Does the weather actually affect your mood?

Gabriele Carrara[1] - 814720 , Alberto Filosa[2] - 815589

[1] *Università degli Studi di Milano Bicocca - Dipartimento di Informatica, Sistemistica e Comunicazione.*
[2] *Università degli Studi di Milano Bicocca - Dipartimento di Informatica, Sistemistica e Comunicazione.*

Sommario

Il fine dello studio è la comprensione della relazione presente tra il clima e l'umore delle persone. Per scoprirlo si è proceduto all'analisi della polarità di un anno di tweets (2019), mettendo questi in rapporto con le temperature locali del giorno. Sono state considerate le città di Milano, Roma, Napoli e Palermo e l'area ad essa circostante in un raggio di 100 km. Sono stati estratti i tweet contenenti almeno una delle seguenti parole: freddo, sole, nuvole, pioggia, neve, temperatura e caldo.

Ci si aspettava di ottenere una *strana correlazione*: tweets positivi che aumentavano all'aumentare della temperatura fino ad un limite, intorno ai 25 °C, da cui sarebbero tornati in numero inferiore.

I software utilizzati per studio sono stati principalmente due: Python Notebook e R Markdown. Per la visualizzazione grafica il software utilizzato è stato Tableau.

Tutto il lavoro è stato ideato ed eseguito insieme.

Parole Chiave *Python, R, Tableau, Twitter, Sentiment Analysis, Meteo.*

Indice

1 Raccolta Dati

1.1 Dati Meteo

La raccolta dei dati meteo è stata effettuata tramite procedure di scraping dal sito `Il Meteo`. Sono state estratte le temperature minime, medie e massime giornaliere delle città di Milano, Roma, Napoli e Palermo per tutto il 2019.

Sul sito però non erano presenti i dati da Agosto a Ottobre per la città di Milano. Per risolvere questo problema, sono stati recupati dal sito `Rp5` le temperature mancanti ed aggiunte al dataset di partenza. Anche in questo caso, vi erano dei valori mancanti in alcuni giorni di ottobre. Per colmare questa ultima lacuna, sono stati raccolti i dati dal sito `Tu Tiempo`.

1.2 Tweets

Per raccolta dei tweets nell'anno 2019 non è stato possibile utilizzzare le API di Twitter in quanto presentavano delle restrizioni in termini di tempo e quantità. Si è perciò optato per l'utilizzo del pacchetto Python `GetOldTweets3` direttamente dalla linea di comando che permette di creare una tabella contenente i tweets per il periodo ed il luogo desiderato.

2 Integrazione Dati

I due dataset ottenuti sono stati integrati in un unico dataset tramite il linguaggio di programmazione Python.

3 Analisi Esplorativa e controllo sulla Qualità dei Dati

In primo luogo, si è verificata la presenza di valori mancanti ed è risultato che il giorno 5 febbraio 2019 non presentava alcuna rilevazione per tutte le città considerate in alcuna fonte dati. Si è quindi optato per imputare i valori rendendoli omogenei a quelli del giorno precedente. Questo è stato l'unico problema di missing values sull'intero dataset.

Successivamente sono stati rappresentati i box plot di temperatura minima, media e massima per verificare la presenta di valori anomali. Non si nota alcuna anomalia.

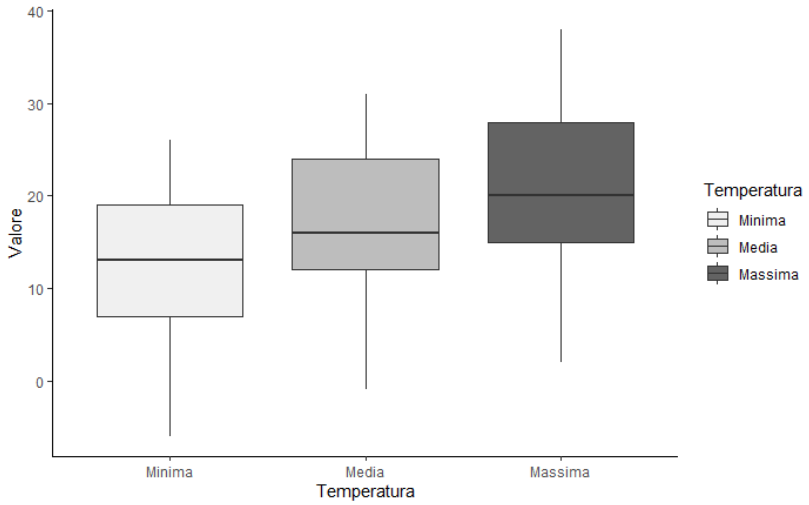


Figura 1: Boxplot Temperature Giornaliere nel 2019

Un ulteriore check sulla coerenza delle temperature è stato fatto verificando che per nessuna riga la temperatura minima fosse superiore alla media e quest'ultima alla temperatura massima. Anche qui non si è riscontrata alcuna anomalia. È stato costruito un grafico `line plot` per osservare meglio questo aspetto.

Si è poi proceduto all'estrazione di un campione di 100 tweet per verificare quanti fossero *off topic* rispetto all'argomento in questione pur contenendo le parole chiave individuate. Il 32% dei tweet è risultato fuori contesto.

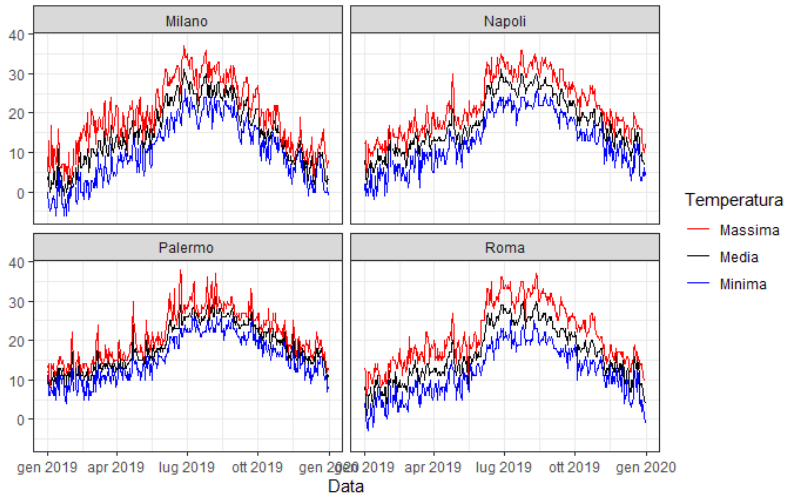


Figura 2: Temperature Giornaliere per ogni Provincia

4 Sentiment Analysis

Per l'analisi della polarità dei tweet sono stati utilizzate le API del sito `MeaningCloud`. I risultati poi sono stati raggruppati in *Positivi* se la risposta era uguale a `P` o `P+`, *Neutrali* se `NEU` o `NONE` e *Negativi* se `N` o `N+`.

È stato effettuato un controllo su un campione di 100 tweet, 50 positivi e 50 negativi, ed è risultato che il 27% è stato classificato non correttamente.

5 Velocity

Tramite il software Kafka sono stati creati un creati un Producer e un Consumer per simulare l'acquisizione dei dati in tempo reale. Su un topic chiamato `tweet_meteo` sono stati caricati i tweet tramite un ciclo in cui 100 tweet per volta vengono trasformati in formato `JSON`, caricati sul broker e dopo aver eseguito questa operazione si aspetta un periodo di tempo pseudo-casuale tra 10 e 20 secondi. In contemporanea, il Consumer *consuma* i dati che da JSON vengono inseriti in un dataframe Pandas.

6 Storage su MongoDB

Ad ogni passaggio, i dati sono stati caricati in formato JSON sul software MongoDB tramite linea di comando in un database chiamato `DataMan`. Per la lettura dei dataset da Mongo, è stato invece utilizzata il pacchetto python chiamato `PyMongo`.

7 Data Visualization

Dapprima, il dataset è stato manipolato per ottenere la variabile rapporto tweet positivi/negativi all'interno della giornata. Una volta calcolato il rapporto in python, sono state create le visualizzazioni grafiche utilizzando il software Tableau.

7.1 Infografica 1: Andamento Temperatura e Numero di Tweet

La prima visualizzazione grafica mostra in parallelo l'andamento della temperatura media nel corso del 2019 ed il rispettivo numero di tweet in un line plot a doppio asse. Questa scelta è stata effettuata nell'ottica di rilevare come valori particolari delle temperatura influenzano il numero di tweet nel corso della giornata. Inoltre, è possibile filtrare in base alla provincia ed alla polarità del tweet.

Il grafico conferma come giornate particolari dal punto di vista metereologico influenzino parecchio l'attività degli utenti sul social network. In particolare, i giorni con un numero di tweet molto alto sono il 27 giugno ed il 13 dicembre 2019.

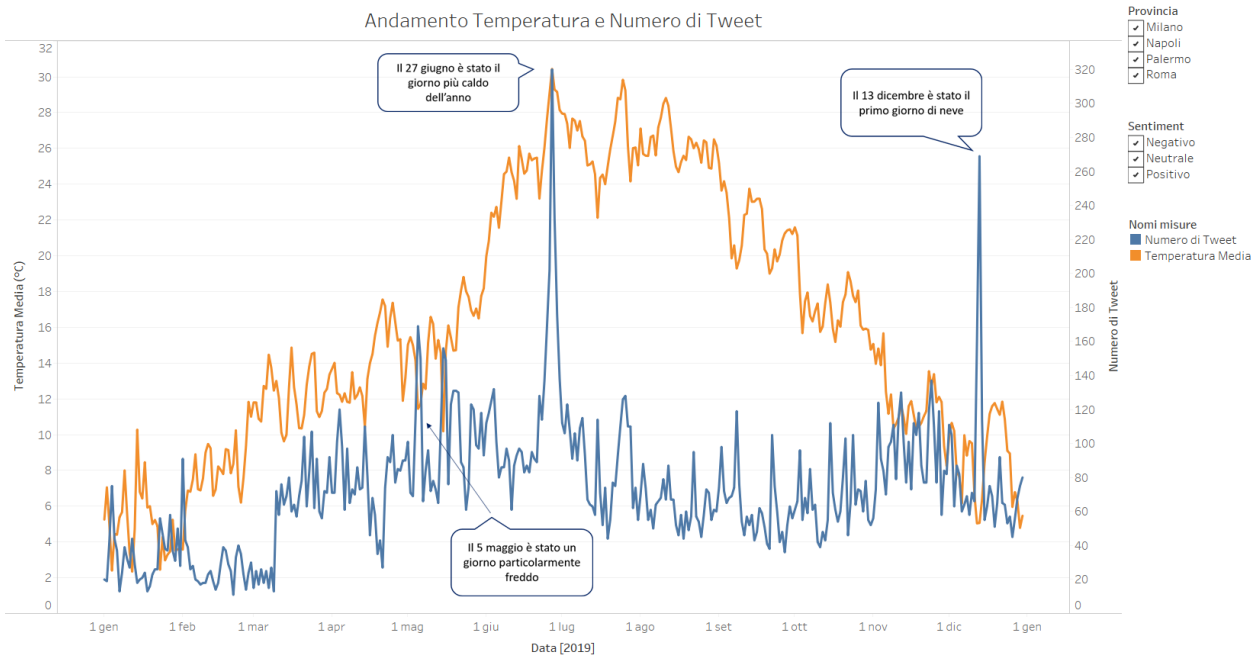


Figura 3: Andamento della Temperatura e Numero di Tweet

7.2 Infografica 2: Rapporto Tweet Positivi/Negativi in Funzione della Temperatura

La seconda visualizzazione grafica mostra il raporto tra il numero di tweet positivi e negativi al variare della temperatura. In parallelo, per mostrare la variabilità del rapporto si è deciso di presentare anche i boxplot stratificati per *contenitori* di temperatura (ogni 3 °C).

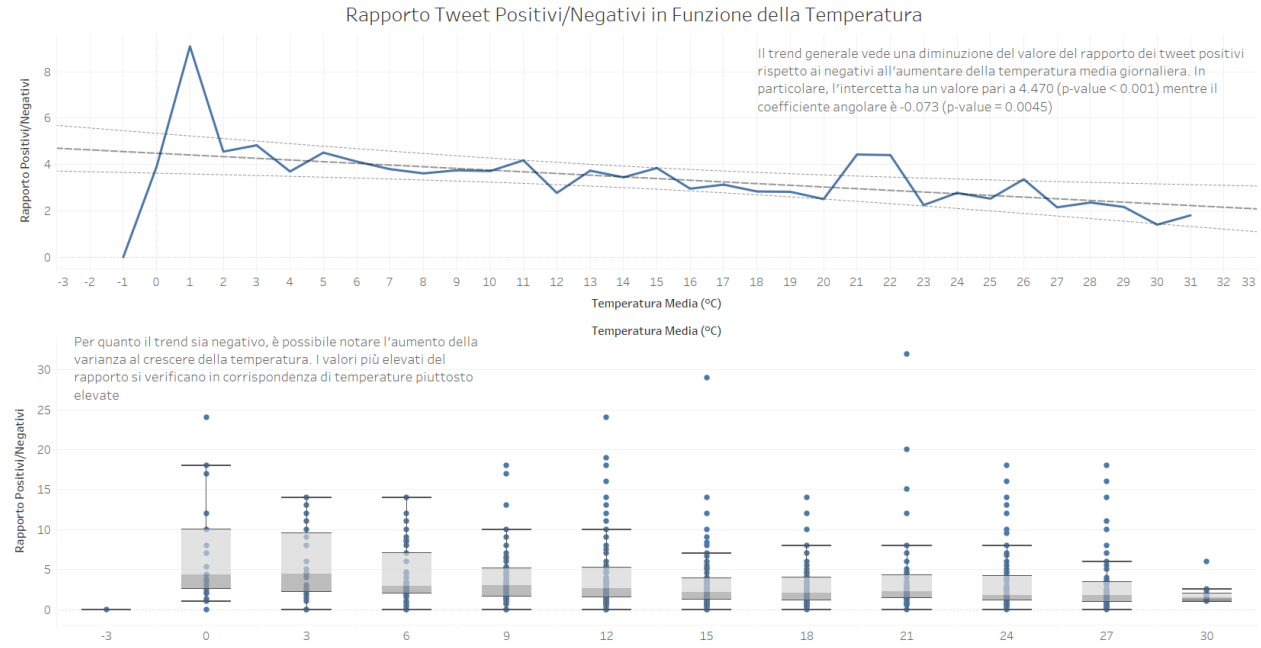


Figura 4: Rapporto Tweet Positivi/Negatini in Funzione della Temperatura

All'aumentare delle temperatura il rapporto del numero di tweet positivi e negativi decresce, per quanto le temperature più apprezzate siano state nelle giornate con temperatura media intorno ai 20 °C.

7.3 Infografica 3: WordCloud

L'ultima visualizzazione è composta da tre grafici: un Pie Chart per la Sentiment Analysis, l'andamento della temperatura tramite un Line Plot ed un grafico WordCloud per mostrare la frequenza delle parole tra quelle da noi cercate. È possibile filtrare il giorno desiderato dal grafico della temperatura, la polarità dei tweet e la provincia d'interesse.

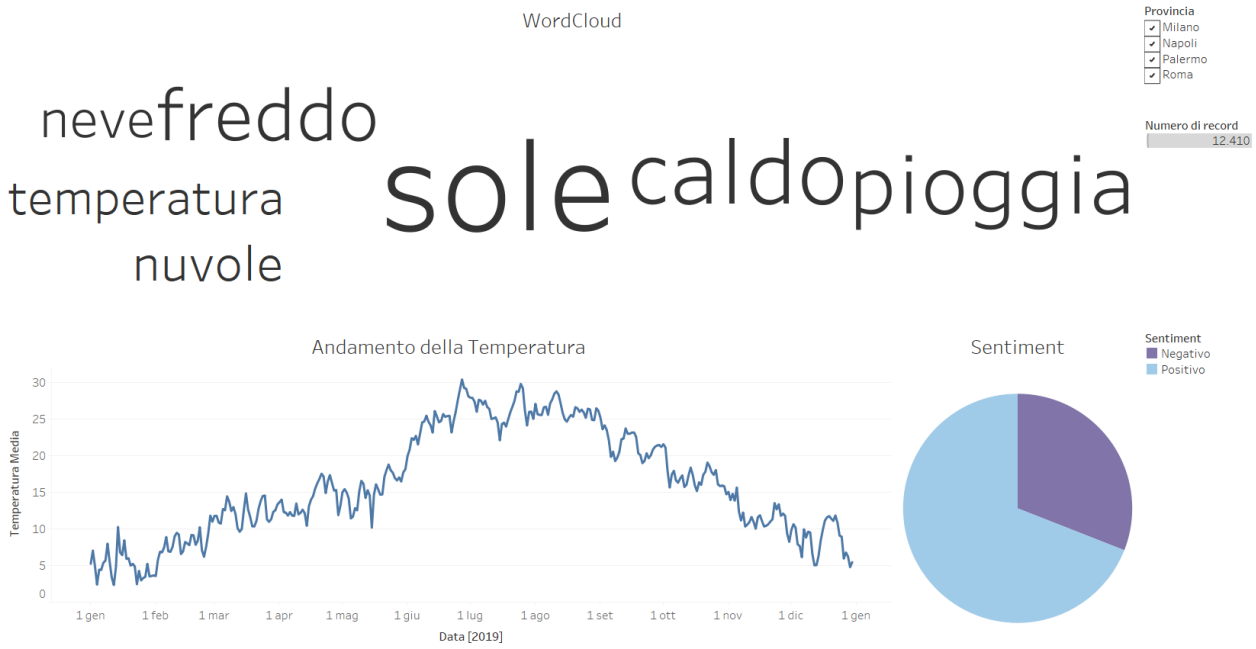


Figura 5: WordCloud con Polarità dei Tweet

In assoluto la parola *Sole* è quella più utilizzata ed appare con maggiore frequenza nei tweet con polarità positiva. Al contrario, le parole *Pioggia* e *Caldo* sono molto più frequenti nei tweet negativi.

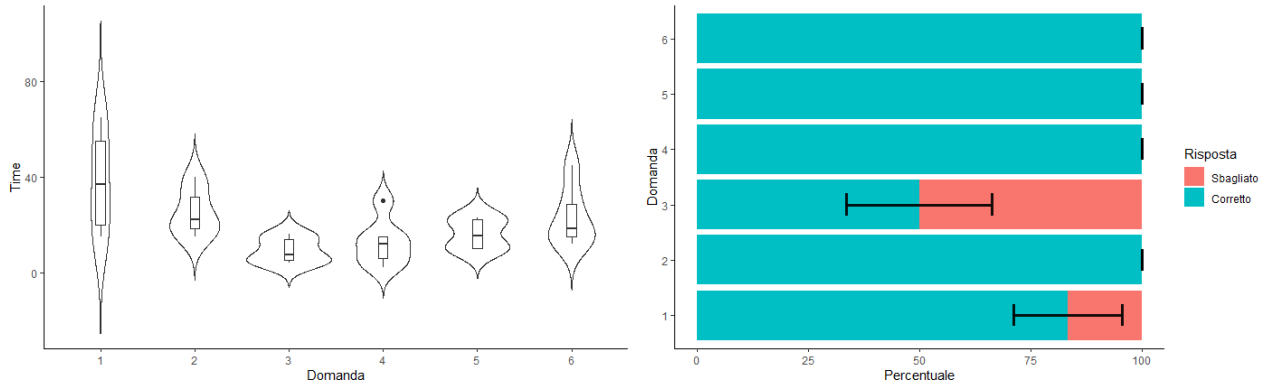
7.4 Assessemt

7.4.1 User Test

Per la parte di Assessement sono state individuate 6 domande (due per ogni infografica) da sottoporre a 6 persone. Le domande in questione sono:

- (1.1) Qual è stata la giornata con più tweet negativi nel corso del 2019? La giornata con più tweet è stato il 27 giugno 2019 con 54 tweet negativi;
- (1.2) Che temperatura media si è verificata a Milano il giorno in cui il numero di tweet totale è stato maggiore? La temperatura media registrata a Milano è stata di 31 °C il 27 giugno 2019;
- (2.1) Quanti gradi c'erano il giorno in cui il rapporto tweet positivi/negativi è stato maggiore? La temperatura media del giorno con il più alto rapporto è stata quella attorno ai 21° C;
- (2.2) È giusto affermare che all'aumentare della temperatura il rapporto tweet positivi positivi/negativi diminuisce? Si, il trend è negativo ed è statisticamente significativo;
- (3.1) Qual è stata la parola più gettonata il giorno 15 maggio 2019? La parola più utilizzata quel giorno è stata Sole;
- (3.2) Quali sono state le due parole più utilizzate all'interno dei tweet, classificati come negativi, nel corso del 2019 nella città di Palermo? Le parole più utilizzate all'interno dei tweet sono state Caldo e Sole.

Di seguito si riportano i violin plot relativi ai tempi di risposta per ciascuna domanda e gli stacked bar chart relativi agli errori compiuti da un campione di 6 intervistati.



(a) Violin Plot del Tempo di Risposta

(b) Percentuale di Risposta alle Domande

Si presentano le principali difficoltà nel rispondere alle domande:

- Ad un primo impatto, difficoltà nella compresione e nell'utilizzo dei filtri, come è possibile vedere dai tempi di risposta alla prima domanda;
- Per rispondere alla prima domanda, alcuni utenti individuavano il valore massimo dei tweet negativi nella parte inferiore del grafico;
- Difficoltà nel comprendere la differenza tra l'andamento generale (il primo grafico della seconda infografica) e quello dei singoli giorni, soprattuto i boxplot sono stati di difficile comprensione.

7.4.2 Think Aloud

Sono state intervistate 3 persone per osservare il loro approccio nel visualizzare le infografiche costruite. Tutti gli intervistati, non esperti di settore, hanno avuto bisogno di tempo per relazionarsi con il software Tableau. In particolare, per quanto riguarda la prima infografica, in un primo momento, non hanno notato sia i filtri che la doppia scala. Nella seconda infografica, invece, i boxplot non sono stati immediatamente compresi, anche perchè per la maggior parte degli intervistati è stata la prima volta ad interfacciarsi con questo tipo di rappresentazione. L'ultima dashboard, invece, è stata definita semplice ed efficace; ciò è dovuto anche al fatto che gli intervistati hanno avuto modo di prendere confidenza con il funzionamento dei filtri.

Dal momento che, in generale, il messaggio rappresentato è stato compreso da tutti, non sono state apportate modifiche ai grafici.

7.4.3 Questionario psicometrico

Sono state intervistate 24 persone per il questionario psicometrico utilizzando la scala `Cabitza - Locoro`. Si presentano i graficamente risultati:

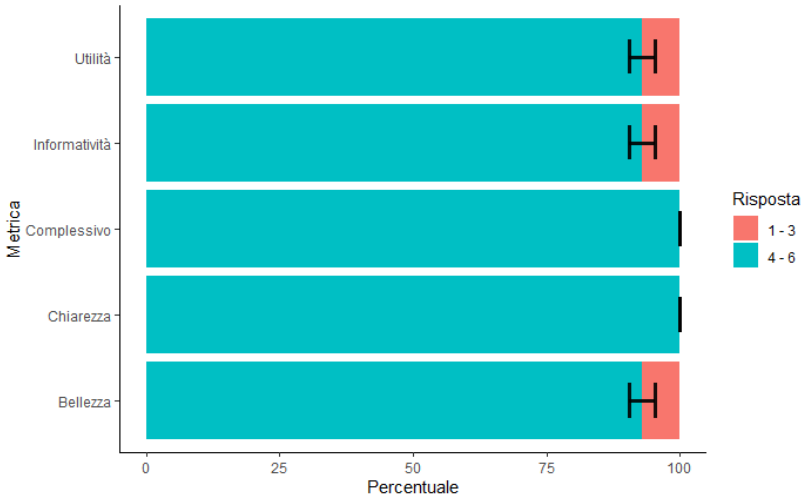


Figura 7: Percentuale di Risposta al Questionario Psicometrico

Le valutazioni sono quasi tutte positive, anche se la maggior parte delle risposte si assestano tra un punteggio di 4 e 5. Le uniche insufficienze vengono comunque valutate con un punteggio di 3. Le valutazioni degli intervistati con conoscenze statistiche sono state quasi sempre superiori per la facilità di compresione dei grafici proposti, soprattutto per quanto riguarda i boxplot. Di seguito, si riporta il corrplot delle metriche utilizzate:

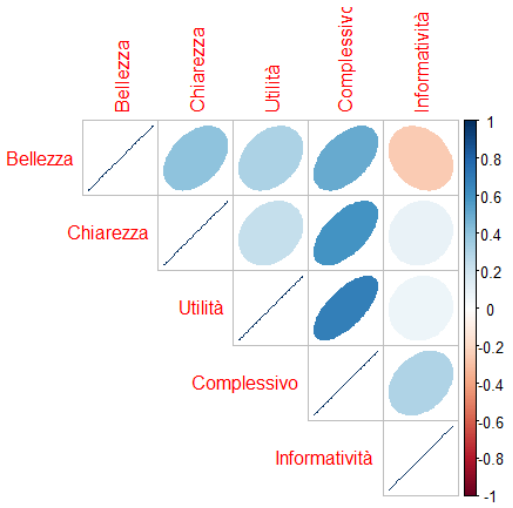


Figura 8: Matrice di Correlazione delle Metrice del Questionario

Si nota che la valutazione della metrica Complessivo ha una maggiore correlazione con la variabile Utilità rispetto alla Bellezza. Inoltre, si osserva che la correlazione tra Bellezza ed Informatività è negativa.

8 Conclusioni

La strana correlazione che ci si aspettava non si è in realtà verificata, bensì la relazione tra il rapporto tweet positivi/negativi e la temperatura è di tipo inversamente proporzionale. Si è poi dimostrato che le condizioni metereologiche influenzano parecchio l'attività degli utenti sui social network (in particolare, le condizioni più estreme), nonchè il loro umore. Si osserva che le condizioni metereologiche più apprezzate sono il tempo soleggiato e la neve, mentre vengono più spesso criticati il caldo e la pioggia.