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

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.

Indice

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 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. Una generica definizione della polarizzazione è la seguente:

> *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.

- *Polarizzazione Politica*: divergenza di opinione su estremi ideologici.

- *Polarizzazione Sociale*: differenza di opinione all'interno delle società che possono nascere da disuguaglianze sociali ed economiche.

La polarizzazione può 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 è 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 processo sociale per cui gli utenti che vi partecipano 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 comunità isolate che non comunicano tra loro, comporta un problema di isolamento delle opinioni cioè un utente che appartiene a quel gruppo difficilmente potrà ricevere informazioni o aderire alle idee del gruppo antagonista. Otteniamo la formazione degli *Echo-Chambers*, definita 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 può formare una forte polarizzazione delle opinioni e delle informazioni sono i *Filter Bubble* ovvero:

> *Uno stato di un isolamento intellettuale che può 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 vorrà vedere l'utente stesso. L'effetto è di isolare l'utente da informazioni che sono in contrasto con il suo punto di vista, isolandolo nella sua bolla culturale o ideologica.*

Come precedentemente anticipato la polarizzazione è uno strumento che può 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 più liberi di poter esprimere le proprie opinioni all'interno di queste piattaforme riguardo problematiche sociali, culturali, politiche ed economiche. Non è 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 può consentire agli amministratori dei social network di individuare i topic più 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 diffuse all'interno della rete sociale. Lo sviluppo di questo strumento è 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 utente che pubblica nella rete un *Tweet* esprimendo la propria opinione, formando un nuovo 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 più 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.

- ***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, così come il social network di riferimento *Twitter*. La differenza principale con la soluzione proposta in precedenza, consiste nell'utilizzare la tecnica dei *Random Walk*. La polarizzazione adottando questo approccio dipende fortemente dalla topologia dell'*endorsement graph*.

Prima di poter effettuare il calcolo vero e proprio della polarizzazione occorre effettuare una prima scrematura,da intendersi 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 semplicità 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'affinità 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 loro vicinanza 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 è 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 polarizzazione, a livello matematico la polarizzazione può 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 è stato effettuato anche uno studio per poter consentire la predizione del valore della polarizzazione in un periodo futuro. In questo modo gli amministratori 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 predirne il valore nell'istante temporale successivo. Sfruttando queste particolarità è stato possibile effettuare una predizione, nel dettaglio le tecniche utilizzate sono tre:

- *Double exponential smoothing*
- *Linear regression*
- *Moving average*

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, attraverso questi due Topic:

- **Elezioni Regionali in Sicilia nel 2017**: analisi in un contesto politico.
- **Biotestamento**: analisi in un contesto 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* più 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 riguarda le elezioni regionali in Sicilia si è deciso 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).

Innanzitutto è 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 più 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, cioè il cambiamento nel tempo della polarizzazione, seguendo il trend riscontrato durante i sondaggi effettuati mensilmente. Nel dettaglio si può 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 capovolgimento 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 è 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, riscontrando 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 cioè quanto un retaggio culturale o religioso possa influenzare il pensiero e le opinioni di una persona. Questa considerazione non è applicabile soltanto nella vita di tutti i giorni ma anche all'interno di un social network e ciò 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 cioè che la polarizzazione è 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.

1.1 Stuttura Tesi

All'interno di questa tesi, verranno illustrati e sviluppati:

- dettagli implementativi
- dettagli teorici
- dettagli sperimentali
- sviluppi futuri

definiti all'interno dell'introduzione.

La parte teorica relativa a tutti gli argomenti precedentemente illustrati verranno trattati all'interno del capitolo 3 *Progetto logico della soluzione del problema*. Nel dettaglio la parte implementativa dei due algoritmi, della raccolta dati, della sentiment analysis e della predizione verrà trattata all'interno del capitolo 4 *Implementazione*. Gli esperimenti e le motivazioni che hanno mosso alla definizione dei due topic scelti per l'analisi della polarizzazione verranno trattate all'interno del capitolo 5 *Realizzazioni sperimentali e valutazione*. Infine gli sviluppi futuri verranno trattati all'interno del capitolo 6 *Direzioni future di ricerca e conclusioni*.

Capitolo 2

Stato dell'arte e Background

2.1 Stato dell'arte

All'interno delle reti sociali sta sempre piú prendendo piede il problema della polarizzazione delle opinioni. Nel linguaggio comune il confronto tra individui ha sempre generato una forte controversia oppure un punto d'incontro. I social network hanno permesso all'utente di poter diffondere attraverso post, messaggi o espressioni audio video le proprie opinioni e/o pensieri all'interno di una comunità sociale. A tal proposito per favorire la diffusione delle diverse correnti di pensiero i social network stanno sempre più sviluppando algoritmi per permettere di identificare le comunità isolate che condividono un unico punto di vista di un argomento. La polarizzazione è un algoritmo matematico che applicato all'interno delle rete sociali consente 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 verrà illustrata una definizione di rete sociale.

2.1.1 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, può 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 modalità di risoluzione di problemi e i sistemi di gestione delle organizzazioni, nonché le possibilità dei singoli individui di raggiungere i propri obiettivi.

2.1.2 Le reti

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

Rete sociale è, 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 comunità di sportivi, attivi o sostenitori di eventi, le comunità unite da problematiche strettamente lavorative e di tutela sindacale del diritto nel lavoro, le confraternite e in

generale le comunità basate sulla pratica comune di una religione e il ritrovo in chiese, templi, moschee, sinagoghe e altri luoghi di culto.

La rete sociale ha una sua versione all'interno di **Internet** definita anche come *Social media*, questa è una delle forme più evolute di comunicazione in rete. La rete delle relazioni sociali che ciascuno di noi tesse ogni giorno, in maniera più o meno casuale, nei vari ambiti della nostra vita, si può così "materializzare", organizzare in una "mappa" consultabile, e arricchire di nuovi contatti. I principali social network sono: *Facebook, MySpace, Instagram, Twitter, Google+, LinkedIn, Ask.fm, Pinterest, Formspring, Bebo, Friendster, Hi5, Ning, Tagged, Meetup, Tumblr.*

2.1.3 Twitter

Twitter è il social network di riferimento utilizzato per la realizzazione di questa tesi di laurea, ne diamo una breve descrizione, è una rete sociale, creata il 21 marzo 2006 dalla *Obvious Corporation* di San Francisco, che fornisce agli utenti, attraverso l'omonima piattaforma, una pagina personale aggiornabile tramite messaggi di testo con lunghezza massima di 140 caratteri; nel 2017 l'azienda ha aumentato la lunghezza dei tweet a 280 caratteri per alcuni paesi. Gli aggiornamenti di stato possono essere effettuati tramite il sito stesso, via SMS, con programmi di messaggistica istantanea, posta elettronica, oppure tramite varie applicazioni basate sulle API di Twitter.

Il nome "Twitter" deriva dal verbo inglese to tweet che significa "cinguettare". Tweet è anche il termine tecnico degli aggiornamenti del servizio. I Tweet che contengono esattamente 140 caratteri vengono chiamati *Twoosh*. Gli aggiornamenti sono mostrati nella pagina di profilo dell'utente e comunicati agli utenti che si sono registrati per riceverli. È anche possibile limitare la visibilità dei propri messaggi oppure renderli visibili a chiunque.

Figura 2.1: Logo di Twitter

Questo social network è molto utilizzato da personaggi famosi, sportivi, ma anche capi di stato, esponenti politici ed economici, perfino il Papa, consentendo a tutti gli utenti iscritti alla piattaforma di poter visualizzare le idee ed opinioni espresse da questi personaggi. Proprio per questo motivo *Twitter* consente una diffusione su larga scala di idee ed opinioni da parte di moltissimi utenti, gli stessi giornalisti utilizzano questa piattaforma per diffondere *scoop* o notizie di vario genere, proprio per la grande visibilità che fornisce questo social network. Vengono generati moltissimi dati ed è proprio per questo motivo che ho deciso di utilizzarlo come strumento per la raccolta dati e cercare di calcolare la polarizzazione di alcuni topic selezionati, permettendomi di visualizzare come le informazioni espresse si propagano all'interno della rete e come queste influenzano il pensiero degli utenti.

Tweet

L'elemento alla base di questo social network è il *Tweet*, che non è altro che un messaggio contenente le opinioni, le idee, i pensieri dell'utente. Tutte queste cose possono essere arricchite utilizzato elementi multimediali come video, immagini oppure link a pagine web. Il tweet una volta pubblicato potrà essere accessibile per tutti gli utenti delle rete che hanno una relazione di "amicizia" con l'utente in questione, potendo anche effettuare delle operazioni aggiuntive quale commentare la notizia, apprezzarla oppure *retwettarla* cioè condividere

queste informazioni a tutti i propri collegamenti. I Tweet possono essere anche etichettati utilizzando due strumenti messi a disposizione da questo social network:

- *Mentions*: sono dei riferimento ad utenti, cioè come assegnare un mittente ad un email, affermando a tutti i membri all'interno della rete, che quel tweet è destinato esclusivamente a quella persona. Può essere utile per condividere più informazioni con gli amici delle rete più stretti, oppure per formulare degli attacchi, riflessioni o considerazioni anche con utenti famosi. Per fare questa operazione è necessario inserire la ed il nome dell'utente in questione per poter collegare al tweet in questione alla persona desiderata.

- *Hashtag*: parole o combinazioni di parole concatenate precedute dal simbolo cancelletto (#). Etichettando un messaggio con un hashtag si crea un collegamento ipertestuale a tutti i messaggi recenti che citano lo stesso hashtag. Molto vengono utilizzati durante *Show Tv*, *comizi*, *pubblicità*, *eventi politici* per poter consentire alle persone che vogliano esprimere la loro opinioni sugli argomenti espressi durante questi eventi.

2.1.4 Polarizzazione

Definiamo in maniera concettuale il significato, le funzionalità della polarizzazione. Citando il paper *A Measure of Polarization on Social Media NetworksBased on Community Boundaries* di *Pedro H. Calais Guerra, Wagner Meira Jr. Claire Cardie, Robert Kleinberg* la polarizzazione viene definita come un processo sociale in cui un gruppo viene diviso: in due sottogruppi aventi un conflitto o una visione differente del problema in questione, e da alcuni individui che rimangono neutrali. Comprendere la polarizzazione e

quantificarla è una sfida a lungo termine per i ricercatori all'interno di diverse aree, proprio perché risulta essere un potente strumento per l'analisi delle opinioni.

Nel paper *Measuring Political Polarization: Twitter shows the two sides of Venezuela* di *A. J. Morales, J. Borondo, J. C. Losada e R. M. Benito*, la polarizzazione assume una connotazione molto simile, vista in chiave politica, infatti all'interno di questo paper viene effettuata tale analisi all'interno di un contesto politico. La definizione che viene data è un fenomeno sociale che avviene quando gli individui contrappongono le proprie credenze ed opinioni in una posizione conflittuale tra loro, mentre alcuni rimangono in una posizione neutrale. Per citare *Jonh Turner*[1] :"Come le molecole polarizzate, i membri di un gruppo allineano il proprio pensiero nella direzione in cui erano originalmente diretti".

Adottare questo strumento all'interno di un contesto politico fornisce una chiara visione della preferenza o meno per i candidati politici, verificando il trend del candidato. Dal punto di vista matematico possiamo definire la polarizzazione tra due gruppi come la distanza che intercorre tra due numeri. Definendo in questo modo quanto due idee siano diverse e contrastanti utilizzando la distanza tra due valori numerici di riferimento. Le problematiche evidenziate dalla polarizzazione vengono trattate all'interno del paper : *Reducing Controversy by Connecting Opposing Views* di *Kiran Garimella ,Gianmarco De Francisci Morales, Aristides Gionis e Michael Mathioudakis.* Quando una popolazione si divide in due gruppi con visioni opposte quello che spesso si può osservare è la creazione degli *echo-chambers* cioè: una si-

[1]Noto psicologo sociale britannico, con altri colleghi ha sviluppato la teoria dell'Autocategorizzazione, che afferma tra l'altro che il sé non è un aspetto fondamentale della cognizione, ma è il risultato di processi cognitivi ed interazioni tra la persona e il contesto sociale.

tuazione dove persone che la pensano allo stesso modo rafforzano le proprie convinzioni a vicenda, senza esporsi verso idee contrastanti. In conclusione la polarizzazione può aiutare ad individuare e a prevenire la formazioni degli *echo-chambers*.

Echo-Chambers

Brevemente illustriamo un gravissimo problema che abbiamo precedentemente riscontrato all'interno della polarizzazione. Gli *Echo-chambers* è una descrizione metaforica di una situazione in cui le informazioni, le idee, o credenze sono rinforzate dalla comunicazione e dalla ripetizione all'interno di un ambiente definito. Praticamente viene creato come un "*eco*" che riproduce lo stesso suono più e più volte all'interno di un ambiente. Nei social media il fenomeno degli echo chambers è un problema piuttosto comune e ciò è dovuto dalla forte presenza di comunità di utenti che condividono sempre le stesse informazioni con un forte tasso di pubblicazione, quindi un nuovo utente che appoggia una di quelle idee entra all'interno di quel circolo vizioso. Ovviamente gli amministratori di questi social network sviluppano algoritmi che possano prevenire la formazione di queste comunità. Le conseguenze di questo problema sono molto evidenti ovvero:

- *Impossibilità della crescita culturale*
- *Impossibilità della crescita sociale*
- *Rendere la comunicazione poco democratica*

La polarizzazione come precedentemente accennato è uno strumento che può aiutare la prevenzione per la formazione di queste comunità.

2.1.5 Sentiment Analysis

La Sentiment Analysis è lo strumento utilizzato per poter catalogare le informazioni degli utenti attraverso un'analisi del testo. Non è possibile citare la sentiment analysis senza menzionare l' "opinion mining" ovvero una branchia del machine learning che studia il comportamento delle opinioni nella rete. Il termine sentiment viene utilizzato in riferimento all'analisi automatica effettuata per valutare il testo ed esprimere un giudizio predittivo sul contenuto. Il compito base di questo processo è quello di classificare la polarità di un testo e quindi il suo contenuto e catalogarlo come *positivo*,*negativo* o *neutrale*. Gli approcci più comunemente individuati rispetto alla sentiment analysis possono suddividersi in tre macro-categorie:

- *rilevamento delle keyword*: questo metodo consente di classificare il testo tramite categorie emotive facilmente riconoscibili, individuate in base alla presenza di parole emotive non ambigue, come felice, triste, e annoiato.

- *affinità lessicale*: questo metodo non rileva solo le keyword emotive, ma assegna anche a parole arbitrarie "un'affinità" probabile a emozioni particolari. Rispetto alla prima metodologia vista, l'affinità lessicale consente di affinare la selezione e l'attribuzione della polarità.

- *metodi statistici*: questi metodi si basano su elementi di apprendimento automatico. Per misurare l'opinione nel contesto e trovare la caratteristica che è stata giudicata, sono usate le relazioni grammaticali delle parole utilizzate. Le relazioni di dipendenza grammaticale sono ottenute attraverso la scansione approfondita del testo. Il processo di apprendimento da parte della macchina (anche detto machine learning) non è immediato, devono infatti essere costruiti dei modelli che asso-

ciano a diverse tipologie di commenti una polarità e se necessario ai fini dell'analisi anche un topic.

Quando parliamo della sentiment analysis non possiamo non esporre una delle grandi limitazioni offerte da questo strumento, cioè l'incapacità della macchina di poter comprendere concetti emotivi complessi come l'ironia. Ad esempio un commento ad un ritardo nel volo:

"Il mio volo è in ritardo. Splendido!"

verranno interpretati e classificati dalla macchina come post dalla polarità positiva mentre invece dovrebbero essere assegnate delle negatività. Quindi in un contesto lessicale pieno di ironia l'attendibilità sarà inferiore rispetto ad un documento con informazioni oggettive. Per migliorare l'accuratezza della predizione è necessario l'intervento dell'uomo, aumentando il volume dei dati di riferimento per l'analisi, banalmente milioni di post, possono alleviare le preoccupazioni sull'affidabilità a livello granulare, ossia di un singolo post. In conclusione la sentiment analysis è uno strumento che può essere molto utile all'interno dei social media in quanto può consentire di catalogare le opinioni degli utenti attraverso l'analisi dei contenuto pubblicati all'interno della rete. Questa potenzialità è ciò che mi ha spinto nell'adottare tale tecnica all'interno della mia tesi, garantendo una classificazione dei diversi *Tweet* collezionati.

Forecasting

Il Forecasting è un processo per fare predizioni del futuro basandosi sui dati passati e presente mediante l'analisi delle tendenze. Questo viene utilizzato molto all'interno delle aziende per poter analizzare i guadagni ed i futuri ricavi. Spesso viene associato ai processi di budgeting e pianificazione, si serve di dati passati e presenti, analisi dei trend e informazioni esecutive per

prevedere la situazione futura di ogni indicatore. Il forecasting prevede tre diverse tecniche:

- *Qualitativa*
- *Analisi e proiezione delle serie temporali*
- *Modelli causali*

La previsione finanziaria non è una scienza esatta e l'incertezza ne è un tipico aspetto. Per limitare gli errori e migliorare l'esattezza delle previsioni è necessario utilizzare precisi dati storici e in tempo reale. Inoltre, la possibilità di combinare previsioni rolling, previsioni a lungo raggio, simulazione di scenari possibili e stress test supporta i risultati delle previsioni e rende le previsioni più agili e reattive ai cambiamenti economici e di business. Ai fini della mia tesi ho deciso di adottare uno studio basato sulle analisi e proiezioni delle serie temporali, perché permette di estrarre statistiche significative ed altre caratteristiche dei dati. Predice i valori futuri attraverso l'osservazione dei dati precedentemente raccolti.

Capitolo 3

Progetto logico della soluzione del problema

In questo capitolo verrà presentato il flusso logico della tesi con la soluzione proposta per la suddivisione dei gruppi partendo dai dati raccolti, il calcolo della polarizzazione ed infine la sua predizione nel tempo.

3.1 Raccolta Dati

La prima parte del flusso logico della mia tesi si basa sulla raccolta dei dati. Questa operazione è stata effettuata utilizzando il social network **Twitter**. Nel dettaglio sono stati raccolti tutti i tweet relativi a due topic, utilizzati per effettuare le analisi,le motivazioni della scelta verranno illustrate più avanti, cioè:

- **La elezioni regionali in Sicilia**
- **Biotestamento**

Come precedentemente illustrato la scelta di questi due topic è dovuta al fatto che sono in primis due argomenti molto recenti e di attualità all'interno del nostro paese, in secundis perché riferiti a due contesti differenti tra loro ovvero quello politico e quello sociale. Prima di effettuare la raccolta di tweet per ognuno dei due topic, è stato effettuato uno studio sugli hashtag, cioè la ricerca veniva effettuata per una serie di hashtag per cui sono stati catalogati i 5 topic più utilizzati dagli utenti per esprimere le loro opinioni sull'argomento in questione. Lo studio di questi hashtag è stato improntato ricercando quelli che non esprimessero un giudizio, bensì che aiutassero l'utente a connotare i loro pensieri sul topic avendo una connotazione generica. La ricerca è stata fatta in maniera del tutto equilibrata sopratutto per quanto concerne le elezioni regionali in Sicilia in quanto è facile cadere in preda in hashtag utilizzati dalle fazioni politiche per attirare gli elettori, ne sono un esempio:

- *#diventeràbellissima*: utilizzato dal centro destra come motto all'interno dei social media per pubblicizzare il proprio piano politico.

- *#impresentabili*: utilizzato dal Movimento 5 Stelle per denunciare i candidati degli altri partiti politici.

Per evitare quindi di raccogliere dati già fortemente polarizzati, si è deciso di adottare una strategia più neutrale cercando 5 hashtags generici che rendessero l'idea del topic in questione. Twitter ha una politica di protezione per i dati, che sono accessibile a qualsiasi utente che abbia effettuato l'abilitazione allo sviluppo attraverso le Api messe a disposizioni, impedendo di effettuare più di 100 richieste ogni 15 minuti al server, impedendo un uso improprio e maligno con i dati pubblicati dagli utenti. Per richieste basta considerare la raccolta di un singolo Tweet. Per ottimizzare i tempi di raccolta si è deciso di utilizzare un'istanza *EC2*, che eseguisse uno script *Python* per la raccolta dei

dati in questione, rimanendo attivo anche durante le ore notturne. I dati in questioni venivano salvati all'interno di file binari, in modo da ottimizzare lo spazio che avrebbero occupato sulla macchina. Il motivo che mi ha spinto ad effettuare tale operazione è dettata dai costi che ha l'istanza EC2 per poter mantenere i dati fisici al suo interno, perché i consumi economici non sono generati soltanto dall'utilizzo delle risorse fisiche della stessa, ma anche dalla quantità di dati presente al suo interno.

Figura 3.1: Servizi AWS

Figura 3.2: Python

3.2 Sentiment Analysis

La sentiment Analysis è stata utilizzata per catalogare i *Tweet* delle persone in riferimento ai due topic d'interesse. Utilizzando questo strumento l'operazione effettuata è stata quella di suddividere in 3 categorie il contenuto pubblicato nella rete. In precedenza sono state illustrate le numerose tecniche per poter utilizzare la sentiment analysis, per questo motivo la tecnica che più si avvicinasse alle mie esigenze è quella basata su metodi statistici, cioè vengono applicati principi del machine learning per poter identificare la vicinanza o meno del testo con il *training set* di riferimento. Come descritto in precedenza maggiore è la quantità dei dati che contengono il training set e maggiore ne risulterà la precisione nella predizione del contenuto desiderato. All'interno del flusso di esecuzione della mia tesi la sentiment analysis viene collocata all'interno della raccolta dati, cioè nel momento in cui viene col-

lezionato un tweet questo viene immediatamente analizzato e collocato nel gruppo di riferimento. Successivamente se il tweet in questione presenta dei *retweet*, ovvero una condivisione del contenuto con altri utenti, poiché c'è la possibilità di poter aggiungere commenti, allora verrà analizzato anche il testo aggiunto dal nuovo utente sul *tweet* in questione.

3.3 Endorsement Graph

L'Endorsement Graph è il grafo di riferimento su cui verranno effettuate le successive operazioni. Concettualmente è un grafo diretto basato sui tweet e sui retweet fatti dagli utenti della rete. Gli elementi costituenti del grafo sono:

- *Tweet*: sono i nodi che formano il grafo.
- *Retweet*: sono assimilabili ad archi e noti, cioè essendo delle estensioni alle pubblicazioni aggiuntive fatte dai nodi, allora all'interno del grafo verranno generati nuovi nodi, rappresentati gli utenti che hanno pubblicato tali informazioni, e nuovi archi che rappresentano il collegamento tra i due. Una menzione particolare su questi archi che sono orientati dal Retweet verso il Tweet.

L'endorsement graph viene popolato da tutti i dati raccolti attraverso la ricerca per hashtag, rendendo in questo modo il grafo molto popolato. Per dare una parvenza grafica della polarizzazione i nodi appartenenti ai due gruppi distinti sono stati colorati in due colori di riferimento:

- **Rossi**
- **Blu**

3.4 Polarizzazione

La polarizzazione viene calcolata subito dopo la creazione del grafo in questione. La realizzazione di questo strumento è stato fatta implementando gli algoritmi proposti all'interno di questi due paper:

- ***Measuring Political Polarization: Twitter shows the two sides of Venezuela***
- ***Reducing Controversy by Connecting Opposing Views***

Per facilitare la dichiarazione all'interno dell'elaborato chiameremo il primo algoritmo come *polarizzazione basata sul grado*, mentre il secondo lo chiameremo *polarizzazione basata sulla topologia*. La polarizzazione basata sul grado analizza il grafo effettuando due operazioni in più passi:

1. Separare tutti i nodi in due categorie:
    - **Elite**: sono quei nodi che pubblicano notizie, in questo caso coloro che postano dei Tweet sul topic esprimendo una loro opinione.
    - **Listener**: sono quei nodi che condividono le notizie altrui attraverso una operazione di retweet.
2. Calcolare la polarizzazione utilizzando il grado in ingresso del nodo, incentrando il calcolo sui nodi *Elite*. Per grado del nodo si intende il numero di archi in entrata verso il nodo di riferimento.
3. Ripetere il passo precedente fino a quando non si stabilizzano i valori della polarizzazione

Per quanto riguarda la polarizzazione basata sulla topologia, come suggerisce il nome scelto, dipende molto dalla forma che assume l'endorsement

graph durante la sua creazione, è soggetta quindi a variazioni dettate dai collegamenti che si vengono a creare tra i diversi nodi presenti. I passi effettuati per la realizzazione di questo algoritmo sono i seguenti:

1. Per ogni nodo verificare se è possibile raggiungere i nodi di grado massimo, appartenenti ai due gruppi distinti che chiameremo *Rossi* e *Blue*.

2. Effettuare dei *Random Walk* [1], utilizzando i pesi sugli archi per calcolare la polarizzazione, per raggiungere quei nodi di grado massimo appartenenti alle due categorie sopra citate.

Queste sono le operazioni effettuate dall'algoritmo basato sulla topologia. Una volta applicati i due algoritmi i risultati vengono salvati per poi essere rappresentati graficamente, mostrando il grafo con le colorazioni in modo da mostrare la polarizzazione a livello visivo. Questa parte è il focus della mia tesi in quanto utilizzando questi algoritmi possiamo visualizzare con i nostri occhi quanto queste informazioni siano polarizzate, quanto le persone (i nodi sono utenti della rete) hanno dibattuto sul quel topic. I risultati ed i dettagli implementativi verranno mostrati in seguito, come anticipazione posso affermare che i topic hanno mostrato attraverso la polarizzazione lo stesso trend del mondo reale.

[1] Si definisce Random Walk la formalizzazione dell'idea di prendere passi successivi in direzioni casuali. Matematicamente parlando, è il processo stocastico più semplice, il processo markoviano, la cui rappresentazione matematica più nota è costituita dal processo di Wiener.

3.5 Predizione

L'ultima fase della tesi consiste nella predizione della polarizzazione in un periodo successivo rispetto a quello considerato durante la fase di raccolta dati. Questa esigenza nasce per poter consentire una prevenzione sul problema della formazione degli *Echo-chambers*, questi molto spesso vengono a formarsi con argomenti fortemente polarizzati, quindi poter prevenire per tempo la formazione di queste comunità isolate può comportare un enorme vantaggio per gli amministratori dei social media. Per la realizzazione di questa parte della tesi si è deciso di adottare una predizione basata sull'analisi dei dati raccolti, nel dettaglio si è deciso di utilizzare algoritmi alla base del *Forecasting* basati sull'analisi delle serie temporali. Nel dettaglio ho effettuato un analisi basata su 3 tecniche differenti tra loro:

- *Double exponential smoothing*: è una tecnica che sfrutta la serie temporale dei dati raccolti, assegnando una crescita esponenziale sui dati nel tempo, tenendo conto anche del trend di crescita o di decrescita nel tempo. In questo modo si cerca di predire il contenuto nel primo instante successivo.

- *Linear regression*: questa tecnica formalizza e risolve il problema di una relazione funzionale tra variabili misurate sulla base di dati campionari estratti da un'ipotetica popolazione infinita. In statistica l'analisi della regressione è associata alla risoluzione del modello lineare.

- *Moving average*: è una tecnica che estende la media aritmetica dei risultati, andando a mediare i risultati all'interno di una finestra mobile, prelevando gli ultimi risultati ottenuti in precedenza mediandoli e definendo come valore futuro questo risultato.

I risultati della predizione verranno trattati in seguito così come i dettagli implementativi. La predizione all'interno di questo contesto può risultare utilissimo per l'analisi futura dei topic nel tempo, perchè ovviamente la formazione degli *Echo-chambers* può essere arginata facendo dei controlli e/o previsioni sul valore della polarizzazione nel futuro.

Capitolo 4

Implementazione

All'interno di questo capitolo verranno illustrate tutte le tecniche implementative per realizzare quanto è stato enunciato nel capitolo precedente. Nel dettaglio verranno spigate, argomentate le implementazioni effettuate su tutte le fasi della tesi, i riferimenti teorici che hanno permesso la realizzazione di questa tesi. Di seguito una breve descrizione grafica del flusso con le operazioni effettuate durante la redazione del lavoro di tesi.(vedi fig:4.1)

4.1 Raccolta dati

La raccolta dati è la fase iniziale della tesi di laurea. Come precedentemente argomentato il social network di riferimento utilizzato è *Twitter*.Prima di spiegare quanto fatto occorre citare le funzionalità della *Twitter Api*. Successivamente verranno illustrate le tecniche adottate per la raccolta dei *Tweet* del passato.

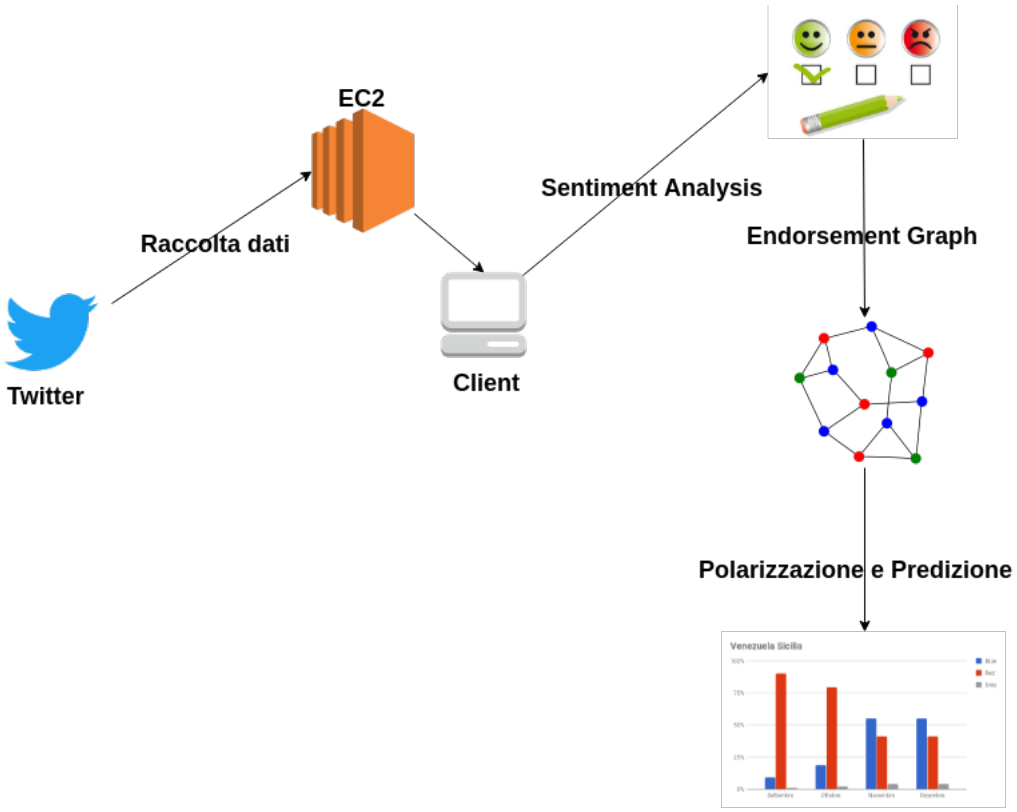

Figura 4.1: Flusso

4.1.1 Twitter Api

Sono delle api messe a disposizione dal social network Twitter, per poterle utilizzare occorre necessariamente effettuare l'iscrizione al reparto sviluppatori di Twitter, mettendo a disposizione il proprio account personale, quindi come prima cosa occorre predisporre di un proprio account. La raccolta dei dati può essere effettuata soltanto dopo aver abilitato l'account. Per raccogliere i dati pubblicati dagli utenti all'interno della rete occorre creare una *Twitter Apps*, la quale rilascerà delle credenziali che dovranno essere utilizzate per poter effettuare le richieste attraverso le Twitter Api al server. Viene inizializzato un processo di streaming tra il client ed il server che consenta la

ricezione dei dati, diamo una definizione delle credenziali, che sono divise in:

- *Token*: identificano il token di accesso ai servizi messi a disposizione da Twitter, a sua volta è composto da:
  - *Access token*
  - *Access secret*
- *Consumer*: è un codice di accesso ai servizi di consumo, ovvero consente lo streaming dei dati dal server, si suddivide in due chiavi:
  - *Consumer key*
  - *Consumer secret*

Queste chiavi di accesso possono essere generate più e più volte, una volta creata un'applicazione per lo sviluppo di Twitter. Le Api sono disponibili in diversi linguaggi di programmazione tra cui il *Python*, utilizzato per lo sviluppo della tesi. Per richiamare tutte le api attraverso il codice è necessario sempre autenticarsi, caricando le credenziali di accesso fornite attraverso i permessi da sviluppatore precedentemente illustrate. Lo streaming per la raccolta dei dati è soggetto a stringenti regole per evitare un uso improprio delle informazioni diffuse dagli utenti all'interno della rete. Infatti c'è un numero massimo di richieste che un utente, con le credenziali da sviluppatore, può effettuare, cioè 100 ogni 15 minuti. Allo scadere di tale tempo potrà ricominciare, altrimenti nel caso in cui un utente non tenesse conto di questa limitazione ed effettuasse una nuova richiesta durante il tempo di pausa le sue credenziali verrebbero bloccate per circa un'ora non potendo più interagire con le api di Twitter. La risposta da parte del server sarà un file *JSON* contenente tutti i metadati dell'utente in questione, come il testo del Tweet, il suo id, lo username dell'utente che ha effettuato il tweet, i mentions gli

hashtag ed il numero di retweet con la lista degli utenti che hanno retwettato la notizia in questione.

4.1.2 Get Old Tweet

Questa sezione illustrerà in che modo sono stati raccolti tutti i tweet del passato. Avendo spiegato nella sezione precedente le problematiche relative al numero di richieste da effettuare nell'arco temporale, è stata seguito un approccio differente che consentiva di limitare il numero di richieste alla volta. L'approccio in questione consiste nell'effettuare una *GET* sulla pagina di ricerca di Twitter, specificando l'argomento ricercato, ed il periodo di validità della ricerca. Twitter suddivide gli account abilitati allo sviluppo in 3 categorie:

- *Standard*: account gratuito soggetto a limitazione temporali e sui contenuti, non è possibile ricercare tweet precedenti a 7 giorni. I dati richiesti non sono completi.
- *Premium*: account a pagamento con sole limitazioni temporali, non è possibile ricercare tweet precedenti a 30 giorni. Non ha nessun problema di completezza sui contenuti.
- *Enterprise*: account a pagamento senza alcuna limitazione temporale e sui contenuti.

Avendo utilizzato un account *Standard* sono stati ricercati dei metodi per poter risolvere le problematiche precedentemente descritte, cercando di ottenere più informazioni possibili.

Tutte queste problematiche sono state risolte la libreria *Get Old Tweet*, il quale crea un indirizzo http con i parametri di ricerca richiesti, nel nostro caso:

- *La query*: la parola chiave, l'hashtag da ricercare all'interno dei Tweet.
- *Data inizio*: la data in cui inizio a raccogliere i dati.
- *Data Fine*: la data in termino la ricerca e la raccolta dei dati.

Una volta definiti questi parametri all'interno della *url*, viene restituita la pagina html contenente tutti i tweet presenti nel periodo indicato. Il risultato verrà convertito in un formato *json*, per estrarre le informazioni basterà parsare la pagina ottenendo i seguenti dati:

- Lo *username* della persona che ha postato il tweet.
- Il numero di *retweet* al tweet.
- Il *testo* del tweet pubblicato.
- La lista degli *hashtag* pubblicati dall'utente all'interno del tweet
- La lista dei *mentions* pubblicati dall'utente all'interno del tweet.
- La *data* di pubblicazione del tweet.

Per far comprendere meglio queste informazioni nell'immagine sovrastante viene presentato un tweet di esempio con le informazioni precedentemente elencate, per far comprendere al meglio gli elementi in questione.

Tutti questi dati sono stati successivamente elaborati ed analizzati. Particolare attenzione è stata posta a due di essi ovvero: il testo ed il numero di retweet. Il primo per effettuare la sentiment analysis e poter definire una prima partizione dei dati, che andranno a generare il grafo. Il secondo perché costituisce la base per i nuovi nodi del grafo e quindi la diffusione del tweet con altri utenti. Per identificare la lista degli utenti che hanno effettuato il retweet sul tweet in questione è stato necessario usare una libreria chiamata

Figura 4.2: esempio Tweet

Tweepy. Questa libreria sfrutta le *Twitter Api*, effettuando delle chiamate rest sul server di Twitter per ricevere le informazioni richieste. Nel nostro caso è stato utilizzato il metodo *retweet* che ricevendo come argomento l'id relativo al tweet pubblicato permette di ricevere la lista degli username che hanno pubblicato quell'argomento all'interno della propria rete. Il problemi presentati in precedente nell'utilizzare queste chiamate al server sono sempre validi, infatti se vengono superate le 100 richieste viene lanciata un'eccezione: *Tweepy Error* che consente di mettere in pausa l'applicazione attraverso una sleep per 15 minuti. Una volta terminato il tempo di attesa la richiesta viene ripresa dal punto richiesto, continuando a collezionare gli username richiesti. Tutti i dati raccolti durante le operazioni di streaming sono stati salvati all'interno di un file binario per poter ottimizzare lo spazio fisico.

EC2 La raccolta dati utilizzando le implementazioni precedentemente citate, è stata eseguita all'interno di una istanza **EC2**. Il motivo di tale scelta è dettata dai tempi di attesa via via sempre più lunghi per la raccolta degli username degli utenti che hanno retweettato i tweet raccolti, per via delle

politiche stringenti dettate da **Twitter**. Utilizzando un'istanza a pagamento, si è dovuto ottimizzare il salvataggio fisico dei dati, perché oltre ad essere un servizio a consumo di risorse di calcolo, vengono pagate anche i dati salvati all'interno dello storage fisico della macchina. Attraverso la libreria *pickle*[1] si è potuto risolvere questo problema, in quanto ottimizza il salvataggio dei dati attraverso una conversione in binario degli stessi.

4.2 Sentiment Analysis

La sentiment analysis è l'approccio utilizzato per la divisione dei diversi tweet, raccolti per un determinato topic, in un determinato periodo, analizzando il "sentimento" espresso nel contenuto del testo pubblicato dagli utenti. Questa fase viene effettuata al momento della raccolta dei dati, cioè una volta collezionati l'insieme dei dati vengono sottoposti alla funzione implementata e salvati in due categorie distinte mediante la libreria *pickle*, precedentemente illustrata.

L'implementazione di questo metodo è stata effettuata seguendo un approccio basato su *metodi statistici*: questi metodi si basano su elementi di apprendimento automatico. Per misurare l'opinione nel contesto e trovare la caratteristica che è stata giudicata, sono usate le relazioni grammaticali delle parole utilizzate. Le relazioni di dipendenza grammaticale sono ottenute attraverso la scansione approfondita del testo. Il processo di apprendimento da parte della macchina (anche detto machine learning) non è immediato, devono infatti essere costruiti dei modelli che associano a diverse tipologie di commenti una polarità e se necessario ai fini dell'analisi anche un topic.

Si può riassumere quanto sviluppato attraverso la seguente immagine:

[1] libreria per il salvataggio dei dati fornita da *Python*

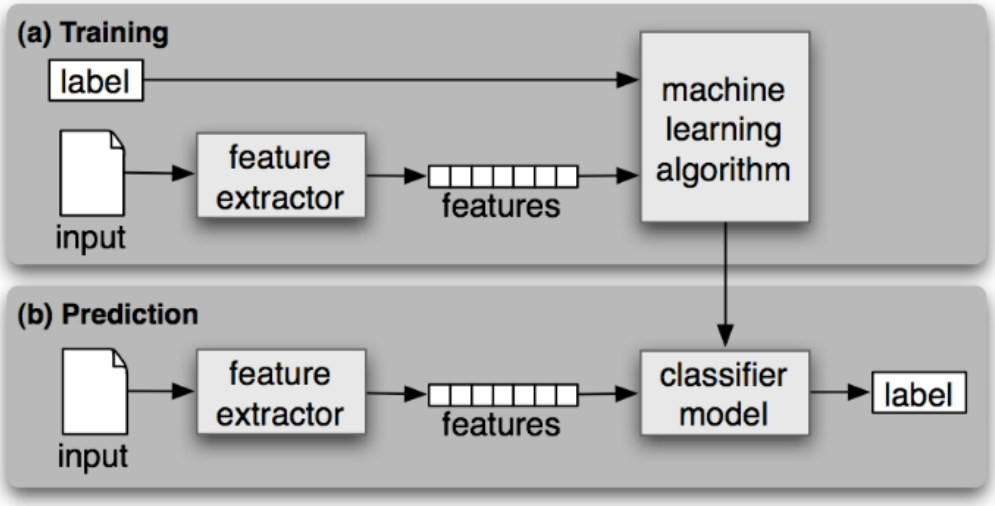

Figura 4.3: schema Sentiment Analysis

Da quanto si evince dall'immagine mostrata risulta evidente quanto un meccanismo di *machine learning* necessiti di un classificatore molto preciso. La precisione può essere ottenuta attraverso dei *training set*, ovvero dei file che istruiscano la macchina permettendo una classificazione ottimale dei dati. Ci sono diversi classificatori, quello utilizzato per l'implementazione della Sentiment Analysis è descritto di seguito.

- **classificatore Naive Bayes**: Un classificatore bayesiano è un classificatore basato sull'applicazione del teorema di Bayes. Richiede la conoscenza delle probabilità a priori e condizionali relative al problema, quantità che in generale non sono note ma sono tipicamente stimabili. Se è possibile ottenere delle stime affidabili delle probabilità coinvolte nel teorema, il classificatore bayesiano risulta generalmente affidabile e potenzialmente compatto. Per costruzione, il classificatore bayesiano minimizza il rischio di classificazione.

Nel gergo della classificazione di testi o *Text Categorization*, con il termine classificatore bayesiano ci si riferisce convenzionalmente al classificatore Naive Bayes, ossia un classificatore bayesiano semplificato con un modello di probabilità sottostante che assume l'indipendenza delle *feature*, ovvero che la presenza o l'assenza di una particolare *feature* in un documento testuale non è correlata alle altre.

L'esperienza dimostra che il metodo funziona in molti problemi pratici, come per esempio il filtraggio antispam adattivo. Un vantaggio del classificatore Naive Bayes è che richiede solo un training set di esigue dimensioni per stimare i parametri necessari per la classificazione.

Tornando allo sviluppo, la prima operazione fatta è stata quella di raccogliere dati per classificare i tweet secondo un determinato topic. Questa operazione è stata effettuata manualmente anche per permettere alla macchina di poter riconoscere l'ironia, cosa che è possibile solo per la mente umana.Per questo motivo la precisione dei classificatori non sarà mai assoluto, conseguentemente le macchine necessitano dell'intervento dell'uomo per risolvere tali problematiche. Il training set in questione è stato redatto attraverso un file csv (vedi tabella: 4.1) costruito con due colonne:

- *label*: ovvero l'identificatore che verrà utilizzato nell'analisi del classificatore per analizzare il contenuto di tweet ed associarlo ad un gruppo piuttosto che ad un altro.

- *Tweet*: il testo del tweet pubblicato, contenente anche caratteri speciali, elementi multimediali, link etc.

Una volta definito un training set occorre definire un *vettore di Feature*:

- Rappresenta l'elemento cardine di un classificatore, maggiore è la sua efficienza e maggiore sarà la sua precisione. Il vettore di feature è utiliz-

Training set	
Label	**Tweet**
5stelle	Chi ama la sua terra non può che votare il #M5S #Regionali #Sicilia pic.twitter.com/gv3CQWhCeI
ForzaItalia	Con elezioni @*matteosalvinimi* e @$Musumeci\_Staff$ regionali 2017 in #Sicilia !!#elezioniregionali2017 #andiamoagovernare #forzalega
Altri	Tutti gli schieramenti alle #elezioniregionali in Sicilia.L'articolo di Pierangelo Bonanno

Tabella 4.1: Esempio di training set

zato per costruire un modello che consenta, attraverso un training set, di poter fare una predizione sui dati che non ha mai analizzato prima. Dovendo analizzare dei dati basati su Twitter si è deciso di adottare degli schemi basati sull'assenza o meno di alcune parole che appaiono nei tweet come feature. Attraverso il training set, composto da tweet suddivisi in 3 gruppi (positivi, negativi e neutrali)[2], suddividiamo ciascun tweet in parole e aggiungeremo ogni parola la vettore di feature. Tale approccio viene definito *unigrams*.Prima di andare ad inserire tali parole all'interno di questo vettore, andremo ad effettuare delle opera-

[2]Questa classificazione può essere modificata in base alle esigenze dell'utente, modificando i label nel file csv del *training set*

zioni di filtraggio scartando quelle parole che non sono necessarie per comprendere il sentimento. Nel nostro caso le congiunzioni, articoli, preposizioni semplici ed articolate, ma anche altri caratteri perché generalmente i tweet contengono caratteri speciali come *#* , *@*, *link*, vengono inseriti all'interno della lista di parole di *Stop*, cioè una lista popolata da tutte quelle parole che potrebbero far saturare il vettore di feature e che non esprimono un sentimento. Il vettore di feature viene popolato a partire da tutte le parole che non sono presenti all'interno della lista di *Stop* e che appartengono a tutti i tweet del training set.

In conclusione una volta definito il classificatore, il vettore di Feature ed il training set non resta che illustrare il calcolo della sentiment Analysis. Per l'implementazione del classificatore *Naive Bayes* si è utilizzata la libreria *Python* **NLTK**[3], la quale una volta istruito il classificatore attraverso il training set, precedentemente popolato in base alle proprie esigenze (come nell'esempio 4.1) utilizza il vettore di feature per ricercare la vicinanza del tweet al training set, restituendo il label corrispettivo (inserito dall'utente nella definizione del csv sopra citato). In questo modo sarà possibile sottoporre ogni testo dei diversi tweet raccolti al sistema il quale restituirà il responso in breve tempo. Per garantire una maggiore accuratezza nelle predizioni sarà necessario arricchire il training set con molteplici tweet. All'interno della sezione dedicata agli esperimenti verranno illustrati alcuni esempi dei label e dei tweet utilizzati per il calcolo dell'analisi illustrata all'interno di questa sezione.

[3]Natural Language Toolkit.

4.3 Endorsement Graph

In questa sezione verranno illustrate le tecniche e gli strumenti utilizzati per la realizzazione dell'*endorsement graph*, cioè un grafo diretto che consenta la diffusione e la pubblicazione delle notizie. Un grafo è un insieme di elementi chiamati nodi connessi tra di loro attraverso degli archi, essendo il nostro un grafo diretto significa che gli archi in questione hanno dei versi di orientamento da un nodo ad un altro. Definiamo gli elementi costituenti:

- *Nodo*: per ogni nodo è stato associato un utente che abbia pubblicato una notizia sul topic analizzato.
- *Arco*: per ogni arco è stata associata una relazione tra il tweet pubblicato da un nodo ed il retweet effettuato da un altro utente. Questi archi sono possono essere pesati, cioè avere una probabilità che una qualsiasi coppia di nodi sia adiacente. Nel nostro caso tale valore viene definito dalla *probabilità di retweet*.

La creazione di questo grafo diretta dipende strettamente dai risultati ottenuti durante la raccolta dati. Nel dettaglio sono stati collezionati tutti i tweet che contenessero i 5 *hashtag* più utilizzati per l'espressione del suddetto topic, una volta raccolta tutti i dati in questione questi sono stati fusi ed utilizzati per la definizione del grafo. I dati sono stati salvati in tre gruppi differenti grazie ai risultati ottenuti durante la *Sentiment Analysis*,questa classificazione risulta necessaria per la colorazione ed il calcolo della polarizzazione sul grafo. Tornando alla creazione è stata applicata una limitazione e cioè che i nodi isolati non venissero considerati all'interno del grafo. La motivazione che ha spinto nell'adottare tale politica è dettata dall'esigenza di ricercare quanto un idea espressa attraverso un tweet venga diffusa in un grafo e quanto questa sia polarizzata, quindi un nodo isolato che non viene

ripubblicato dagli utenti risulta inutile ai fini dell'analisi in questione. In precedenza nella definizione di Arco si è definita la probabilità di retweet:

> Si definisce tale il rapporto che intercorre tra il numero dei retweet che l'utente ha effettuato su un nodo[4] con il numero di retweet totali effettuati dall'utente stesso.
>
> $$P(retweet)_{ij} = \frac{\#retweet_{ij}}{\#retweet_j} \tag{4.1}$$
>
> con i,j= nodi adiacenti appartenenti al grafo.

Avendo raccolto dati da *hashtag* differenti è possibile che alcuni utenti potessero retwettare lo stesso utente che avesse pubblicato nuovi tweet, per mantenere questa relazione si è incrementato il numero totale di retweet effettuati complessivamente così come il numero di volte che l'utente ha retwettato le opinioni di quello specifico utente.

Prima di popolare il grafo sono state effettuate tutte le precedenti operazioni. La realizzazione del grafo è stata ultimata attraverso la libreria **Networkx**[5], scritta in *Python*. Il grafo in questione può essere modificato graficamente, nel nostro caso per meglio far comprendere la polarizzazione si è deciso di adottare i colori: *Rosso* e *Blu*; per rappresentare le due partizioni raccolte attraverso la sentiment analysis, successivamente calcolarne la polarizzazione.(Vedi fig:4.4)

[4]Per nodo ci si riferisce ad un tweet pubblicato da un utente

[5]NetworkX è una libreria Python per la creazione, manipolazione e studio delle strutture, dinamiche, e delle funzioni di una rete complessa.

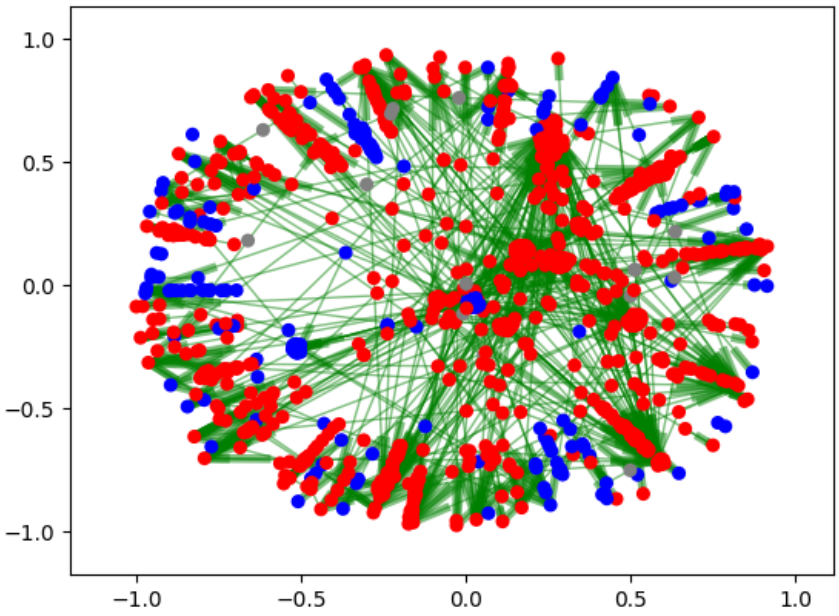

Figura 4.4: esempio di Endorsement Graph

4.4 Polarizzazione

La polarizzazione viene definita come un processo sociale in cui un gruppo viene diviso: in due sottogruppi aventi un conflitto o una visione differente del problema in questione, e da alcuni individui che rimangono neutrali. Dal punto di vista matematico la polarizzazione è una assegnazione di un valore compreso all'interno di un range:

$$[-1, 1] \tag{4.2}$$

I valori in questione identificano la vicinanza ad un gruppo piuttosto che ad un altro, in conclusione occorre assegnare alle due visioni contrastanti, che si vogliono analizzare, due valori di riferimento ovvero gli estremi. Di

seguito verranno illustrate le due implementazioni realizzate per il calcolo della polarizzazione all'interno di un endorsement graph.

4.4.1 Polarizzazione basata sul grado

All'interno di questa sezione verrà illustrata l'implementazione della polarizzazione attraverso l'algoritmo illustrato all'interno del paper: *Measuring Political Polarization: Twitter shows the two sides of Venezuela*.

Come precedentemente spiegato all'interno di questa tesi, il primo passo consiste nell'individuare due tipi di nodi:

- *Elite*: sono i nodi che hanno pubblicato una notizia all'interno del grafo, sul topic selezionato.
- *Listener*: sono i nodi che seguono le informazioni pubblicate dai nodi *Elite*.

Questa suddivisione è indipendente dai risultati ottenuti dalla sentiment analysis, cioè la prima partizione effettuata analizzando i contenuti, in questo modo si analizza la diffusione delle opinioni. Illustriamo ora i passi algoritmici implementati per la realizzazione di questa polarizzazione:

1. Per prima cosa individuiamo tutti i nodi *Elite* e quelli *Listener*, per poter assegnare il primo valore della polarizzazione.

$$-1 \leq X_s \leq 1 \tag{4.3}$$

   La formula precedentemente illustrata indica il range dei valori che la polarizzazione può assumere all'interno del grafo.

2. Il passo successivo, definisce le condizioni iniziali dell'algoritmo, consiste nell'assegnare ai nodi Elite e Listener rispettivamente:

$$\begin{cases} X_e = \pm 1 \\ X_l = 0 \end{cases}$$

Il valore dei noti elite dipende dall'appartenenza o meno ad i gruppi ottenuti attraverso la classificazione effettuata dalla sentiment analysis, per esempio Rossi = +1, Blu = -1.

3. I nodi elite propagheranno le loro opinioni verso i nodi listener, tale operazione verrà effettuata iterativamente fino al verificarsi di alcune condizioni, cioè diffonderà le proprie notizie ai propri vicini. Calcoliamo la polarizzazione dell'opinione per ogni listener appartenente al grafo:

La polarizzazione all'istante temporale t, di un dato listener i, è data dalla seguente espressione:

$$X_i(t) = \frac{\sum_j A_{ij} X_j(t-1)}{k_i^{in}} \tag{4.4}$$

Dove A_{ij} definisce gli elementi della matrice di adiacenza del grafo, il cui valore è pari a 1 se esiste un collegamento da j a i, e k_i^{in} corrisponde al proprio *indegree*[6]. Tale formula è stata modificata poiché è stata modificata la topologia del grafo, cioè all'interno del paper gli archi congiungeva i nodi elite verso i nodi listener, mentre nella tesi i collegamenti sono stati creati nel verso opposto. L'endorsement graph collega i nodi in base alla relazione di retweet, quindi si è deciso di creare gli archi dai retweet verso il tweet. Alla luce di questa considerazione

[6]Indica il numero di archi entrati nel nodo selezionato

la formula in questione viene modificata in questo modo:

$$X_i(t) = \frac{\sum_j A_{ij} X_j(t-1)}{k_i^{out}} \tag{4.5}$$

Dove k_i^{out} corrisponde all'*outdegree*[7] del nodo.

4. La formula illustrata nel passo precedente deve essere eseguita fino a quando avremo una stabilizzazione della polarizzazione.

Brevemente ora verrà illustrato un esempio dell'applicazione dell'algoritmo del grafo. Come prima cosa sono stati impostati i pesi ai nodi *elite*, nel nostro caso rappresenti dai nodi : A, B, C (vedi fig:4.5). Successivamente è stata applicata la formula precedentemente illustrata (vedi 4.5), quindi utilizzando i valori della polarizzazione nel passo iniziale dell'algoritmo insieme alla seguente matrice di adiacenza del grafo:

		Verso			
		A	B	C	D
Da	A	0	1	0	0
	B	0	0	0	0
	C	0	1	0	0
	D	1	1	1	0

Tabella 4.2: Matrice di adiacenza

Come possibile notare dall'esempio assistiamo ad un cambiamento di polarizzazione poiché mentre in un primo momento i nodi C e D che erano di un colore Blu (risultato ottenuto attraverso la sentiment analysis) possiamo notare come il valore cambi drasticamente in funzione del grado del nodo

[7]Indica il numero di archi uscenti dal nodo selezionato

e agli archi,conseguentemente cambia il colore del nodo stesso. In questo modo risulta evidente come un grafo cambi drasticamente la propria polarizzazione. L'algoritmo come si evince dall'esempio termina nel momento in cui l'algoritmo converge stabilizzando i valori della polarizzazione nei nodi (vedi fig:4.8).

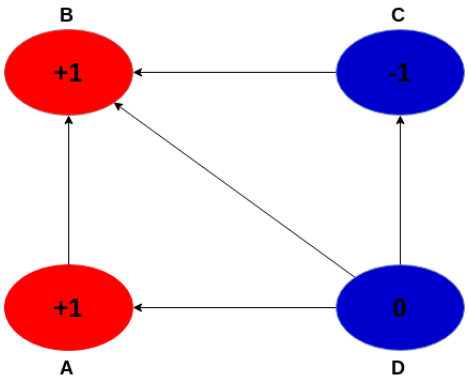

Figura 4.5: Passo 0

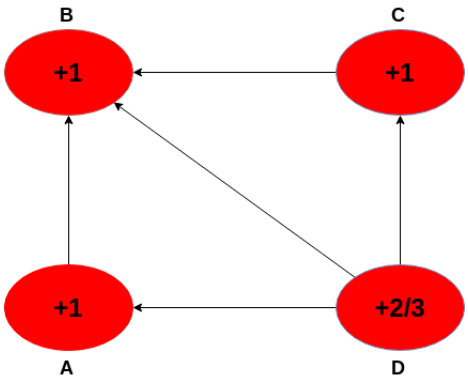

Figura 4.6: Passo 1

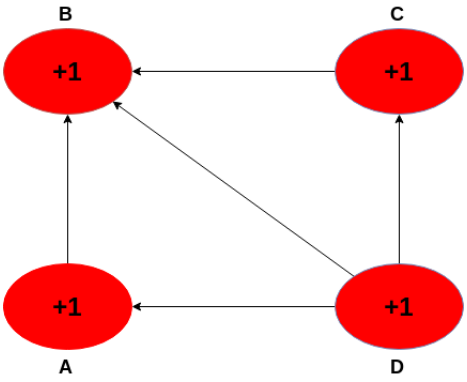

Figura 4.7: Passo 2

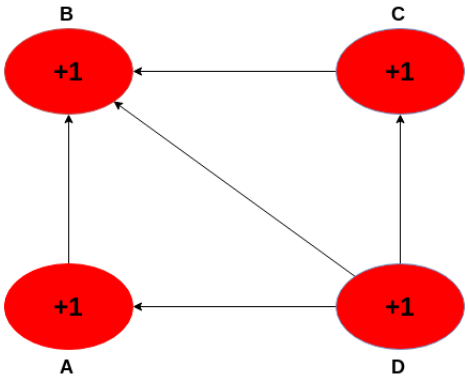

Figura 4.8: Passo 3

Figura 4.9: Esempio algoritmo basato sul grado

Dal punto di vista implementativo l'algoritmo trattato è stato implementato mediante le funzionalità contenute all'interno della libreria *NetworkX*, la quale garantisce ottime prestazione nel caricamento dei dati relativa alla matrice di adiacenza, senza far saturare la memoria volatile del sistema.

4.4.2 Polarizzazione basata sulla topologia

La polarizzazione basata sulla topologia del grafo è un algoritmo definito all'interno di questo paper:*Reducing Controversy by Connecting Opposing Views*. L'algoritmo in questione sfrutta la topologia del grafo per poter valutare la polarizzazione di ogni singolo nodo a fronte dell'opinione espressa all'interno del tweet pubblicato. Per prima cosa una volta costruito l'*endorsement graph* un ruolo fondamentale lo svolge la scelta delle due partizioni che compongono il grafo. Nel paper in questione le partizioni venivano effettuate attraverso una catalogazione degli hashtags, cioè degli elementi testuali che esprimessero in poche parole, o una loro concatenazione, un parere su una determinata notizia. Per fare un esempio se il topic dell'analisi fosse stata una partita di calcio ed un utente avesse pubblicato un tweet con un hashtag come *#ForzaBlu* questo sarebbe stato associato come un parere positivo verso la squadra blu. Però se nel testo ci fosse stata una frase negativa insieme all'hashtag precedente, sarebbe stato un errore perché non più a favore della squadra blu bensì per il suo avversario. Per risolvere questa problematica la suddivisione è stata effettuata attraverso la *Sentiment Analysis*(vedi sezione n:4.2, per ulteriori spiegazioni). Tornando alla definizione della polarizzazione anche in questo caso tale strumento può assumere un valore contenuto nel seguente range:

$$[-1, 1] \tag{4.6}$$

La polarizzazione applicando quanto illustrato in questo paper risulta essere:

> Il tempo atteso per un random walk per raggiungere un nodo di grado massimo appartenente ad un insieme X ed Y partendo

rispettivamente da un nodo u. Otteniamo così:

$$\begin{cases} \rho^X(u) \in [0,1] \\ \rho^Y(u) \in [0,1] \end{cases}$$

Rispettivamente i risultati dei Random Walk per raggiungere i nodi di grado massimo di X ed Y[8]. Infine definiamo la polarizzazione come:

$$P_u = \rho^X(u) - \rho^Y(u) \ \in [-1,1] \tag{4.7}$$

In caso di nodi che non hanno uscite, cioè quei nodi che non hanno alcun arco uscente che gli permetta di comunicare con altri nodi, assumono il comportamento di nodi di *dangling*, cioè quei nodi che non hanno possibilità di comunicazione e quindi per dar loro un valore della polarizzazione viene assegnato loro il valore massimo a seconda della partizione a cui sono associati. Un discorso analogo è valido anche per i nodi di grado massimo, questi al passo iniziale hanno una polarizzazione pari al valore massimo del loro insieme di appartenenza, però se hanno degli archi in uscita che li metta in contatto con nodi di opinione opposta allora tale valore verrà aggiornato utilizzato l'algoritmo precedentemente illustrato.

All'interno della tesi sono state affrontate diverse problematiche per la realizzazione di questo algoritmo, per prima cosa è stata inserita la *probabilità di retweet* per calcolare la polarizzazione, in questo modo è stata assegnata una forte relazione tra i nodi che retwettano notizie, quindi nel calcolo del random walk il numero di passi dipende dalla probabilità che relaziona i due nodi, rendendo il calcolo più verosimile alla realtà. Per completezza si definisce *probabilità di retweet*:

[8]Per X e Y sono da intendere come le due partizioni ottenute attraverso la sentiment analysis.

Si definisce tale il rapporto che intercorre tra il numero dei retweet che l'utente ha effettuato su un nodo[9] con il numero di retweet totali effettuati dall'utente stesso.

$$P(retweet)_{ij} = \frac{\#retweet_{ij}}{\#retweet_j} \tag{4.8}$$

con i,j= nodi adiacenti appartenenti al grafo.

Detto ciò illustriamo un esempio dell'applicazione di questo algoritmo, per completezza consideriamo lo stesso grafo utilizzato in precedenza (vedi fig:4.8):

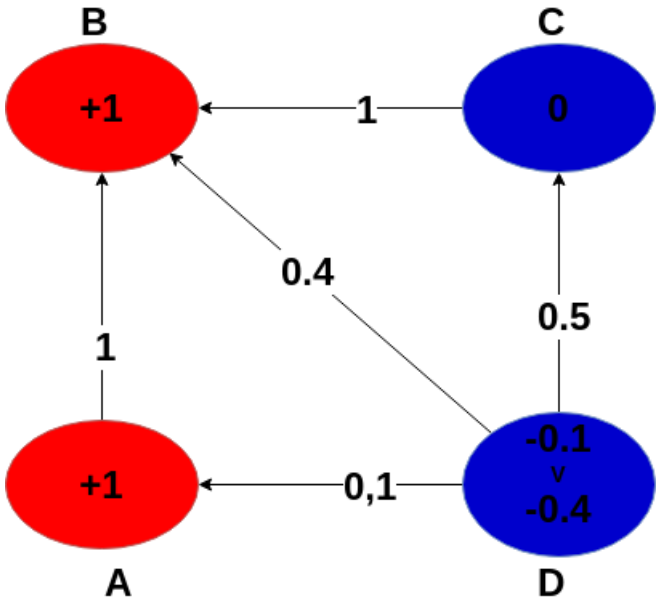

Figura 4.10: esempio di polarizzazione basata sulla topologia

Come possiamo notare i nodi di grado massimo della partizione *Rossa* e *Blu* sono rispettivamente i nodi: B e C, perché quelli aventi *indegree* massima per le rispettive partizioni. Le colorazioni sono il risultato della sentiment

[9]Per nodo ci si riferisce ad un tweet pubblicato da un utente

analysis. Analizzando i risultati presentati nell'immagine 4.10, possiamo notare come la probabilità di retweet giochi un ruolo cruciale nell'analisi in questione. Possiamo notare un comportamento particolare nel nodo B il quale assume un polarizzazione pari a zero. Essendo un nodo di grado massimo appartenente alla partizione Blu, ha pubblicato notizie a favore di quel gruppo, però ha anche condiviso notizie a favore del gruppo opposto modificando la sua natura originale. In conclusione applicando la formula 4.7, otteniamo un annullamento della sua polarizzazione Un altro comportamento particolare è osservabile nel nodo D, questi avrà due possibili valori a seconda del percorso scelto durante il Random Walk. Se consideriamo il percorso per raggiungere il nodo di grado massimo appartenente alla partizione Blu è facile notare che tale valore è pari a:

$$\rho^Y(D) = (D, C) = 0.5$$

Per quanto riguarda la partizione rossa il discorso non è semplice infatti:

$$\begin{cases} \rho^X(D) = (D, B) = 0,4 \\ \rho^X(D) = (D, A) \times (A, B) = 0,1 \times 1 = 0,1 \end{cases}$$

In conclusione la sua polarizzazione può assumere questi due valori:

$$\begin{cases} P_B = \rho^X(D) - \rho^Y(D) = 0,4 - 0,5 = -0.1 \\ P_B = \rho^X(D) - \rho^Y(D) = 0,1 - 0,5 = -0.4 \end{cases}$$

Per concludere dal punto di vista implementativo essendo molto oneroso dal punto di vista della memoria volatile, l'esecuzione di tutti questi random walk per ogni nodo, sono state ottimizzate le operazioni. Piuttosto che utilizzare la matrice di adiacenza si è utilizzato un metodo definito nella libreria *NetworkX* che restituisce i primi vicini dell'algoritmo. Dovendo mantenere in memoria una lista di nodi visitati durante l'esecuzione, al raggiungimento di un nodo

di grado massimo viene effettuata una operazione di *free* che consente al processo di deallocare le risorse occupate da tale lista.

4.5 Predizione

4.5.1 Double Exponential Smoothing

4.5.2 Linear Regression

4.5.3 Moving Averaqe

Capitolo 5

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 6

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.

Bibliografia

[1] Computer Emergency Response Team. Come funziona il ransomware WannaCry e cosa fare per proteggersi. https://www.certnazionale.it/news/2017/05/15/come-funziona-il-ransomware-wannacry-e-cosa-fare-per-proteggersi/, May 2017.