UNIVERSITÀ
DEGLI STUDI
DI MILANO

Twitter Sentiment Analysis

Sentiment e Time Series Analysis sul conflitto Tra Russia e Ucraina

Visualizzazione Scientifica

Marco Molinati, 923530

Scopo dell’analisi

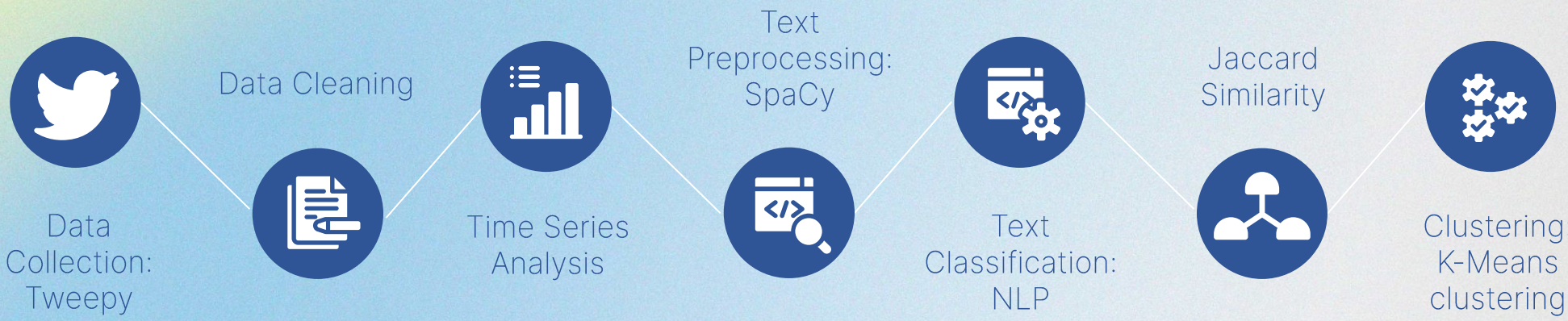
Lo scopo di questa analisi è quello di capire, attraverso i tweet del social network, i sentimenti contenuti nel testo e fare una classificazione tra positivi, negativi e neutrali, oltre che analizzare, tramite la componente temporale, l’andamento giornaliero/settimanale della pubblicazione dei tweet.

Un altro punto dell’analisi è quello del Clustering, ovvero classificare i tweet all’interno di contenitori tematici e mettere poi questi ultimi in relazione tra loro per verificare un possibile legame tra di essi.

Come tema centrale del progetto è stato scelto il conflitto tra Russia e Ucraina, in quanto è un argomento attuale e ci sono molte fonti e dati a disposizione da verificare, valutare. Il progetto rimane ugualmente scalabile e adattabile a qualsiasi altra tematica sfruttando lo stesso flusso, in quanto è solo necessario cambiare il criterio per la ricerca e raccolta dati.

Scala colore utilizzata

Flusso seguito per l’analisi



Data Collection e timeline eventi

Sono stati raccolti i dati relativi al conflitto Russo-Ucraino tramite i tweet degli utenti contenenti come parola chiave «Ukraine» tramite la libreria Python «Tweepy», la quale sfrutta le API di Twitter per automatizzare la raccolta/creazione di dati.

Il periodo considerato per la raccolta dati va dal 31/12/2021 al 05/03/2022, in quanto si ha modo di fare un'analisi pre-conflitto e subito dopo l'invasione

Data Cleaning e Text Preprocessing

Data Cleaning: sono stati ripuliti i dati contenuti dal csv dei tweet estratti, in particolare:

- Date in formato [datetime]
- Username utenti ripuliti da bio, handle e id
- Eliminati caratteri speciali e numeri dai tweets

Text Preprocessing: Tramite la libreria SpaCy sono stati rimossi i caratteri di punteggiatura e le «stopwords», successivamente sono stati applicati i processi di «Lemmatizzazione» e «Tokenizzazione»

> Lemmatizzazione: è il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma. Nell'elaborazione del linguaggio naturale, la lemmatizzazione è il processo algoritmico che determina automaticamente il lemma di una data parola.

> Tokenizzazione: è il processo di divisione di una frase o singola parola in «token», ovvero un elemento che racchiude parole dello stesso significato semantico

Tweet text before cleaning and preprocessing:

There was an anti war protest for #Ukraine as well but it’s not getting any headlines. \n\nSerbia has a good relationship with both Ukraine and Russia, both counties have supported Serbian territorial integrity over Kosovo. \n\nStop pitting Orthodox Christians against each other. https://t.co/CTulO2hTsA https://t.co/3EpcGaW6Oo

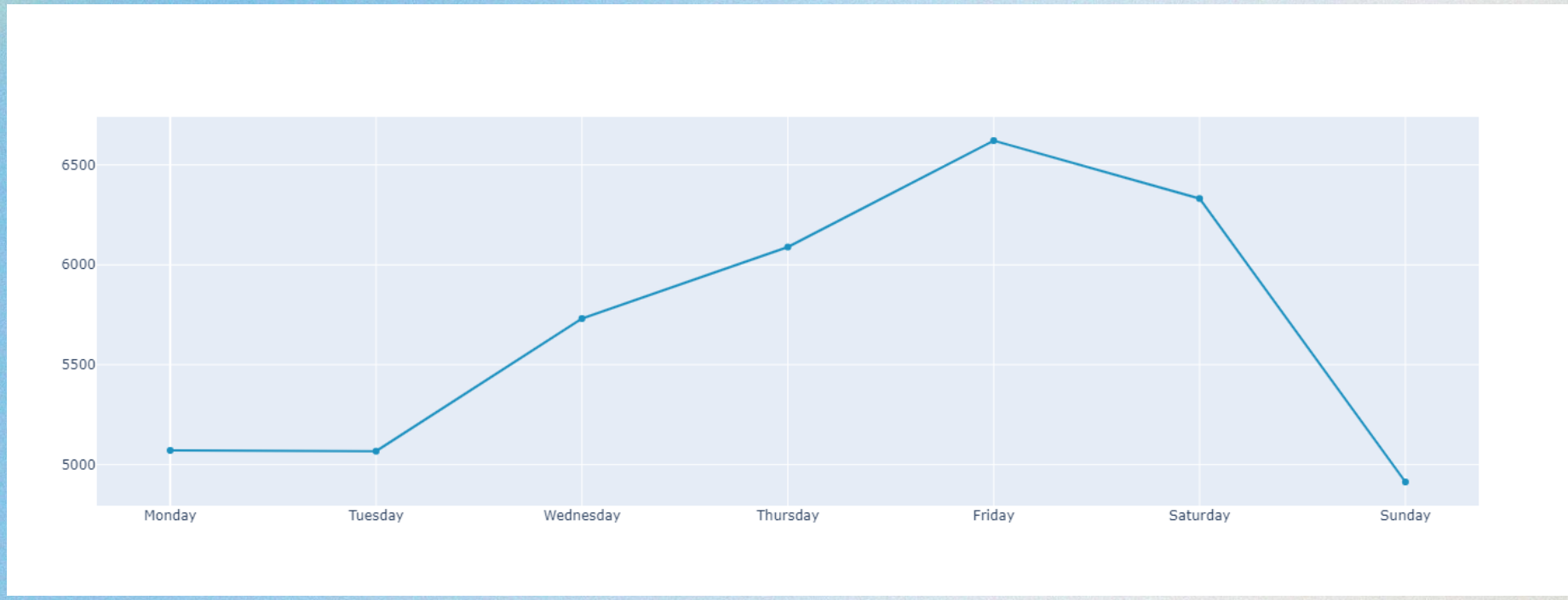
Tweet text after cleaning and preprocessing:

anti war protest ukraine well getting headline serbia good relationship ukraine russia county supported serbian territorial integrity kosovo stop pitting orthodox christian

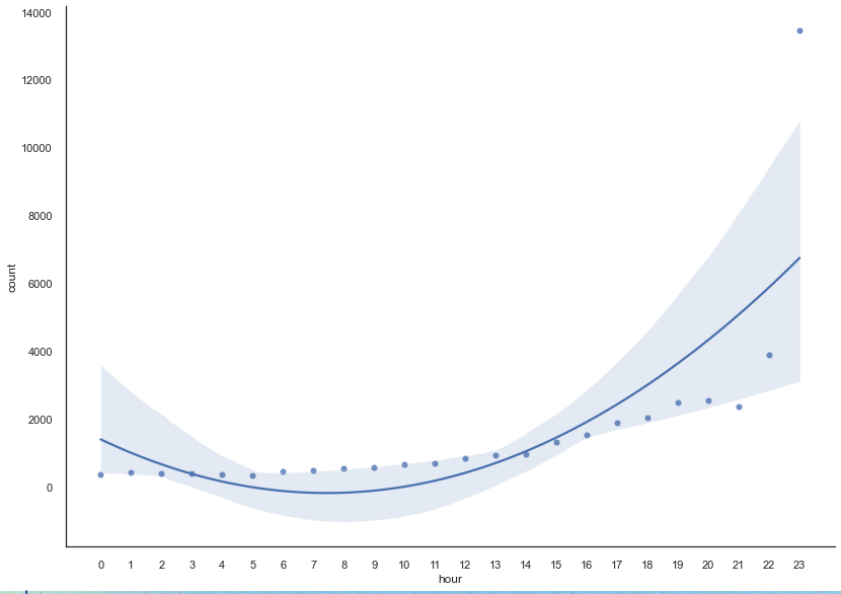
Time Series Analysis

É stata sfruttata la componente temporale presente nella colonna «date» del dataset.

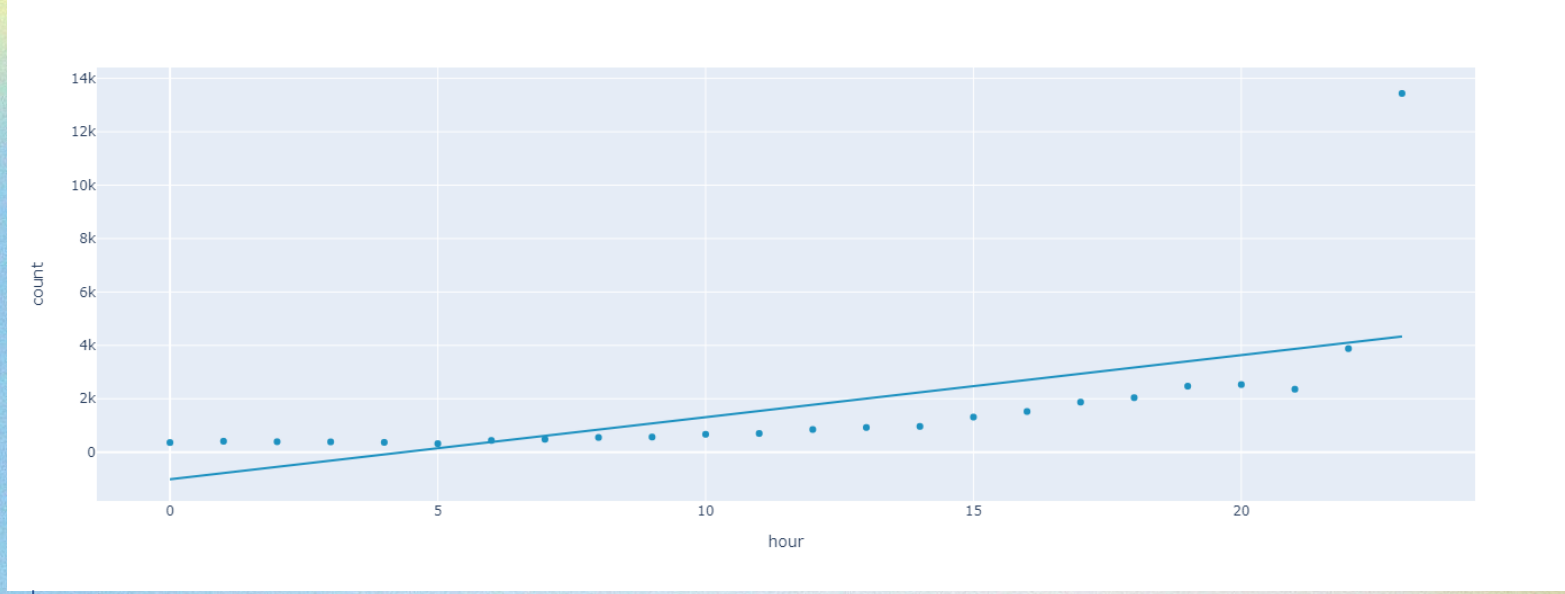
L'obiettivo è stato quello di verificare l'andamento dei tweet in funzione del tempo e capire se essi avessero una distribuzione uniforme. Accompagnano le analisi dei grafici relativi all'andamento settimanale dei tweet e dei giorni con più attività su Twitter.



Time Series Analysis



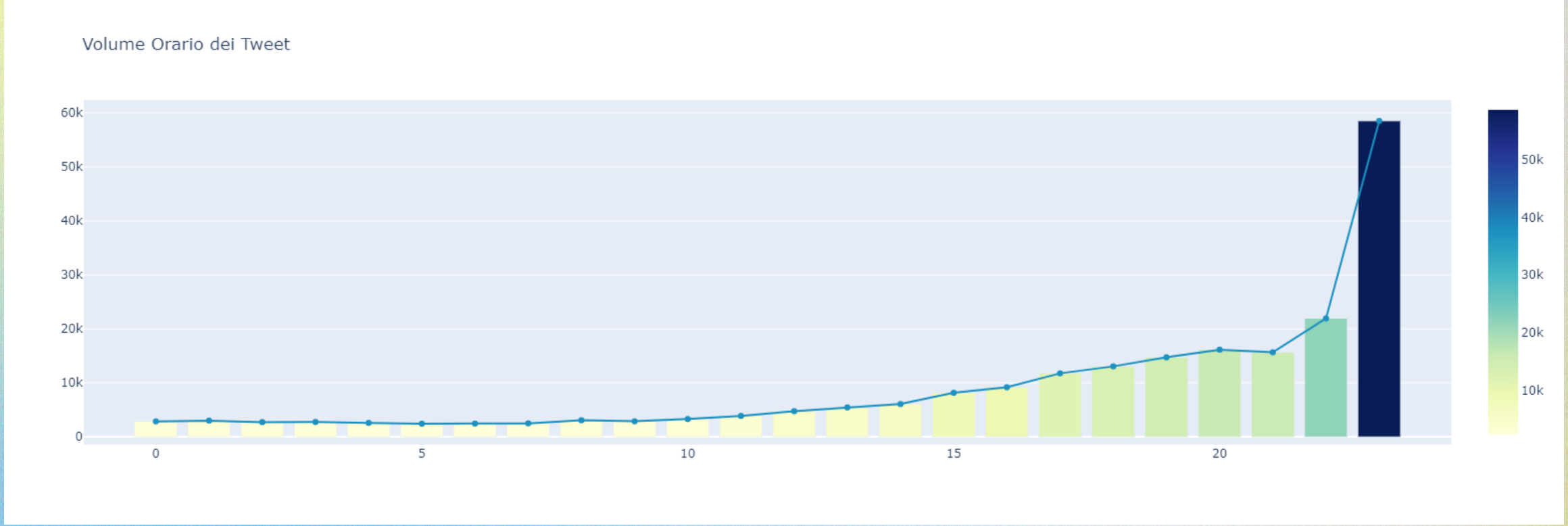
LmPlot ottenuto con Seaborn
(Regressione Polinomiale)



Ordinary Least Squares plot (Regressione Lineare)

> Ordinary Least Squares (OLS): è una tecnica di regressione che permette di trovare una funzione, rappresentata da una curva ottima, che si avvicini il più possibile ad un insieme di dati. La funzione trovata deve essere quella che minimizza la somma dei quadrati delle distanze tra i dati osservati e quelli della curva che rappresenta la funzione stessa.

Time Series Analysis



Text Classification

Sono state sfruttate due librerie per lo studio del linguaggio naturale, TextBlob e NLTK, le quali mettono già a disposizione delle pipeline addestrate per la classificazione dei sentimenti

TextBlob è utilizzato per calcolare «polarità» e «soggettività», mentre NLTK, in particolare il modulo «vader» serve per ottenere dei punteggi per sentimenti positivi/neutrali/negativi, oltre al «compound»

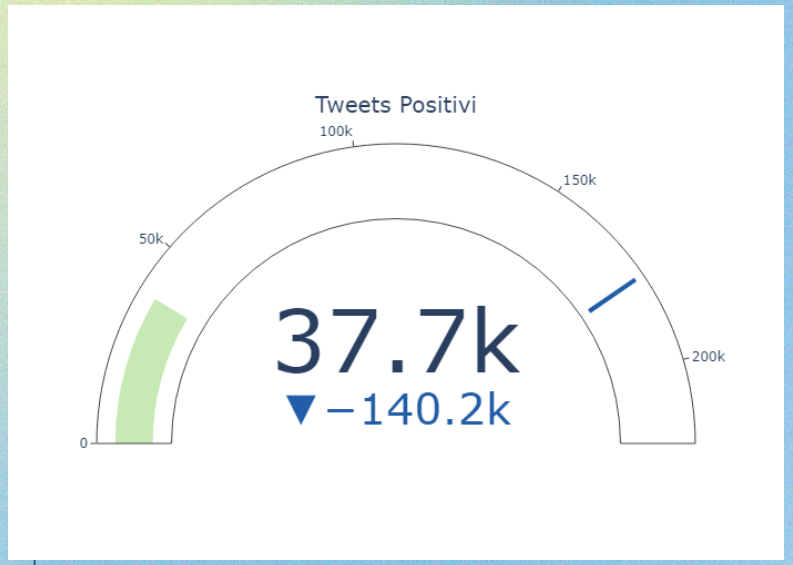
Polarità: valore numerico compreso nell'intervallo [-1, 1], dove i valori negativi indicano dei sentimenti negativi e quelli positivi l'opposto.

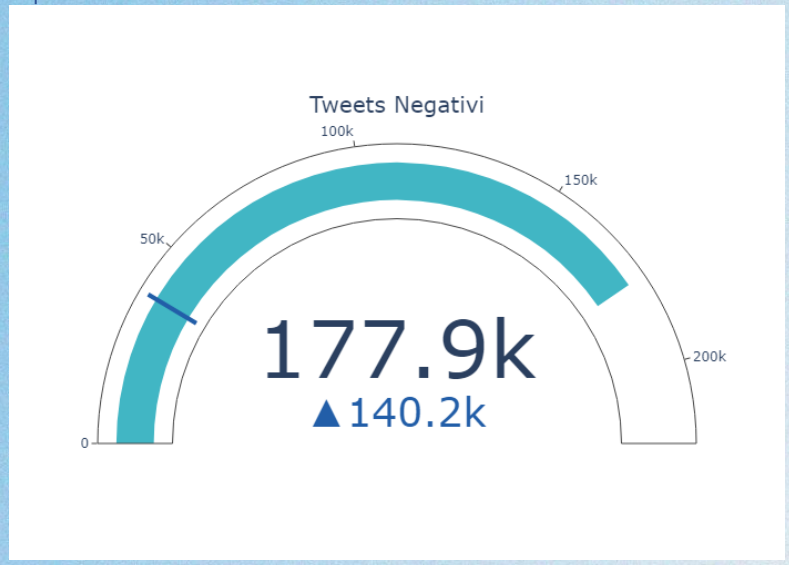
Soggettività: valore numerico che quantifica quanto di personale e non oggettivo ci sia nel testo del tweet

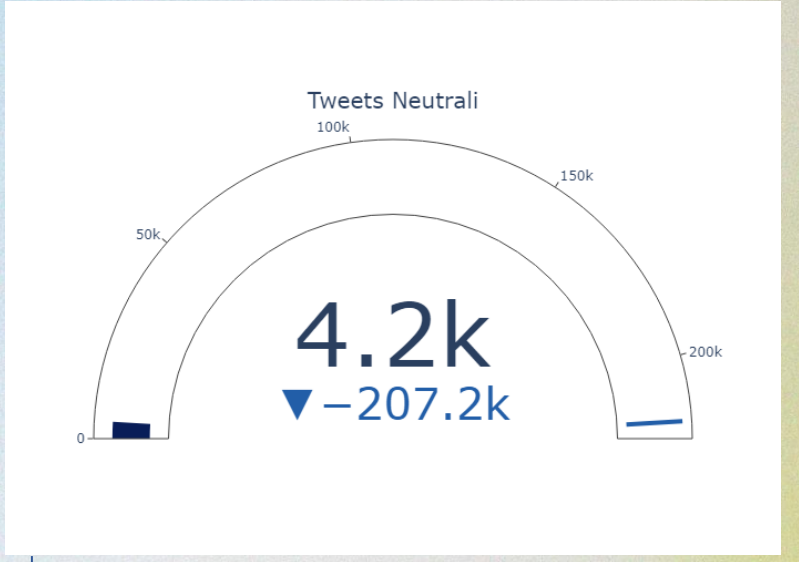
Compound: valore numerico che, se pari a zero indica che il sentimento è neutrale, mentre se ≥ 0.05 è positivo e se ≤ -0.05 è negativo.

Grafici a misuratore radiale

Tweets Negativi in relazione a quelli Positivi



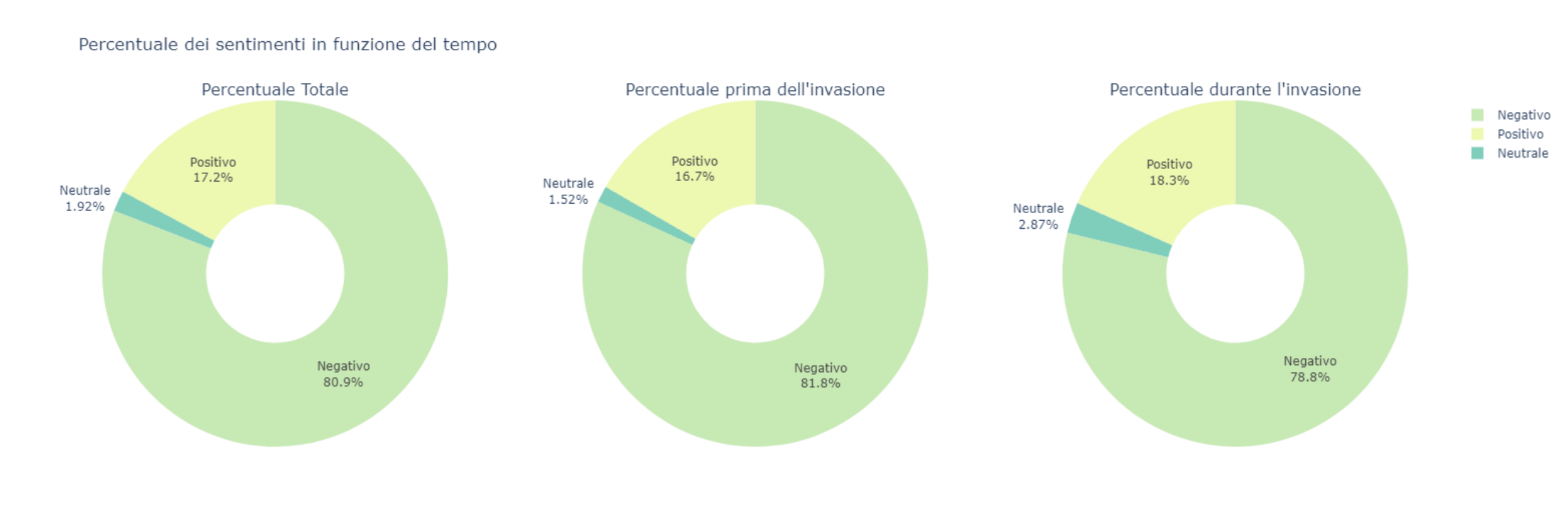




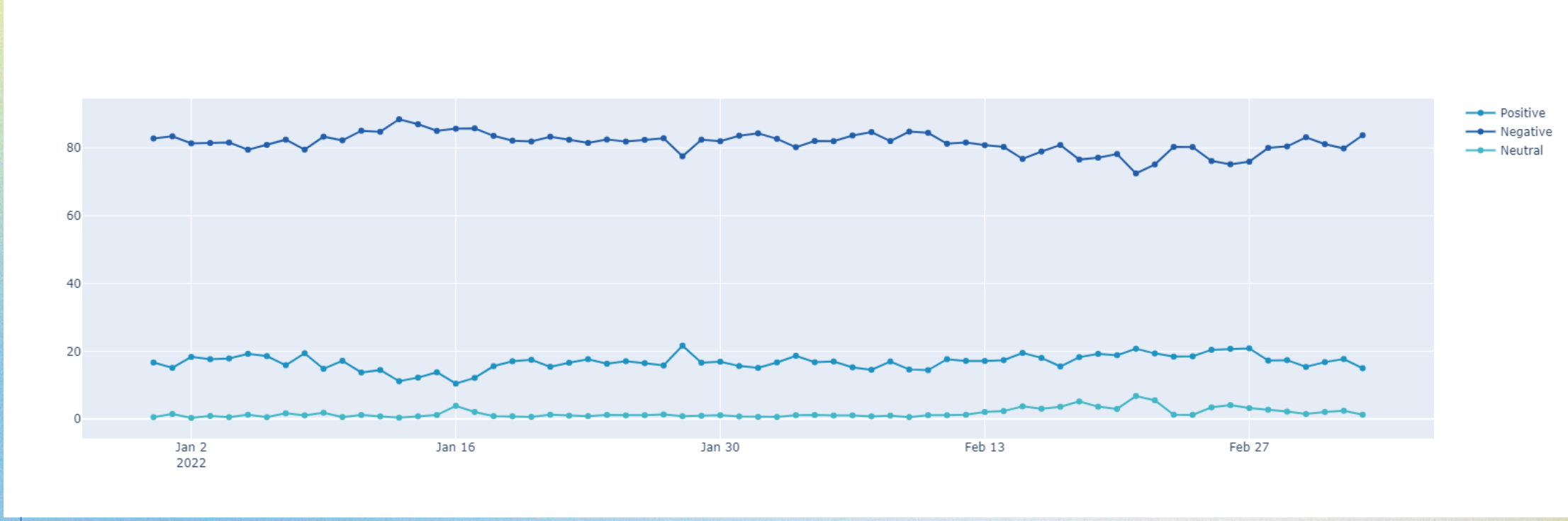
Tweets Positivi in relazione a quelli Negativi

Tweets Neutrali in relazione a Negativi+Positivi

Sentiment Analysis antecedente e durante il conflitto

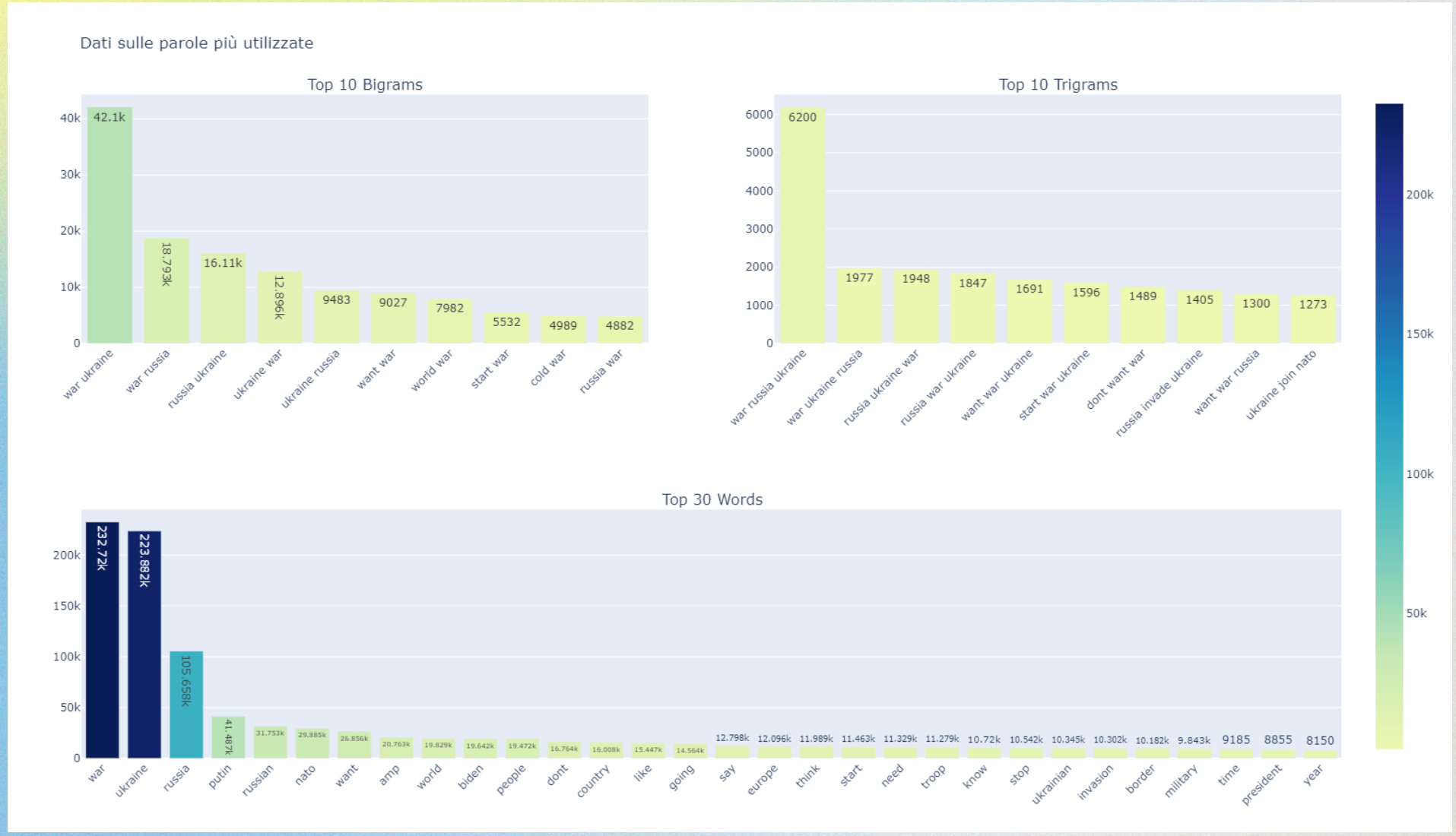


Sentiment Analysis antecedente e durante il conflitto

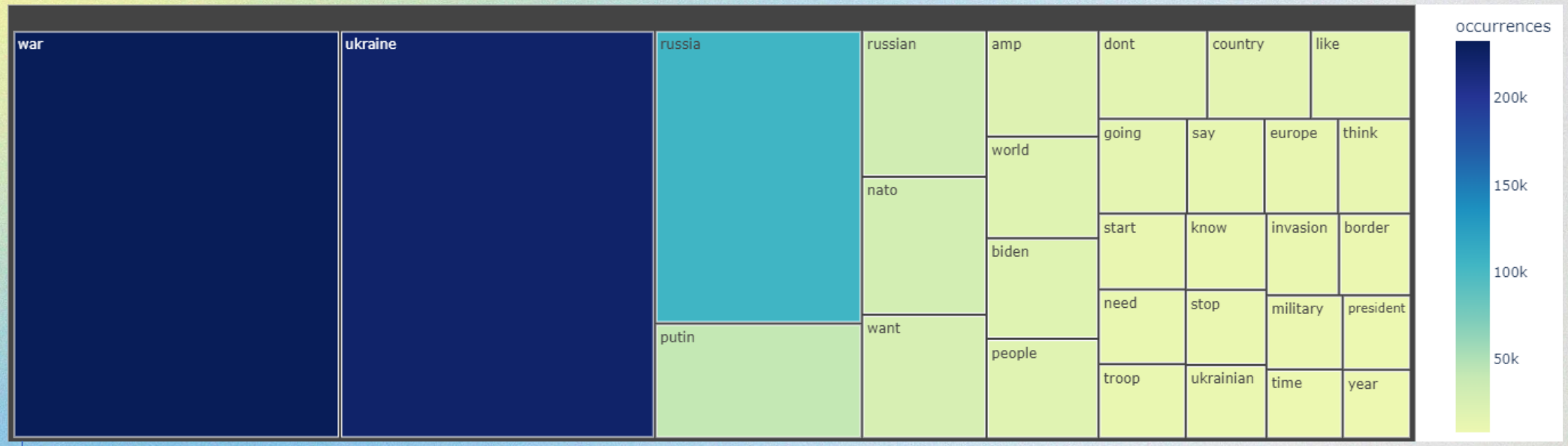


Percentuale dei sentimenti in funzione del tempo. Versione animata nel notebook

Analisi sul testo dei tweets



Analisi sul testo dei tweets



Tree Map delle parole più utilizzate nei tweets

Con Bigrams, Trigrams si intendono rispettivamente le coppie e terne di parole adiacenti nella frase. Nei grafici sono state rappresentate le 10 combinazioni più utilizzate

Jaccard Similarity

Per questa sezione e quella successiva di Clustering, sono stati definiti degli insiemi di parole relative a diversi temi, in particolare Economy, Social, Culture, Health.

L'obiettivo della Jaccard Similarity è trovare un valore numerico, uno score, utile a capire quanto due insiemi siano simili tra di loro; in questo caso gli insiemi sono i set di parole per tematica e i diversi tweet.

Una volta associato uno score, vengono classificati i tweet secondo una delle categorie tematiche presentate in precedenza.

Jaccard Index: Il coefficiente di Jaccard misura la similarità tra insiemi campionari, ed è definito come la dimensione dell'intersezione divisa per la dimensione dell'unione degli insiemi campionari:

$$jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Clustering

La suddivisione dei tweets nelle quattro categorie è abbastanza equa, in quanto le parole possono essere sovrapposte ed intese con più significati.

Per la parte di Clustering l'obiettivo è stato quello di utilizzare l'algoritmo K-Means per cercare di individuare dei cluster, o classi di appartenenza, tra due diversi attributi e per verificare se ci fosse un legame tra i dati. Le categorie scelte per l'analisi sono:

- Economic - Social
- Social - Culture
- Economic - Health
- Economic - Culture

> Algoritmo K-Means: è un algoritmo di analisi dei gruppi partizionale che permette di suddividere un insieme di oggetti in *k* gruppi sulla base dei loro attributi.
> L'obiettivo che l'algoritmo si prepone è di minimizzare la varianza totale intra-gruppo; ogni gruppo viene identificato mediante un centroide o punto medio.

Clustering

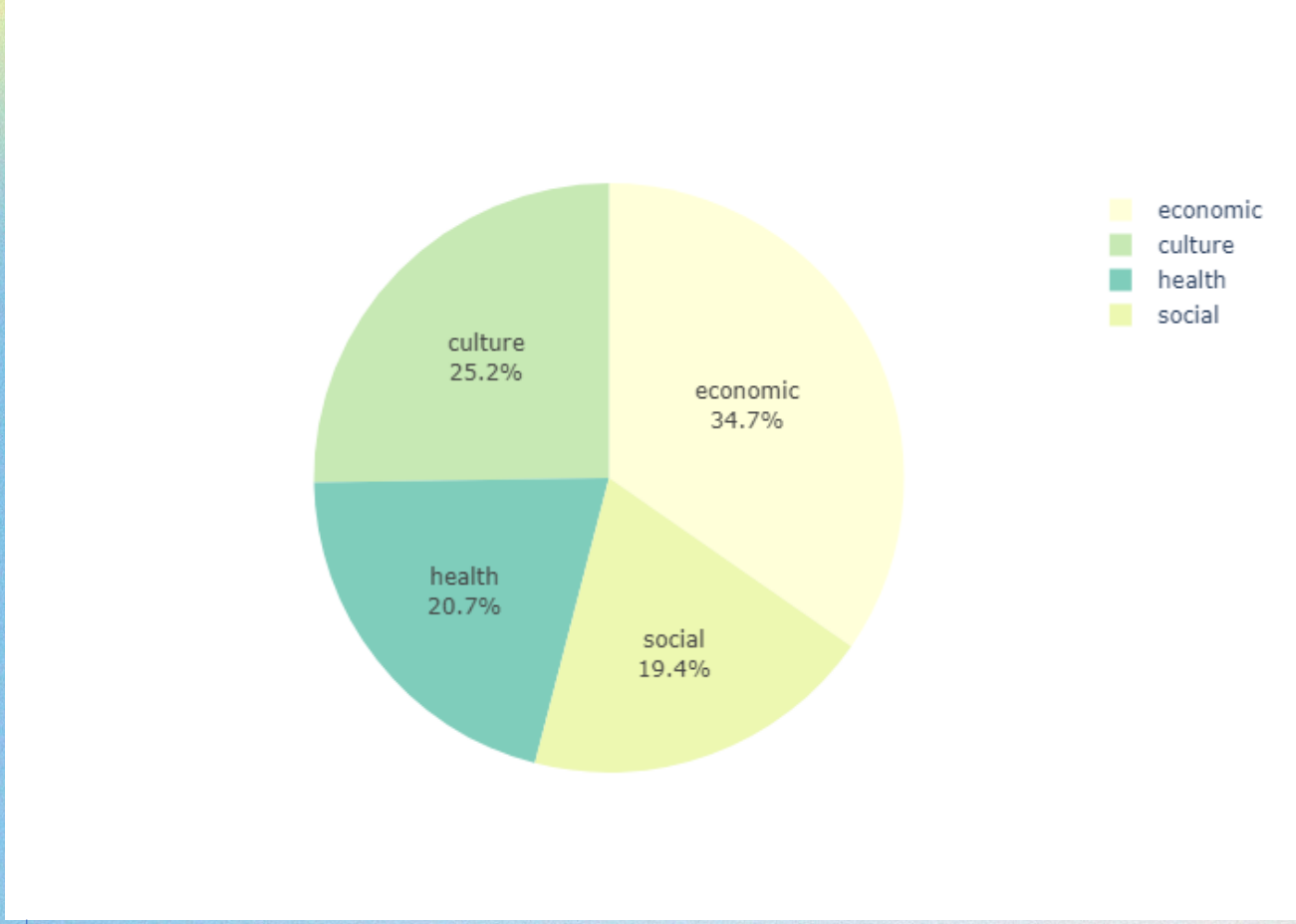
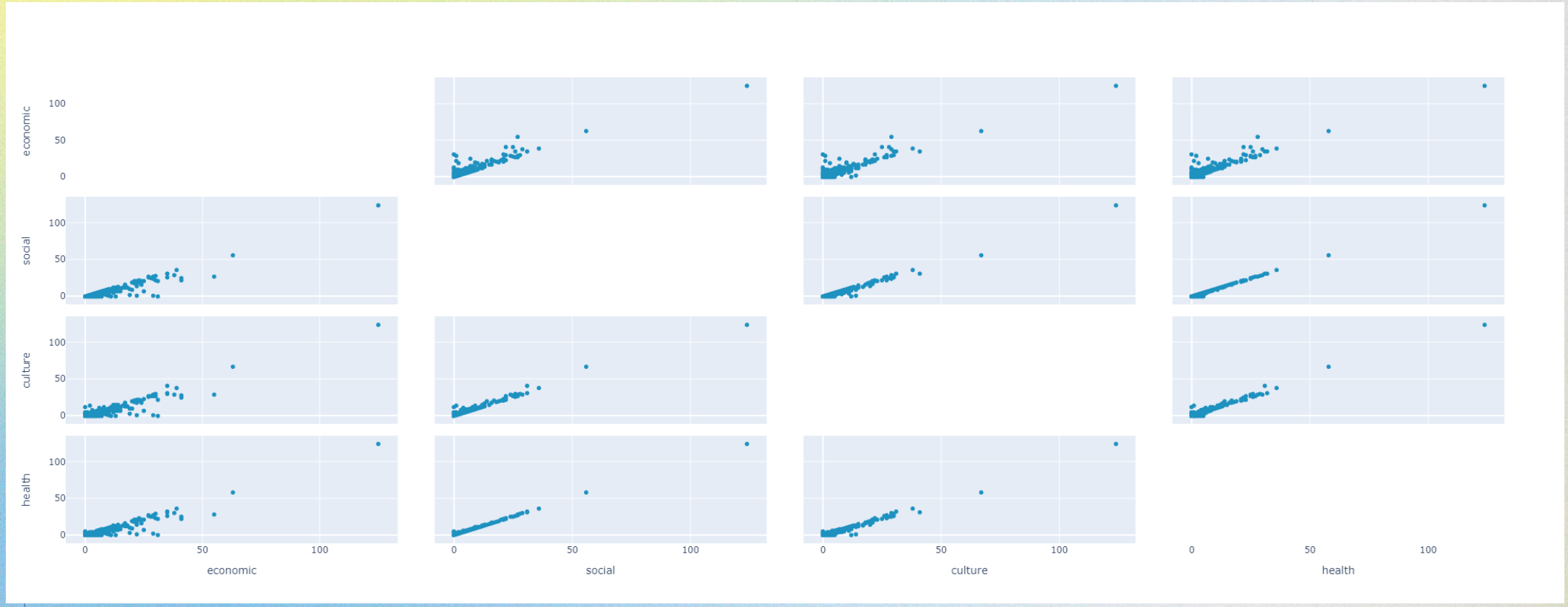


Grafico a torta per rappresentare la divisione dei tweets all’interno dei contenitori tematici

Clustering



«Scatter matrix», utilizzata per avere una visione generale sulle relazioni dei contenitori tematici.
La diagonale è omessa in quanto si confronta un attributo con se stesso

Clustering

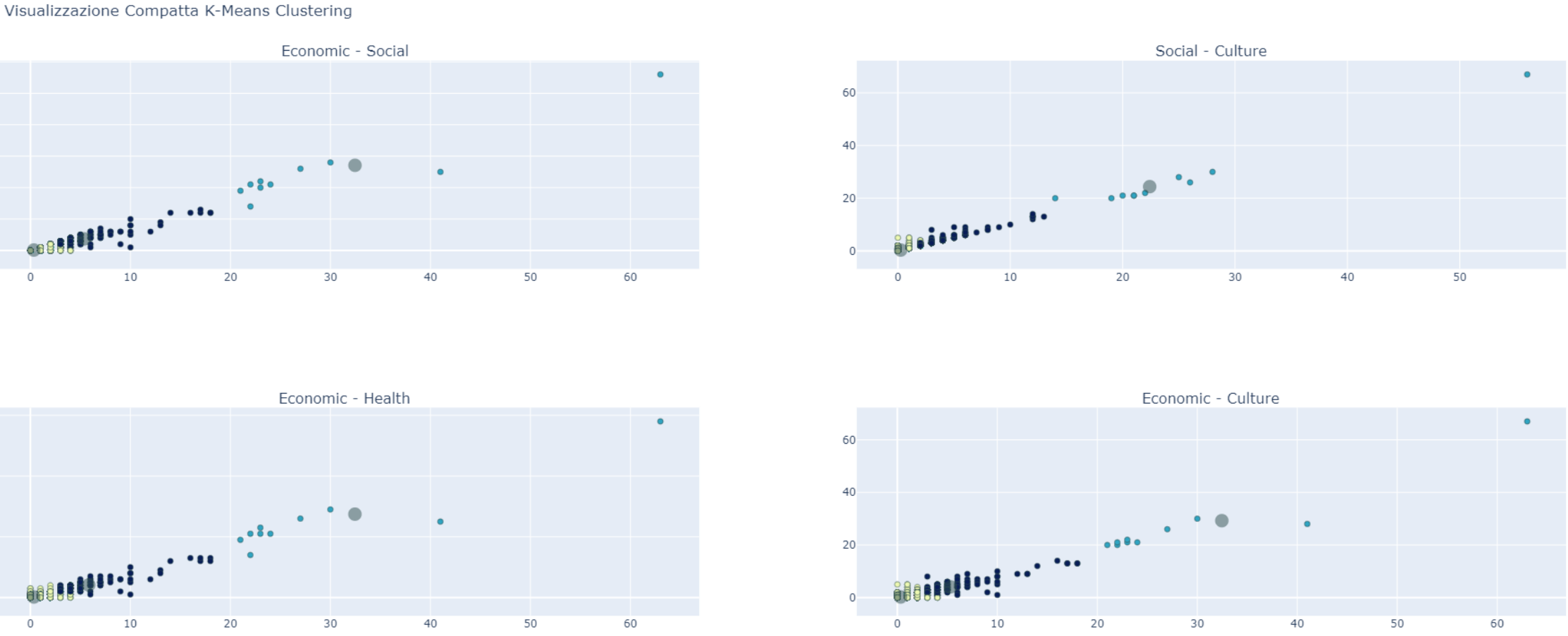
Per individuare il numero di cluster da utilizzare per l'algoritmo si usa il cosiddetto «Elbow Method», in quanto visualizzando uno scatter plot per gli attributi non è sufficiente per stabilire una netta divisione in clusters; esso consiste nel calcolare dei valori relativi alle distanze dei campioni dal centro dei clusters.

In questo caso per l'elbow method sono stati rappresentati i valori usando la Distorsione e l'Inerzia

Essendo la divisione nelle classi tematiche abbastanza equa, negli scatter plot si nota una relazione lineare tra gli attributi. Per una questione di performance sono stati presi per rappresentare i grafici i primi 5000 valori della lista (1000 per la visualizzazione compatta).

> Distorsione: viene calcolata come la media delle distanze al quadrato dai centri dei cluster dei rispettivi cluster. Tipicamente, viene utilizzata la metrica della distanza euclidea.
>
> Inerzia: è la somma delle distanze al quadrato dei campioni dal centro del cluster più vicino.

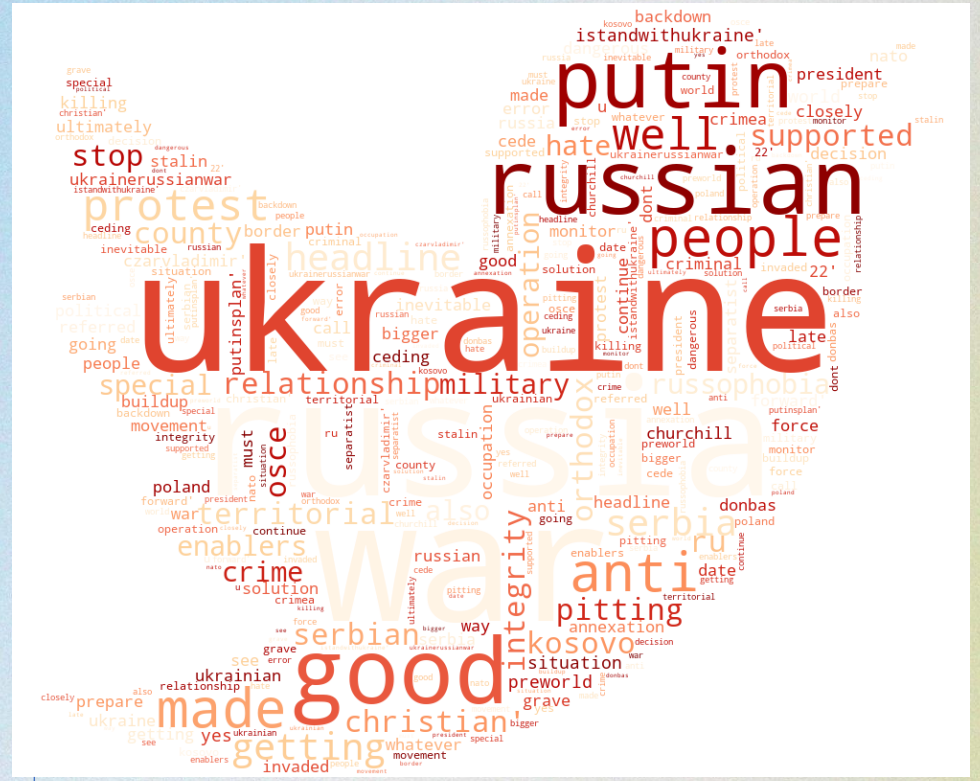
Clustering: K-Means



Word Clouds

Sono state infine realizzate delle Word Clouds per rappresentare in modo diretto un sottoinsieme delle parole Positive e Negative

Tweets Positivi



Tweets Negativi