SOCIAL MEDIA E SOCIAL NETWORK ANALYSIS
UN'ANALISI DEL FESTIVAL DI SANREMO SU TWITTER E SPOTIFY

A PREPRINT

Stefano Daraio
718443
Università degli studi di Milano Biocca
s.daraio@campus.unimib.it

Andrea Armando Tinella
771399
Università degli studi di Milano Biocca
a.tinella@campus.unimib.it

Federico Maria Viotti
785867
Università degli studi di Milano Biocca
f.viotti@campus.unimib.it

July 2, 2019

ABSTRACT

Il festival della canzone italiana o più comunemente noto come festival di Sanremo è una manifestazione canora che ha luogo ogni anno in Italia, a Sanremo, nella provincia di Imperia, in Liguria, a partire dal 1951.

In particolare, l'edizione di quest'anno si è svolta in 5 prime serate dal 5 al 9 febbraio 2019. La competizione ha visto concorrere ad un unico premio 24 cantanti, 22 dei quali di chiara fama e 2 provenienti da Sanremo Giovani 2018, in qualità di vincitori.

Scopo di questo elaborato lo studio e l'analisi di due aspetti, riassumibili in specifici dati, relativi a questo evento: la rete formata dagli artisti partecipanti a questa 69sima edizione e dai loro artisti simili (mediante i dati del ben noto servizio "Spotify") e l'analisi del flusso di *tweet* generati dagli utenti durante la settimana del Festival.

1 Introduzione

Eventi storici e importanti per i mass media come il Festival della canzone italiana sono da sempre seguiti da un gran numero di persone e con il nascere di nuovi strumenti di comunicazione quali i social media la possibilità di espressione degli utenti, spettatori e non, è di gran lunga aumentata rispetto al passato.

E' quindi indubbiamente interessante il monitoring del comportamento di questi soggetti che esprimono il loro parere, le loro convinzioni e sentimenti riguardanti questi eventi, con lo scopo di trarne trend, pattern insoliti e altre informazioni altrimenti non accessibili.

Il canale che consente più facilmente accesso a informazioni di questo tipo è sicuramente il ben noto social network Twitter, che mette a disposizione la sua API da cui poter ricavare una quantità di dati importante, senza la quale non sarebbero possibili tali analisi.

Un altro fenomeno da sempre studiato e che sta assumento sempre più rilevanza negli ultimi anni e in un'ampia gamma di settori, è quello delle reti sociali o reti complesse. Al fine di soddisfare le richieste di tale analisi è stata quindi costruita e analizzata una "rete sociale" basata però non sui dati ottenibili dalla API di Twitter, bensì dalle informazioni rilasciate pubblicamente (sempre tramite API) da Spotify, il noto servizio di streaming musicale.

2 Obiettivo

L'obiettivo di questa analisi è di tipo descrittivo ed analitico e presenta una duplice natura. Da un lato si sono voluti analizzare i tweet relativi al Festival di Sanremo, mentre dall'altro si è cercato di costruire da zero e analizzare una rete di "artisti simili", ponendo come base della costruzione di tale rete la lista dei cantanti partecipanti all'evento.

2.1 Obiettivo: analytics & sentiment analysis

Per quanto riguada la parte di progetto relativa alla sentiment e agli analytics dei social media gli obiettivi perseguiti sono stati:

- analizzare il testo dei tweet relativi all'intera settimana del Festival attraverso la sentiment analysis degli stessi, sia in termini di n di tweet positivi/negativi che in termini di sentimenti (emozioni) espresse all'interno degli stessi. In questo caso si è studiata l'evoluzione del sentiment giorno per giorno, lungo l'intera settimana.
- generare degli analytics, a partire dalle informazioni ottenute tramite l'API di Twitter, che possano rispondere a quesiti quali:
  - come si distribuisce il numero di tweet generati dagli utenti durante la settimana del Festival;
  - quali sono gli utenti di Twitter più menzionati e gli hashtag più utilizzati;
  - quali fonti (devices) più utilizzati per "twittare" in relazione al n di followers, di retweet, e mi piace ottenuti dai singoli utenti;
  - qual è la distribuzione geografica dei tweet effettuati realtivamente all'Italia e al resto del mondo;

2.2 Obiettivo: social network analysis

Per quanto riguarda l'analisi delle reti sociali si sono perseguiti due obiettivi principali:

- in primis si è voluta sperimentare la generazione di una rete non basata su relazioni tra utenti, bensì sulle relazioni esistenti tra un certo numero di musicisti, relazioni derivanti dal diretto utilizzo di Spotify da parte dei suoi utenti.
- una volta generata la rete questa è stata analizzata mediante l'utilizzo di specifiche metriche tra le quali centralità e community detection.

3 Data

Come già scritto, le fonti dati da cui si sono ottenute le informazioni necessarie sono le rispettive API di Twitter e Spotify rispettivamente utilizzate per i dati necessari alla sentiment/analiytics e all'analisi della rete.

3.1 Twitter

Dopo alcuni primi tentativi di analisi dei tweet italiani non andati a buon fine, si è deciso di scaricare due informazioni distinte: una specifica per la sentiment analysis e una per la parte più generale degli analytics.

In particolare, sono stati scaricati tutti i tweet postati in lingua inglese nella settimana del festival con l'hashtag ufficiale dell'evento "#Sanremo2019" così da poter utilizzare i lessici e dizionari necessari all'analisi testuale; succesivamente, sono stati scaricati per lo stesso arco temporale tutti i tweet effettuati dagli utenti con gli hashtag ufficiali relativi alle 24 canzoni in gara, sui quali sono state effettuate analisi statistiche descrittive mediante il software "Tableau" (in allegato alla cartella della consegna sono presenti i workbooks di Tableau con all'interno i relativi progetti).

Di seguito un elenco riportante gli hashtag tramite i quali sono state effettuate le chiamate all'API di Twitter:

"RollsRoyce", "LeNostreAnimeDiNotte", "MiSentoBene", "PerUnMilione", "Argentovivo", "ParoleNuove", "Nonno-Hollywood", "SoloUnaCanzone", "senzafarloapposta", "AspettoCheTorni", "RoseViola", "MusicaCheResta", "LaRagazzaColCuoreDiLatta", "cosatiaspettidame", "Dov'è l'Italia", "Soldi", "IRagazziStannoBene", "MiFaràTrovarePronto", "UnAltraLuce", "LultimoOstacolo", "UnPoComeLaVita", "AbbiCuraDiMe", "ITuoiParticolari", "LamoreèUnaDittatura", "Sanremo2019".

Nonostante il numero dei tweet non fosse proibitivo da gestire con i limiti imposti dall'API di Twitter, per questioni di correttezza nelle politiche di gestione dei dati si è deciso di pianificare un'estrazione giornaliera retroattiva di un

giorno, così da non raggiungere mai il limite imposto di 18 mila tweet ogni 15 minuti e non incorrere nella possibilità di escludere informazioni, anche se poche, dall'analisi.

Una volta ottenuti i dataset relativi alle canzoni in gara, questi sono stati uniti tramite la creazione di una variabile identificatore (l'hashtag, che ha la funzione di chiave univoca) così da ottenere un'unico dataset ordinato secondo il timestamp (si fa notare che twitter permette di ottenere infatti la variabile "createdat" che contiene la data e l'ora in cui il tweet è stato pubblicato). Infine sono state eliminate le variabili non utili al fine delle analisi successive.
Le stesse operazioni sono state inoltre effettuate sui tweet in lingua inglese.

3.2 Spotify

La sezione di progetto relativa alle reti è stata effettuata, come precedentemente scritto, tramite i dati della piattaforma musicale Spotify. La sua API (che è stata utilizzata tramite la libreria "spotifyR" in ambiente R) permette infatti di ottenre tantissime informazioni relative agli artisti iscritti al servizio come tali, informazioni riguardanti le loro discografie i loro indici di popolarità e le caratteristiche audio delle loro canzoni, oltra ai testi completi dei brani.
Un'informazione molto interessante che Spotify permette di ottenere è quella realtiva agli artisti simili ad un musicista. L'API infatti, tramite il comando "get related artists" restituisce l'elenco dei 20 artisti (Spotify limita l'output a 20 per artista) più simili a quello dato in pasto alla funzione; la somiglianza tra gli artisti inoltre viene calcolata sulla base di quanto la coppia di artisti venga ascoltata all'interno della stessa playlist degli utenti.

Ciò significa che due musicisti, per quanto dissimili nel genere e/o nello stile, vengono classificati come simili se entrambi vengono inseriti nelle playlist di molti utenti.
Questa metodologia di Spotify nel definire la similarità degli artisti è stata ritenuta interessante per la generazione di una rete in quanto la somiglianza (e dunque il collegamento) tra gli artisti è strettamente collegato all'utilizzo del servizio da parte degli utenti e dai loro gusti in fatto di musica.

Per la costruzione della rete si è quindi creata una lista dei 27 artisti partecipanti al festival e per ognuno di essi sono stati cercati i 20 artisti più simili (in particolare il loro nome, la popolarità e il numero di follower) così da ottenere un primo livello della rete.

Successivamente si sono cercati gli artisti simili a tutti quelli che sono risultati dal passaggio precedente, andando a creare un secondo livello di profondità della rete.

Infine si è ricercato un terzo livello di profondità, andando ad ampliare ancora la grandezza della rete.
I dati così ottenuti sono stati uniti in un unico dataset formato da tutti i collegamenti e successivamente è stato effettuato il preprocessing eliminando tutti i collegamenti duplicati in modo da ottenre al massimo due collegamenti per ogni coppia di artisti: uno che indica la somiglianza di un artista, ad esempio A, ad un altro, B e l'altro che indica la somiglianza di B ad A.

Oltre alle informazioni relative ai collegamenti tra artisti simili sono state scaricate successivamente altre informazioni generali di ognuno dei cantanti partecipanti al festival (numero di album, popolarità degli album e delle singole tracce, date di pubblicazione e atichette con cui sono stati pubblicati gli album) così da poter effettuare in ambiente Tableau alcuni grafici descrittivi.

I file di output generati sono i seguenti

- **mercati_tutti_artisti.csv** : ci sono tutti gli artisti che hanno cantato a Sanremo con tutti gli album pubblicati, la relativa popolarità e le etichette discografiche di ogni album; inoltre, c'è l'elenco di tutti i mercati in cui l'album è disponibile.

- **tutti_artisti.csv** : i dati sono simili alla tabella precedente, considerando però un livello di granularità a livello di canzone.

- **tabella_archi.csv** : sono presenti le relazioni tra coppie di artisti e la relativa frequenza con cui compare tale relazione

- **tabella_nodi.csv** : è presente l'elenco di tutti gli artisti fino al terzo grado, ovvero 2793, con l'informazione sul numero di follower e la popolarità.

4 Modelli e Visualizzazioni

4.1 Social Analysis Tweet

Dopo aver ottenuto i *tweet* ed effettuate le operazioni di preprocessing abbiamo effettuato delle prime analisi descrittive in Tableau per vedere la distribuzione nei vari giorni e considerando le diverse ore della giornata per tutto il periodo del festival analizzando tutti gli hashtag in esame.

Vediamo nella [Fig. 1] che il numero di hashtag presenta tendenzialmente un aumento con il passare dei giorni, in particolare qualli con un andamento più interessanti sono :

- **Abbi Cura di Me** : Cristicchi ha pubblicato un album nel periodo in cui è presente un maggior numero di tweet.
- **Rolls Royce** : la risonanza mediatica dovuta al servizio di Striscia la notizia è uno degli elementi che ha portato al picco che vediamo.
- **Soldi** : Il vincitore presenta un andamento piuttosto basso di tweet, con un aumento durante il solo giorno della vittoria.

per quanto riguarda la rappresentazione per fasce orarie invece vediamo che i picchi in cui sono stati postati i tweet corrisponde alle ore di diretta del festival.

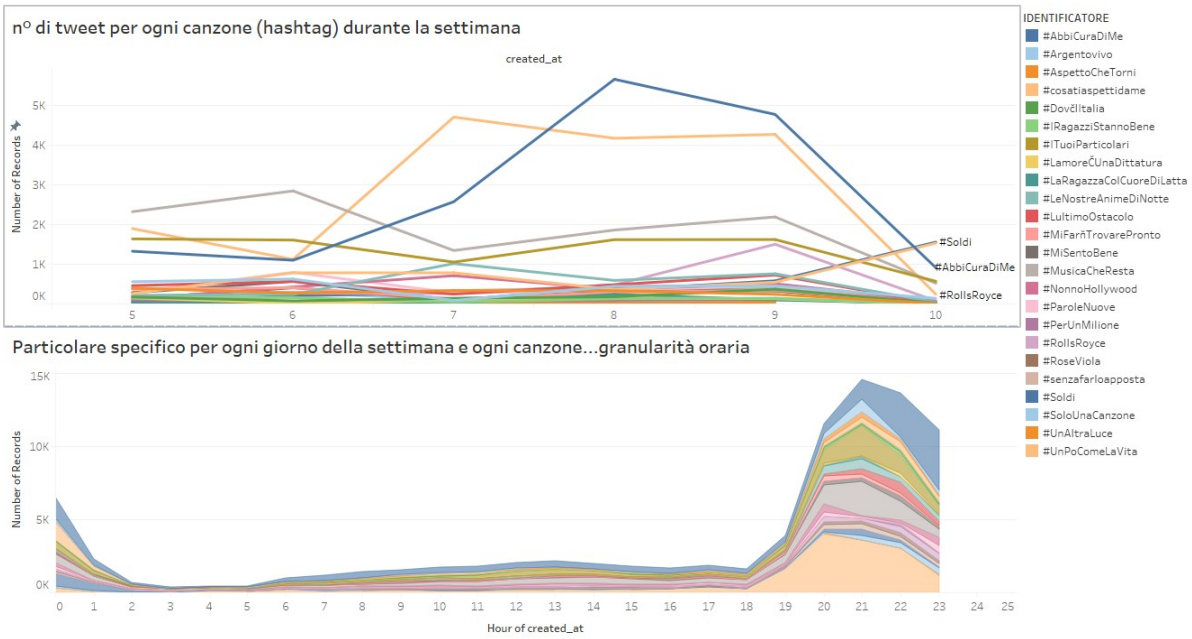

Figure 1: .

considerando gli account che hanno pubblicato i tweet, vediamo come non sono presenti solo quelli direttamente collegati con l'organizzazione del festival ma anche altri che hanno avuto un importante numero di like. E' interessante vedere ad esempio come ci siano tweet da parte di importanti marchi di moda come Dolce e Gabbana e Moschino o altri di tipo ironico ad esempio trash italiano. Questi presentano un importante numero di follower e infatti sono quelli che hanno anche il maggior numero di retweet e quindi una maggiore risonanza come possiamo vedere dal grafico [Fig. 2].

Dalla distribuzione geografica dei tweet è evidente che il maggior numero di tweet e quelli con il maggior numero di mi piace si concentrano nelle principali città italiane, in particolare Roma e Milano, seguiti da Torino e Roma come da [Fig. 3]

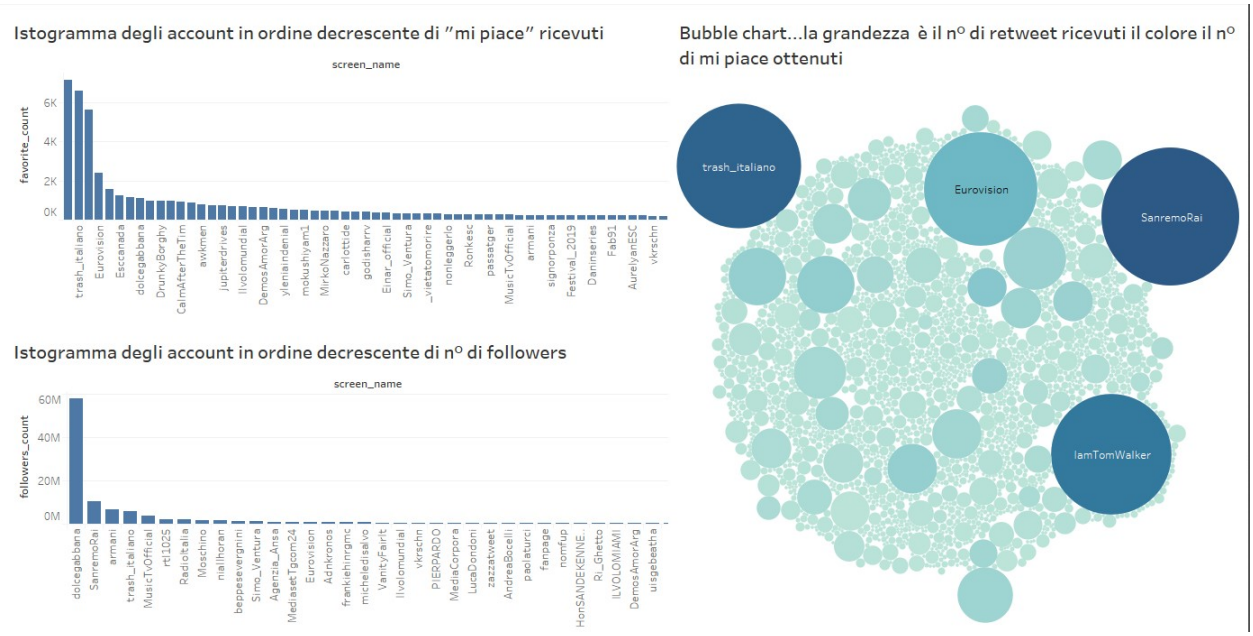

Figure 2: .

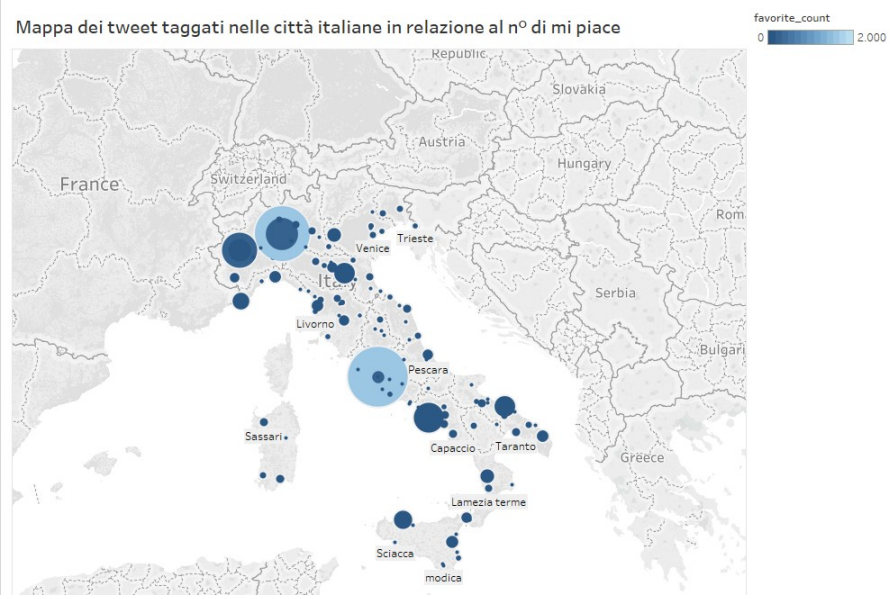

Figure 3: .

4.1.1 Sentiment Analysis

Abbiamo quindi calcolato la sentiment analysis sui tweet in inglese usando il sofware statistico "R" ed inparticolare il pacchetto suzynet.Questo in quanto per quelli in italiano non è possibile trovare dei dizionari che risultino completi.

Per effettuare questa analisi ci siamo basati sul **NRC Word-Emotion Association Lexicon (aka EmoLex)**, il quale consiste in una lista di parole inglesi e la relativa associazione con le 8 emozioni di base (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) e due sentimenti (negative and positive). L'annotazione è stata fatta manualmente tramite crowdsourcing.

Considerando l'intera settimana vediamo come ci sia un rapporto tendenzialmente paritario di sentimento positivo e negativo con una leggera tendenza a quello positivo [Fig. 4], e questo lo vediamo anche nei singoli giorni.

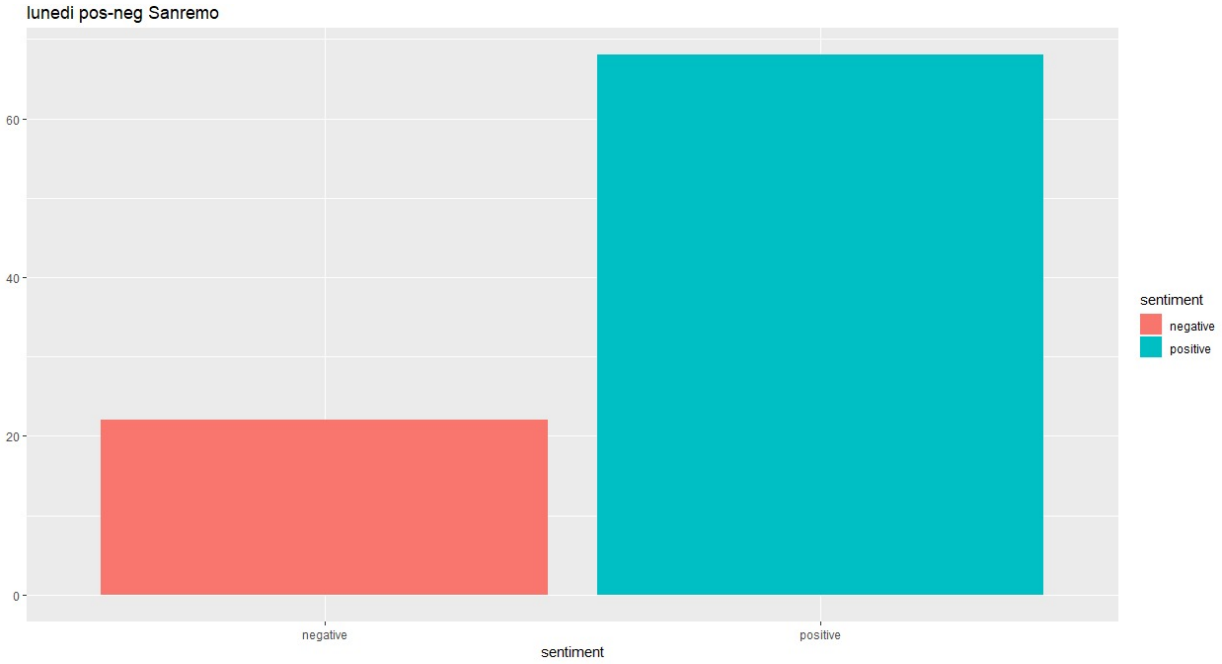

Figure 4: .

L'unico che differisce è l'ultimo giorno, ovvero Lunedì nelle prime ore della mattina, dove vediamo una preponderanza di tweet positivi rispetto a quelli negativi. Questo indica una accoglienza positiva in concomitanza alla conslusione del festival [Fig. 5].

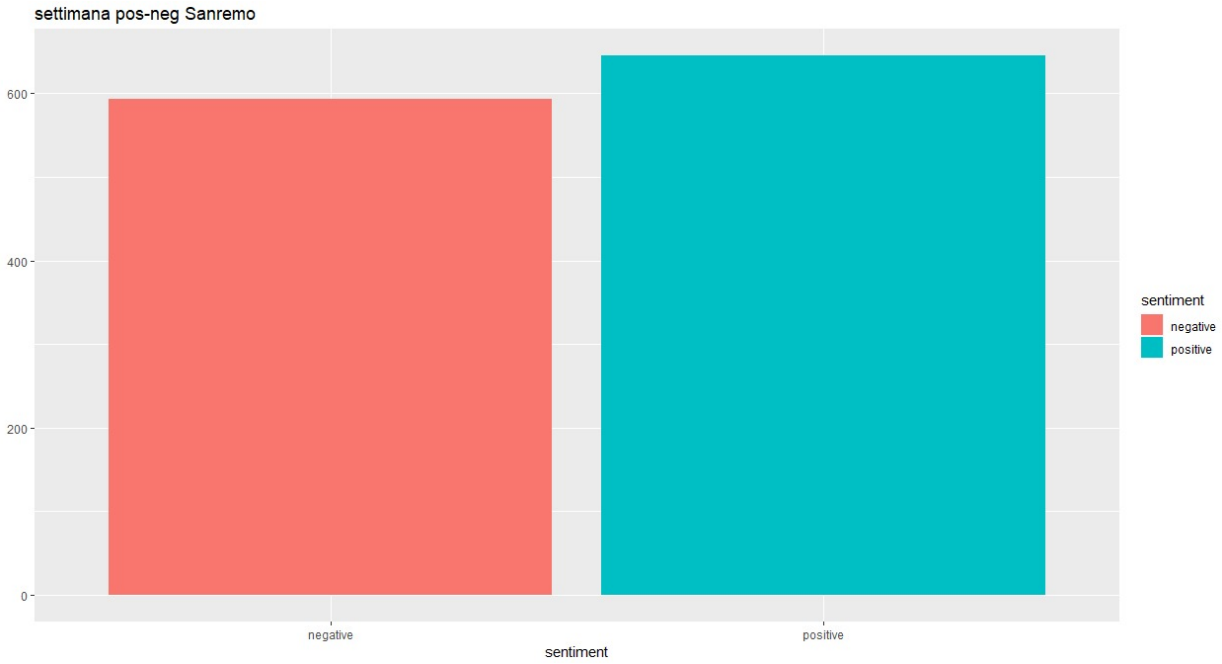

Figure 5: .

Dall'analisi basata sulla NRC Word-Emotion Association Lexicon sull'intera settimana di tweet in lingua inglese si evince che il sentimento dominante sia quello di fiducia, questo possiamo imputarlo all'importanza e alla valenza storica dell'evento ed evidenzia la percezione di sentirsi ascoltati e parte della manifestazione da parte delle persone. Vediamo

d'altro canto come ci siano stati pochi commenti associati ad un sentimento di sopresa, questo in quanto non ci sono stati eventi straordinari o una sostanziale differenziazione da quelle che erano le aspettative [Fig. 6].

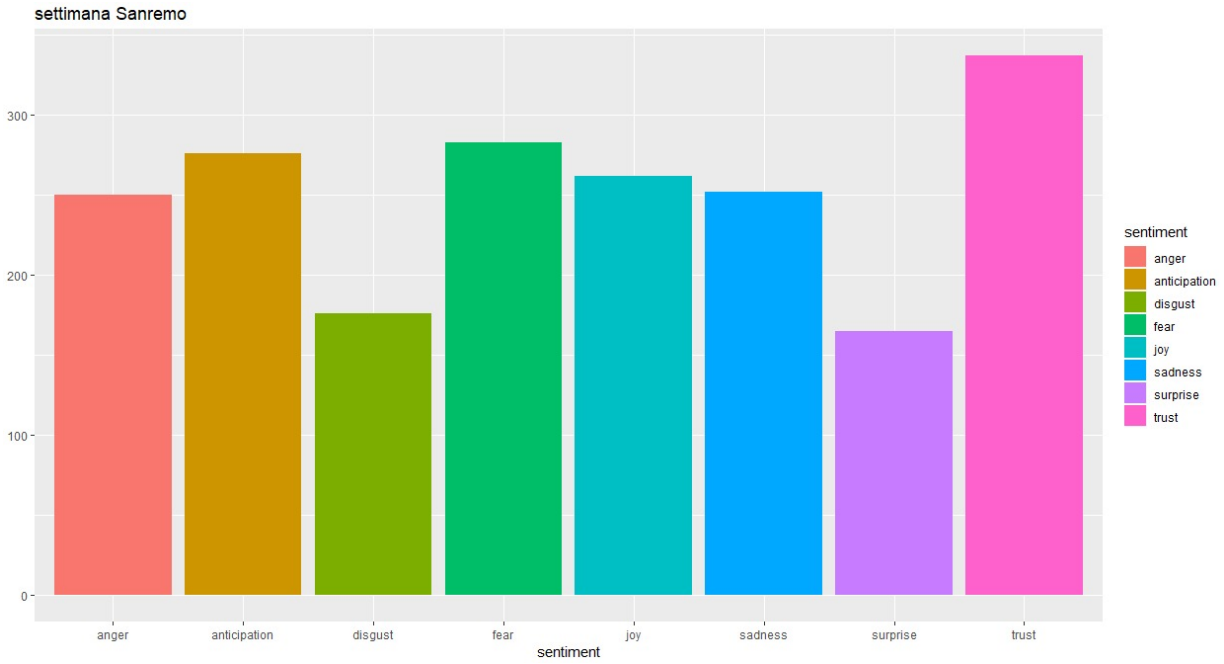

Figure 6: .

Le reazioni contestuali alla proclamazione del vincitore sono principalmente di aspettativa (anticipation) come è lecito immaginare mista a felicità per quello che è stato l'esito finale e la conclusione dell'evento [Fig. 7]

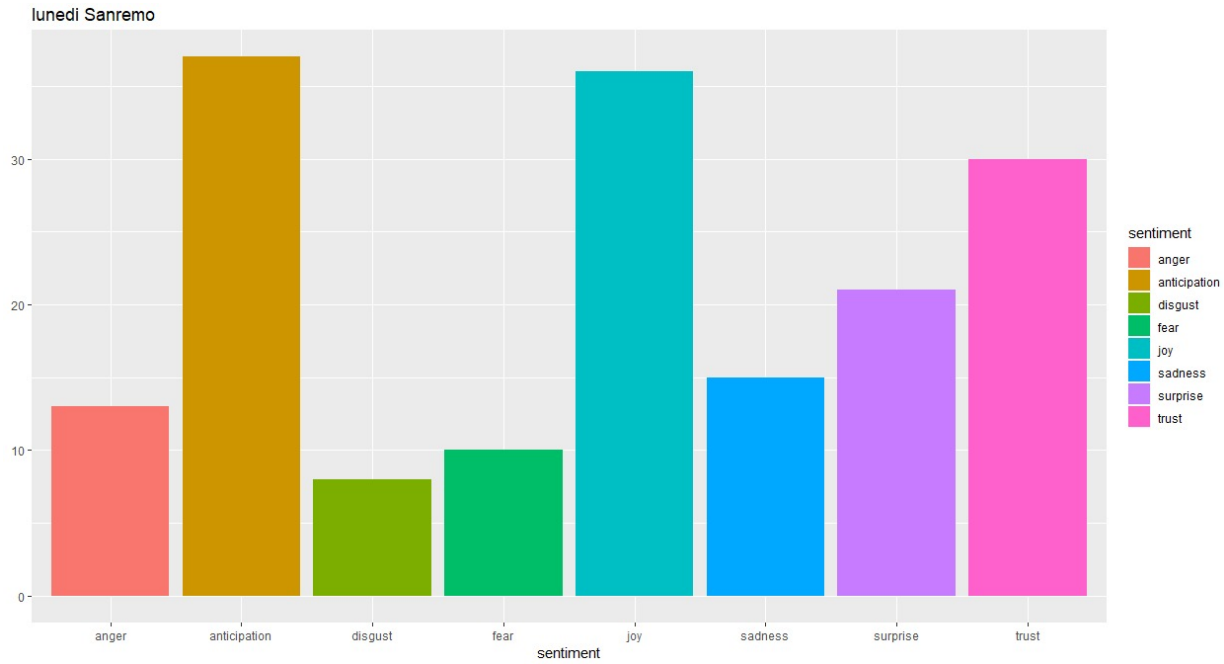

Figure 7: .

4.2 Reti

Per quanto riguarda l'analisi della rete, si è scelto di utilizzare il software statistico "r", con IDE "rstudio" e avvalendosi della libreria "igraph", particolarmente adatta nella gestione e analisi di reti complesse.

Dopo la fase di caricamento dei due file in formato .csv contenenti gli schemi dei nodi e degli archi, si è proceduto a un leggero preprocessing dei dati, eliminando i nodi isolati (aventi quindi grado nullo) e non considerando gli attributi non rilevanti ai fini dell'analisi.

Data la tipologia e la struttura della rete, non c'è stato bisogno di valutare come trattare i cappi o i pesi degli archi, in quanto la rete era sprovvista di cappi e i pesi avevano tutti valore unitario: è stato quindi effettuato un primo plot della rete, che ha mostrato le prime difficoltà da parte del programma nel gestire e visualizzare una rete di tale complessità.

Si è quindi proceduto all'analisi delle caratteristiche descrittive della rete e dei nodi:

- La densità, cioè il rapporto tra gli archi della rete e gli archi massimi possibili, uguale a 0.002392785, un valore notevolmente basso
- La reciprocità, equivalente alla percentuale delle diadi con legami reciproci tra di loro, pari a 0.4319244. Si è inoltre ricavato il numero di diadi reciproche, asimmetriche e nulle
- La transitività, o coefficiente di clustering globale, data dal rapporto di triadi chiuse rispetto al numero totale delle triadi, pari a 0.4307229.
- Il diametro, data dalla distanza geodesica più lunga presente nella rete considerando la direzione degli archi, pari a 25, e uguale nello specifico al cammino tra i cantanti Canova e Fabio Lepore.
- Il grado dei nodi, sia rispetto ai nodi entranti, che ai nodi uscenti, che alla somma dei due, cioè il numero degli archi connessi a ogni nodo della rete.

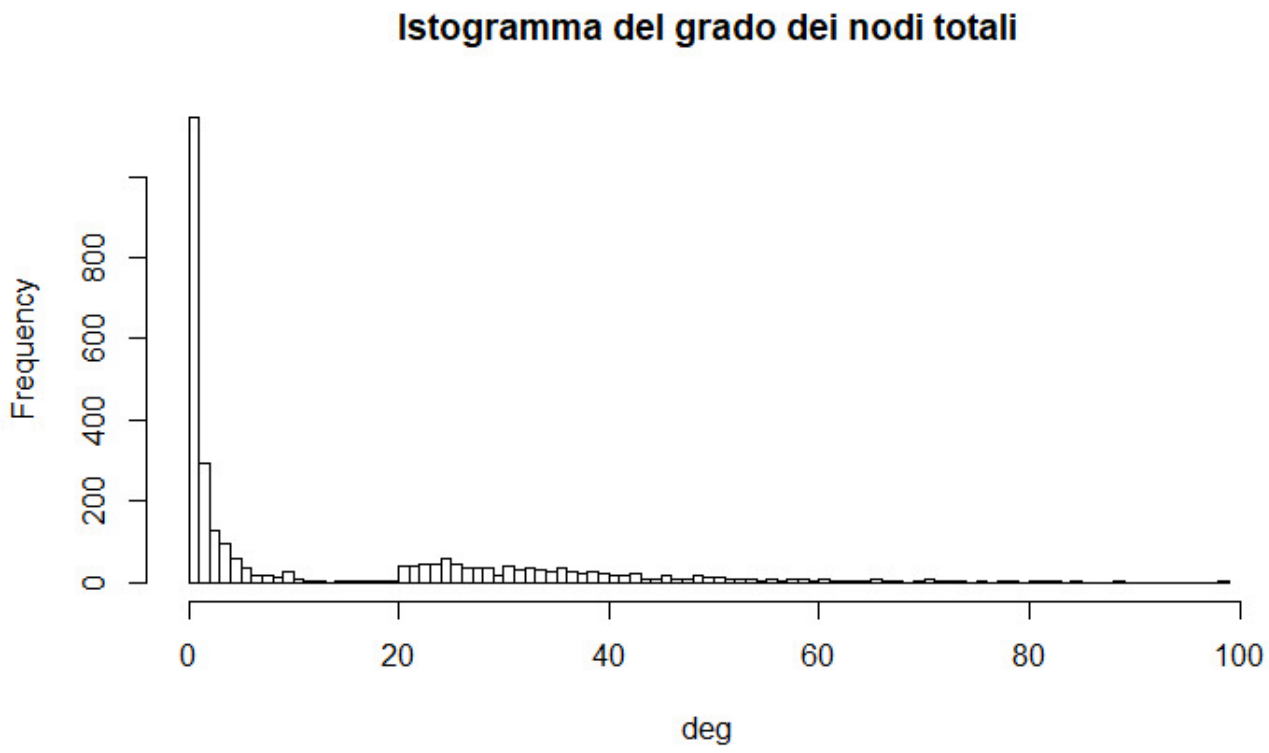

Figure 8: .

Si nota come la stragrande maggioranza dei nodi ha grado unitario, mentre i nodi a grado maggiore arrivano fino ad avere grado vicino al 100

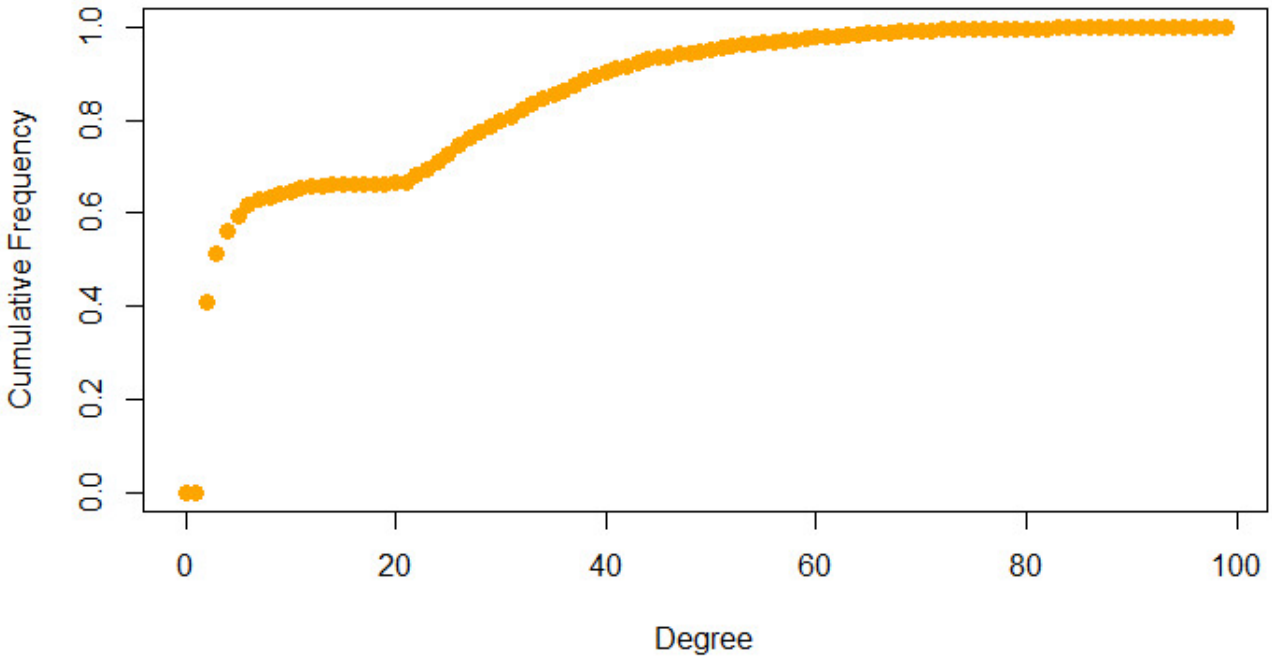

Figure 9: .

- La centralità, cioè la differenza tra il numero di collegamenti per ogni nodo diviso per la massima somma possibile di differenze, pari a 0.02593278, e con numero massimo di collegamenti pari a 7781310.
- La vicinanza, o closeness centrality, cioè la lunghezza dei cammini più brevi dato un vertice, con un valore medio pari a 0.1606205.
- La betweeness centrality, pari al numero di volte in cui un nodo viene attraversato durante un cammino breve tra due nodi, con un valore medio per nodo pari a 0.05906594.
- La definizione di hub e autorità, che permette di trovare i nodi più importanti della rete valutati rispetto agli archi entranti e uscenti nella rete. Nella rete in questione, gli hub più rilevanti sono cantanti come Gianluca Grignani, Alex Baroni, Nek e Francesco Renga, mentre le autorità più influenti sono Raf, Alex Britti, Francesco Renga e Fabrizio Moro.

Una seconda parte dell'analisi della rete ha riguardato i cammini e i percorsi della rete. Si sono perciò ottenuti i seguenti risultati:

- Lunghezza del cammino medio diretto e non diretto, rispettivamente pari a 8.48802 e 6.08452.
- Distanze dei cammini più brevi in forma grafica, processo non eseguito con successo in quanto la visualizzazione della rete ottenuta era troppo esigente in termini di risorse, avendo a disposizione risorse informatiche nella media
- Distanze di un nodo da qualsiasi altro nodo nella rete, dopo aver definito il nome del nodo preso in esame. Anche in questa circostanza, la visualizzazione grafica del risultato non è stata soddisfacente, mentre la visualizzazione tabellare è corretta ma poco intuitiva.
- Cammino più breve data una coppia di nodi, che ha mostrato anche i nodi intermedi attraversati nel cammino, solo in forma testuale per i ricorrenti problemi computazionali

La terza fase dell'analisi ha riguardato la community detection, con risultati però non soddisfacenti data la complessità della rete e la carente potenza computazionale degli strumenti informatici in possesso del gruppo per gestirla.

Gli unici output degni di nota sono stati la ricerca delle cricche con numero massimo di nodi nella rete, con quattro cricche composte da 17 nodi, cricche però molto simili per varietà di nodi le une alle altre.

Si è reso necessario, dunque, utilizzare un altro software per andare incontro alle esigenze analitiche, che fosse in grado di gestire il carico di lavoro richiesto dalla rete.

La scelta è ricaduta su Gephi, programma creato ad-hoc per la gestione di reti complesse e più peformante rispetto a R. Nella parte relativa alla community detection, si è applicato l'algoritmo della modularity maximixation, appartenente alla famiglia delle community detection network-centric.

L'algoritmo, che presenta un valore variabile da -1 a 1, esprime la tendenza della rete ad essere randomica o meno. In caso i collegamenti tra i nodi risultano essere non randomici, ci sono buone probabilità che la rete sia molto simile alle reti nel mondo reale, e avrà perciò un valore di modularità alto.

Impostando l'algoritmo con i parametri suggeriti, la rete ha dato risultati molto incoraggianti, con un indice di modularità pari a 0.822, e suddividendo la rete in 21 comunità, valore anch'esso molto buono, mostrando una suddivisione ottimale considerando la grandezza della rete.

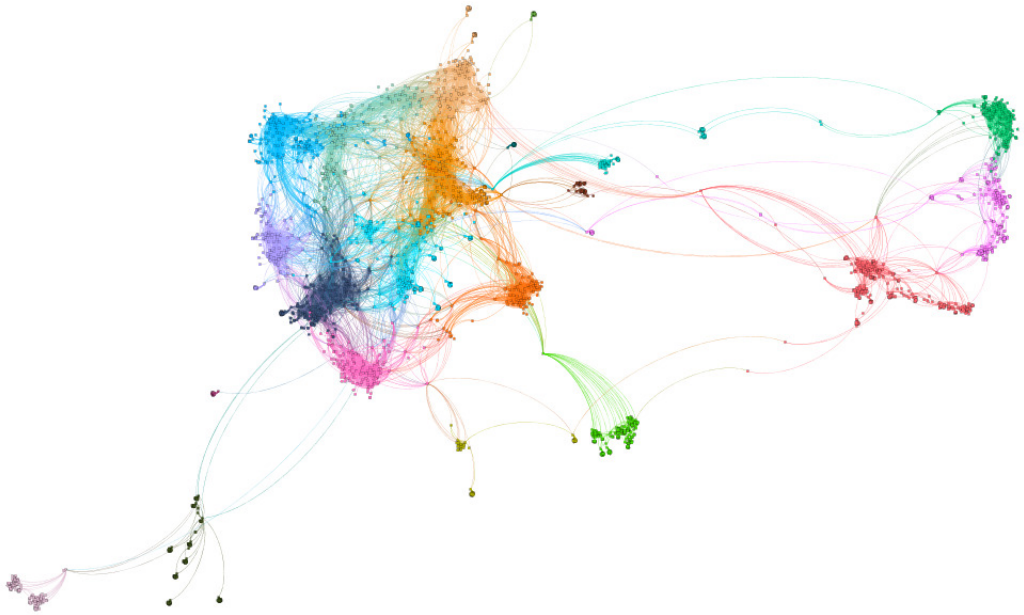

Figure 10: .

5 Conclusioni: criticità e possibli sviluppi

Le maggiori criticità riscontrate nello svolgimento del progetto sono state di diverso tipo per le due parti :

- **Twitter** : tramite le API di twitter non possiamo avere la garanzia riguardo la totalità dei tweet raccolti, inoltre risulta difficile individuare i tweet "genuini" degli utenti da quelli generati da bot o di carattere pubblicitario, di spam o in generale non direttamente inerente all'hashtag citato. Un altro problema riguardante la sentiment analysis è la penuria di dizionari completi e performanti per questo ambito in lingua italiana. Bisogna inoltre considerare come twitter possa non rappresentare un campione significativo ed esaustivo riguardo l'effettivo interesse delle persone per l'evento trattato ma è l'unico strumento che ci permette di avere queste informazioni.
- **Grafo** : anche la parte relativa alla rete non è stata esente da problematiche di diversa natura. In particolare, dovendosi basate su API non ufficiali di Spotify ma create dalla community, è mancata la possibilità di personalizzazione delle funzioni. Questo fenomeno è risultato evidente nella costruzione della rete, in quanto il valore di default di 20 artisti simili non poteva essere modificato.

  Inoltre, non è stato possibile trasferire tutte le conoscenze in materia nell'analisi delle reti, in quanto la grandezza della rete generata non ha permesso di testare diversi algoritmi, in particolare quelli di community detection, ponendo dei limiti alla varietà dell'analisi.

Si possono inoltre osservare alcuni aspetti comuni alle due parti di progetto. Uno dei più importanti riguarda l'assenza di un esperto di dominio che possa interpretare in modo più competente i dati raccolti e dare supporto sullo stato dell'arte del mercato della musica italiana: un supporto del genere avrebbe apportato una maggior consistenza all'analisi svolta.

Un altro aspetto importante è che, nonostante le due parti del progetto avessero obiettivi e metodologie differenti, si è spesso giunti a conclusioni simili e correlate. Ad esempio si poteva dedurre che gli artisti più famosi sono importanti sia all'interno della rete che a livello di presenza social, legittimando la loro importanza da due fonti diverse.

In conclusione possibili sviluppi relativi a questo progetto potrebbero essere il confronto tra i dati ottenuti da Twitter e quelli di altri social media come Facebook o Instagram, per effettuare una comparazione del sentiment e dei comportamenti degli utenti sulla base dei diversi media, rispetto l'evento considerato nell'analisi.
Inoltre, per quanto riguarda l'analisi della rete, si potrebbe generare una rete sociale a partire dai dati relativi ai tweet e porre quindi l'attenzione sugli utenti e non sugli artisti partecipanti.

References

[1] Rstudio

[2] Twitter API

[3] Spotify API

[4] Tableau software

[5] Gephi

[6] Link all'analisi Tableau dei tweet