VINCITORI E VINTI, MA PER CHI?

Miryam Bartoli 867107, Claudio Maffi 875789, Federico Campanella 793652 and Irene Lupino 866912

Corso di Data Management & Data Visualization

Abstract: Il seguente elaborato ha l'obiettivo di illustrare se gli utenti della piattaforma social Twitter e l'applicazione di musica Spotify sono rappresentativi del pubblico votante del Festival di di Sanremo. A questo scopo sono stati analizzati i tweet postati nelle giornate dell'evento contenenti la parola *#Sanremo2021* e gli ascolti di Spotify nello stesso periodo. Sono quindi state stilate, per entrambe le piattaforme, classifiche giornaliere successivamente confrontate con quella ufficiale del concorso. Inoltre, attraverso l'analisi di indicatori specifici relativi alle canzoni degli ultimi undici Festival (2010-2021) e ai loro generi si è cercato di determinare eventuali caratteristiche volte ad individuare la possibile canzone vincitrice.

INTRODUZIONE

Il Festival di Sanremo è da sessantacinque anni la manifestazione televisiva più seguita dal pubblico italiano, superata, in termini di ascolti, solamente dalle partite della Nazionale Italiana di calcio. Nasce nel 1951, inizialmente trasmesso in diretta radiofonica (1951-1954), successivamente in diretta televisiva. Riscuote grande successo nella scena musicale italiana ispirando altre manifestazioni a livello europeo, come l'Eurovision song contest. Nella storica sede del concorso, il palco del Teatro Ariston, si sono esibiti alcuni tra i più famosi cantanti della storia italiana. Ogni anno, artisti affermati e giovani promesse, presentano un brano inedito che viene giudicato da un misto tra giuria scelta e popolare. Il Festival ha accompagnato l'evolversi della musica italiana, rimanendo un evento mediatico di grande portata, nonostante l'eterogeneità del panorama musicale odierno.
La formula del concorso è stata modificata negli anni, e la manifestazione solleva puntualmente dibattiti e polemiche che portano all'evento risonanza mediatica nazionale e internazionale. La rivoluzione informatica ha portato alla ribalta l'importanza dei social network, e il Festival di Sanremo è riuscito nel tempo a confermarsi anche online l'evento più atteso. Considerato che attraverso i social è possibile indagare le opinioni delle persone, nella nostra analisi abbiamo voluto testare quanto le piattaforme Twitter e Spotify possano essere rappresentative degli spettatori votanti di Sanremo. Dopodichè abbiamo cercato di estrarre le caratteristiche principali delle canzoni in gara negli ultimi 10 anni, analizzando quelle più decisive al fine di arrivare alla vittoria.

1 DATA MANAGEMENT

1.1 ACQUISIZIONE DATI

1.1.1 Twitter

Il primo passo è stato l'acquisizione dei dati provenienti dalla piattaforma Twitter. La raccolta è avvenuta in tempo reale tramite l'utilizzo di *Apache Kafka* in Jupyter Notebook di Python: prima di tutto è stato creato il topic "sanremotweet", dopodichè sono stati creati i relativi producer e consumer in grado di archiviare in coda tutti i dati e successivamente salvarli su *Mongodb*.
La libreria *Tweepy*, utilizzata per lo scraping dei tweet, ha permesso di ottenere informazioni complete quali:

- testo;
- username;
- data;
- like;
- follower

Nonostante Twitter risulti una delle piattaforme social più accessibili allo scraping, non è stato possibile ottenere per tutti le coordinate e il luogo di pubblicazione del post. Il periodo di acquisizione va dal 3 Marzo 2021 (data della seconda puntata) fino al 6 Marzo 2021 (data dell'ultima puntata). Sono stati esclusi i tweet del 2 Marzo 2021 in quanto la prima serata ha visto la partecipazione della metà dei big in gara. Inoltre sono stati esclusi i tweet del 7 Marzo 2021 dal momento che l'ultima serata si è conclusa oltre la mezzanotte: la riapertura della votazione del podio, e la successiva proclamazione del vincitore, avrebbe influenzato esageratamente l'analisi del sentiment.

1.1.2 Spotify

La raccolta dei dati di Spotify è avvenuta tramite scraping su Python grazie alla libreria *spotipy*. Sono state estratte le playlist contenenti le canzoni di Sanremo nell'arco temporale 2010-2021. La libreria *spotipy* ha permesso di ottenere, oltre al titolo della canzone e l'artista, indici di performance (Figura 1) come:

- **danceability**: quanto una canzone risulta adatta a essere ballata, tramite una combinazione di caratteristiche musicali come tempo, ritmo, beat e regolarità;
- **valence**: positività di una canzone. Più una canzone ha una valence alta, più trasmetterà un sentimento positivo (felicità, amore, euforia). Al contrario, se bassa, rispecchierà un sentimento negativo (tristezza, rabbia, depressione);
- **energy**: intensità della canzone. Le tracce più energiche suonano veloci, forti e rumorose (esempio heavy metal);
- **tempo**: media stimata di una traccia in beat al minuto;
- **loudness**: rumorosità media di una traccia in decibels;

- **speechness**: quantità di parole in una canzone. Più una traccia risultata parlata, più il suo valore sarà vicino a 1;
- **instrumentalness**: quantità di parti strumentali;
- **liveness**: presenza di un audience durante la registrazione. Un alto valore di liveness rappresenta una probabilità maggiore che la traccia sia stata suonata in live;
- **acousticness**: è una misura di confidenza compresa tra 0 e 1, dove maggiore è il valore e maggiore è l'acustica;
- **key**: chiave media musicale di una canzone;
- **mode**: modalità di una canzone. La modalità può essere maggiore (se è pari a 1) o minore (se è pari a 0);
- **duration**: durata della traccia in millisecondi;
- **time signature**: tempo in chiave che specifica i beat presenti in una frazione

Anno	danceability	loudness	acousticness	duration_ms	energy	instrumentalness	key	liveness	mode	speechiness	tempo
2010	0.619	-7.015	0.4950	238080	0.536	0.000000	6	0.1170	1	0.0407	124.059
2011	0.479	-3.325	0.6360	253987	0.663	0.000000	9	0.1310	1	0.0413	74.501
2012	0.569	-5.853	0.0420	225253	0.731	0.000000	6	0.1290	1	0.0342	126.884
2013	0.556	-4.438	0.1960	218893	0.632	0.000000	6	0.1140	1	0.0359	126.030
2014	0.703	-6.173	0.5080	219200	0.555	0.000000	1	0.1340	1	0.0284	95.010
2015	0.466	-5.853	0.0586	224947	0.617	0.000092	1	0.1470	1	0.0290	142.121
2016	0.411	-4.615	0.0402	244747	0.753	0.000000	9	0.1160	1	0.0287	174.041
2017	0.669	-4.014	0.0424	217625	0.832	0.000004	10	0.1110	1	0.0750	114.046
2018	0.568	-4.612	0.3090	208080	0.890	0.000000	6	0.2070	1	0.0554	90.025
2019	0.690	-9.832	0.0404	195476	0.581	0.000000	3	0.1110	0	0.0466	94.904
2020	0.417	-6.055	0.2180	216140	0.626	0.000000	6	0.0805	1	0.0366	143.358
2021	0.620	-3.082	0.0014	192655	0.944	0.000000	4	0.7330	0	0.0863	103.024

Figure 1: Indicatori di spotify

1.1.3 Sanremo

Le classifiche di ogni serata sono state scaricate dal sito ufficiale del Festival di Sanremo.

1.2 PREPROCESSING E INTEGRAZIONE DATI

Sono stati analizzati i tweet raccolti per individuare eventuali osservazioni doppie o nulle. In seguito è stata eseguita un'operazione di normalizzazione relativa al nome del gruppo "maneskin", in quanto presente con diverse diciture.
Anche per il dataset di Spotify sono stati analizzati eventuali valori nulli o anomali. Sono stati riscontrati errori grammaticali all'interno dei titoli delle canzoni che hanno portato alla necessità di correggerli e normalizzarli sulla base delle classifiche ufficiali di Sanremo. Inoltre, comprendeva anche brani relativi a "Sanremo giovani" che è stato deciso di isolare. Successivamente si è deciso di integrare il dataset di Spotify con due fonti di dati:

- gli ascolti giornalieri per ogni canzone di Sanremo 2021 nello stesso periodo dello streaming dei tweet. Questi dati sono stati presi dal sito *Spotifycharts*;
- il genere di tutte le canzoni partecipanti a Sanremo dal 2010 al 2021 estratti dal sito *Rollingstones Italia*.

1.3 DATA EXPLORATION

1.3.1 Twitter

Da una prima esplorazione dei tweet sulle parole maggiormente presenti nel testo è subito risultato che i più chiacchierati sono stati 'Fiorello' e 'Achille Lauro', entrambi parte dello spettacolo e non dei cantanti in gara, Figura 2. Si è dunque deciso di concentrare l'analisi

Figure 2: Parole maggiormente ripetute nei tweet

sui tweet contenenti i nomi dei cantanti in gara al fine di estrapolare i giudizi delle persone in merito.

Sentiment Analysis

Per poter condurre la sentiment analysis si è proceduto alla pulizia dei tweet. Inizialmente si è cercata la presenza di eventuali, bot che avrebbero potuto rendere la sentiment troppo sbilanciata. In un secondo momento sono stati estratti username contenenti la parola "bot", poi controllati manualmente.
É stato deciso di rimuovere tweet derivanti da società di informazione (come giornali, telegiornali e riviste) perchè avrebbero potuto falsare i valori della sentiment. Si è poi proceduto ad effettuare una ricerca manuale di ulteriori telegiornali e giornali italiani, eliminando i loro tweet.
A questo punto sono state rimosse dal testo dei tweet le stopwords, parole valutate neutre dalla libreria *vader*. Grazie a questo passaggio si è potuto constatare che il punteggio, positivo o negativo, dei testi dei tweet ripuliti aveva un valore più marcato. La libreria utilizzata assegna ad ogni tweet un valore compreso tra -1 (sentiment negativo) e +1 (sentiment positivo).
Al fine di approfondire l'analisi è stata aggiunta una colonna in cui era indicato il nome del cantante al quale il tweet si riferiva, in modo da poter osservare l'andamento del sentiment medio riferito al solo cantante durante il Festival di Sanremo.
Nonostante l'utilizzo di una libreria Python piuttosto rinomata nell'ambito statistico, si è notato che i tweet non sono sempre stati classificati correttamente, ottenendo un'accuratezza non molto elevata. Si è deciso di non adottare una soluzione manuale a questo problema dato che il controllo e la correzione di migliaia di tweet sarebbe risultato troppo dispendioso e poco scalabile.

Figure 3: Sentiment media

1.3.2 Spotify

Generi

Da un'analisi iniziale sui generi delle canzoni (Figura 4) è risultata una grande preponderanza del genere 'Pop' con una frequenza percentuale pari al $37,5\%$ sul totale. Considerato questo dato si è ritenuto più interessante concentrarsi sulle caratteristiche specifiche delle canzoni.

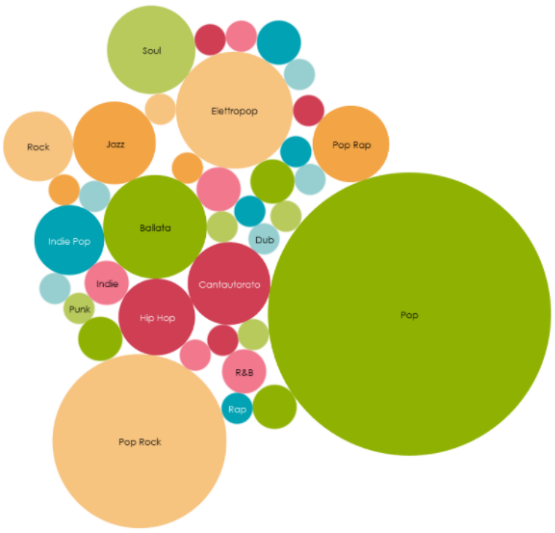


Figure 4: Generi delle canzoni dal 2010 al 2021

Analisi descrittive degli indici

In primo luogo si è deciso di analizzare la presenza di valori nulli o doppi nella tabella. In seguito ci si è orientati sull'eventuale correlazione tra le variabili osservate (Figura 5). La correlazione risulta medio-alta tra le variabili energy e loudness (pari a $0,69$). Dal momento che il progetto non prevede l'utilizzo di modelli di regressione si procede con le analisi descrittive. Si segnala comunque che in caso di futuri sviluppi in questa direzione si consiglia la rimozione di una delle due. Al fine di trovare eventuali caratteristiche rilevanti per raggiungere la vittoria, si è deciso di rappresentare il valore medio delle variabili per anno, confrontato col valore della stessa variabile della canzone vincitrice. Da questa analisi non si è potuto individuare con certezza le variabili più significative. Anche in questo caso, una possibile soluzione potrebbe essere lo studio di un modello di regressione multiplo, confrontando le variabili più significative negli ultimi 10 anni, al fine di prevedere

Figure 5: Correlogramma tra gli indici di Spotify

il vincitore. Di seguito alcuni esempi dei grafici ottenuti(Figura 6 e Figura 7).

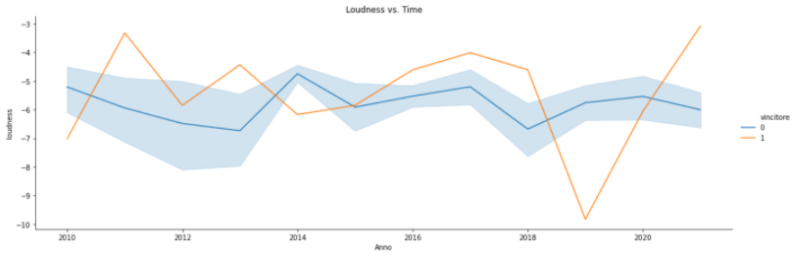


Figure 6: Confronto canzone vincitrice con le altre per la variabile loudness

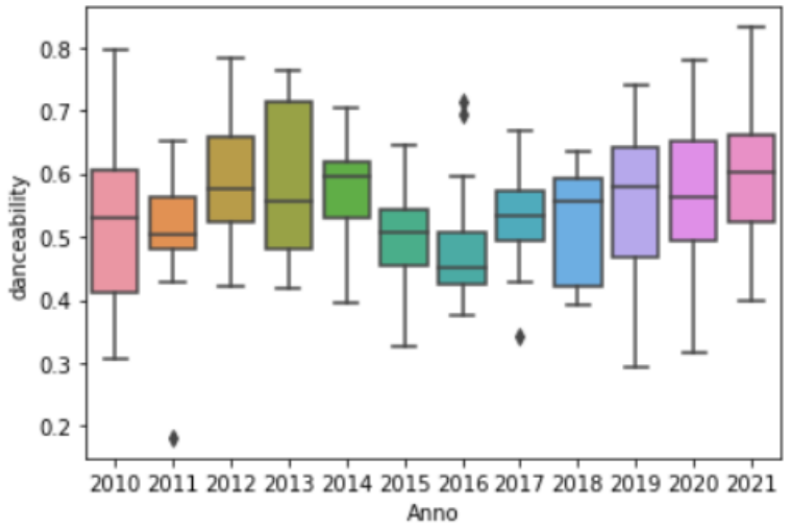


Figure 7: Variabile danceability dal 2010 al 2021

2 DATA VISUALIZATION

Si è deciso di utilizzare quattro infografiche sviluppate sulla piattaforma Tableau Public. Le stesse sono consultabili al seguente link:
https://public.tableau.com/app/profile/claudio.maffi/viz/Sanremo2021_16308662361360/Story?publish=yes

2.1 INFOGRAFICHE

2.1.1 Vincitori e vinti, ma per chi?

Nella prima infografica è possibile confrontare le classifiche di Sanremo, Spotify e Twitter, dal giorno di inizio del Festival fino alla classifica finale, filtrando per cantante. La rappresentazione delle classifiche è riconducibile a quella di una serie temporale. É stato di utilizzare i colori ufficiali del Festival, dell'applicazione musicale e del social network, al fine di renderli facilmente riconoscibili. L'infografica consente, attraverso filtri ad hoc, la visualizzazione simultanea di più cantanti e classifiche. É possibile notare quanto la classifica ufficiale di Sanremo di scosti da quelle relative agli utenti di Twitter e Spotify.

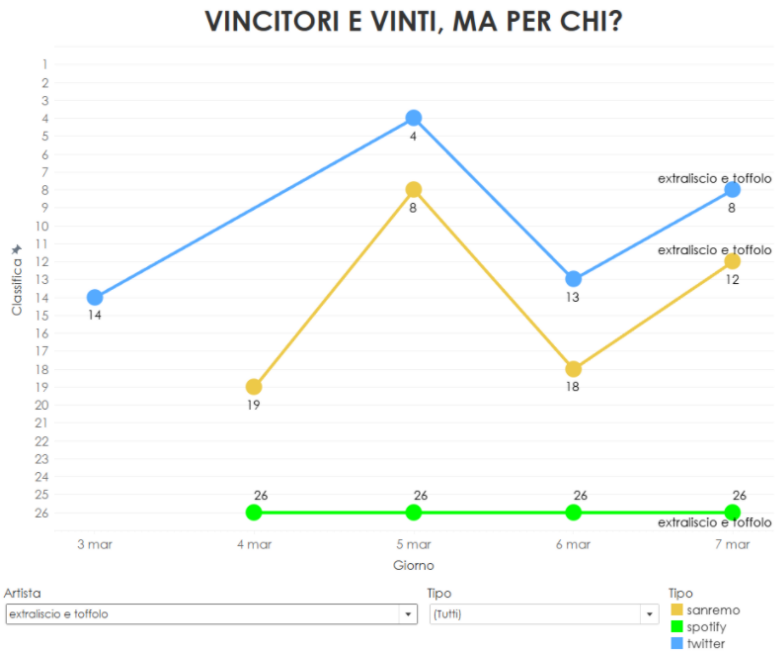


Figure 8: Prima infografica: Vincitori e vinti, ma per chi?

2.1.2 Il mondo è bello perchè... è vario!

In questa infografica si è voluto rappresentare la variabilità relativa alle posizioni assunte da ogni cantante (attraverso un boxplot verticale), e complessiva, nelle differenti classifiche. L'infografica offre la possibilità di selezionare solo un concorrente, al fine di sottolineare la differenza tra le tre classifiche. Questo grafico è stato integrato rappresentando la deviazione standard complessiva per classifica. Da quest'ultima rappresentazione si evince che le classifiche hanno una variabilità molto differente tra loro, dimostrandosi particolarmente eterogenee e difficili al confronto.

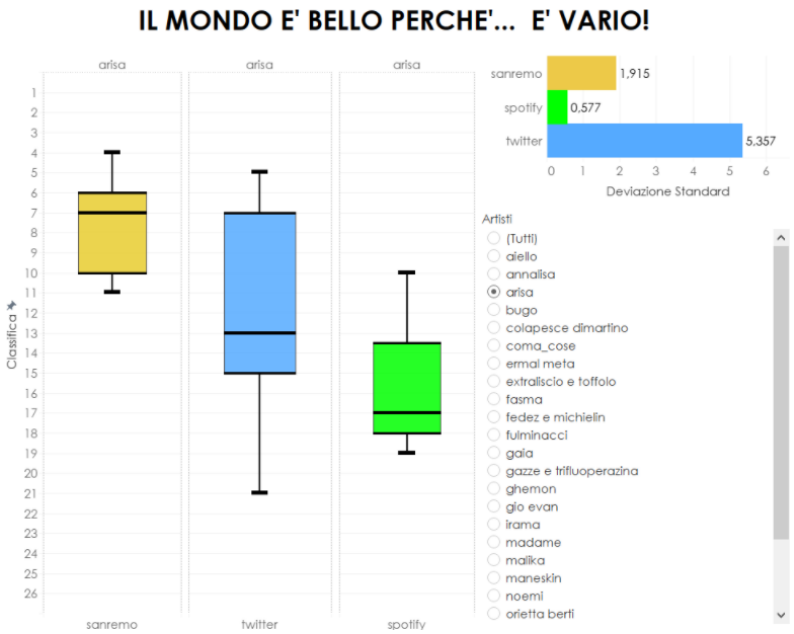


Figure 9: Seconda infografica: Il mondo è bello perchè... è vario!

2.1.3 Se sei felice e tu lo sai...vincerai?

La terza infografica si concentra sull'indice di positività delle canzoni partecipanti al Festival dal 2010 ad oggi. L'indice di positività è dato dalla somma delle variabili "valence", "energy" e "danceability". Grazie all'indicazione della positività media (indicata per anno, tra le canzoni partecipanti), è possibile osservare, in modo immediato, dove si posiziona la canzone vincitrice in base a questo parametro. Attraverso il grafico è possibile, quindi, notare come, dal 2010 al 2016, le canzoni vincitrici presentano un indice di positività inferiore alla media. Dal 2017 in poi si osserva un andamento più altalenante.

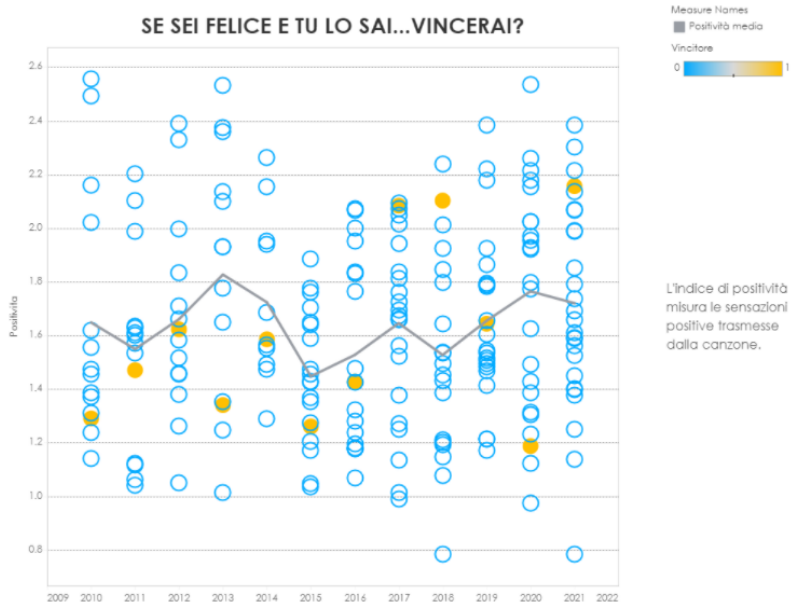


Figure 10: Terza infografica: Se sei felice e tu lo sai...vincerai?

2.1.4 L'importante è che se ne parli!

Nella quarta infografica si è rappresentato sia il numero di tweet che la sentiment media. Per quest'ultima, la positività è rappresentata dal colore blu, mentre la negatività dal colore arancione. I cantanti sono ordinati sia in base alla sentiment media (dal valore più positivo a quello più negativo), che in base al numero di tweet. Per quest'ultimo grafico, è stata utilizzata una scala di colore che varia dal rosso scuro al rosso chiaro in base al numero di tweet.

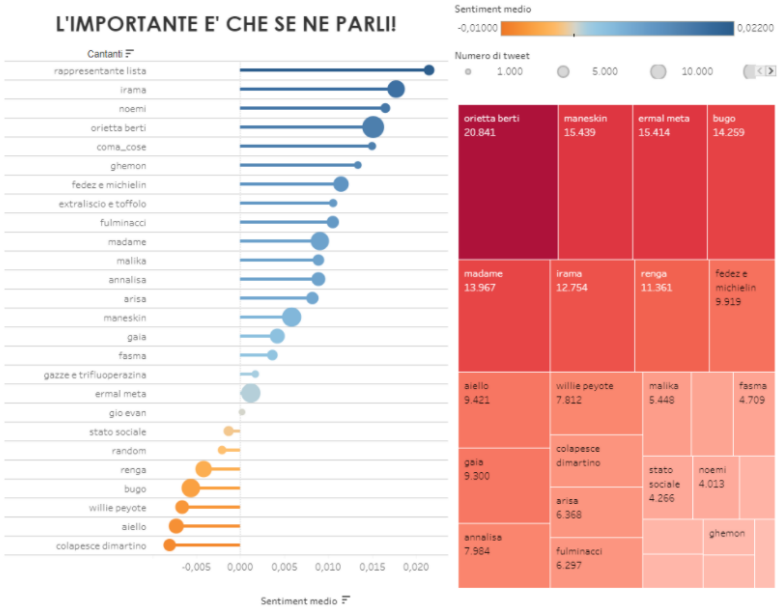


Figure 11: Quarta infografica: L'importante è che se ne parli!

2.2 VALUTAZIONE INFOGRAFICHE

2.2.1 USER TEST

Al fine di valutare la qualità delle infografiche sono state individuate 8 domande (due per ogni infografica), poi sottoposte a 35 persone.

PRIMA INFOGRAFICA

- *In che posizione finale si sono classificati i Maneskin secondo la classifica di Spotify?*
- *Per quale classifica Madame ha raggiunto la posizione più alta il 4 di marzo?*

SECONDA INFOGRAFICA

- *Per quale classifica la posizione di Fedez e Michielin varia di meno?*
- *Quale classifica ha la deviazione standard complessiva maggiore?*

TERZA INFOGRAFICA

- *Nell'anno 2011 la canzone vincitrice ha un indice di positività sopra o sotto alla media?*
- *Quante volte i vincitori hanno avuto un indice di positività sopra la media?*

QUARTA INFOGRAFICA

- *Quale artista ha ricevuto il maggior numero di tweet?*
- *Quale artista si è classificato terzo come indice di gradimento?*

Per ciascuna domanda è possibile osservare, grazie ai violin plot (Figura 12), i relativi tempi di risposta. Inoltre, gli stacked bar chart (Figura 13) mostrano la percentuale di errori compiuti dal campione intervistato.

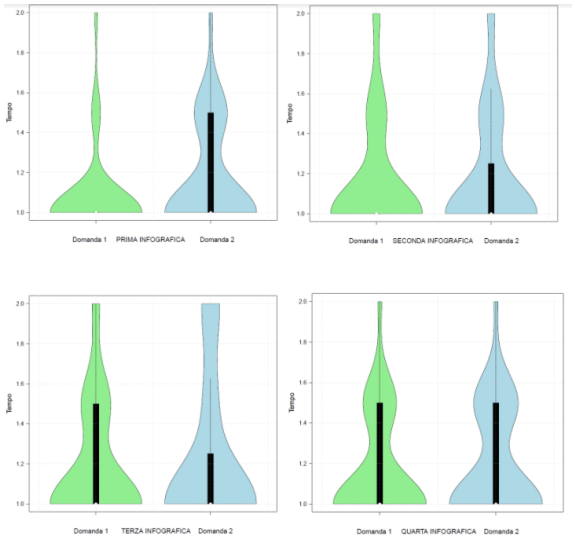


Figure 12: Violin plot per ogni domanda

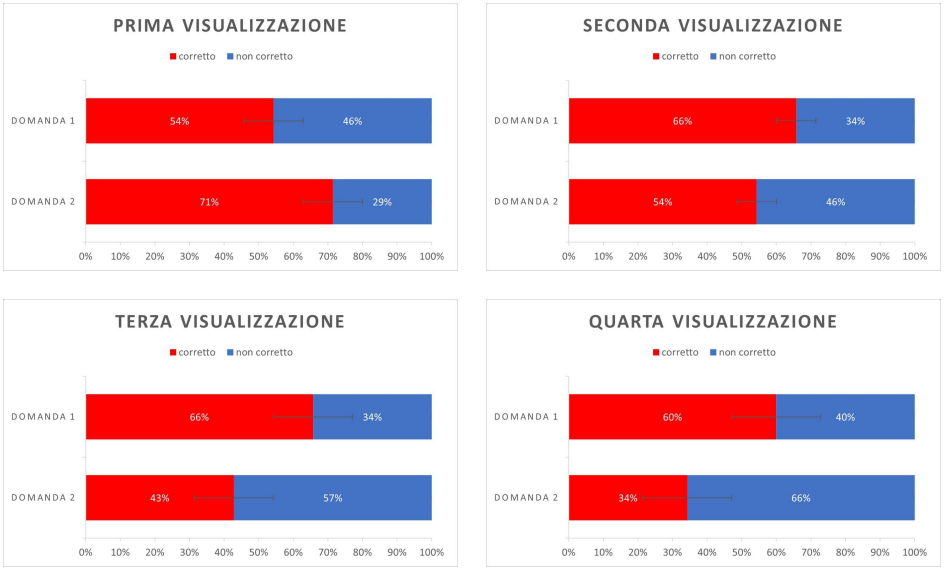


Figure 13: Stacked bar per ogni domanda

2.2.2 VALUTAZIONE EURISTICA

Sono state intervistate 35 persone per valutare le infografiche e interagire con esse così da individuare eventuali criticità. Complessivamente i risultati sono stati positivi, anche se la maggior parte degli intervistati ha espresso leggere difficoltà nell'interpretazione della seconda infografica. La più apprezzata, in termini di semplicità e chiarezza, risulta essere la prima dashboard, mentre l'interpretazione della terza infografica non è stata immediata. Molti hanno segnalato difficoltà nell'individuare la canzone vincitrice rispetto alle altre. La quarta dashboard è risultata invece la più efficace. Gli intervistati hanno apprezzato la possibilità di osservare, simultaneamente, sia il numero di tweet che la sentiment relativa ad ogni cantante.
A seguito dei feedback ricevuti, si è convenuto di apportare alcune modifiche alla terza infografica. In particolare si è optato per un contrasto di colore più marcato così da evidenziare maggiormente la canzone vincitrice rispetto alle altre.

2.2.3 QUESTIONARIO PSICOMETRICO

Alle 35 persone intervistate è stato, inoltre, sottoposto il questionario psicometrico. Di seguito i risultati ottenuti rappresentati tramite la scala Cabitza-Locoro:

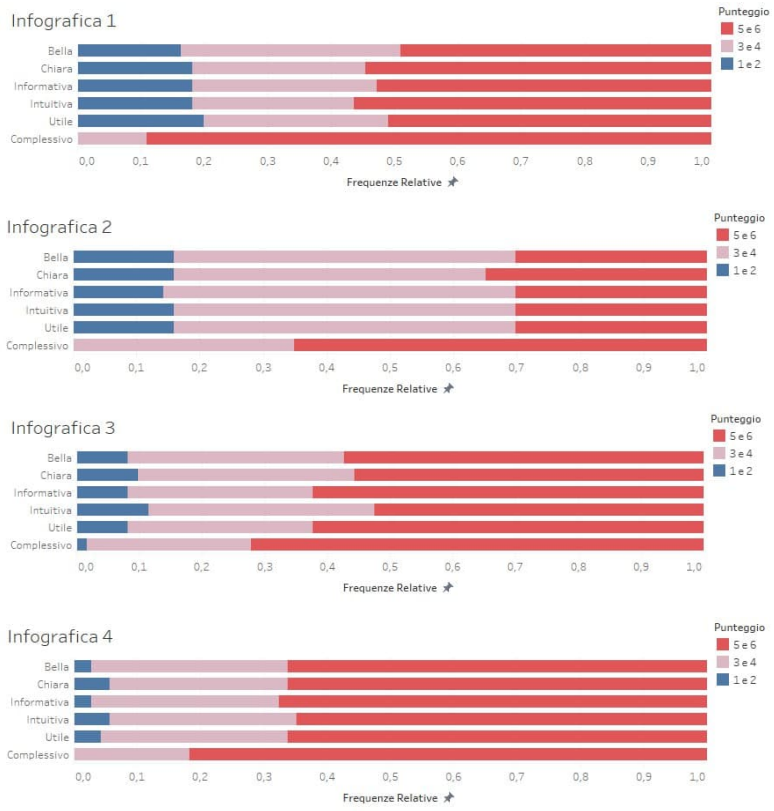
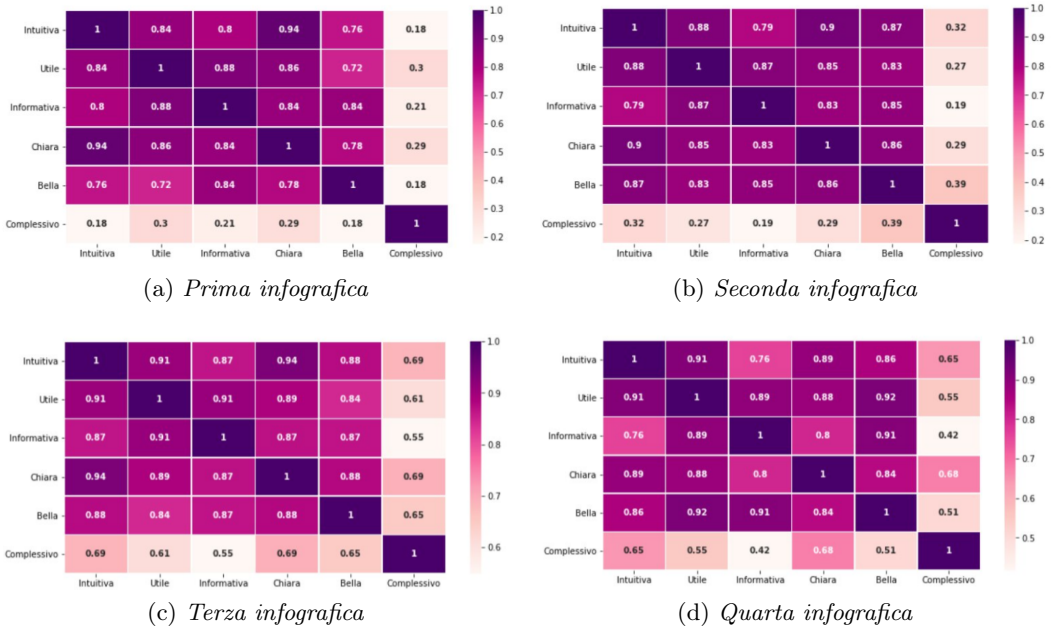


Figure 14: Scala Cabitza-Locoro per ogni infografica

Attraverso i grafici sopra, è possibile notare come le valutazioni complessive siano per la maggior parte positive. La prima e la quarta risultano le infografiche più apprezzate; per la seconda, ad esclusione del giudizio complessivo, si osservano per le restanti metriche valutazioni con punteggio "3 e 4".
Di seguito, i correlogrammi:



(a) *Prima infografica* (b) *Seconda infografica*

(c) *Terza infografica* (d) *Quarta infografica*

Figure 15: Correlogrammi

3 CONCLUSIONI

Dall'analisi effettuata sono emersi diversi risultati. In primo luogo è stata riscontrata una grande differenza in termini di variabilità tra classifiche. La classifica di Spotify risulta la meno variabile. Si è ipotizzato che questo risultato sia riconducibile al bacino di utenza, probabilmente non incline a seguire l'evento mediatico. Gli indici di Spotify non hanno permesso di iindividuare caratteristiche comuni relative alla canzone vincente, non si esclude che l'implementazione di un modello di regressione possa fornire risultati differenti. L'unico dato che si è ritenuto interessanti è il cambiamento nell'andamento dell'indice di positività delle canzoni vincitrici dal 2017 ad oggi.
Per quanto riguarda la classifica di Sanremo si nota che la variabilità risulta maggiore rispetto a quella di Spotify ma minore di Twitter. Questa differenza di valore può essere riconducibile al sistema misto di votazioni del Festival che assegna pesi differenti in base alla giuria di riferimento. I risultati più interessanti sono stati ottenuti dall'analisi dei tweet. Su 633.208 tweet soltanto 206.100 sono effettivamente relativi ai cantanti. Questo significa che le persone hanno ritenuto più interessante commentare l'evento in termini di spettacolo piuttosto che valutare le canzoni in gara. La sentiment ha rilevato infatti tweet, più o meno favorevoli, sui cantanti, indipendentemente dalla performance musicale, confermando che Sanremo più che essere considerato una gara canora viene visto da chi usa Twitter come uno spettacolo.
Per concludere non si ritengono Spotify e Twitter rappresentativi degli utenti votanti al Festival di Sanremo.