Analisi della polarizzazione di Endorsement Graph, attraverso sentiment analysis

Alessandro Valenti

La diffusione delle informazioni e di opinioni sin dai tempi antichi ha generato conflitti di ogni genere. Tali problematiche sono sempre più evidenti all'interno delle reti sociali che si vengono a creare mettendo in contatto individui con pensieri ed idee differenti tra loro. I conflitti vengono generati in base al tipo di argomento e quanto tale è "caldo" per gli utenti in questione. La polarizzazione è un utilissimo strumento per lo studio e l'analisi delle opinioni in differenti aree di ricerca all'interno di una rete sociale. Generalmente, la polarizzazione può essere applicata all'interno di contesti politici, sociali e culturali permettendo di comprendere al meglio quali siano gli schieramenti delle persone riguardo tali argomenti. La possiamo definire come:

> *Divisione in due gruppi fortemente contrastanti per una serie di opinioni o credenze.*

Questo processo di analisi può assumere diversi significati a seconda dello scenario studiato. Come ad esempio: *Polarizzazione Politica* (divergenza di opinione su estremi ideologici) o *Polarizzazione Sociale* (differenza di opinione all'interno delle società che possono nascere da disuguaglianze sociali ed economiche).

Un problema che sta affliggendo i social media è la formazione degli *Echo-chambers*, cioè quelle comunità che condividono le stesse opinioni rafforzando il proprio punto di vista senza cercare di verificare la presenza o meno di altre opinioni. La polarizzazione è uno strumento che può essere utilizzato per identificare questi gruppi di utenti, ma può assumere anche altri compiti come lo studio delle opinioni e la loro diffusione all'interno della rete. Questa tesi è stata sviluppata per cercare di studiare il comportamento delle informazioni espresse nei social media, in questo caso si è scelto *Twitter*. I dati, ovvero i *Tweet*, sono stati raccolti attraverso una ricerca per *hashtag* e classificati in due gruppi di pensiero distinti rispetto all'argomento analizzato. La raccolta

dei dati è stata effettuata all'interno di una finestra temporale per poter studiare la polarizzazione e la sua evoluzione nel tempo.

Il primo problema che si è cercato di risolvere è stato quello di classificare le notizie attraverso una comprensione testuale. Tale operazione è stata resa possibile attraverso la tecnica della *Sentiment Analysis*, che consente di classificare i messaggi degli utenti attraverso un'analisi del sentimento. Una volta suddivise le opinioni all'interno della rete è stato costruito un grafo contenente i tweet ed i relativi retweet, questo viene definito *endorsement graph*. Partendo dal grafo sono stati applicati due algoritmi per il calcolo della polarizzazione definiti come segue:

- *Algoritmo basato sul grado del grafo*: Una volta generato il grafo vengono catalogati i nodi in due categorie:
  - *Elite*: l'utente che ha *tweettato* un'opinione.
  - *Listener*: l'utente che ha *retwettato* il tweet di uno o più nodi *Elite*.

  Partendo da queste categorie viene calcolata la polarizzazione sfruttando il grado di ogni nodo, assegnando un valore numerico, in base al gruppo di appartenenza (*Elite*) e poi calcolare la polarizzazione attraverso la formula espressa nel paper *Measuring Political Polarization: Twitter shows the two sides of Venezuela*.

- *Algoritmo basato sulla topologia*: La differenza principale con il precedente algoritmo, consiste nell'utilizzare la tecnica dei *Random Walk*. La polarizzazione dipende fortemente dalla topologia dell'*endorsement graph* e dalla probabilità sugli archi. Questa è definita probabilità di *retweet*, ed è pari al rapporto tra il numero di retweet fatti da un utente su un certo tweet sul numero totale di retweet che l'utente ha effettuato su quel topic. Anche questo algoritmo si basa su un paper: *Reducing Controversy by Connecting Opposing Views*

Una volta calcolata la polarizzazione sono stati utilizzate tecniche di *Forecasting* come: *Double exponential smoothing*, *linear regression* e *moving average*. Queste tecniche consento di poter predire il comportamento della rete nel futuro e quindi potrebbe essere utile in ottica di prevenzione di *Echo-chamber*. Tutte queste funzionalità sono state realizzate mediante librerie *Python*, mentre la raccolta dei dati è stata effettuata all'interno di una istanza *EC2*.

I topic utilizzati per sperimentare il comportamento della polarizzazione sono stati due cioè: *elezioni regionali in Sicilia* ed il *Biotestamento*. Il motivo

di tale scelta è dettata dalla curiosità di testare queste funzionalità per due argomenti che ricoprissero due contesti differenti tra loro; anche perché molto dibattuti. I risulti ottenuti per entrambi i topic sono stati molto soddisfacenti perché la polarizzazione ha assunto un comportamento in linea con la realtà. Per il primo il trend della diffusione delle opinioni ha rispecchiato quanto osservato nei risultati delle elezioni, infatti attraverso la raccolta dei tweet è stato possibile notare come nel mese di Novembre (mese in cui ci sono state le elezioni), prima del giorno delle elezioni gli utenti che si schieravano verso la coalizione del centro-destra fosse simile a quelle del Movimento 5 Stelle. Subito dopo il giorno delle elezioni in conformità con i risultati ottenuti si è potuto constatare come gli utenti aderenti al centro-destra superassero quelli del Movimento 5 Stelle. Per sperimentare tale topic attraverso la polarizzazione è stato necessario considerare solo due forze politiche, in questo caso quelle che hanno ottenuto il maggior numero di voti.

Per quanto riguarda il biotestamento si è potuto notare come il retaggio culturale e religioso di un paese avessero una forte influenza anche nei social-media, infatti la polarizzazione della coalizione degli utenti contrari era di gran lunga superiore a quella dei favorevoli. La studio di questo topic si può definire *hot*, perché ancora di attualità e lo si è potuto riscontrare raccogliendo i dati nel mese successivo al 14 Dicembre (giorno in cui è stata emanata la legge sul Biotestamento) notando come il numero dei tweet aumentava. Da questo progetto è possibile effettuare nuove ricerche per risolvere il problema delle comunità fortemente polarizzate attraverso una analisi basata sulla *controversia* cioè cercare di far comunicare le comunità isolate in qualche modo. Altri spunti potrebbero essere quello di raffinare la classificazione dei tweet attraverso uno studio delle immagini, media e link pubblicati all'interno dei tweet, perchè la sentiment analysis per quanto possa essere raffinata non consente di poter percepire l'ironia all'interno del testo, ironia che può essere espressa attraverso elementi multimediali.