THE SUPERBOWL NETWORK

Laura Nembrini 819059, Gabriele Strano 866563
Università degli Studi di Milano-Bicocca, CdLM Data Science

12 Gennaio 2022

Sommario

Il Super Bowl è un evento sportivo popolare che si svolge ogni anno per determinare la squadra del campionato della National Football League (NFL). Trasmesso in più di 170 paesi, il Super Bowl è uno degli eventi sportivi più seguiti al mondo. Non si tratta solamente di un evento a livello sportivo ma di un vero e proprio spettacolo che ospita diversi artisti affermati a livello mondiale.
In questo elaborato, viene utilizzato il social network Twitter per estrarre i vari tweet prodotti dagli utenti in merito all'argomento. Verranno utilizzati tre hashtag per la ricerca dei tweet: #Bengals, #Rams e #SuperBowl.
Nella prima parte dell'analisi ci si è concentrati su alcuni aspetti di livello descrittivo come la provenienza dei tweet, gli hashtag più utilizzati, gli autori più attivi e gli account maggiormente menzionati. Successivamente abbiamo posto l'attenzione alla struttura della rete sociale andando a verificare, tramite metriche e grafi, la composizione delle community riguardanti gli argomenti da noi considerati. In questa fase è stata inoltre individuata la presenza di cluster naturali all'interno della rete sociale.
Le ultime due fasi della nostra analisi hanno riguardato i contenuti prodotti, all'interno della rete sociale analizzata. Sono state utilizzate tecniche di analisi quali LDA e sentiment analysis.

Keywords: Social Content Analysis, Social Network Analysis, Sentiment Analysis, Commununity detecton, Super Bowl, Rams, Bengals.

Indice

1 Introduzione

Il Super Bowl è l'evento sportivo più seguito negli Stati Uniti ed ha un impatto mediatico molto forte anche nel resto del mondo. Dal punto di vista sportivo rappresenta il match finale del

campionato della National Football League, la massima serie di football americano. Parlarne però solo dal punto di vista sportivo sarebbe riduttivo. Si parla infatti di un evento mediatico di grande impatto; basti pensare che nel 2019 l'emittente "Cbs" che ha trasmesso l'evento ha incassato in pubblicità 336 milioni di euro. Caratteristica di questo evento sono poi gli spettacoli di artisti affermati al livello mondiale durante il famosissimo halftime show, da cui dipendono ingenti quote di ricavi ed interessi. L'accordo per i diritti tv firmato dalla lega NFL, con i network che ne trasmettono negli Usa le partite vale complessivamente circa 5 miliardi di dollari all'anno. I numeri sopracitati permettono di avere un'idea in merito all'impatto economico ma anche sociale di questo evento. Proprio il secondo aspetto, quello sociale, è stato scelto come focus della nostra analisi. Si è cercato di analizzare la struttura sociale ed i contenuti prodotti dagli utenti, in relazione a 3 hashtag: #Rams, #Bengals, #SuperBowl. I primi due erano gli hashtag più utilizzati per riferirsi alle due squadre partecipanti al Super Bowl di questa stagione. Dal punto di vista della struttura di queste reti sociali, si è cercato di individuare ed analizzare eventuali community presenti all'interno di queste reti, analizzandone gli argomenti trattati ed identificandone i nodi più influenti. Guardando invece ai contenuti prodotti da queste reti si è deciso di effettuare una sentiment analysis per cercare di catturare differenze di approccio o posizionamento sia tra le due squadre che tra le squadre e l'evento.

2 Dataset e domanda di ricerca

La domanda che ci siamo posti come base per lo sviluppo di questo progetto è stata la seguente: come vengono percepite le due squadre da parte della community? e l'evento, come viene percepito? Per cercare di rispondere a questa domanda, sono stati analizzati i contenuti pubblicati dagli utenti all'interno del social network Twitter monitorando la presenza di almeno uno dei seguenti hashtag all'interno del tweet: #bengals, #rams, #superbowl. Gli hashtag si riferiscono rispettivamente alle due squadre che competono nella finale e al nome dell'evento. La raccolta dei contenuti è stata effettuata tramite l'utilizzo delle API messe a disposizione da Twitter e il linguaggio di programmazione Python.
Sono stati raccolti in totale circa 15mila tweet in un range temporale che va dal 02/02/2022 al 10/02/2022. Oltre al contenuto del tweet sono state raccolte altre variabili come:

1. *Id*: identificativo univoco del tweet raccolto;
2. *date*: data di creazione del tweet;
3. *text*: parte del contentuto pubblicato dall'utente;
4. *like*: numero di like ricevuti al tweet;
5. *n_rt*: numero di retweet;
6. *author*: autore del tweet;
7. *location*: luogo in cui l'utente ha pubblicato il tweet;
8. *full_text*: testo del tweet;
9. *hashtag*: uno dei tre hashtag usati come filtro per la ricerca presente nel testo del tweet;
10. *mentions*: utenti menzionati all'interno del tweet;
11. *h*: hashtag presenti all'interno del tweet.

Successivamente alla raccolta dati si è proceduto con un'analisi preliminare dei dati ottenuti, così da riuscire ad estrarre le prime informazioni.

3 Analisi preliminare

Successivamente alla raccolta dati, è stata effettuata un'analisi preliminare dei dati in modo da poter estrarre alcune informazioni dai dati raccolti.
In prima istanza si è effettuato un plot della distribuzione dei tweet suddivisi per autore per andare a vedere quali sono gli autori che hanno twittato maggiormente in merito all'argomento. E' possibile visualizzare il grafico ottenuto nella

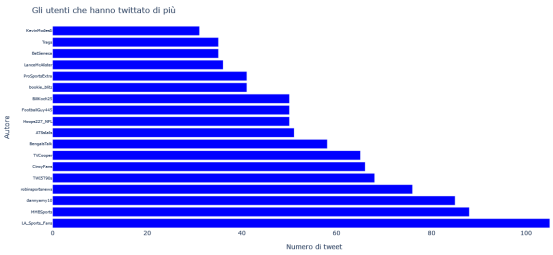

Figura 1: Autori più attivi in merito al Super Bowl.

Figura 1.
Attraverso la Figura 2 andiamo a visualizzare invece gli account Twitter più menzionati dagli utenti.

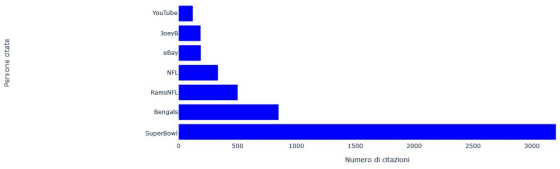

Figura 2: Account più menzionati dagli utenti.

Dal grafico è possibile notare come gli account Twitter più menzionati sono quelli delle squadre che partecipano al superbowl (overro Bengals e Rams) e quello della National Football League, identificata anche dall'acronimo NFL, che è la maggiore lega professionistica nordamericana di football americano.
Infine le stesse operazioni sono state effettuate sugli hashtag presenti. I risultati ottenuti è possibile osservarli nella Figura 3

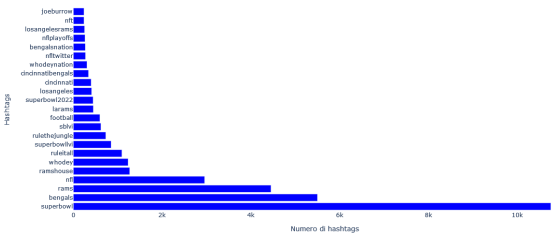

Figura 3: Hashtag più usati dagli utenti.

Senza considerare gli hashtag monitorati per la raccolta dati, emerge dal grafico che quelli più utilizzati dagli utenti sono nfl, ramshouse che si riferisce a una delle due squadre partecipanti, whodey ovvero la mascotte dei Bengals. E' possibile inoltre notare che, oltre a molti hashtag riguardanti in modo specifico le due squadre, sono presenti anche tanti altri riguardanti l'evento in generale.

4 Social Network Analysis

Parlando di social media e in particolare di Twitter, cosa la cosa più interessante è che crea connessioni; queste connessioni formano Social Network che possono essere studiati per capire come interagiscono gli utenti o come vengono diffuse opinioni e argomenti. Per rispondere a ciò, sono state utilizzate le proprietà della teoria dei grafi.

4.1 Costruzione grafi

Per la fase di implementazione del grafo è stata usata la libreria Python denominata *networkx*. In particolare, per lo scopo di questa analisi, sono stati messi in relazione gli autori dei diversi tweet e gli hashtag da loro utilizzati creando un arco tra il nodo rappresentante l'utente e il nodo rappresentante l'hashtag utilizzati. E' stata quindi generata la rete degli hashtag che è possibile osservare nella seguente figura.

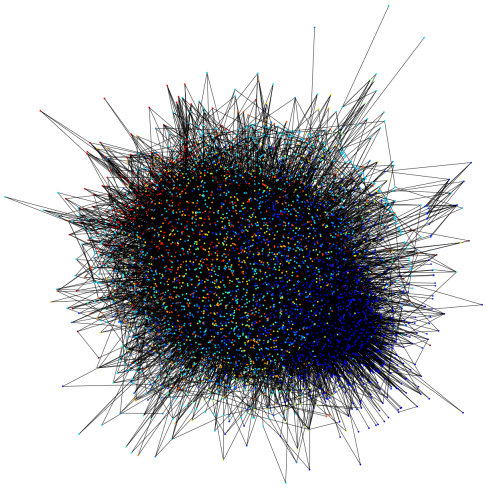

Figura 4: Rete degli hashtag

Dal dataset sono quindi stati considerati:

- l'autore del tweet
- gli hashtag utilizzati dall'utente all'interno del tweet.

Come operazione preliminare, è stato necessario estrarre gli hashtag in modo tale da creare un dataframe contenente solamente l'autore e i vari hashtag da lui utilizzati. E' noto che l'hashtag è preceduto da # quindi è stato estratto dal testo del tweet attraverso una espressione regolare.

Il grafo presente in Figura 4 è non orientato, non pesato e con sottografi sconnessi. Inoltre la struttura della rete degli hashtag risulta essere ego-centrica, formata principalmente da sotto-grafi a forma di stella centrati in nodi rappresentanti degli hashtag.

4.1.1 Metriche

Per la rete precedentemente generata, sono state analizzate delle metriche riguardanti la struttura, più precisamente le metriche di centralità, ovvero misure che permettono di calcolare l'importanza di un nodo all'interno della rete.
In particolare le metriche utilizzate sono state:

- *Betweenness centrality*: misura usata per studiare il ruolo di un nodo nella propagazione dell'informazione. In particolare permette di individuare il nodo che, se eliminato, non permette il fluire di informazioni tra gli altri nodi ad esso collegati.
- *Degree centrality*: calcola il numero di vertici adiacenti a un dato vertice. Viene rappresentata tramite valori percentuali.

I risultati ottenuti sulla rete degli hashtag per le 2 misure descritte sono riportate nelle seguenti tabelle.

Dalla Tabella 1 e 2 notiamo come i nodi con valori più alti delle due metriche sono superbowl, bengals e rams come ci si poteva aspettare in quanto sono gli hashtag usati durante la ricerca dei tweet ai quali si aggiungono nfl, ramshouse, whodey che, come precedentemente accennato, sono appunto tra gli hashtag più usati dagli utenti.

Nodo	**Betweennes centrality**
superbowl	0.65
bengals	0.16
rams	0.13
nfl	0.07
ramshouse	0.014
whodey	0.013

Tabella 1: Valori betweennes centrality.

Nodo	**Degree centrality**
superbowl	0.67
bengals	0.34
rams	0.30
nfl	0.26
ramshouse	0.12
whodey	0.11

Tabella 2: Valori degree centrality.

Generalmente i nodi aventi alta betweeness sono quelli aventi anche alta degree centrality, in quanto sono i cosiddetti hub.

4.2 Community detection

Uno dei compiti più importanti quando si studiano le reti è quello di identificare le community. Questo consente di scoprire gruppi coesi tra di loro e individuare le loro relazioni.
Per la Community Detection è stato applicato l'algoritmo gerarchico di Louvain, che massimizza la modularità entro i gruppi. Questo significa massimizzare la densità di connessioni entro il cluster rispetto a quella verso l'esterno del cluster. La misura di modularità quantifica quindi la qualità di una assegnazione di nodi alla community.
I cluster ottenuti sono i rappresentati nella seguente tabella.

Cluster	**Numerosità**
Cluster 0	2172
Cluster 1	2801
Cluster 2	2461
Cluster 3	929
Cluster 5	914

Tabella 3: Cluster selezionati dalla rete

In particolare il valore di modularità ottenuto per il grafo degli hashtag è pari a 0.38.

5 Social Content Analysis

Per l'analisi dei Social Content è stata effettuata una Sentiment Analysis basata sull'approccio lessicale descritto di seguito.

5.1 LDA

Attraverso l'algoritmo Latent Dirichlet Allocation, usato per il topic modeling, si è cercato di costruire un modello di linguaggio applicato ad ogni cluster in modo da identificare la tipologia di linguaggio utilizzato in ciascun gruppo. L'analisi LDA si basa su due concetti principali: ogni documento può essere descritto da una distribuzione di topic e ogni topic può essere descritto da una distribuzione di parole. Per poter stimare le varie probabilità condizionate *P(word w | topic t)* e *P(topic t | document d)*, si è passati ad una rappresentazione delle parole attraverso il calcolo dell'indice Tf-Idf, in modo tale da misurarne la rilevanza e non la frequenza. Nella fase di pre-processing dei dati si è deciso di rimuovere tutte le stopword, per poi passare alla forma lemmatizzata di ciascuna parola. Inoltre, sono stati rimossi la punteggiatura, gli indirizzi url e gli hashtag, in modo tale da poter eliminare il rumore causato da termini non utili ai fini dell'analisi. Questo ha portato a migliorare la fase di identificazione delle varie distribuzioni. Prima di ottenere i vari argomenti per ogni cluster, la dimensionalità del dizionario è stata ridotta, rimuovendo sia le parole che compaiono in meno di 10 documenti che le parole che compaiono in più della metà dei documenti. Sono stati considerati quindi solo i 5000 termini più frequenti. Per permettere al modello di captare una buona rappresentazione del linguaggio, sono stati selezionati 3 topic.

5.2 Sentiment Analysis

L'analisi del sentimento è stata applicata tramite un approccio basato sul lessico utilizzando la libreria Vader. VADER (Valence Aware Dictionary for Sentiment Reasoning) è una libreria utilizzata per l'analisi del sentimento sensibile ad entrambe le polarità del testo (positivo / negativo) e l'intensità delle emozioni. In genere, quantifichiamo Essa può variare da -1 (tweet estremamente negativo) a +1 (tweet estremamente positivo).
Prima di applicare la Sentiment Analysis sono state effettuate delle operazioni preliminari di pulizia del testo. Sono stati eliminati eventuali numeri, spazi multipli, indirizzi url, hashtag, menzioni, caratteri di escape, diversi segni di punteggiatura e stop words, ovvero parole non utili all'estrazione del significato della frase.
In questo caso si è scelto di andare a capire il sentiment dei 3 hashtag usati come filtro di ricerca dei tweet per vedere se l'opinione risultava prevalentemente positiva, negativa o neutrale. E' possibile osservare i risultati emersi nella figura seguente. Dalla Figura 5 è possibile notare

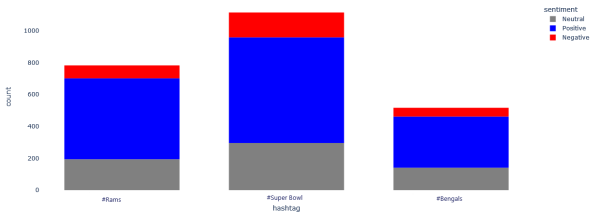

Figura 5: Distribuzione del sentiment per topic.

come il sentiment degli autori è prevalentemente positivo sia nei confronti delle due squadre sia nei confronti dell'evento Super Bowl in generale.
Inoltre è stata effettuata una distribuzione della sentiment dei tweet per le diverse date di raccolta dei tweet.

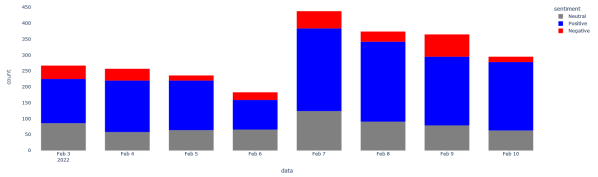

Figura 6: Distribuzione del sentiment per data.

Anche attraverso la Figura 6 è possibile dedurre che i tweet con sentiment positivo sono nu-

merosi. Anche i tweet con sentiment neutrale sembrano essere molteplici.

6 Risultati

Per ogni cluster identificato tramite community detection sulla rete degli hashtag sono state effettuate tutte le precedenti analisi per capire le caratteristiche di ogni singolo cluster. Sono emersi principalmente 5 cluster nei quali vengono affrontati differenti punti di vista in merito all'evento Superbowl.
Nel **Cluster 0**, si discute principalmente di una delle due sfidanti: Los Angeles Rams. Di fatti, è possibile notare dal wordcloud successivo che le parole che emergono maggiormente dai tweet appartenenti a questo cluster sono: Rams, NFL.

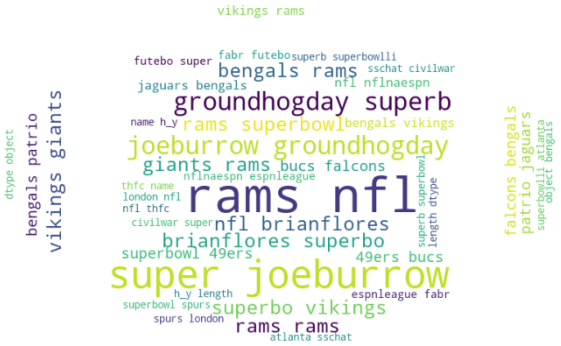

Figura 7: Wordcloud cluster 0

Nel **Cluster 1** invece si parla principalmente dell'altra squadra partecipante alla 56esima edizione del Superbowl: Cincinnati Bengals. Le parole principali che emergono infatti sono: Cincinnati, Superbowl, playoff.

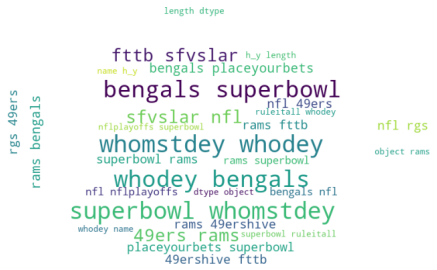

Figura 8: Wordcloud cluster 1

Il terzo cluster, ovvero **Cluster 2** parla principalmente dell'evento in generale senza focalizzarsi su nessuna delle due squadre. Le parole che emergono dal Wordcloud sono: NFL, Superbowl e ramsvsbengals.

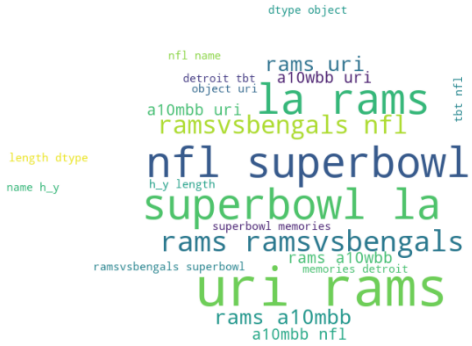

Figura 9: Wordcloud cluster 2

Gli ultimi due cluster, ovvero il **Cluster 3** e **Cluster 5** trattano invece altri argomenti affini, principalmente di interesse economico. Infatti vediamo menzionati nei due wordcloud alcuni principali portatori di interesse, come alcune aziende sponsor dell'evento ed altri termini che richiamano il mondo delle scommesse sportive.

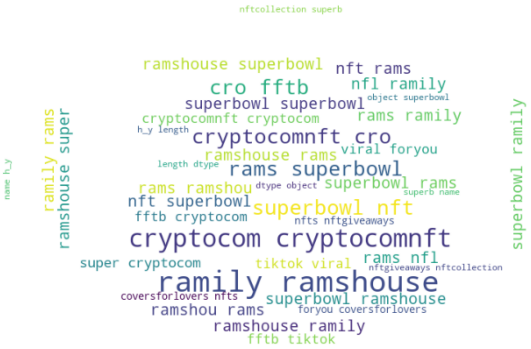

Figura 10: Wordcloud cluster 3

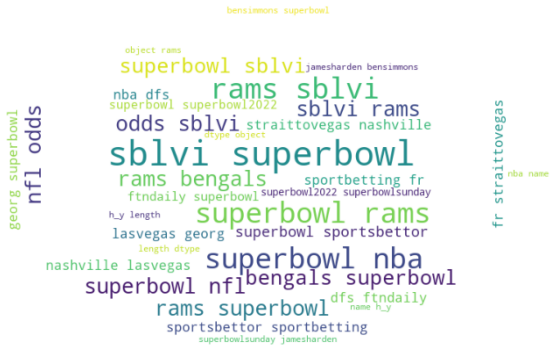

Figura 11: Wordcloud cluster 5

7 Conclusioni e sviluppi futuri

Dalla social network analysis è emersa una struttura ego-centrica, dove superbowl è l'hashtag che può essere identificato come nodo ego, essendo quello con grado maggiore (6869). Sono stati inoltre individuati vari cluster ognuno con caratteristiche ed argomenti diverso dagli altri. I primi 3 cluster avevano infatti come principale argomento rispettivamente uno dei 3 hashtag utilizzati in fase di stream dei tweet. Gli altri cluster trattavano invece l'evento in relazione a tematiche differenti come gli artisti che si sarebbero esibiti nell'halftime show o le scommesse sull'evento. Dalla social content analysis, tramite l'utilizzo della tecnica di sentiment analysis non sono emerse differenze significative tra lo score delle due squadre, per entrambe comunque è stato individuato un sentiment positivo da parte della community di Twitter. Quello che è emerso invece un po' a sorpresa è stato il sentiment score dell'evento, anch'esso positivo ma minore rispetto alle singole squadre. Questo è possibile che possa dipendere dalla differenza di enfasi nel commentare che può avere uno spettatore "disinteressato", che guarda l'evento attirato dalla spettacolarità, rispetto ad un tifoso che commenta le prestazioni della propria squadra del cuore o esprime le sue sensazioni rispetto ad essa. Questo tipo di analisi è certamente migliorabile, sia aumentando il numero di Tweet ed il periodo di streaming utilizzati, sia aumentandone la complessità delle tecniche di analisi utilizzate.
È sicuramente un'analisi utile soprattutto se utilizzata per finalità di marketing, estendere un'analisi del genere a periodi lunghi permette di capire il piazzamento di un'azienda sportiva di elevate dimensioni come i club della lega NFL, nelle menti degli appassionati e dei portatori di interesse, di un'azienda sportiva di elevate dimensioni come i club della lega NFL. Può inoltre essere utilizzata come metodo per capire del medio termine come stiano funzionando e che tipo di riscontro stiano avendo le campagne e le azioni intraprese dalla società.

Riferimenti bibliografici

[1] https://www.ilsole24ore.com/art/tutti-numeri-super-bowl-ACgZk1FB.

[2] https://www.sciencedirect.com/topics/computer-science/community-detection.