella collection corrispondente. L'oggetto finale che ogni volta viene salvato nella collection presenta la seguente struttura:

```
{
  _id: <str>,
  n_tweet: <int>,
  time_tweet: <dt.time>,
  date_tweet: <dt.date>,
  source: <str>,
  retweet_count: <int>,
  placeTweet: <str>,
  placeUser: <str>,
  retweet_id: <str>, (non sempre presente),
  text: <str>,
  hashtag: [<str>, <str>, ...],
  stats: [event_date: <date>,
    event_key: <str>,
    time: <dt.time>,
    event_status: <str>,
    final_score: <str>,
    scores: [1stQuarter: [
      {score_home: <int>,
       score_away: <int>}],
      ...
      4thQuarter: [
        {score_home: <int>,
         score_away: <int>}],
      ],
    statistics: [Match:
      [type: <str>,
       home: <int>,
       away: <int>, ...],
      1stQuarter:
      [type: <str>,
       home: <int>,
       away: <int>, ...],
      ...
      4thQuarter:
      [type: <str>,
       home: <int>,
       away: <int>, ...]
    ]
  ]
}
```

**Listing 1.** Struttura finale oggetto salvato in MongoDB

## 2. Dati aggiuntivi

Oltre ai tweet e alle stats live ottenute durante le partite, per il progetto si è reso necessario ricavare dati da altre due fonti.

### 2.1 Statistiche giocatori

La prima fonte ([Basketball Reference](#)) è servita per ottenere le statistiche dei giocatori delle due franchigie nella Regular Season 2018-19 e nelle Finals 2019; non si è considerato tutto il roster<sup>2</sup> per la valutazione ma solamente i 6 giocatori più rappresentativi per squadra:

- **Golden State Warriors:** DeMarcus Cousins, Stephen Curry, Draymond Green, Andre Iguodala, Klay Thompson, Kevin Durant
- **Toronto Raptors:** Marc Gasol, Serge Ibaka, Kawhi Leonard, Kyle Lowry, Pascal Siakam, Fred Vanvleet

Le statistiche sono state ottenute scaricando due file csv per ogni giocatore: il primo contiene le stats di tutte le partite di Regular Season (si veda Figura 3) mentre il secondo quelle delle Finals (si veda Figura 4). I campi più rilevanti sono *G* (numero della partita), *PTS* (punti segnati), *AST* (assist), *TRB* (rimbalzi totali), *3P%* (percentuale dei tiri da 3), *FG* (canestri segnati) e *FGA* (tiri tentati).

	G	PTS	AST	TRB	3P%	FG	FGA
0	1.0	32	9	8	0.556	11	20
1	2.0	31	8	4	0.556	13	24
2	3.0	30	6	4	0.375	10	23
3	4.0	29	8	4	0.462	11	18
4	5.0	51	3	4	0.688	15	24

**Figura 3.** Overview statistiche di un giocatore nella Regular Season

	G	PTS	AST	TRB	3P%	FG	FGA
0	1	34	5	5	0.444	8	18
1	2	23	4	3	0.300	6	17
2	3	47	7	8	0.429	14	31
3	4	27	6	4	0.222	9	22
4	5	31	7	8	0.357	10	23

**Figura 4.** Overview statistiche di un giocatore nelle Finals

<sup>2</sup>roster: lista dei giocatori che fanno parte di una squadra

## 2.2 Posizioni geografiche

La seconda sorgente ([simplemaps](#)) è servita sia per integrare informazioni aggiuntive sulla posizione dei tweet che per risolvere problemi di sintassi sulle città contenute nel campo *PlaceUser*. Da questa fonte si sono ottenuti 3 csv:

1. citiesUS.csv

2. citiesCA.csv

3. worldcities.csv

Il primo csv (si veda Figura 5) contiene 37842 città americane per cui sono associati ad ognuna 16 campi, tra tutti sono considerati più rilevanti *state\_id* (sigla stato americano), *state\_name* e *population*.

	city	state_id	state_name	population
0	Prairie Ridge	WA	Washington	NaN
1	Edison	WA	Washington	NaN
2	Packwood	WA	Washington	NaN
3	Wautauga Beach	WA	Washington	NaN
4	Harper	WA	Washington	NaN

**Figura 5.** Overview citiesUS.csv

Il secondo (si veda Figura 6) contiene 247 città canadesi con 9 campi, tra questi i più importanti risultano essere *iso2* (codice a 2 lettere che indica lo stato), *admin* (provincia) e *population*.

	city	admin	iso2	population
0	Toronto	Ontario	CA	5213000
1	Montréal	Québec	CA	3678000
2	Vancouver	British Columbia	CA	2313328
3	Ottawa	Ontario	CA	1145000
4	Calgary	Alberta	CA	1110000

**Figura 6.** Overview citiesCA.csv

Il terzo csv (si veda Figura 7) invece contiene 12959 città del resto del mondo dove ad ognuna sono associati 11 campi. I più importanti risultano essere *country*, *iso2*, *iso3* (codice a 3 lettere che indica lo stato), *admin\_name* (provincia/regione) e *capital* (principale/amministrativa).

	city	country	iso3	iso2	admin_name	population
0	Malishevë	Kosovo	XKS	XK	Malishevë	NaN
1	Prizren	Kosovo	XKS	XK	Prizren	NaN
2	Zubin Potok	Kosovo	XKS	XK	Zubin Potok	NaN
3	Kamenicë	Kosovo	XKS	XK	Kamenicë	NaN
4	Viti	Kosovo	XKS	XK	Viti	NaN

**Figura 7.** Overview worldcities.csv

## 3. Operazioni effettuate

Dopo aver memorizzato i tweet e le stats di tutte le partite disputate, si è deciso di "appiattire" i documenti nel db, eliminare gli attributi ritenuti superflui o ridondanti e infine creare un csv per ciascuna collezione. Il motivo di questa scelta è legato a due fattori: il primo è che Tableau (software usato per visualizzare i risultati del progetto) richiede che i dati siano preferibilmente in formato csv, il secondo è che la struttura di un file csv è più facile da gestire e manipolare in Python rispetto a JSON. Per eseguire tali operazioni si è utilizzato lo script Python *Preprocessing* che ha permesso di generare 3 file: *tweet\_gsw.csv*, *tweet\_toronto.csv* e *tweet\_neutro.csv*. Successivamente si è effettuata l'operazione di sentiment analysis.

### 3.1 Sentiment Analysis

La *Sentiment Analysis* [12] è il campo del *NLP*<sup>3</sup> che permette l'identificazione, l'estrazione e la valutazione di opinioni da un testo. Lo strumento utilizzato per calcolarla in questo progetto è *VADER* [13] una libreria python "lexicon e ruled-based" particolarmente utile per l'analisi dei social in quanto considera anche le emoji che spesso sono utilizzate per esprimere un sentimento. Applicando l'algoritmo di VADER a un testo si ottengono 4 risultati: *pos*, *neg* e *neu* che stanno ad indicare quanto sia positivo, negativo o neutro la stringa analizzata (i valori sono compresi tra 0 e 1) mentre il *compound* è un indicatore più generale della valutazione che varia da -1 (molto negativo) a +1 (molto positivo). Nel progetto si è scelto di usare quest'ultimo come metrica di valutazione del sentimento. Durante le prime analisi per verificare la veridicità dell'algoritmo sono sorti due problemi. Il primo è relativo all'utilizzo di parole provenienti dal gergo ceстistico con accezione positiva, ad esempio "fire" (es. "Leonard tonight is on fire!!"), che nel vocabolario di VADER hanno associato un punteggio negativo. Il secondo problema riguarda invece le emoji: VADER può gestirle, ma non ha un vocabolario apposito in cui ad ognuna di esse è associato un valore, bensì

<sup>3</sup>NLP: Elaborazione linguaggio naturale

	text	vader	blob
Draymond Green has a 69.2% chance to see more rebounds than any player during the #NBAFinals - he also has a 68.3% chance to see more assists than any other player. \n\nHe has a 12.5% chance to win the Finals MVP Award as well.\n\n#StrengthInNumbers \n#NBAPlays	Let's fucking go #DubNation	0.0000	-0.600000
Steph Curry walks into Scotiabank Arena ahead of @warriors/@Raptors Game 1! #StrengthInNumbers \n\n: GSW/TOR, Game 1\n\n: 9:00pm/et \nus: ABC ca: Sportsnet https://t.co/NttQJq3NGJ		0.9246	0.300000
I'm shimmying for #TeamSteph! Do you think Steph can steal the show vs Kawhi? Choose your team for a reminder to watch Game 1 during the NBA Finals on ABC at 9 PM ET. @ESPN NBA #NBAFinals #NBAPlays #DubNation		-0.5411	-0.400000
STEPHEN CURRY: ORIGIN STORY\n\n"It's like a dream come true..." \n\nGame 1 of the #NBAFinals presented by @YouTubeTV tips TONIGHT at 9pm/et on ABC & Sportsnet! #StrengthInNumbers https://t.co/nQzHMDepCT		0.7644	-0.075000
The legend of the Warriors dynasty grows. I'm taking the Golden State Warriors to be #NBAFinals champs. #DubNation Make your pick 🎲 with @budweiserusa for a chance to win a limited-edition Budweiser Championship Bottle. #sweepstakes		0.8885	0.550000

**Figura 8.** Confronto tra VADER e TextBlob

traduce il simbolo con la sua descrizione testuale (es. : red heart) e poi calcola il punteggio. Se però le parole, come nell'esempio "red heart" non hanno corrispondenza nel dizionario o sono valutate neutre non influiscono sul valore del sentiment. Per questo motivo si è deciso di arricchire il vocabolario o modificare il valore di alcune parole che con alcune analisi esplorative sembravano più influenti. Prima di modificarlo però si è controllato anche che nei tweet la parola presa in considerazione non fosse stata usata sia con accezione negativa che positiva. Riferendosi all'esempio precedente, "heart" e "fire", si è constatato che ogni volta appaiono solo con accezione positiva e quindi il loro valore è stato modificato.

Un'altra libreria presa in considerazione è stata **TextBlob**, ma come si può osservare nella Figura 8 i risultati ottenuti con VADER si rivelano migliori. Il sentiment è calcolato nel notebook *Sentiment\_final* e prevede 4 fasi:

1. *pulizia tweet*: vengono rimossi tutti i simboli e parti di frasi che non influenzano il calcolo del sentiment (es. numeri e link a siti web)
2. *tokenization*: divisione dei tweet in singole parole (token) e rimozione delle *stopwords*<sup>4</sup>
3. *ricomposizione*: il tweet pronto per l'analisi viene ricomposto dopo la fase precedente
4. *calcolo compound*

### 3.2 Merging

Una volta effettuata la sentiment analysis, si ottengono 3 file csv: *final\_gsw.csv*, *final\_tor.csv* e *final\_neutro.csv*.

A questo punto con lo script *Tweet\_merger* si effettuano tre ulteriori operazioni. La prima consiste nel concatenare i tre file csv in modo da ottenerne uno solo (*final\_totali.csv*). La seconda operazione prevede di aggiungere per ogni riga al csv ottenuto in precedenza, il team con cui è stato catalogato il tweet e la gara. Infine la terza operazione include sia la sistematizzazione del campo *source* (estrarre la stringa contenuta nel tag HTML) che la modifica della posizione del punteggio e delle stats in maniera tale che i dati relativi a Golden State precedano sempre quelli di Toronto. Il motivo per cui si svolge questo iter è sempre legato alla esemplificazione della gestione dei dati in Tableau che risulta molto "rigido" nella struttura del dato richiesto per la visualizzazione.

A seguito del processo appena descritto ne è risultato un file contenente 1 305 789 tweet (*final\_totali.csv*), a cui è stata applicata una procedura di data integration per risolvere i problemi di sintassi delle città contenute nel campo *placeUser*. Da questo file sono stati creati tutti i csv necessari per la fase di Data Visualization.

### 3.3 Data Integration

La data integration è un processo che consiste nel combinare dati provenienti da sorgenti diverse e fornire una visione unitaria del dato [14]. In altre parole dati n record che si riferiscono allo stesso oggetto si vuole ottenere un'unica istanza che contenga informazioni aggiuntive (Data Quality Improvement). Quando si parla di data integration in generale ci si riferisce a due operazioni: schema matching e instance matching. Nel primo, dati due db che hanno schemi diversi, si cerca di unificare le strutture per ottenerne una sola; il secondo invece è a livello di istanza (in generale si

<sup>4</sup>stopwords: articoli, preposizioni...

parla di record linkage). Esistono 4 diverse tecniche di record linkage:

1. empirical
2. probabilistic
3. knowledge based
4. mix

Il primo prevede che il matching avvenga quando due parole sono molto simili in termini di simboli alfabetici (Es. Carlo e Crlo), il secondo si basa su tecniche statistiche (campioni e funzione di distanza), il terzo applica delle regole di dominio per il matching e infine l'ultimo è una combinazione tra tecniche probabilistic e knowledge based. Nel progetto lo script utilizzato per eseguire il task si chiama *sistemazione\_pos* e la tecnica adottata per il record linkage è di tipo mix. L'obiettivo preposto è individuare le città contenute nel campo *placeUser* e aggiungere informazioni aggiuntive come provincia, nazione, iso3, iso2. Prima di effettuare qualsiasi operazione alle istanze si sono applicate delle tecniche di preprocessing sulle stringhe contenute nel campo *placeUser*. La prima operazione è stata quello di dividere le stringhe in 4 parti usando come separatore la virgola, in questo modo ogni volta si crea un nuovo campo ad ogni occorrenza trovata (si veda Figura 9).

	place_user	info1	info2	info3	info4	city
0	Tehran	Tehran	nan	nan	nan	None
1	United States	United States	nan	nan	nan	None
2	Washington, DC	Washington	DC	nan	nan	Washington
3	Noisy-le-Roi, France	Noisy-Le-Roi	France	nan	nan	None
4	San Diego, CA	San Diego	CA	nan	nan	San Diego

**Figura 9.** Splitting stringhe.csv

Successivamente si sono eliminati gli spazi inutili presenti prima/dopo le parole e poi si è normalizzata la stringa contenuta nel campo *info1* riscrivendola in un unico formato con la funzione *str.title()* (Esempio San DIEGO → San Diego, LOS ANGELES → Los Angeles). Il motivo di tale scelta è che nella maggior parte dei casi le stringhe presenti nel campo *placeUser* sono costituite da due parole separate da una virgola dove la prima stringa indica sempre la città mentre l'altra contiene la provincia/regione (se la città è americana o canadese) oppure la nazione. A questo punto si procede con il riconoscimento delle città; per far ciò si utilizza una libreria Python *GeoText* che applicata al campo *placeUser* consente di estrarre le città presenti. Una volta applicata la libreria, si procede ad

integrare le informazioni per le città riconosciute. Per far ciò si utilizzano in sequenza 3 file csv già definiti nel Capitolo 2: *citiesUS.csv*, *citiesCA.csv* e *worldcities.csv*. Con il primo si effettua un inner join tra le città individuate da *GeoText* e quelle presenti nel csv (si veda Figura 10), dopodiché il risultato viene filtrato controllando il campo *state\_id* (sigla stato americano) e *info2*: se i due valori coincidono il matching è corretto e vengono inserite le informazioni aggiuntive, altrimenti viene scartato.

	place_user	info1	info2	city	state_id	state_name	iso3
0	Washington, DC	Washington	DC	Washington	VA	Virginia	USA
1	Washington, DC	Washington	DC	Washington	DC	District of Columbia	USA
2	Washington, DC	Washington	DC	Washington	WV	West Virginia	USA
3	Washington, DC	Washington	DC	Washington	NJ	New Jersey	USA
4	Washington, DC	Washington	DC	Washington	TX	Texas	USA

**Figura 10.** Risultato join con citiesUS.csv

Successivamente si effettua un inner join tra le città escluse dal merge precedente e quelle presenti nel csv *citiesCA.csv* (si veda Figura 11).

	place_user	info1	info2	city	state_name	iso3
0	Edmonton, Alberta Canada	Edmonton	Alberta Canada	Edmonton	Alberta	CAN
1	Edmonton, Alberta	Edmonton		Alberta	Edmonton	Alberta
2	Edmonton, Alberta	Edmonton		Alberta	Edmonton	Alberta
3	Edmonton, Alberta	Edmonton		Alberta	Edmonton	Alberta
4	Edmonton, Alberta Canada	Edmonton	Alberta Canada	Edmonton	Alberta	CAN

**Figura 11.** Risultato join con citiesCA.csv

Alla fine di questa seconda operazione si realizza un ulteriore inner join tra le città rimanenti e quelle presenti nel csv *worldcities.csv* (si veda Figura 12).

	place_user	info1	info2	city	country	state_name	iso3
104669	Kuching	Kuching	nan	Kuching	Malaysia	Sarawak	MYS
104670	Laurinburg, NC	Laurinburg	NC	Laurinburg	United States	North Carolina	USA
104671	Volgograd	Volgograd	nan	Volgograd	Russia	Volgogradskaya Oblast	RUS
104672	Sahuarita, AZ	Sahuarita	AZ	Sahuarita	United States	Arizona	USA
104673	Odense, Danmark	Odense	Danmark	Odense	Denmark	Syddanmark	DNK

**Figura 12.** Risultato join con citiesWORLD.csv

In questa situazione per validare il matching si effettuano prima due operazioni: la prima è creare una lista contenente tutte le informazioni aggiuntive della città ottenute dal join con il csv mentre la seconda è controllare se all'interno di *placeUser* esiste almeno una parola contenuta nella lista creata: se è così il matching viene considerato corretto altrimenti viene scartato. Il motivo è dato dal fatto che una città non necessariamente si trova in una sola nazione (si veda Figura 13) di conseguenza il join restituisce più di un risultato.

	place_user	city	country	iso2	iso3	state_name
2681	San Francisco, California	San Francisco	Argentina	AR	ARG	Córdoba
2682	San Francisco, California	San Francisco	United States	US	USA	California

**Figura 13.** Esempio disambiguazione città

A questo punto si effettua un ultimo inner join tra le città rimanenti e quelle contenute nell'unione dei csv *citiesUS* e *citiesCA*. Il risultato finale è che a buona parte delle città individuate da GeoText si è riusciti ad aggiungere le informazioni, quelle rimanenti invece vengono ignorate.

Nello step successivo si procede ad individuare le città contenute nel campo *placeUser* che non sono state riconosciute da GeoText in quanto mancanti nel vocabolario oppure scritte in modo errato. Per far ciò si applicano delle regole di dominio, una funzione di distanza e una misura di similarità. La regola di dominio è che il secondo campo (*info2*) contiene lo stato americano abbreviato (es: CA, MO, AZ) mentre la funzione di distanza è la *jaccard distance*. Si è scelta questa misura in quanto è ritenuta una delle migliori per soddisfare il task e tra tutti i campioni creati risulta essere quella con più alte performance. A questo punto ad ogni istanza viene applicata una funzione (i parametri passati sono i campi *info1* e *info2*) che estrae dal csv *citiesUS* tutte le città che hanno valore *state\_id* identico a *info2*. Dopo di che si calcola per ogni città sia la jaccard distance che la edit distance normalizzata.

La prima è una funzione di distanza che prevede di dividere le stringhe T in token, la formula è la seguente:

$$1 - \frac{T1 \cap T2}{T1 \cup T2} \quad (1)$$

dove  $T_i$  è il token della stringa iesima.

La seconda prende spunto dalla distanza di Levenshtein (edit distance ED) ma rispetto ad essa misura l'accuratezza di  $T_1$  rispetto a  $T_2$ , la formula è la seguente:

$$1 - \frac{ED}{n} \quad (2)$$

dove ED è il numero minimo di inserimenti, cancellazioni e sostituzioni per trasformare  $T_1$  in  $T_2$ . Per il record linkage non viene effettuata nessuna operazione di blocking in quanto i confronti da effettuare sono in numero limitato. A questo punto per validare il matching si effettuano due operazioni: la prima è estrarre dal risultato della funzione solamente le stesse città che le due misure considerano migliori; la seconda operazione è definire una soglia (0.3) entro il

quale considerare i match validi (si veda Figura 14): se la jaccard distance assume un valore  $\leq 0.3$  viene considerato un match altrimenti viene ignorato.

	best_edit	best_jacc	city_originale	val_jacc	city
7	East Oakdale	East Oakdale	East Oakland	0.181818	East Oakdale
24	Williston Park	Williston Park	Clifton Park	0.285714	Williston Park
30	Saint Louis	Saint Louis	St Louis	0.2	Saint Louis
32	Saint Louis	Saint Louis	St Louis	0.2	Saint Louis
35	Coal City	Coal City	Choppa City	0.3	Coal City

**Figura 14.** Risultato applicazione indici di distanza

Per aumentare il numero di matching delle città si è deciso per ultimo di effettuare un inner join tra le istanze rimaste (sempre sul campo *info1*) e le capitali presenti nel file *citiesWorld.csv*. In questo modo tutte le istanze che contengono nel campo *info1* solo la nazione (es. Canada, Ghana etc.) viene aggiunta la capitale di quel paese. Alla fine si ottiene un file csv che contiene per ogni tweet la città dell'utente, la provincia (non sempre presente), la nazione e i codici iso3 e iso2 oltre alla gara e al team con cui è stato catalogato il tweet.

Country	city	gara	id	iso2	iso3	state_name	team
662923	Hong Kong	Hong Kong	6	HK	HKG	NaN	Toronto Raptors
662924	Hong Kong	Hong Kong	6	HK	HKG	NaN	Toronto Raptors
662925	Hong Kong	Hong Kong	6	HK	HKG	NaN	Toronto Raptors
662926	Equatorial Guinea	Calatrava	6	GQ	GNQ	NaN	Neutro
662927	Yemen	Rida	6	YE	YEM	Al Bayda'	Toronto Raptors

**Figura 15.** Csv finale con posizioni note

### 3.4 Conteggio Hashtag

Un attributo molto interessante all'interno di un tweet è rappresentato dagli hashtag. Queste componenti testuali sono usate per indicare il contesto del discorso o per comunicare un messaggio in pochissimi caratteri, per tale motivo è importante capire quali sono i più utilizzati. Per far ciò è stato utilizzato il notebook *hashtag* in cui sono stati contati gli hashtag presenti in tutti i tweet divisi per tifoseria e in totale.

### 3.5 Conteggio Retweet

Un secondo attributo interessante all'interno del meccanismo di Twitter è il retweet. Esso consiste nella ricondivisione pubblica di un messaggio twittato da un altro utente con o senza aggiunta di un commento al testo originale. In questo senso, è importante analizzare i retweet, per scoprire l'argomento che possiede più visibilità in quel determinato momento. Per far ciò è stato utilizzato il notebook *retweet* che ha permesso di trovare i tweet più retwittati ed associarli alla tifoseria e gara di provenienza.

## 4. Risultati

Una volta effettuata la sentiment analysis e sistemata la posizione dei tweet, si è deciso di visualizzare i risultati con 4 infografiche:

1. Twitter sentiment e team stats
2. Tweet map
3. NBA Finals hashtag
4. Confronto giocatori

La prima infografica (si veda il [link](#)) contiene tre tipologie di plot: linechart, barchart e un bubble chart. Il linechart mostra l'andamento del sentiment medio delle due tifoserie in ogni partita in funzione del punteggio del match (aggiornato ogni 30 secondi), il barchart permette di confrontare le statistiche complessive delle due squadre ed il bubble chart mostra i tweet che hanno avuto il maggior numero di retweet. La seconda infografica (si veda il [link](#)) contiene due tipi di plot: dot map e barchart. La dot map mostra la posizione da cui provengono gli utenti che pubblicano e condividono i tweet per ogni partita, i barchart invece presentano due tipi di informazioni diverse: i primi due collocati in basso a sinistra indicano rispettivamente la distribuzione dei tifosi nel mondo (sinistra) e nelle città (accanto a destra), mentre quello situato a destra e colorato di azzurro mostra i dispositivi usati dagli utenti per interagire con il match.

La terza infografica (si veda il [link](#)) contiene un barchart che mostra per ogni hashtag utilizzato durante le NBA Finals la sua distribuzione nelle tifoserie.

La quarta infografica (si veda il [link](#)) mostra la media (con intervallo di confidenza al 95%) di 4 statistiche relative ai 6 giocatori più rappresentativi per franchigia durante le Finals 2019 e la Regular Season 2018-19.

### 4.1 Risultati generali

L'edizione 2019 delle NBA Finals non è stata una delle edizioni più viste negli ultimi anni come molti potrebbero pensare: stando alle statistiche di ABC<sup>5</sup>, lo share medio negli USA è stato pari a 8.8 (uno dei più bassi negli ultimi 10 anni [15]). Il valore sorprende parecchio perché quest'anno per la prima volta nella storia i Toronto Raptors (unica franchigia canadese) sono arrivati a contendere il titolo con gli evergreen Golden State (negli ultimi 4 anni sono sempre arrivati in finale): sorprende quindi che l'ascolto sia stato inferiore rispetto agli anni precedenti e abbia lasciato spiazzati i media. Una delle cause, come detto dallo stesso NBA Commissioner Adam Silver al termine

di gara 1, è che nelle Finals dopo 8 anni manca sul parquet il giocatore più rappresentativo della lega: la stella dei Los Angeles Lakers Lebron James [16]. La situazione opposta invece si è registrata in Canada, dove si sono raggiunti numeri altissimi di share. Basti pensare che per gara 1 si è frantumato il record del numero di telespettatori all-time di una partita NBA. Inoltre il 56% della popolazione, ovvero 20.5 milioni di telespettatori, ha visto tutte o almeno una parte delle 6 gare, garantendo il sesto posto all'evento nella classifica dei più visti in Canada nel 2019 [17].

Nel progetto, durante l'orario di attività di raccolta dati in real time (si veda Capitolo 1), si sono ottenuti in totale 1 305 789 tweet: una media di circa 215 000 a partita; osservando la distribuzione dei tweet si evince che la maggior parte appartengono alla tifoseria Neutro, a seguire Toronto Raptors e poi Golden State Warriors (si veda Figura 16).

	team	#Tweet
1	Neutro	624603
2	Toronto Raptors	569144
0	Golden State Warriors	112042

**Figura 16.** Distribuzione tweet.csv

Un'altra curiosità riguarda la gara che ha avuto il numero più alto di tweet, corrispondente a gara 6, e quella che ha avuto visibilità minore, ossia gara 3 (si veda Figura 17)

	Game	#Tweet
0	1	226103
1	2	153897
2	3	139553
3	4	165585
4	5	258618
5	6	362033

**Figura 17.** Numero tweet raccolti per gara.csv

I motivi di questi due risultati possono essere legati a due fattori: il primo è che in gara 6 si è chiusa la serie 4-2 con la vittoria non scontata dei Toronto Raptors e quindi Twitter è stato invaso dai fans della squadra canadese (e non solo). Per gara 3 invece il motivo potrebbe essere che gli appassionati abbiano pensato che anche quest'anno Golden State, dopo aver riportato la serie 1-1, mostrasse la sua superiorità.

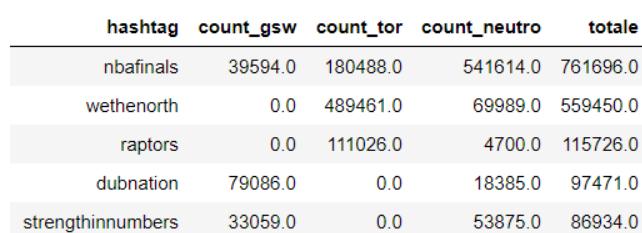
<sup>5</sup>storica emittente televisiva statunitense

tà conquistando per la terza volta consecutiva il titolo NBA rendendo le Finals più noiose e meno sorprendenti. Per quanto riguarda i tweet che hanno avuto il maggior numero di RT (condivisioni dagli altri utenti) si può osservare la Tabella 2, in cui si evince che i primi 4 sono scritti dai tifosi di Toronto mentre il quinto da un tifoso di Golden State.

Testo	N. Retweet	Tifoseria	gara
CANADA, THE @NBA TITLE IS YOURS! #WeTheNorth	10450	Toronto Raptors	6
The @Raptors are the 2019 NBA Champions! #WeTheNorth	7686	Toronto Raptors	6
BALL GAME. @NBA CHAMPIONS!!! #WeTheNorth	7347	Toronto Raptors	6
Presidential walk. #WeTheNorth   @BarackObama	7043	Toronto Raptors	2
Drake had some words for Draymond after the Raptors' Game 1 #NBAFinals win	6248	Golden State Warriors	1

**Tabella 2.** Tweet con maggior numero di retweet

Per quanto riguarda gli hashtag più utilizzati dagli utenti sono stati in ordine *nbafinals*, *wethenorth* e *raptors* (si veda la colonna "totale" nella Figura 18).



**Figura 18.** Hashtag più utilizzati



**Figura 19.** Wordcloud hashtag escludendo quelli di ricerca

Considerando solo i tweet dei tifosi di Toronto Raptors gli hashtag più utilizzati sono *wethenorth* e *nbafinals* (si veda Figura 20).

hashtag	count_top
wethenorth	489461.0
nbafinals	180488.0
raptors	111026.0
warriors	25251.0
toronto	16417.0

**Figura 20.** Hashtag più utilizzati dai tifosi di Toronto

Osservando invece i tweet dei tifosi dei Golden State Warriors gli hashtag più utilizzati sono *dubnation* e *nbafinals* (si veda Figura 21).

hashtag	count_gsw
dubnation	79086.0
nbafinals	39594.0
strengthinnumbers	33059.0
warriors	6593.0
goldenstatewarriors	5366.0

**Figura 21.** Hashtag più utilizzati dai tifosi di Golden State

È stato verificato, eliminando gli hashtag utilizzati per la raccolta di tweet, se potessero emergere nuovi argomenti di interesse non strettamente legati all’Nba, ma così non è stato.

Country	#Tweet
United States	325445
Canada	229640
Philippines	14383
Brazil	10318
United Kingdom	6273

**Figura 22.** Top 5 Nazioni alto numero di tweet

Considerando invece le città in cui si ha avuto l'affluenza maggiore sono in ordine *Toronto*, *Ottawa* e *Los Angeles* (si veda Figura 23).

city	#Tweet
Toronto	128569
Ottawa	36140
Los Angeles	20616
Washington	19921
New York	12284

**Figura 23.** Top 5 città alto numero di tweet

Osservando il sentiment medio delle due tifoserie si può constatare che nella serie il trend di Golden State è molto più altalenante rispetto a Toronto: l'intervallo comprendente il sentiment medio di Golden State include quello di Toronto (per esempio in gara1 il sentiment di Toronto va da [0.1;0.3] mentre quello di Golden State da [0.0;0.4]). Il picco di sentiment medio più alto registrato dalla squadra californiana (0.45) è avvenuto, nonostante l'esito, in gara 6 così come per la squadra canadese (0.53). Il picco più basso (-0.17) invece è stato registrato per Golden State in gara 4 quando i tifosi realizzano che gli avversari stanno aumentando la differenza di prestazioni e punteggio; per Toronto il valore minimo è registrato in gara 3 (-0.02) quando i tifosi si lamentano dell'arbitraggio dei direttori di gara.

Dall'analisi dei trend nelle varie partite emerge una certa dipendenza tra punteggio e prestazioni squadra rispetto al valore del sentiment. Per esempio in gara 2, nell'intervallo di tempo in cui Curry e Cousins stavano tirando con basse percentuali di tiro, il sentiment medio di Golden State è risultato negativo (-0.09), e in maniera opposta nello stesso match (quando Golden State ha realizzato un parziale 18-0) il sentiment è risultato molto positivo (0.21).

Oltre a questo rapporto tra sentiment e punteggio si evidenzia una certa dipendenza tra sentiment ed

eventi salienti della partita (sia positivi che negativi). Per esempio in gara 3 quando VanVleet ha segnato una tripla allo scadere del tempo il sentiment medio di Toronto è aumentato da 0.10 a 0.20 ed è rimasto tale per un intervallo di tempo consistente.

## 4.2 Risultati specifici

Dopo aver presentato risultati di carattere generale, si è proceduto ad analizzare le tre gare ritenute più determinanti della serie: gara 1, 5 e 6 concentrando-si principalmente sulle due tifoserie e tralasciando i tweet neutrali.

**Gara 1** Gara 1 è la prima partita delle Finals, quella in cui le due franchigie si studiano maggiormente. Per i tifosi è l'inizio della "vera" sfida: il momento che aspettano da tutto l'anno. Contro i pronostici a Est è riuscita a vincere Toronto, mentre come era prevedibile a Ovest l'ha spuntata Golden State riconfermandosi una delle migliori squadre dell'epoca moderna. I bookmaker vedono favorita quest'ultima per la vittoria del titolo, ma giocatori simbolo come Curry e Thompson non sono nella migliore forma fisica, mentre Durant (altra stella della franchiglia) è addirittura infortunato da qualche settimana. Per i Toronto Raptors la forma fisica è praticamente ottimale, tutte le loro stelle, in particolare Leonard, sono pronti ad affrontare l'avversario al meglio. Al termine di una partita emozionante Toronto si aggiudica il primo round riuscendo anche a distaccare Golden State nel finale. Osservando le statistiche, in particolare le percentuale ai tiri (eFG%), Toronto risulta migliore, ma non emerge una grande differenza (40.6% - 46.9%).

I tifosi hanno reagito su Twitter in maniera abbastanza prevedibile a favore di Toronto, è stata una delle gare con più tweet registrati (226103 seconda solo alle ultime due partite). I primi tweet più ritwittati provengono dalla tifoseria di Toronto (si veda Tabella 3) ed entrambi sottolineano la strabiliante vittoria evidenziando che la franchigia canadese è un progetto giovane (6 anni fa è stato iniziato un importante restyling) e alla prima apparizione alla serie delle Finals ha vinto la partita.

Per i tifosi di Golden State invece entrambi i tweet più condivisi sono ad inizio partita: il primo riguarda l'annuncio della tanto attesa formazione titolare, il secondo invece parla dell'incredibile record di Curry; sembra, dai dati, che i tifosi dopo la sconfitta non siano in vena di commentare la gara.

Testo	N. Retweet	Tifoseria
First #NBAFinals. First W. #WeTheNorth	3213	Toronto Raptors
Six years ago we started a movement. Starting tonight, we make history with a nation behind us. #WeTheNorth	2159	Toronto Raptors
#DubNation, say hello to tonight's first 5	1269	Golden State Warriors
Congrats to Stephen Curry of the @warriors for becoming the first player in #NBAFinals history to make 100 3-point field goals! #StrengthInNumbers	1037	Golden State Warriors

**Tabella 3.** I due tweet più retwittati per franchigia di gara 1

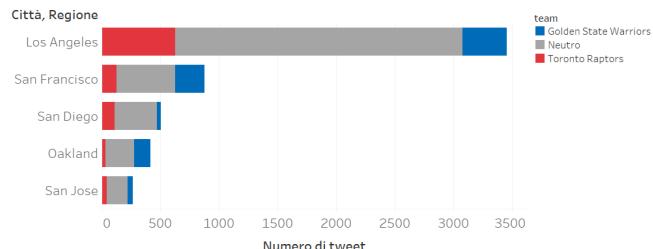
Il record di Curry è stato l'argomento più scottante durante la partita, si nota infatti che nel momento in cui è avvenuto il centesimo canestro si è registrato il valore più alto di sentimento medio (0.42) dei tifosi di Golden State.



**Figura 24.** Centesima tripla di Curry nelle Finals NBA

In gara 1 un altro fatto interessante è la strepitosa prestazione di Siakam, eletto *man of the match*, che ha superato il suo record personale di punti in carriera (32). Quando accade (si veda al solito [link](#)) il sentimento dei tifosi di Toronto si alza, ma a fine partita non è uno degli argomenti più discussi perché i fans sono tutti concentrati sulla prima vittoria nelle Finals.

Osservando la provenienza dei tweet ([link](#)) e la loro distribuzione, come era prevedibile, si nota che Stati Uniti e Canada superano per la stra gran maggioranza il resto del mondo. Ciò che sembra strano invece è che negli Stati Uniti i Toronto Raptors sono tifati in maggioranza. Perfino in molte città della California è così. Le poche a favore di Golden State sono San Francisco e Oakland le due città simbolo della franchigia.



**Figura 25.** Distribuzione dei tifosi nelle maggiori città della California

**Gara 5** Gara 5 è la partita del tanto atteso ritorno di Kevin Durant (soprannome KD), il grande assente di Golden State delle prime partite. Tutti i tifosi stanno aspettando lui nella speranza che il suo aiuto possa riaprire la serie che per ora è a favore della squadra canadese per 3-1. La partita giocata in casa da Toronto termina con la vittoria di un punto di Golden State ma con una brutta notizia: Durant si è nuovamente infortunato.



**Figura 26.** Durant si infortuna il tendine d'achille

Le statistiche finali mostrano una forte discrepanza in termini di percentuali di tiro da 3 tra le due franchigie: Golden State ha tirato da 3 con il 50% mentre Toronto solamente il 20%. Un dato molto importante per una franchigia come Golden State che è diventata famosa grazie ai suoi fantastici tiratori. Questo match, complice l'"hype" per il ritorno di KD, ha generato un numero considerevole di tweet (258618), il più alto registrato nelle prime 4 partite. I 3 messaggi più retweetati provengono tutti dai tifosi di Toronto (si veda Tabella 4) e il primo sorprende parecchio perché il tweet mostra in un video KD che sorretto dai medici torna negli spogliatoi e tutto il palazzetto gli tributa un lungo applauso. Questo fa emergere il rispetto e il fair-play dei tifosi di Toronto. Il secondo tweet più condiviso invece è un incitamento ai giocatori di Toronto che vincendo la partita si aggiudicherebbero la serie.

Testo	N. Retweet	Tifoseria
What the media doesn't show #nbafinals #raptors <a href="https://t.co/eMiEc9QnQK">https://t.co/eMiEc9QnQK</a>	3921	Toronto Raptors
Let's do this! #WeTheNorth	2350	Toronto Raptors
"I hope people see why Canada is so special." @Raptors Superfan Nav Bhatia makes us proud to be Canadian. #WeTheNorth #ThisIsCanada	1322	Toronto Raptors

**Tabella 4.** Tweet più retwettati in gara 5

Per quanto riguarda la sentiment anche in questa partita emerge una relazione con le prestazioni della squadra: quando Curry e Thompson segnano 36 punti in maniera rapida e quasi automatica il valore di sentiment si alza parecchio raggiungendo il culmine più alto (0.33) (è possibile verificare utilizzando l'infografica 1 al [link](#) e selezionando gara 5 tramite il filtro). Osservando la provenienza dei tweet anche qui la situazione rispecchia quella di gara 1 sia per quanto riguarda le nazioni che le città da cui proviene la maggior parte di tweet.

**Gara 6** Gara 6 è l'ultima partita della serie e soprattutto l'ultima di sempre alla Oracle Arena (l'anno prossimo i Golden State Warriors torneranno a San Francisco). Da una parte Toronto vuole chiudere la serie, dall'altra Golden State non vuole abbandonare il suo amato stadio con una sconfitta, alla fine però i Toronto Raptors contro i pronostici vinceranno. Qui il popolo di Twitter si scatena: il numero di tweet registrati è il più alto di tutti (362033); la partita è la più combattuta della serie: Curry e Thompson spinti dalla dell'infortunio di KD vogliono portare la serie in parità e ci provano fino alla fine, ma Toronto gioca più da "vera squadra" e grazie alle triple di VanVleet e Leonard riesce a conquistare il titolo NBA per la prima volta nella storia.

**Figura 27.** Esultanza di Leonard alla termine della partita

Le statistiche della partita dicono che tra le due squa-

Testo	N. Retweet	Tifoseria
CANADA, THE @NBA TITLE IS YOURS! #WeTheNorth	10450	Toronto Raptors
The @Raptors are the 2019 NBA Champions! #WeTheNorth	7686	Toronto Raptors
BALL GAME. @NBA CHAMPIONS!!! #WeTheNorth	7347	Toronto Raptors

**Tabella 5.** Tweet più retwettati in gara 6

dre c'è stato equilibrio, solamente la percentuale di tiro di Golden State è leggermente più alta di Toronto (42% contro 37%). Come ci si può facilmente aspettare i tweet più condivisi sono tutti postati da tifosi di Toronto, la loro gioia è incontenibile (si veda Figura 5).

Per quanto riguarda la sentiment anche qui emerge la relazione con le prestazioni della squadra e gli eventi salienti: quando Thompson si infortuna al ginocchio il sentiment medio dei tifosi di Golden State scende, diversamente quando Curry sbaglia la tripla decisiva allo scadere il sentiment dei tifosi di Toronto ha un picco verso l'alto evidenziando quindi che un momento saliente influenza i sentiment di entrambe le tifoserie. Nonostante la sconfitta il sentiment dei tifosi dei Golden State Warriors cresce tantissimo, come per voler condividere ugualmente apprezzamento nei confronti dei giocatori.

Osservando la posizione di pubblicazione dei tweet, le nazioni e le città da cui provengono maggiormente sono le stesse della gara precedente; si sottolinea però che i numeri registrati sono decisamente superiori, basta considerare il Canada dove rispetto a gara 5 sono aumentati della metà (da 40 mila si è passati più di 60 mila).

#### 4.3 Confronto stats giocatori

Una volta analizzati i risultati ottenuti, si è cercato di capire se effettivamente i giocatori migliorano le loro prestazioni tra la Regular Season e le Finals. Molte persone, tra i quali gli esperti, pensano che le loro performance siano completamente diverse nelle due fasi della stagione e che questo avvenga in modo netto per due motivi: per primo nelle Regular Season si giocano 82 partite ed è quindi è normale che un giocatore non abbia sempre alte prestazioni, inoltre alcune squadre riescono ad avere la certezza matematica di partecipare ai playoff molto tempo prima e quindi stelle come Curry, Leonard si riposano o comunque giocano a basse intensità durante le partite rimanenti per preservare la forma fisica; l'altro motivo

è che le Finals rappresentano la fase conclusiva della stagione in cui un giocatore è più motivato ad avere alte prestazioni per aiutare la squadra a conquistare il titolo. Questo influenza in maniera positiva lo stesso spettacolo offerto, infatti nelle Finals (in generale i playoff) si vedono sempre grandi testa a testa tra franchigie, cosa completamente diversa in Regular Season dove a meno delle partite di forte appeal (per esempio Los Angeles Lakers vs Los Angeles Clippers oppure Houston Rockets vs Golden State) lo spettacolo è nettamente inferiore e meno seguito.

Da queste considerazioni si è deciso quindi di capire se i giocatori abbiano performance differenti nelle due fasi della stagione. Per far ciò si sono confrontate 4 stats dei 6 giocatori più rappresentativi delle due franchigie nelle Finals 2019 e nella Regular Season 2018-19. Le statistiche considerate sono:

- 3P%: Percentuale dal tiro da 3
- AST: Assist
- eFG%: Percentuale dei canestri effettuati dove si dà maggior peso a quelli da 3
- PTS: Punti
- TRB: Rimbalzi

Prima di procedere all'analisi dei risultati bisogna dire che Durant non è stato considerato per il confronto delle prestazioni in quanto nelle Finals ha disputato soltanto mezza partita (gara 5). Dai confronti effettuati emergono alcune cose interessanti. La prima è che pochi giocatori sono migliorati mentre altri contro il pensiero comune sono addirittura peggiorati. L'unico giocatore che statisticamente è migliorato nelle Finals in termini di percentuali di tiro da 3 è stato Thompson, la guardia<sup>6</sup> dei Golden State; ulteriori analisi basate sui dati relative a Toronto indicano che nessuno dei Raptors è migliorato statisticamente, ma anzi alcuni giocatori sono riusciti a peggiorare le proprie medie come Gasol e Siakam (si veda Figura 28).

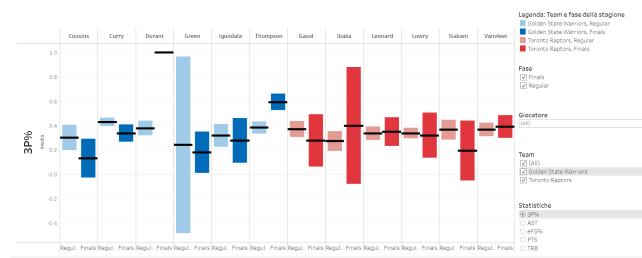


Figura 28. Statistica %3P

<sup>6</sup>guardia: uno dei 4 ruoli del basket, gli altri sono ala, playmaker e centro

Per quanto riguarda i punti realizzati (PTS), un giocatore è statisticamente migliorato (Green) ed è ancora di Golden State mentre Cousins, un altro giocatore della stessa franchigia, è peggiorato. Considerando la squadra di Toronto invece nessun giocatore è statisticamente migliorato o peggiorato. Come media punti nelle Finals Leonard e Curry sono i "cecchini" delle rispettive franchigie, per il playmaker di Golden State il valore è 30.5 mentre per la guardia di Toronto 28.5 (si veda Figura 29).

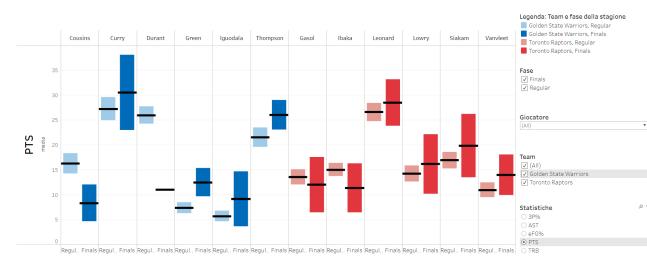


Figura 29. Statistica PTS

Per quanto riguarda la statistica della percentuale di tiro emerge una cosa sorprendente: Curry statisticamente è peggiorato mentre Thompson è migliorato. Guardando Toronto invece emerge che nessun giocatore è statisticamente migliorato/ peggiorato; bisogna dire però che VanVleet a sorpresa è quello che ha avuto una media percentuale di tiro più alta tra tutti i compagni maggiore perfino della stella Leonard, il suo valore è 59% (si veda Figura 30).

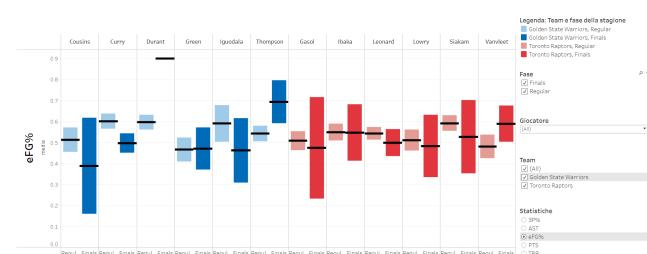


Figura 30. Statistica %eFG

Considerando invece la statistica AST (assist), nessun giocatore di Golden State è statisticamente migliorato/ peggiorato, mentre per Toronto Gasol e VanVleet sono statisticamente peggiorati. Un'altra cosa interessante che emerge è che Thompson ha tenuto una media assist identica alla Regular Season (2.4) risultando il più costante.

Per ultima si effettua un'analisi sulla statistica TRB (rimbalzi), da questa emerge che nessun giocatore di Golden State ha modificato le sue prestazioni medie significativamente, mentre per Toronto il centro Ibaka è peggiorato; in questo caso un solo giocatore ha

tenuto la stessa media rimbalzi sia nella regular che nelle Finals ed è Vanvleet, il suo valore è 2.6.

## 5. Conclusioni

Il progetto si è focalizzato prima di tutto sull'analisi dei tweet registrati durante le NBA Finals 2019 e sul sentimento delle tifoserie, dopodichè si sono confrontate le stats dei giocatori nelle due fasi conclusive della stagione. Per quanto riguarda il primo task i risultati ottenuti della sentiment non sono perfetti principalmente per due motivi: per primo la libreria utilizzata (VADER) non riesce a valutare se una frase è ironica o meno con la stessa capacità di un umano, quindi se un tweet è per esempio "Cousins is playing very well" accompagnato da una emoji che piange dal ridere viene catalogato come positivo quando in realtà non lo è. Il secondo motivo è legato alle emoji, che non sempre vengono interpretate correttamente: anche dopo le modifiche apportate solamente una piccola parte è inclusa nel vocabolario su cui fa affidamento VADER. Nonostante le due considerazioni appena fatte bisogna apprezzare però che la libreria utilizzata rispetto a quelle concorrenti risulta essere una delle migliori, soprattutto per l'analisi dei social.

Per quanto riguarda i risultati si è riscontrata una certa relazione tra il sentimento medio e le prestazioni della squadra: soprattutto quando una franchigia riesce a distaccare nel punteggio l'altra o a scavalcarla. Un altro legame interessante emerso è quello tra sentimento ed eventi salienti o insoliti nella partita (per esempio quando i tifosi di Golden State in gara 4 abbandonano il palazzetto prima della fine del match il sentimento si abbassa).

Per quanto riguarda il secondo task realizzato (confronto statistiche giocatori) i risultati mostrano che non è vero che la maggior parte dei giocatori migliora le proprie prestazioni tra la Regular Season e le Finals, anzi alcuni peggiorano in maniera considerevole. Per quanto riguarda gli sviluppi futuri si potrebbero effettuare altre analisi interessanti. Una di queste potrebbe essere la valutazione del sentimento non più sulla squadra ma sul singolo giocatore, in modo da capire se esiste o meno un rapporto tra il sentimento e le sue statistiche.

Un'ultima considerazione va lasciata agli eventuali miglioramenti apportabili al progetto: si potrebbe aumentare la proporzione delle città riconosciute adottando tecniche diverse di record linkage in ambito di geolocalizzazione dei tifosi e si potrebbe implementare un nuovo algoritmo di sentiment personalizzato che migliori i risultati ottenuti e li renda ancor più veritieri.

## Riferimenti bibliografici

- [1] Wikipedia. Twitter. URL: <https://it.wikipedia.org/wiki/Twitter>.
- [2] Jacob Kastrenakes. *Twitter keeps losing monthly users, so it's going to stop sharing how many*. URL: <https://www.theverge.com/2019/2/7/18213567/twitter-to-stop-sharing-mau-as-users-decline-q4-2018-earnings>.
- [3] Bryan Anderson. «Winning over Fans: How Sports Teams Use Live-Tweeting to Maximize Engagement». In: *ELON JOURNAL* (2018), p. 54.
- [4] Wikipedia. NBA Finals. URL: [https://it.wikipedia.org/wiki/NBA\\_Finals](https://it.wikipedia.org/wiki/NBA_Finals).
- [5] Twitter. Twitter API. URL: <https://developer.twitter.com>.
- [6] All Sport API. All Sport API. URL: <https://allsportsapi.com>.
- [7] Jay Kreps, Neha Narkhede, Jun Rao et al. «Kafka: A distributed messaging system for log processing». In: *Proceedings of the NetDB*. 2011, pp. 1-7.
- [8] Kafka. Kafka-python documentation. URL: <https://kafka-python.readthedocs.io/en/master/index.html>.
- [9] Twitter. Tweepy. URL: <https://www.tweepy.org>.
- [10] MongoDB. MongoDB. URL: <https://www.mongodb.com/it>.
- [11] MongoDB. PyMongo. URL: <https://api.mongodb.com/python/current/>.
- [12] Theresa Wilson, Janyce Wiebe e Paul Hoffmann. «Recognizing contextual polarity in phrase-level sentiment analysis». In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 2005.
- [13] C.J Hutto. VaderSentiment. URL: <https://github.com/cjhutto/vaderSentiment>.
- [14] AnHai Doan, Alon Halevy e Zachary Ives. *Principles of data integration*. Elsevier, 2012.
- [15] Wikipedia. NBA Finals television ratings. URL: [https://en.wikipedia.org/wiki/NBA\\_Finals\\_television\\_ratings](https://en.wikipedia.org/wiki/NBA_Finals_television_ratings).
- [16] Tim Reynolds. «NBA Finals ratings released, Canada sets records». In: *The Associated Press* (2019).

- [17] Ross Lager. *Ratings Roundup: Toronto's First NBA Championship Shatters, Viewership Numbers in Canada.* URL: <https://www.sportsvideo.org/2019/06/19/ratings-roundup-torontos-first-nba-championship-shatters-viewership-numbers-in-canada/>.