UNIVERSITÀ DI PISA

Tesi di Laurea Triennale in Ingegneria Informatica

VALUTAZIONE DI METRICHE PER L'ANALISI DELLA POLARIZZAZIONE SU TWITTER

CANDIDATO

Stefano Agresti

RELATORI

Prof. Marco Avvenuti

Dott. Stefano Cresci

Dott. Leonardo Nizzoli

Il problema della polarizzazione sui social network

IL PROBLEMA

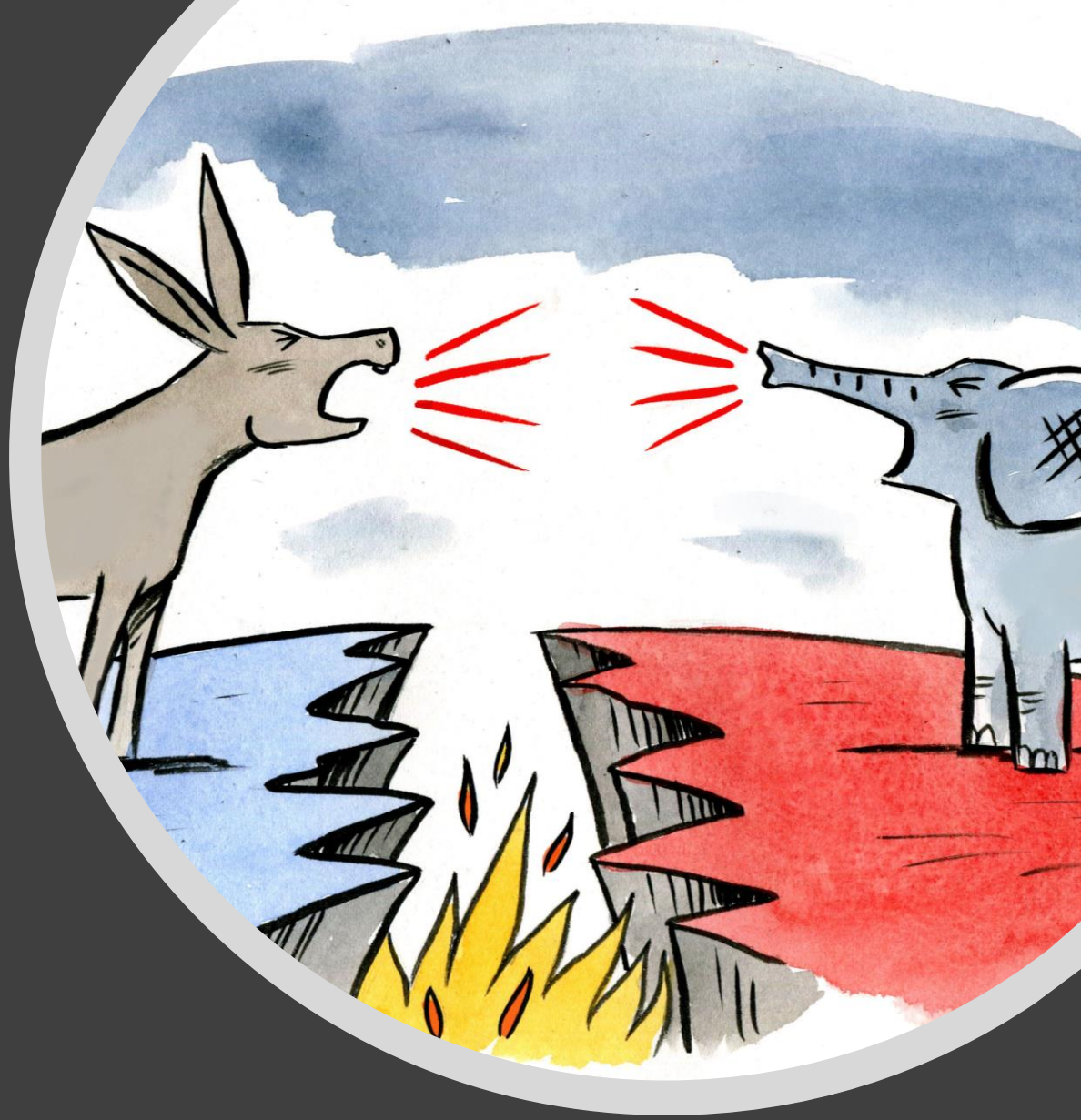
- La nascita di numerosi e variegati social network (come *Facebook*, *Twitter*, *Instagram* …) ha rivoluzionato il mondo delle comunicazioni interpersonali, portando con sé nuove e complesse sfide da affrontare.

IL PROGETTO DEL CNR

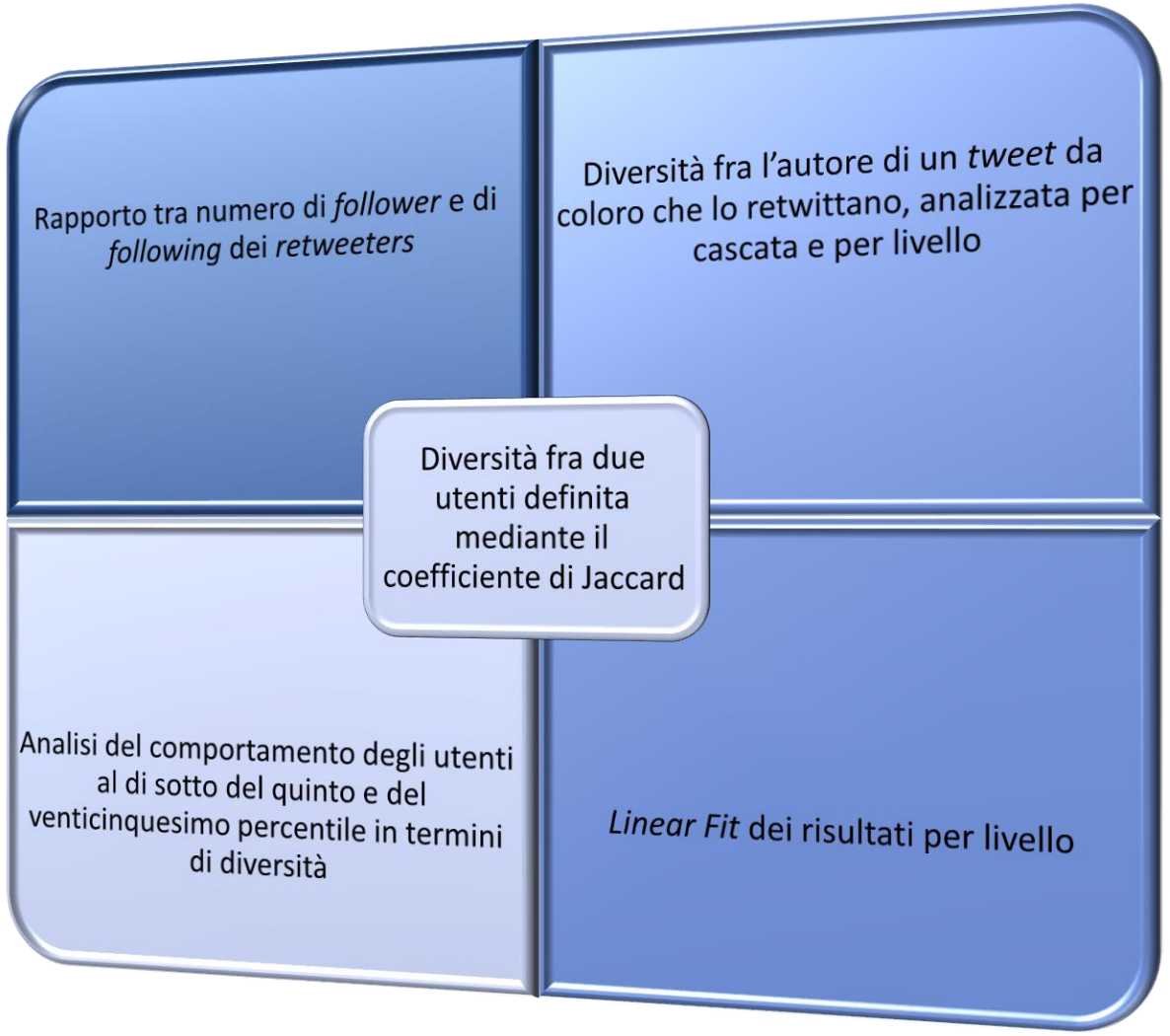
- Studiare la sempre maggior polarizzazione degli utenti su *Twitter* nell'ambito di discussioni su temi delicati come etica e politica.

LO SCOPO DELLA TESI

- Valutare la qualità di una serie di metriche con cui poter analizzare il problema in futuro.



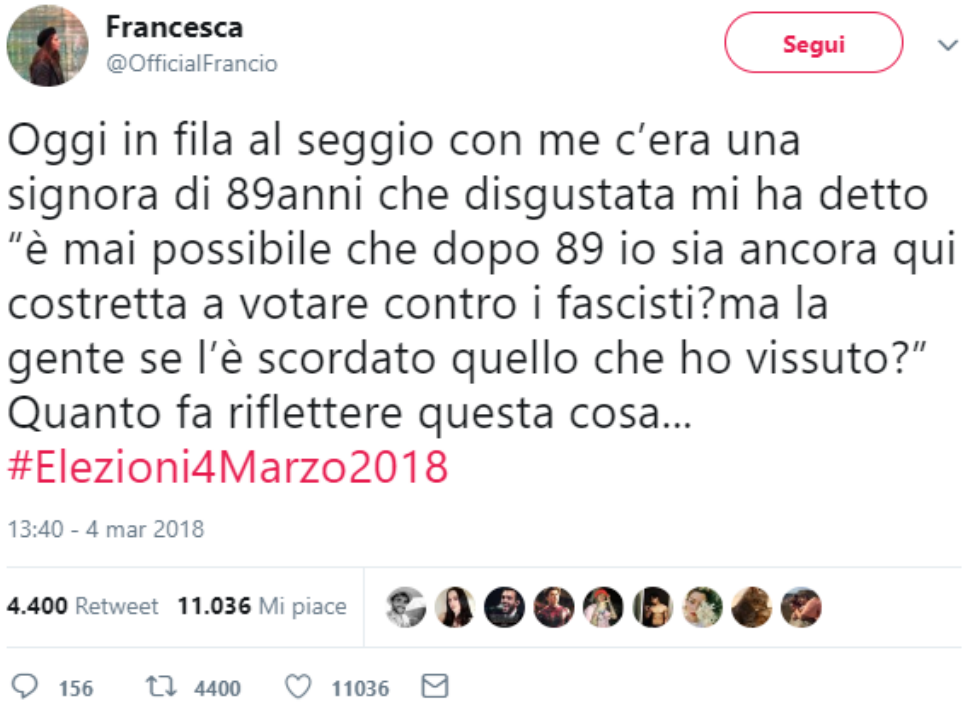
Metriche da valutare



Dataset utilizzato

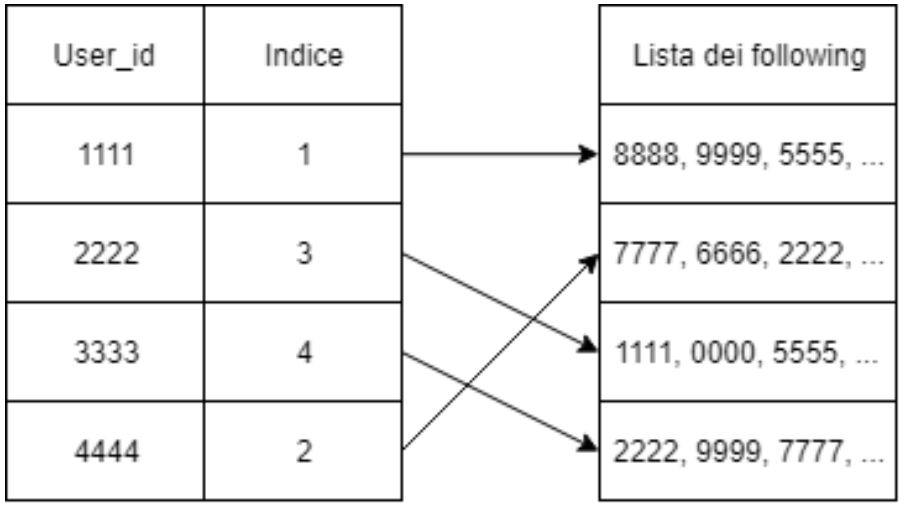
Il dataset utilizzato è stato realizzato raccogliendo le informazioni riguardanti 100 *tweet* reali relativi alle elezioni politiche del 2018.

- Tabella TWEETS ~ 100000 record
- Tabella USERS ~ 60000 record
- Tabella RETWEET_TREE ~ 100000 record
- Tabella LINKS ~ 91000000 record
- Tabella METRICS ~ 100 record



Raccolta dei *following* di un utente

- Uno dei principali problemi “tecnici” da affrontare è stata la gestione della tabella LINKS.
- Con oltre 91 milioni di record, sono richiesti più di tre minuti semplicemente per scorrerla.
- Per ridurre le tempistiche di accesso, viene indicizzata.



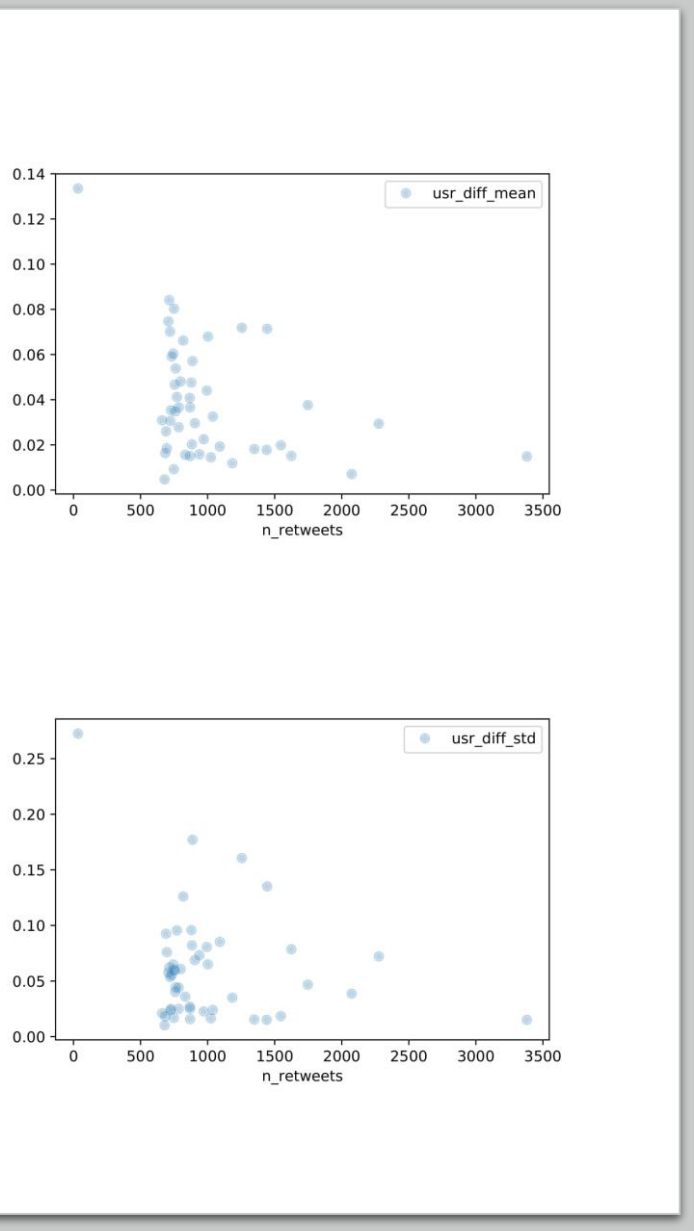
Calcolo del coefficiente di Jaccard

- La soluzione più immediata, basata sull'utilizzo di liste, non è molto efficiente in *Python.*
- Per migliorare le prestazioni si sfruttano i *set*, nettamente superiori per le operazioni sugli insiemi.
- Sia in termini di tempistiche, sia in termini di complessità, questo approccio viene premiato.

$$Coefficiente\ di\ Jaccard = \frac{Cardinalità(A \cap B)}{Cardinalità(A \cup B)}$$

Operation	Average case	Worst Case
x in s	O(1)	O(n)
Union s|t	O(len(s)+len(t))	
Intersection s&t	O(min(len(s), len(t))	O(len(s) * len(t))
Multiple intersection s1&s2&..&sn		(n-1)*O(l) where l is max(len(s1),..,len(sn))
Difference s-t	O(len(s))	
s.difference_update(t)	O(len(t))	
Symmetric Difference s^t	O(len(s))	O(len(s) * len(t))
s.symmetric_difference_update(t)	O(len(t))	O(len(t) * len(s))

	list_timing	set_timing
elements		
100	0.000370	0.000011
1000	0.008075	0.000082
10000	0.477722	0.001216
100000	49.045367	0.016954

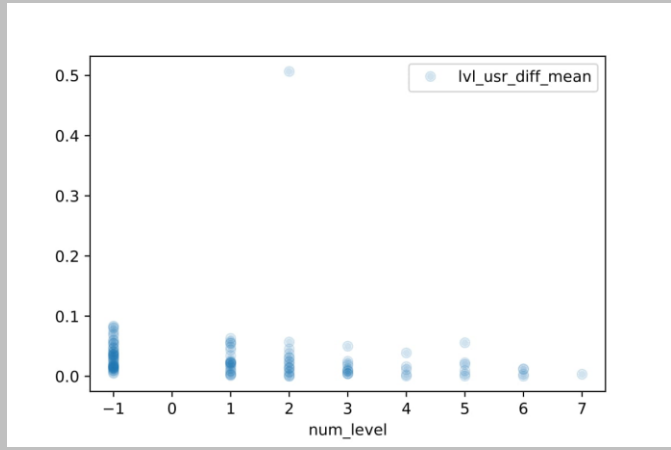


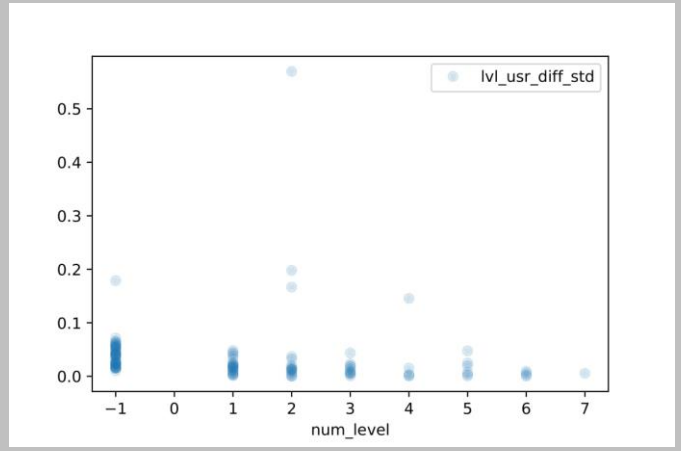
Risultati - Divisione per nodo

- Non è possibile individuare una netta diminuzione nella media del coefficiente di Jaccard all'aumentare dei *retweet.*
- È interessante notare la diminuzione del valore della deviazione standard riferito alle distribuzioni del coefficiente di Jaccard.

Risultati - Divisione per livello

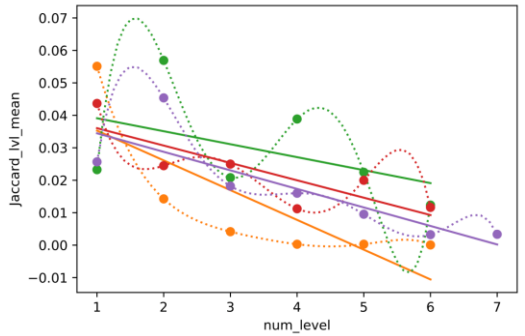
- Si può già osservare una diminuzione del coefficiente di Jaccard all'aumentare del livello.
- Come prima, si può vedere una diminuzione del valore della deviazione standard.

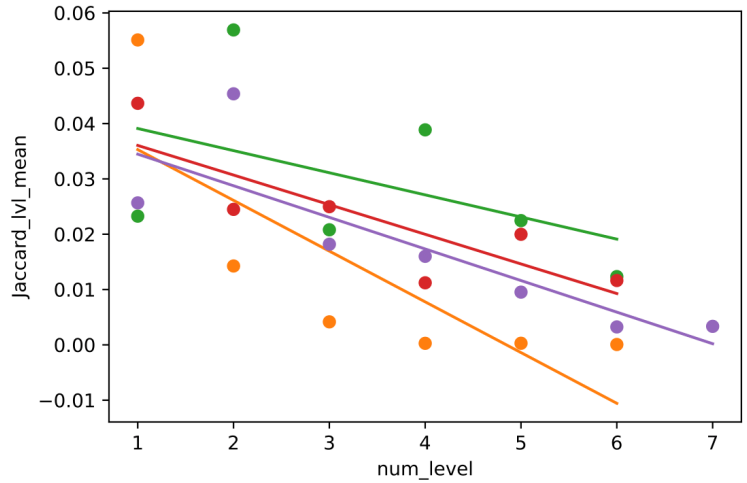


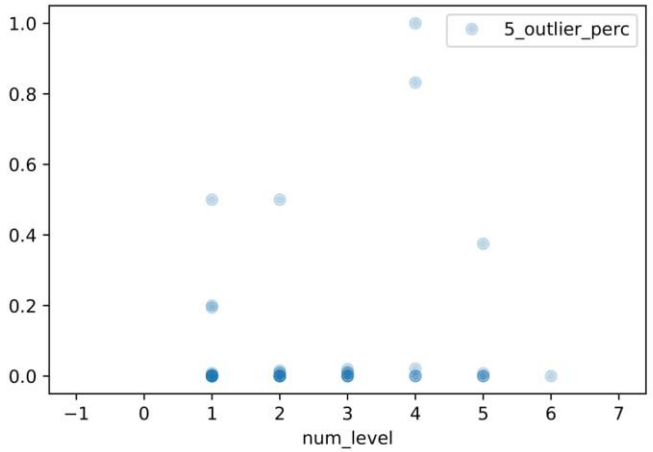


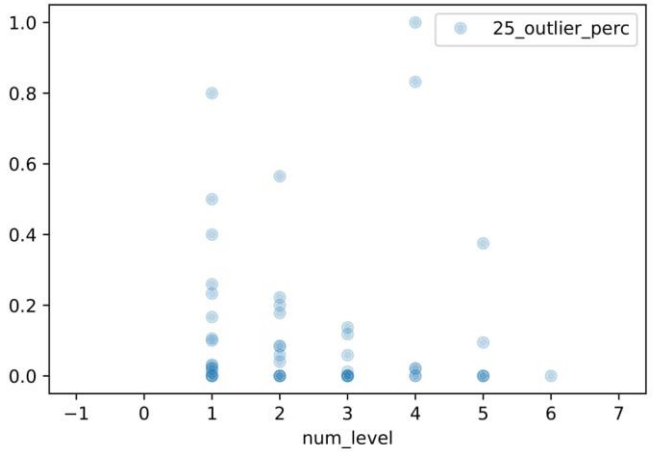
Risultati – *Linear Fit*

- Si conferma il trend osservato nei grafici precedenti.
- Si avvalora l'ipotesi secondo cui a cascate più lunghe corrispondono utenti più diversi.







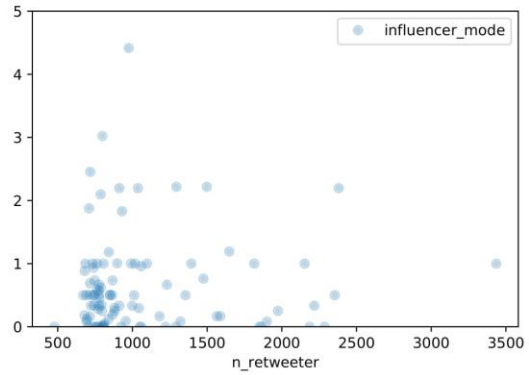


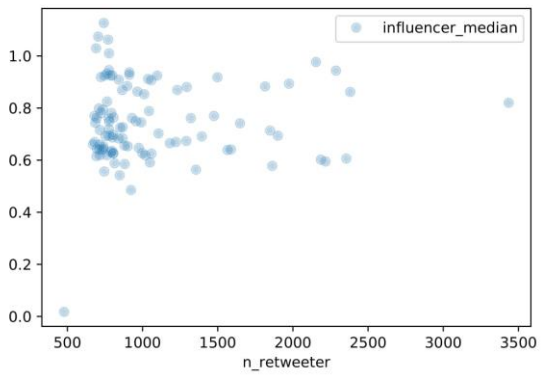
Risultati – Analisi dei percentili

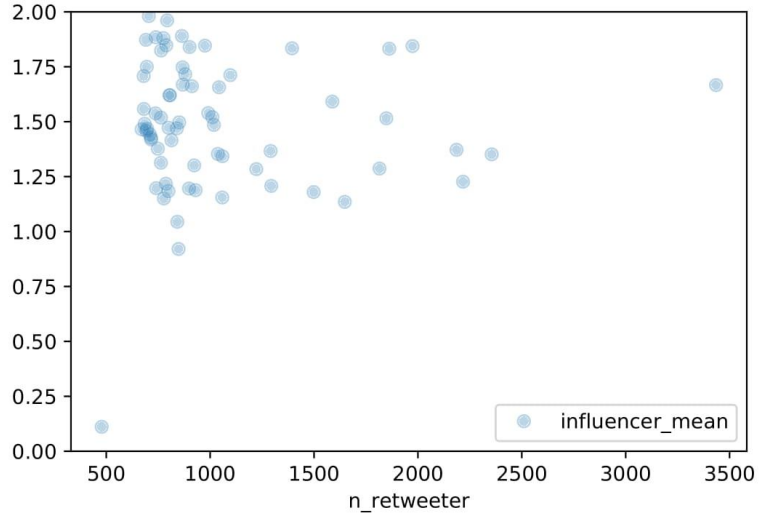
- In questi grafici è mostrata la percentuale di *retweet* ottenuti dagli utenti al di sotto del quinto e del venticinquesimo percentile (ovvero dal 5% e dal 25% degli utenti più diversi).
- Si evidenzia che i *retweet* non sono distribuiti equamente.

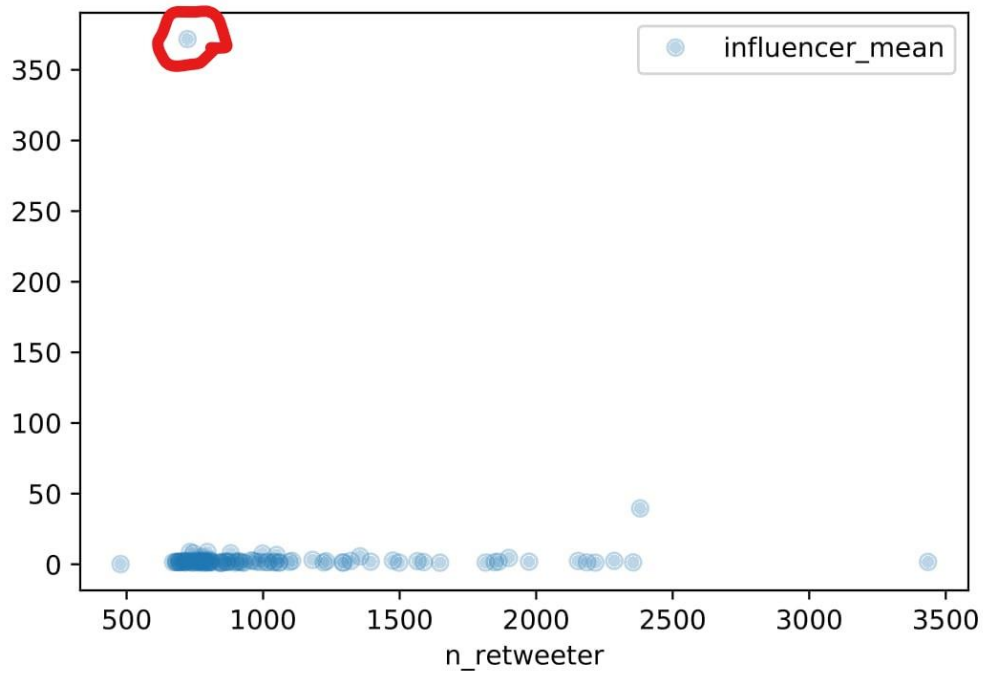
Risultati – Rapporto *follower/following*

- L’informazione più importante che si ricava da questi grafici è che la media dei rapporti tende ad essere superiore all’unità.
- Al contrario, moda e mediano sono generalmente inferiori a questo valore.
- Si può dedurre che pochi utenti molto popolari influenzino notevolmente questa metrica.









Caso particolare interessante

Un *tweet* della *BBC* spicca notevolmente dal punto di vista della metrica appena descritta

Aree di studio future

Questo lavoro può essere ampliato in moltissimi modi. Ecco alcune idee.

- Utilizzare un dataset di dimensioni più considerevoli per confermare i trend osservati finora.
- Analizzare l'andamento del valore medio del coefficiente di Jaccard con un insieme di cascate più variegato.
- Approfondire il discorso sugli *influencer*, quantificando il loro contributo alla diffusione di un *tweet*.