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'attivitàï¿½ 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 dovràï¿½ essere preceduto da Ing. oppure Dott., a seconda dei casi.

Ringraziamenti

Ringrazio .....

Capitolo 1

Introduzione

La polarizzazione é un utilissimo strumento per lo studio e l'analisi delle rete sociali sulle opinioni all'interno di differenti aree di ricerca. Generalmente la polarizzazione puó essere tranquillamente applicata all'interno di contesti politici, sociali e culturali permettendo di comprendere al meglio quali siano le vere opinioni 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 essere scaturite da disuguaglianze sociali ed economiche.

La polarizzazione puó comportate 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.

Il problema che si puó facilmente evincere risulta essere la formazione di due comunitá isolate che non comunicano tra loro, ció 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 opposto. Otteniamo un problema che genera 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 problamatiche all'interno dei moderni Social Network come Facebook e Twitter e molti altri. Questo perché sempre piú si stanno facendo largo nella vita di tutti i giorni e le problematiche relativi a contesti sociali, culturali e politici vengono sempre piú affrontati all'interno di queste piattaforme, in cui gli utenti si sentono sempre piú liberi di poter espimere le proprie opinioni. Il problema é che non é sempre possibile uscire dalle Filter Bubble perché gli stessi social network tendono a indirizzare l'utente a visualizzare informazioni che potrebbero interessarli senza fargli confrontare con opinioni divergenti. Alla luce di questo grande problema il calcolo di una polarizzazione puó consetire agli amministratori dei social network di individuare i topic piú polarizzati e consentire una diffusione democratica delle informazioni.

L'obiettivo della mia tesi consiste nell'utilizzare la polarizzazione per poter individuare quegli argomenti fortemente polarizzati e comprendere come tali informazioni vengono prodotte all'interno della rete sociale. Lo sviluppo di questo strumento é stato effettuato sfruttando 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 sfruttando la diffusione delle informazioni, 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.(Per una piú dettagliata spiegazione si rimanda al Capitolo??)

- *Reducing Controversy by Connecting Opposing Views*: Identifica la polarizzazione sfruttando la struttura del grafo. Il grafo viene generato utilizzando la medesima tecnica precedentemente illustrata, inoltre anche questo algoritmo é stato studiato sul social netwrok *Twitter*. La differenza principale é che non vengono catalogati i nodi in due gruppi in base al loro comportamento nel grafo. Adotta la tecninca dei *Random Walk* sfruttando la probabilitá di retweet per ottenere il valore della polarizzazione per ogni nodo appartenente all' *endorsment graph*. Per la spiegazione relativa al calcolo del valore assoluto della polarizzazione attraverso la tecnina dei *Random Walk* si rimanda al capitolo ??.

Prima di poter effettuare il calcolo vero e proprio della polarizzazione occorre effettuare una prima scrematura, 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'altra, 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**. Per concludere é stato effettuato anche uno studio per poter consentire di predirre il valore della polarizzazione in un periodo futuro, in questo modo gli amministratore dei social network possono effettuare

eventuali 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 particoliritá é stato possibile predirre il valore della polarizzazione nell'istante temporale successivo, 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 utilizzato. Per lo sviluppo della mia tesi ho deciso di analizzare due argomenti che presentano due contesti differenti della polarizzazione:

- **Elezioni Regionali in Sicilia nel 2017**: permettendo di analizzare la polarizzazione in un contesto politico.
- **Biotestamento**: permettendo di analizzare la polarizzazione in un constesto sociale.

I dati relativi a questi due topic sono stati raccolti sul social network *Twitter*, in un periodo temporale che andava dal 01/09/2017 al 20/12/2017, sfruttando diversi *hashtags* nel dettaglio per ogni topic sono stati scelti i 5 hashtag piú utilizzati dagli utenti per esprimere la loro opinioni su questi differenti argomenti. Una volta collezionati diversi dati é stato fatto quanto precedentemente illustrato. 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*.

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 per quanto riguarda questo topic sono stati scartati i dati relativi ai candidati politici degli altri partiti politici e coalizione, 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, infatti in un primo istante c'é una totale polarizzazione verso il *Movimento 5 stelle* per poi terminare con una polarizzazione tendente verso *Forza Italia*.

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 una valutazione positiva o negativa riguardo questa legge, infatti si é riscontrata una fortissima polarizzazione verso i contrari all'attuazione di tale legge. Ció puó essere facilmente addidato al contesto religioso e sociale presente in Italia, ovvero la forte presenza di una societá cattolica che é contraria all'attuazione del biotestamento. I risultati hanno confermato il contesto sociale cioé che la presenza relegiosa ha letteralmente dominato

anche all'interno di *Twitter* favorendo la creazione di un *Echo-chambers* che non permette agli utenti di poter visualizzare opinioni discordanti rispetto alla loro opinioni.

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 le comunitá e consentire una possibile risoluzione degli *Echo-chambers*. Eventuali sviluppo futuri relativi a questa tesi possono essere la possibilitá di eliminare la creazione degli *Echo-chambers* utilizzando algoritmi per eliminare la controversia delle informazioni all'interno dei social network favorendo una facile diffusione delle informazioni.

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 comunitÃ isolate che condividono un unico punto di vista di un problema. La polarizzazione Ã¨ 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 comunitÃ 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: Perchï¿½ $\frac{1}{2}$, altrimenti vi arrabbiate?
Bud e Terence: Siamo giï¿½ $\frac{1}{2}$ 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 ï¿½ 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, macchï¿½ sei sordo? Ti ho detto che il pistacchio non ce l'ho!
Terence: Ok ok, non c'ï¿½ 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.