Universitá di Roma Tor Vergata
Corso di Laurea magistrale in Ingegneria Informatica
Dipartimento di Ingegneria Informazione

Analisi della polarizzazione di Endorsement Graph, attraverso sentiment analysis

Relatore:
Giuseppe F. Italiano
Correlatore:
Ing. Nikos Parotsidis

Candidato:
Alessandro Valenti
matricola 0228709

Anno Accademico 2016-2017

A qualcuno...

Sommario

Il sommario deve contenere 3 o 4 frasi tratte dall'introduzione di cui la prima inquadra l'area dove si svolge il lavoro (eventualmente la seconda inquadra la sottoarea più specifica del lavoro), la seconda o la terza frase dovrebbe iniziare con le parole "Lo scopo della tesi è ..." e infine la terza o quarta frase riassume brevemente l'attività svolta, i risultati ottenuti ed eventuali valutazioni di questi.

NB: se il relatore effettivo è interno al Politecnico di Milano nel frontesizo si scrive Relatore, se vi è la collaborazione di un altro studioso lo si riporta come Correlatore come sopra. Nel caso il relatore effettivo sia esterno si scrive Relatore esterno e poi bisogna inserire anche il Relatore interno. Nel caso il relatore sia un ricercatore allora il suo Nome COGNOME dovrà essere preceduto da Ing. oppure Dott., a seconda dei casi.

Ringraziamenti

Ringrazio .....

Capitolo 1

Introduzione

La diffusione delle informazioni e di opinioni sin dai tempi antichi ha generato conflitti di ogni genere, per contrapposizioni sociali, culturali, religiosi ed economici. Tali problematiche sono sempre piú 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 argomento é "caldo" per gli utenti in questione. La polarizzazione é 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 puó essere applicata all'interno di contesti politici, sociali e culturali permettendo di comprendere al meglio quali siano gli schieramenti delle persone riguardo tali argomenti. Una generica definizione della polarizzazione é la seguente:

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

Questo processo di analisi puó assumere diversi significati a seconda dello scenario studiato.

- *Polarizzazione Politica*: divergenza di opinione su estremi ideologici.
- *Polarizzazione Sociale*: differenza di opinione all'interno delle societá che possono nascere da disuguaglianze sociali ed economiche.

La polarizzazione puó comportare diversi cambiamenti sullo scenario in questione, in quanto mette in luce come la formazione di due grandi gruppi non consenta una diffusione democratica delle opinioni. A tal proposito é interessante notare come la divisione in queste due grandi partizioni generi alcune problematiche quali:

- La frammentazione della rete stessa.
- L'isolamento delle opinioni.

In conclusione potremmo definire la polarizzazione come un un processo sociale per cui gli utenti che vi partecipiano vengono divisi in due grandi sottogruppi aventi visioni, punti di vista ed opinioni differenti del problema in questione, con alcuni individui che rimangono neutrali tra i due grandi gruppi.

La formazione di due comunitá isolate che non comunicano tra loro, comporta un problema di isolamento delle opinioni cioé un utente che appartiene a quel gruppo difficilmente potrá ricevere informazioni o aderire alle idee del gruppo antagonista. Otteniamo la formazione degli *Echo-Chambers*. Si definiscono *Echo-Chambers* come:

> *Una situazione in cui le informazioni, le idee e le credenze vengono rinforzate e amplificate perché espresse all'interno dello stesso ambiente, rimanendo isolato.*

Un'altro problema che puó formare una forte polarizzazione delle opinioni e delle informazioni sono i *Filter Bubble* ovvero:

> *Uno stato di un isolamento intellettuale che puó essere ottenuto a partire da risultati di ricerche su siti che registrano la storia del comportamento dell'utente. Questi siti sono in grado di utilizzare informazioni sull'utente per scegliere selettivamente tra tutte le risposte quelle che vorrá vedere l'utente stesso. L'effetto é di isolare l'utente da informazioni che sono in contrasto con il suo punto di vista, effettivamente isolandolo nella sua bolla culturale o ideologica.*

Come precedentemente anticipato la polarizzazione é uno strumento che puó essere facilmente utilizzato per individuare tutte queste problematiche all'interno dei moderni Social Network come *Facebook*, *Twitter* e molti altri. Questo perché gli utenti si sentono sempre piú liberi di poter esprimere le proprie opinioni all'interno di queste piattaforme riguardo problematiche sociali, culturali, politiche ed economiche. Non é sempre possibile poter uscire dalle *Filter Bubble*, perché gli stessi social network tendono a indirizzare l'utente a visualizzare informazioni che potrebbero interessargli senza farli confrontare con opinioni divergenti. Alla luce di questo grande problema il calcolo di una polarizzazione puó consentire agli amministratori dei social network di individuare i topic piú polarizzati garantendo una diffusione democratica delle informazioni, facendo comunicare gli utenti con opinioni divergenti.

L'obiettivo della mia tesi consiste nell'utilizzare la polarizzazione per poter individuare gli argomenti fortemente polarizzati e comprendere come tali informazioni vengono prodotte all'interno della rete sociale. Lo sviluppo di questo strumento é stato effettuato attraverso due algoritmi, presentati nei seguenti paper:

- ***Measuring Political Polarization: Twitter shows the two sides of Venezuela***: Studia la diffusione delle informazioni all'interno

di un *endorsement graph* collezionando i dati relativi alle elezioni in Venezuela all'interno del *social network Twitter*. Viene effettuato uno studio della polarizzazione all'interno di un contesto politico attraverso la diffusione delle opinioni sui candidati politici durante le ultime elezioni presidenziali, l'*endorsement graph* viene costruito partendo da un nodo che pubblica nella rete un *Tweet* esprimendo la propria opinione, formando un nodo, mentre eventuali follower di quell'utente che *retwettano* tale notizia sono nuovi nodi all'interno del grafo con archi uscenti verso il nodo che hanno *retwettato*. In questo modo viene generato un grafo basato sul *retweet*. 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 piú nodi *Elite*.

Partendo da queste categorie viene calcolata la polarizzazione sfruttando il grado di ogni nodo, tale operazione viene eseguito iterativamente fino ad ottenere una stabilizzazione della polarizzazione.(Per una piú dettagliata spiegazione si rimanda al Capitolo??)

- ***Reducing Controversy by Connecting Opposing Views***: Identifica la polarizzazione sfruttando la topologia del grafo. Il grafo viene generato utilizzando la medesima tecnica del precedente paper, cosí come il social network di riferimento *Twitter*. La differenza principale con la soluzione proposta in precendenza, consiste nell'utilizzare la tecnica dei *Random Walk*. La polarizzazine adottando questo approccio dipende fortemente dalla topologia dell'*endorsement graph*. Per la spiegazione tecnica si rimanda al capitolo??.

Prima di poter effettuare il calcolo vero e proprio della polarizzazione occorre effettuare una prima scrematura,da interdersi come una prima classificazione delle opinioni in due gruppi contrastanti, nel dettaglio attraverso la *Sentiment Analysis*. Questa particolare tecnica consente di partizionare il grafo in due gruppi che per semplicitá chiameremo **Rossi** e **Blu**, nel dettaglio viene analizzato il testo contenuto in un tweet o in un post (a seconda del social network adottato) catalogandolo per un gruppo piuttosto che un altro a seconda del contenuto e all'affinitá col topic in questione. Per meglio comprendere cosa viene effettuato presentiamo la definizione di *Sentiment Analysis*:

> *L'Analisi del sentiment o Sentiment analysis (ma anche opinion mining) é la maniera a cui ci si riferisce all'uso dell'elaborazione del linguaggio naturale, analisi testuale e linguistica computazionale per identificare ed estrarre informazioni soggettive da diverse fonti.*

In conclusione l'analisi semantica consente di poter catalogare le informazione in base alla lora vicininanza alle opinioni di un gruppo piuttosto che ad un altro, ed eventualmente scartare quelle informazioni che non sono di alcun interesse per l'utente. Tale operazione é possibile soltanto se la macchina é stata precedentemente istruita sul topic in questione, infatti si definisce *training set* l'insieme delle informazioni di riferimento che consentono alla macchina di poter distinguere le opinioni a seconda del loro contenuto.

Dopo aver effettuato questa separazione o catalogazione delle informazioni é possibile identificare quali utenti siano piú o meno vicini ai due poli di un'opinione. Ricapitolando partendo da *post* o *topic* viene effettuata la *Sentiment analysis*, viene costruito l'*endorsement Graph* ed infine calcolata la **polarizzazione**. Diamo ora qualche informazione in piú sulla polarizza-

zione, a livello matematico la polarizzazione puó assumere valori compresi tra [-1,1] definendo in questo modo due poli opposti. I nodi che avranno una polarizzazione pari a 0 sono da ritenersi nodi neutri ovvero che non sono soggetti ad una forte polarizzazione, ma sono l'emblema della democrazia, in quanto ricevono informazioni da entrambi i gruppi.

Per concludere é stato effettuato anche uno studio per poter consentire la predizione del valore della polarizzazione in un periodo futuro. In questo modo gli amministratore dei social network possono effettuare degli accorgimenti alla rete consentendo una democratica diffusione delle opinioni, senza creare *Echo-Chambers* e *Filter Bubble*. La predizione é stata realizzata attraverso tecniche di *Forecasting* molto utilizzate in contesti economici, in quanto consentono, attraverso delle serie numeriche, di poter predirre il valore nell'istante temporale successivo. Sfruttando queste particolaritá é stato possibile effettuare una predizione, nel dettaglio le tecniche utilizzate sono tre:

- *Double exponential smoothing*
- *Linear regression*
- *Average window*

Per i fondamenti teorici si rimanda al Capitolo???.

Terminiamo questa sezione presentando i casi studio utilizzati. Per lo sviluppo della mia tesi ho deciso di analizzare la polarizzazione all'interno di due contesti differenti, attreverso questi due Topic:

- **Elezioni Regionali in Sicilia nel 2017**: analisi in un contesto politico.
- **Biotestamento**: analisi in un constesto sociale.

I dati relativi a questi due topic sono stati raccolti attraverso il social network *Twitter*, dal 01/09/2017 al 20/12/2017. Il periodo indicato é stato scelto per analizzare l'evoluzione della polarizzazione nel tempo comprensiva della conclusione di questi due topic. Il 05/12/2017 si sono svolte le elezioni regionali in sicilia ed il 14/12/2017 il parlamento italiano ha approvato la legge sul biotestamento. I dati sono stati raccolti effettuando una ricerca attraverso i 5 *hashtags* piú utilizzati per entrambi i topic. Questi *hashtag* non esprimono nessuna opinione o parere presi singolarmente ma sono delle parole chiavi necessarie per catalogare il contesto del *tweet*. Per catalogare i dati raccolti e poi valutare la polarizzazione sono state adottate le tecniche precedentemente illustrate.

Per quanto rigurda le elezioni regionali in sicilia si é decisono di raccogliere i tweet relativi alle due grandi fazioni che hanno dominato la scena politica siciliana:

- Il *Movimento 5 Stelle*.
- *Forza Italia* (la coalizione del centro destra).

Innanzi tutto é stata una scelta dettata dai risultati conseguiti durante le suddette elezioni e dal fatto che in Italia non sono presenti soltanto due fazioni politiche, come in molti altri paesi del mondo, quindi sarebbe risultato impossibile definire un valore polarizzato se avessimo considerato piú di due fazioni politiche. A tal proposito sono stati scartati i dati relativi a quei candidati politici appartenenti ad altri partiti politici e coalizione rispetto a quelli sopra elencati, utilizzando la *Sentiment Analysis*. I risultati ottenuti da questo topic hanno un comportamento interessante, cioé il cambiamento nel tempo della polarizzazione, seguendo il trend riscontrato durante i sondaggi effettuati mensilmente. Nel dettaglio si puó facilmente assistere ad un

cambiamento di trend col passare del tempo. Si comincia con una polarizzazione del 79% a favore del *Movimento 5 Stelle* per concludere alla fine del suddetto periodo con un capolgimento di fronte con il 59% della polarizzazione a favore della coalizione di *Forza Italia*, mostrando un cambiamento radicale nel tempo conforme con quanto accaduto nei sondaggi.

Per quanto riguarda il *Biotestamento* si é deciso di raccogliere i tweet relativi alla legge approvata il 14 Dicembre 2017 dal parlamento italiano, per analizzare la polarizzazioni in un contesto sociale. La polarizzazione riguardava l'adesione o meno a questa legge, infatti si é riscontrata una fortissima polarizzazione verso i contrari all'attuazione di tale legge. Questi risultati sono conformi al contesto sociale e religioso presente in Italia, confermando quanto spiegato in precedenza e cioé quanto un retaggio culturale o religioso possa influenzare il pensiero e le opinioni di una persona. Questa considerazione non é applicabile soltanto nella vita di tutti i giorni ma anche all'interno di un social network e ció viene dimostrato dai risultati ottenuti all'interno di questo topic.

In conclusione questi due Topic hanno contribuito a confermare quanto precedentemente spiegato all'interno di questo capitolo e cioé che la polarizzazione é un potentissimo strumento che consente di poter individuare gli *Echo-chambers* presenti nella rete. Eventuali sviluppi futuri possono riguardare l'eliminazione degli *Echochambers* abbassando la controversia tra i due gruppi ottenuti attraverso la polarizzazione, effettuando dei congiungimento tra quei gruppi di nodi che condividono sempre le stesse opinioni.

Capitolo 2

Stato dell'arte

2.1 Stato dell'arte

All'interno delle reti sociali sta sempre piú prendendo piede il problema della polarizzazione delle opinioni. Nel liguaggio comune il confronto tra individui ha sempre generato una forte controversia nelle opinioni oppure una situzione di neutralitá nelle opinione oppure una visione comune nelle opinione. I social netwrok hanno permesso all'utente di poter diffondere attraverso post, messaggi o espressioni audio video le proprie opinioni eo pensieri all'interno di una comunitá sociale. A tal proposito per favorire la diffusione delle diverse correnti di pensiero i social network stanno sempre piú sviluppando algoritmi per permettere di identificare le comunità isolate che condividono un unico punto di vista di un problema. La polarizzazione è un algoritmo matematico che applicato all'interno delle rete sociali permette di capire quanto un utente che accede per la prima volta all'interno di una rete sociale venga influenzato dagli altri utenti e quanto una news o un giudizio si propaga all'interno di una rete sociale. Prima di poter illustrare questo algoritmo con le relative problematiche verrá illustrata una definizione di rete sociale.

Rete Sociale Una rete sociale consiste in un qualsiasi gruppo di individui connessi tra loro da diversi legami sociali. Per gli esseri umani i legami vanno dalla conoscenza casuale, ai rapporti di lavoro, ai vincoli familiari. Le reti sociali sono spesso usate come base di studi interculturali in sociologia, in antropologia, in etologia.

L'analisi delle reti sociali, ovvero la mappatura e la misurazione delle reti sociali, puó essere condotta con un formalismo matematico usando la teoria dei grafi. In generale, il corpus teorico ed i modelli usati per lo studio delle reti sociali sono compresi nella cosiddetta social network analysis.

La ricerca condotta nell'ambito di diversi approcci disciplinari ha evidenziato come le reti sociali operino a piú livelli e svolgano un ruolo cruciale nel determinare le modalitá di risoluzione di problemi e i sistemi di gestione delle organizzazioni, nonché le possibilitá dei singoli individui di raggiungere i propri obiettivi.

Le reti La diffusione del web e del termine social network ha creato negli ultimi anni alcune ambiguitá di significato. La rete sociale é infatti storicamente, in primo luogo, una rete fisica.

Rete socialeé, ad esempio, una comunità di lavoratori, che si incontra nei relativi circoli dopolavoristici e che costituisce una delle associazioni di promozione sociale. Esempi di reti sociali sono inoltre le comunitá di sportivi, attivi o sostenitori di eventi, le comunitá unite da problematiche strettamente lavorative e di tutela sindacale del diritto nel lavoro, le confraternite e in generale le comunitá basate sulla pratica comune di una religione e il ritrovo in chiese, templi, moschee, sinagoghe e altri luoghi di culto.

Una rete sociale si puó inoltre basare su di un comune approccio educativo come nello scautismo, o nel pionierismo, di visione sociale, come nelle reti

segrete della carboneria e della massoneria.

Capitolo 3

Impostazione del problema di ricerca

"Bud: Apri!
Cattivo: Perch, altrimenti vi arrabbiate?
Bud e Terence: Siamo gi arrabbiati!"

Altrimenti ci arrabbiamo

In questa sezione si deve descrivere l'obiettivo della ricerca, le problematiche affrontate ed eventuali definizioni preliminari nel caso la tesi sia di carattere teorico.

Capitolo 4

Progetto logico della soluzione del problema

"Bud: No, calma, calma, stiamo calmi, noi siamo su un'isola deserta, e per il momento non t'ammazzo perché mi potresti servire come cibo ..."

Chi trova un amico trova un tesoro

In questa sezione si spiega come è stato affrontato il problema concettualmente, la soluzione logica che ne è seguita senza la documentazione.

Capitolo 5

Architettura del sistema

"Terence: Ma scusa di che ti preoccupi, i piedipiatti hanno altro a cui pensare, in questo momento stanno cercando due cadaveri scomparsi
Bud: Se non spegni quella sirena uno di quei due cadaveri scomparsi lo trovano di sicuro!"

Nati con la camicia

Si mostra il progetto dell'architettura del sistema con i vari moduli.

Capitolo 6

Realizzazioni sperimentali e valutazione

"Bambino: Questo l'ultimo avviso per voi e i vostri rubagalline
Il pistolero si alza: Che avete detto?
Bambino: RUBAGALLINE
Il pistolero si risiede: Aaah."

Lo chiamavano Trinità ...

Si mostra il progetto dal punto di vista sperimentale, le cose materialmente realizzate. In questa sezione si mostrano le attività sperimentali svolte, si illustra il funzionamento del sistema (a grandi linee) e si spiegano i risultati ottenuti con la loro valutazione critica. Bisogna introdurre dati sulla complessità degli algoritmi e valutare l'efficienza del sistema.

Capitolo 7

Direzioni future di ricerca e conclusioni

"Terence: Mi fai un gelato anche a me? Lo vorrei di pistacchio.
Bud: Non ce l'ho il pistacchio. C'ho la vaniglia, cioccolato, fragola, limone e caffé.
Terence: Ah bene. Allora fammi un cono di vaniglia e di pistacchio.
Bud: No, non ce l'ho il pistacchio. C'ho la vaniglia, cioccolato, fragola, limone e caffé.
Terence: Ah, va bene. Allora vediamo un po', fammelo al cioccolato, tutto coperto di pistacchio.
Bud: Ehi, macch sei sordo? Ti ho detto che il pistacchio non ce l'ho!
Terence: Ok ok, non c' bisogno che t'arrabbi, no? Insomma, di che ce l'hai?
Bud: Ce l'ho di vaniglia, cioccolato, fragola, limone e caffé!
Terence: Ah, ho capito. Allora fammene uno misto: mettici la fragola, il cioccolato, la vaniglia, il limone e il caffé. Charlie, mi raccomando il pistacchio, eh."

Pari e dispari

Si mostrano le prospettive future di ricerca nell'area dove si è svolto il lavo-

ro. Talvolta questa sezione può essere l'ultima sottosezione della precedente. Nelle conclusioni si deve richiamare l'area, lo scopo della tesi, cosa è stato fatto,come si valuta quello che si è fatto e si enfatizzano le prospettive future per mostrare come andare avanti nell'area di studio.

Appendice A

Documentazione del progetto logico

Documentazione del progetto logico dove si documenta il progetto logico del sistema e se è il caso si mostra la progettazione in grande del SW e dell'HW. Quest'appendice mostra l'architettura logica implementativa (nella Sezione 4 c'era la descrizione, qui ci vanno gli schemi a blocchi e i diagrammi).

Appendice B

Documentazione della programmazione

Documentazione della programmazione in piccolo dove si mostra la struttura ed eventualmente l'albero di Jackson.

Appendice C

Listato

Il listato (o solo parti rilevanti di questo, se risulta particolarmente esteso) con l'autodocumentazione relativa.

Appendice D

Il manuale utente

Manuale utente per l'utilizzo del sistema

Appendice E

Esempio di impiego

Un esempio di impiego del sistema realizzato.

Appendice F

Datasheet

Eventuali Datasheet di riferimento.