POLITECNICO DI MILANO

Corso di Laurea Specialistica in Ingegneria Informatica
Dipartimento di Elettronica e Informazione

Studio empirico delle dinamiche di diffusione dell'informazione in Twitter

Relatore: Prof.ssa Chiara Francalanci
Correlatore: Ing. Leonardo Bruni

Tesi di laurea Specialistica di:
Stefano POLVARA
Matricola 765681

Anno accademico 2011/2012

Sommario

Questo lavoro di tesi viene realizzato con lo scopo di definire una metodologia ed i relativi strumenti per effettuare un'analisi automatica riguardante le modalità e le cause preponderanti che permettono una diffusione dell'informazione online; nello specifico si vogliono studiare le dinamiche che contraddistinguono Twitter.

Uno degli obiettivi del lavoro di tesi è quello di definire e sviluppare uno strumento che riesca, in maniera del tutto automatica, ad associare ad ogni tweet i rispettivi retweet; operazione fondamentale per poi valutare, sempre attraverso lo strumento realizzato, le ragioni che permettono ad alcune informazioni di trovare una migliore propagazione.

Sempre più, nell'ultimo anno, sono nati nella rete siti web che asseriscono di poter misurare la capacità di influenza che hanno gli utenti dei social network; questo lavoro vuole osservare se è realmente vero che, per avere una buona diffusione dell'informazione, come alcuni studiosi affermano [30], è più importante la qualità delle relazioni che un determinato utente possiede, piuttosto che il contenuto intrinseco dell'informazione stessa.

Questo tipo di risultato risulta fondamentale per le aziende, le quali vogliono sapere se per convincere il proprio target di marketing, devono maggiormente puntare sulla pubblicazione di contenuti con determinate caratteristiche rispetto ad affidare le proprie campagne pubblicitarie ad utenti che risultano particolarmente influenti.

Abstract

This thesis work has been developed with the scope of defining a methodology and related tools necessary to carry out an automated analysis on the procedure and the prevailing causes that allow online information spreading; specifically, the analysis deep dives on the typical dynamics of Twitter.

One of the goal of this thesis work is to define and develop a tool that would allow, in a completely automated manner, to associate to each tweet its related retweets; this operation is necessary to evaluate, trough the instrument developed, the rationale behind the wider propagation of specific facts compared to others.

During the last year more and more websites claim to be able to measure the influence exercised by social network users; this work wants to prove if, as some authors allege [30], to obtain a capillary diffusion of a piece of information the relations that a certain user acknowledges are more important than the intrinsic content of the information itself.

The results of this experiment would be extremely useful for corporations willing to know whether, to attract their targeted customers, it is better to concentrate the marketing effort on the publication of certain contents or instead to assign the marketing campaign to users proven to be particularly influent.

Indice

Elenco delle figure

Elenco delle tabelle

TESI

Il ruolo dell'informazione su Twitter.

1.0 Introduzione

E' innegabile lo sviluppo che hanno avuto negli ultimi tempi i Social Network, basti pensare che attualmente Facebook ha in media 600mila iscritti al giorno, e il loro straordinario impatto sociale su una fetta così ampia della popolazione.

Se inizialmente la dinamica di costruzione delle reti sociali si era sviluppata attorno a tre grandi filoni tematici come l'ambito professionale, l'amicizia e le relazioni amorose, successivamente il panorama è cambiato profondamente.

L'avvento del web 2.0 ha radicalmente modificato il modo di comunicare sia da parte di chi invia si da parte di chi riceve il messaggio e questo è dovuto al fatto che questo tipo di uso del mezzo permette di poter veicolare una quantità maggiore di informazioni.

I Social Network rappresentano dunque un'evoluzione molto recente del web ed un fenomeno sociale molto diffuso che produce una grossa quantità di dati, per nulla strutturati, che, se opportunamente "ripuliti", risultano fondamentali per poter effettuare delle operazioni di *data mining* al fine di ottenere conoscenza.

Conoscenza che viene integrata dalle aziende con la propria business intelligence e, grazie alla quale quest'ultime riescono ad effettuare scelte strategiche mirate durante, per esempio, la fase di pianificazione delle attività future.

In questo lavoro di tesi, nello specifico, è stato effettuato uno studio empirico sulle modalità e sul comportamento con cui le informazioni si propagano sul servizio di microblogging Twitter.

Obiettivo finale è quello di osservare e capire se il contenuto presente nei vari tweet, come il sentiment e gli argomenti del messaggio stesso, possa in qualche modo influenzare la diffusione del post in rete, ovvero l'impatto che esso suscita sugli utenti.

In ultimo, si vuole inoltre osservare se le relazioni di amicizia, nel social network sopra citato, possano essere anch'esse causa di una diffusione maggiore di un determinato tweet.

Il lavoro di tesi è strutturato come segue:

- Il capitolo 2 espone una panoramica della letteratura riguardo l'argomento trattato, vengono introdotti i concetti base delle reti sociali e della *social network analysis*; viene data una definizione di rete sociale online trattando in particolar modo il social network Twitter ed il concetto di influence.
- Il capitolo 3 presenta una panoramica generale del progetto sviluppato, vengono definiti i dataset necessari, le tecniche utilizzate e le considerazioni assunte per poter effettuare le analisi, vengono definite le ipotesi che verranno poi dimostrate nel capitolo successivo ed infine si espongono i problemi incontrati dovuti ai limiti delle Twitter API.
- Il capitolo 4 espone tutti i risultati ottenuto con i vari test statistici effettuati sui dati sperimentali raccolti durante il lavoro, al fine di validare le ipotesi precedentemente effettuate.

  Si cercherà quindi di effettuare uno studio empirico sulla diversa propagazione che hanno i messaggi in base al sentiment, alla categorie ed ai diversi rapporti di amicizia tra gli utenti che li effettuano.
- Infine, il Capitolo 5 presenta una panoramica conclusiva del lavoro di tesi, analizzando i risultati raggiunti e descrivendo eventuali possibilità di ulteriori sviluppi futuri.

2.0 Stato dell'arte

2.1. Introduzione

Negli ultimi anni, l'utilizzo di Twitter come strumento di comunicazione personale ed aziendale ha avuto un forte incremento.

L'adozione da parte di alcuni VIP ha sicuramente aiutato ad attirare l'attenzione verso questo servizio; se poi aggiungiamo i mass media, i giornalisti, gli attori ed i politici, ecco che si raggiungono, finalmente, livelli di attenzione ed utilizzo che meritano l'avvio di una serie di analisi mirate.

E' del tutto lecito pensare, senza una accurata analisi, che per diffondere un messaggio, Twitter sarebbe poco utile o, in ogni caso, meno vantaggioso, rispetto ad altri social network, come Facebook. Almeno stando ai dati di Google Ad Planner (giugno 2012) [1], dato che i visitatori unici mensili di Twitter Italia sarebbero 2,8 milioni, un decimo dei 28 milioni di Facebook Italia. Ma come molte volte accade, un dato nudo e crudo esprime poco rispetto alla complessità di un fenomeno. In questo caso ci dice poco riguardo al potere di influenza e all'effetto moltiplicatore che l'utente Twitter può avere rispetto a quello di un utente Facebook; disuguaglianze che dipendono dalle caratteristiche peculiari dei diversi network e dal differente modo di viverli e di usarli.

L'era del 2.0 non annulla i meccanismi tipici della formazione delle opinioni nella società; non nega lo sviluppo di piccole "piramidi dell'influenza".

Al contrario, le sottoreti si sviluppano, si influenzano e interagiscono. Ritroviamo quindi anche nelle dinamiche della rete e dei social network le caratteristiche della formazione delle élite e dei gruppi di influenza studiate dai grandi sociologi del Novecento. Anche sui social network le élite godono infatti di una valutazione positiva a livello sociale, che permette loro di avere una posizione di

preminenza, per quanto settoriale e specifica. A questo va aggiunto il fatto, assolutamente non secondario, che un *Twitter influencer* in buona parte dei casi cura un blog o scrive per una importante testata giornalistica [2]; ha dunque a disposizione una vasta scelta di canali per la diffusione dei messaggi. E tramite quei canali, spesso e volentieri, stimola scientificamente il “brusio delle conversazioni”.

2.2. Le reti sociali

Il concetto di rete sociale, nato in sociometria intorno agli anni Trenta del Novecento per meglio descrivere e misurare le occorrenze delle relazioni sociali, si è mostrato utile nell'analisi sociale, soprattutto quando, a partire dagli anni Sessanta, si cominciò a notare l'inadeguatezza dell'apparato concettuale tradizionale nel definire certe nuove realtà sociali. Ecco allora che, in presenza di nuove situazioni in cui concetti come, per esempio, quello di gruppo tribale o familiare andavano perdendo importanza a causa della sempre crescente mobilità sociale o dei nuovi modelli relazionali urbani, il concetto di rete sociale si mostrava più adatto a definire di volta in volta:

- Le strutture oggettive dei rapporti interpersonali composte da persone con vincoli affettivi.
- Le strutture di persone legate da modalità di interazione caratteristiche.
- Le strutture di relazioni interpersonali che si trovano in condizioni di equilibrio.
- Le relazioni di interdipendenza tra persone che condividono valori, atteggiamenti e orientamenti ecc.. .

Questa prima spiegazione di rete sociale [3], grazie alla sua capacità di definire relazioni con persone provenienti da qualsiasi categoria strutturale (vicini di casa, parenti, colleghi ecc..) ed essendo più duttile ed aperta del concetto di gruppo, fornisce un'immagine più completa dell'ambiente sociale in cui è immerso un individuo.

Gli elementi costitutivi di una rete sociale sono:

- I soggetti ("**unità**" o attori) che appartengono alla rete;
- I **nodi** che compongono la rete. I nodi della rete possono essere costituiti da singoli individui, da gruppi, da istituzioni;
- Le **relazioni** che legano tra loro i soggetti della rete sociale, che possono essere monodirezionali, bidirezionali o multidirezionali. Le relazioni vengono solitamente rappresentate da linee e frecce.

Le reti sono quindi strutture relazionali tra attori ed in quanto tali costituiscono una forma sociale rilevante che definisce il contesto in cui si muovono quegli stessi attori. La rete sociale risulta essere allora la struttura di relazioni le cui caratteristiche possono essere usate per spiegare, almeno in parte, il comportamento delle persone che costituiscono la rete.

Con riferimento al contenuto della relazione è possibile cogliere ed individuare alcune particolari reti che, per il tipo di legami che le costituiscono, si caratterizzano per essere *reti di sostegno* (supporto sociale), *reti formali*, costituite dalle istituzioni sociali, *reti informali*, reti che non presentano una veste istituzionalmente definita, *reti primarie* costituite da relazioni "faccia a faccia" in virtù dei legami naturali che accomunano gli individui come i rapporti familiari, parentali, amicali, di vicinato, *reti secondarie*, costituite da relazioni di conoscenza indiretta e *reti personali* (reti ego-centrate).

Una caratteristica della rete sociale molto usata è la **densità***,* cioè la misura in cui gli individui nella rete sono collegati l'uno all'altro. Avremo quindi reti sociali a maglie larghe (*loose knit*) se tutti gli individui sono collegati a un individuo "focale" e non l'uno all'altro, e reti sociali a maglie strette (*close knit*) se tutti o la maggior parte degli individui sono in relazione reciproca tra loro. Una rete sociale può essere misurata anche in base alla durata, alla direzione*,* alla raggiungibilità (la misura in cui un individuo è accessibile ad altri individui della rete), al grado di connessione (il numero medio di relazioni che ogni individuo ha con gli altri della rete), ed infine ai clusters (i segmenti ad alta densità della rete).

2.2.1. Social Network Analysis

L'analisi delle reti sociali (***Social Network Analysis***) è una prospettiva teorica e metodologica che si occupa dello studio delle reti sociali.

Essa presenta due caratteri principali: in primo luogo veicola l'idea in base alla quale la società può essere considerata come un intreccio complesso di relazioni sociali variamente strutturate, ed è proprio questo "intreccio" nel suo complesso a costituire il focus centrale dell'analisi; ogni fenomeno sociale può, dunque, essere letto in termini relazionali e strutturali: la condizione è che la struttura del fenomeno possa essere espressa in termini di attori sociali e di interconnessioni di varia natura tra quegli stessi attori; in secondo luogo si tratta di una prospettiva fondata metodologicamente e tecnicamente.

Questa prospettiva nasce e si sviluppa dalla confluenza di due principali filoni di ricerca: il primo è rappresentato dalla scuola antropologica di Manchester che pone una attenzione preponderante rivolta alla processualità "in situazione". Il secondo filone di pensiero, l'analisi strutturale americana, si sviluppa a partire dagli anni '70 ad Harvard e si caratterizza per l'interesse prioritario rivolto alla forma delle reti più che al loro contenuto.

Secondo gli esponenti di quest'ultimo filone di ricerca, la forma delle relazioni sociali determina ampiamente i loro contenuti; il comportamento individuale è interpretato in termini di vincoli strutturali sulle azioni piuttosto che in termini di forze interne che agiscono a partire dall'attore (da cui la critica ad esso rivolta di eccesso di determinismo strutturale) e si sostanzia in un forte rigore matematico e in una elevata sofisticazione delle tecniche di analisi.

Attraverso i contributi della scuola di Harvard si consolida l'apparato tecnico della *network analysis*.

Il gruppo di Harvard elabora i concetti matematici dell'analisi strutturale; l'obiettivo è quello di modellizzare strutture sociali dotate di differenti proprietà, partendo dalla teoria matematica dei grafi e dall'utilizzo dell'algebra delle matrici.

Ciò di cui si fa portatrice la *social network analysis* è sicuramente una prospettiva teorica che accentua una particolare dimensione della realtà sociale,

quella della sua struttura reticolare, dell'insieme complesso di interdipendenze e interconnessioni cercando di comprendere le condizioni della reciproca chiamata in causa tra comportamenti sociali e tali sistemi di interdipendenze.

La *social network analysis*, sebbene non abbia ancora conseguito uno statuto epistemologico definito, costituisce una prospettiva teorica affidabile e coerente strettamente collegata con una metodologia di ricerca pertinente e distinta dalle metodologie di tipo convenzionale.

L'enfasi posta sulle relazioni differenzia l'approccio reticolare rispetto alla ricerca tradizionale che viceversa privilegia gli aspetti attributivi degli attori sociali coinvolti; questa enfasi sui legami e di conseguenza sulla struttura generata dalle molteplici relazioni nelle quali le unità di osservazione sono coinvolte, comporta l'impossibilità di trattare gli attori di un campione come osservazioni indipendenti, precludendo così l'uso di tecniche statistiche convenzionali di stima parametrica.

La caratteristica fondamentale che contraddistingue l'analisi di rete rispetto alle modalità di ricerca più tradizionali (*surveys*) è lo spostamento dell'obiettivo da spiegazioni atomistiche in termini di attributi di casi indipendenti, alla spiegazione dei fenomeni in termini di relazioni tra un sistema di attori interdipendenti.

L'unità di base nella *network analysis* non è il soggetto preso singolarmente (attributi degli individui) ma è costituita dal legame tra i soggetti, definito individuando la coppia di individui tra i quali si stabilisce la relazione (attributi di coppie di individui).

I dati relazionali hanno una natura intrinsecamente diversa dai dati attributo; questa diversità risiede non soltanto nella forma e nella modalità di costruzione, ma soprattutto nella loro natura e nel ruolo giocato all'interno dei modelli descrittivi ed esplicativi, per cui nella prospettiva di rete la struttura delle relazioni in cui gli attori sono inseriti è considerata responsabile del fatto che certi attributi acquistano significato sociale e contribuiscono a differenziare comportamenti, credenze ed atteggiamenti.

2.3. Le reti sociali online

La nascita dei Social Media ha accelerato profondamente il processo di coinvolgimento dell'utente in rete. Oggi, nell'era del 2.0, è possibile affermare che Internet è innanzitutto una realtà umana dove le relazioni e le interazioni tra gli utenti sono alla base di tutto il suo potenziale.

Le reti sociali online si sono sviluppate grazie a software e piattaforme che consentono l'interazione tra utenti, permettendo loro di scambiarsi varie tipologie di contenuti (testuali, audio e video).

Anche online le reti sociali sono costituite da un gruppo di persone unite da legami specifici.

Sebbene le comunità virtuali siano un fenomeno nuovo ed in continuo cambiamento, tale da non consentire una lettura legata troppo rigidamente alle definizioni classiche di comunità, è comunque possibile rinvenire in esse alcune delle connotazioni che la comunità classica assume come: il sentire comune, il senso di appartenenza al gruppo, le coesione e la vicinanza spirituale.

Al centro di ogni social network c'è infatti un "interesse" che accomuna chi vi prende parte; si parla spesso di Facebook, Linkedin, Twitter, MySpace, Badoo, ecc.. , ma queste non sono le reti sociali, ma gli strumenti che permettono alle persone di ampliare le proprie reti sociali online.

Ogni social network permette agli utenti iscritti di:

- Creare un profilo pubblico o semi-pubblico entro un sistema limitato: ogni utente è generalmente profilato con una serie di dati, alcuni definiti dall'utente a sua discrezione, altri organizzati automaticamente in base all'utilizzo della piattaforma. A seconda delle norme vigenti è possibile renderlo più o meno fruibile alla rete in maniera pubblica, semi-pubblica o privata;
- Avere una lista di utenti con cui condividere una connessione. La creazione di specifici gruppi di utenti all'interno del network consente una condivisione di contenuti e conoscenze più mirati. Inoltre, se la rete è di grandi dimensioni è

fondamentale il senso di appartenenza ad un sottogruppo per incrementare l'engagement di un utente (amico o semplice conoscente, ecc..);

- Visualizzare la lista delle proprie connessioni e quelle fatte da altri entro il sistema;
- Cercare altri utenti dalle liste degli amici oppure dalla lista pubblica per aumentare le proprie conoscenze.

In altre parole, un social network permette di condividere dei contenuti con una community più o meno vasta di persone, creando delle connessioni con altri utenti. Infatti, anche se si ha l'opportunità di incontrare persone estranee o di stabilire nuovi contatti, l'utilizzo principale di tali piattaforme è, senza dubbio, quello di mantenere le relazioni esistenti.

Sauer e Coward [4] individuano come obiettivo dei social network quello di soddisfare le esigenze personali degli utenti, e come funzione quella di estendere i propri rapporti interpersonali.

Negli ultimi anni sono nati diversi luoghi d'incontro virtuali, ma solamente alcuni di questi sono riusciti a farsi largo ed avere un impatto, più o meno notevole, sulla vita della gente. Sotto, una mappa dei principali social media, divisi per categoria.

Figura 1: Social Media Landscape Gennaio 2011

Per quanto concerne la diffusione, il social network di gran lunga più utilizzato al mondo è Facebook, come evidenziato dall'immagine seguente. Degni di nota, sono anche i suoi principali concorrenti Qzone e Twitter, che stanno cercando di ritagliarsi uno spazio importante in questo vasto mercato. Mentre Qzone è prettamente utilizzato in Cina, Twitter è in forte espansione in tutto il Globo.

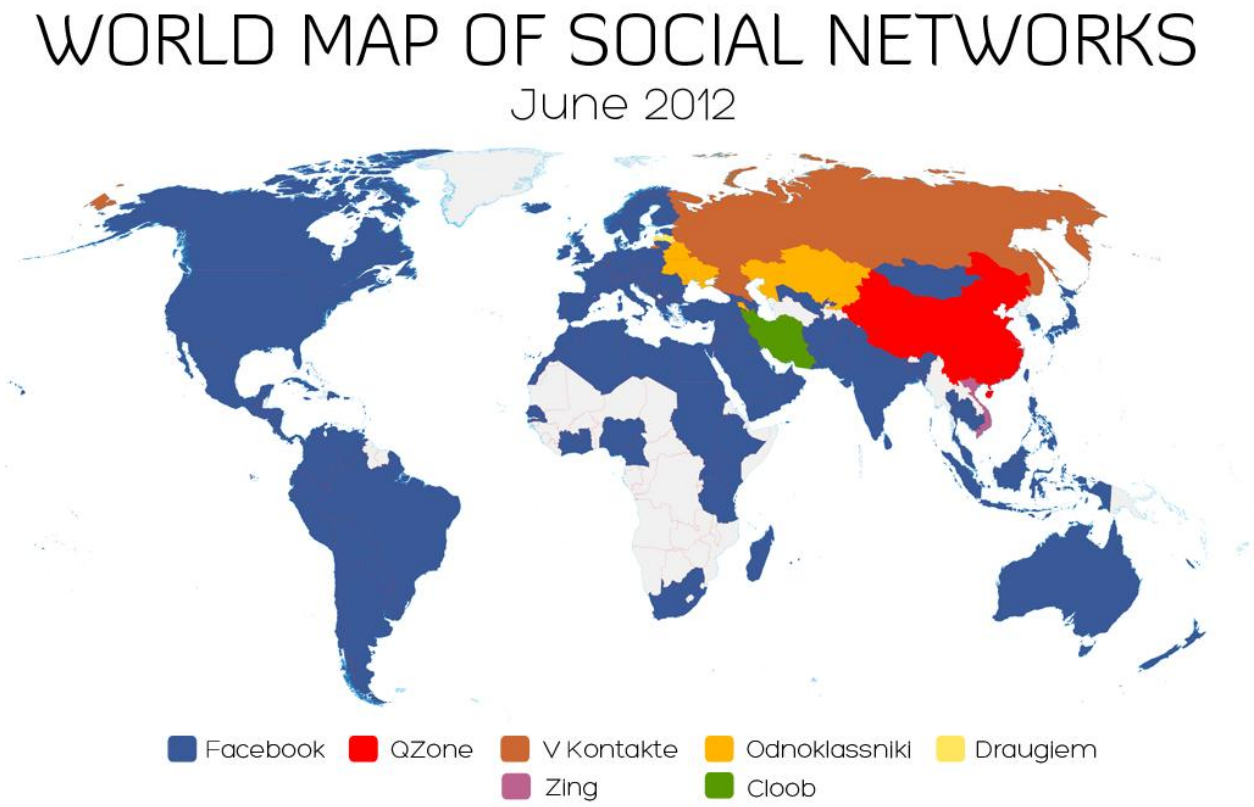

Figura 2: Mappa di utilizzo dei social network nel mondo, Giugno 2012

Le reti sociali online sono quindi delle fonti Web molto importanti, in quanto consentono di raccogliere in modo automatico le informazioni relative ai comportamenti sociali degli individui per effettuare analisi su larga scala [5].

Un social network è quindi una rete sociale supportata da uno strumento web-based che permette alle persone principalmente di:

- Creare un profilo utente;
- Creare dei gruppi;
- Fare operazioni di ricerca.

2.3.1. Twitter ed il micro-blogging

Il micro-blogging è una forma di comunicazione emergente che consente agli utenti di pubblicare dei brevi messaggi, i quali possono essere inoltrati su differenti

canali. È dunque una pubblicazione costante di piccoli contenuti in Rete, sotto forma di messaggi di testo (fino a 140 caratteri). Si tratta di un servizio molto simile all'invio di sms, con la differenza sostanziale che il destinatario non è una sola persona ma un'intera comunità formata da (potenzialmente) milioni di persone.

Uno dei punti di forza di questo tipo di blogging risiede proprio nella brevità del messaggio, che consente al lettore di acquisire le informazioni molto più velocemente rispetto ad esempio ad un articolo su un comune blog. Dato il ristretto numero di caratteri messo a disposizione dalla piattaforma, si è costretti a scrivere solo l'essenziale. Questo ne facilita la lettura permettendo a chi segue le discussioni su queste piattaforme di rimanere sempre aggiornato.

Uno dei più famosi ed apprezzati servizi di micro-blogging è Twitter. Esso presenta funzionalità da social network, ma a differenza degli altri utilizza un sistema chiamato "following"; si distingue infatti radicalmente per l'approccio che offre ai suoi utenti, esso non prevede rapporti di mutua "amicizia", dove una persona risulta amica di un'altra solo se il rapporto è reciproco, ma piuttosto la possibilità di seguire "passivamente", senza bisogno di alcuna autorizzazione, altri utenti, essi siano persone fisiche, associazioni, aziende, marchi ed altro.

Twitter ha conosciuto la popolarità fin dalla prima apparizione su Internet (Ottobre 2006), ed ha saputo raccogliere il frutto di quello che inizialmente sembrava solo una moda o una tendenza. Questo ormai noto social network, è stato il primo esempio a livello internazionale di tale tipo di comunicazione ed ha confermato nel tempo la leadership nel settore, dopo facebook, soprattutto negli Stati Uniti, grazie alla semplicità d'uso della piattaforma, rimasta ancora oggi il punto di forza del servizio [20].

L'utente ha a disposizione una pagina web personale, su cui può inviare gli aggiornamenti tramite il sito stesso, via SMS oppure tramite varie applicazioni basate sulle API di Twitter.

Gli aggiornamenti sono mostrati istantaneamente nella pagina di profilo dell'utente e comunicati agli utenti che si sono registrati per riceverli (follower). È anche possibile limitare la visibilità dei propri messaggi oppure renderli visibili a chiunque.

Una particolarità di Twitter sono proprio le sue open API, rese disponibili liberamente a tutti gli sviluppatori. Questo è stato un altro fattore chiave del suo successo, molti utenti accedono al social network tramite applicazioni di terze parti anziché passare per il sito ufficiale.

Twitter combina gli elementi tipici di un social network con quelli di un comune blog, aggiungendo qualche piccola differenza. Come accade nei social network, i profili sono uniti da un'articolata rete di connessioni, più dirette che indirette; ispirandosi ai blog invece, le pagine degli utenti mostrano i tweet in ordine cronologico inverso, anche se non c'è possibilità di commentare il singolo post.

2.3.2. Le tipologie di contenuti su Twitter

Uno dei tratti caratteristici di Twitter è la possibilità di esprimere la propria opinione su un qualsiasi argomento, senza alcun vincolo. Questo avviene anche grazie alla modalità di comunicazione *many-to-many* tipica di questo social network. Un utente inizia una discussione, un altro risponde e via via si forma una catena di messaggi che coinvolgono potenzialmente un gran numero di altri follower e non.

Le tipologie di contenuti principali presenti in Twitter, sono rappresentati con la relativa distribuzione percentuale nella Figura seguente [6].

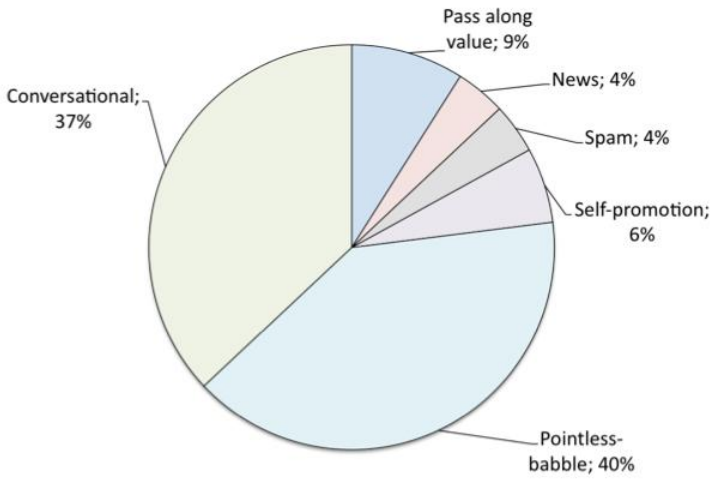

Figura 3 - Distribuzione dei contenuti per tipologia

Come si evince dall'immagine, poco più di un terzo è occupato dalla conversazioni tra coppie o gruppi di utenti, mentre i retweet rappresentano circa un decimo del totale. Da notare l'importanza della categoria *pointless babble,* che contiene proprio quelle opinioni ed espressioni personali che sono di forte interesse ai fine della social media analysis.

Risulta doveroso sottolineare che Twitter presenta un forte grado di volatilità ed è un social network che punta forte sui contenuti real-time, in quanto i post sono costantemente twittati [21]. Come è facile intuire, questo va a discapito dell'attendibilità.

Ciò è dovuto al fatto che non esiste alcun tipo di revisione dei contenuti.

2.4. Social influence: gli influencer nelle reti sociali

Estremamente importante è il concetto di **influence**, ovvero il ruolo che il contenuto stesso del messaggio assume nel processo di diffusione dell'informazione tramite il passaparola.

Un primo elemento chiave che determina l'impatto del passaparola è rappresentato dal contenuto effettivo del messaggio.

Il passaparola sta diventando una forma tangibile di influenza dei consumatori.

Per influence si intende quindi la capacità di generare un effetto, di indurre un cambio di comportamento, di ispirare azioni (retweet, mention).

Negli ultimi anni sono nati molti servizi online che asseriscono di poter misurare l'influenza sociale (Figura 4); in realtà questi tipi di servizi misurano il capitale sociale di un determinato individuo, ovvero la loro possibilità teorica di produrre influenza digitale, e non la sua reale capacità di influenzare.

Figura 4 - Social Influence Measurement Tools

I cosiddetti **influencer** nei social network risultano essere utenti con uno status sociale degno di nota e un focus all'interno di una community; il punto di forza di tali utenti è che hanno la capacità di causare un effetto in termini di cambio di comportamento da parte delle persone a lui connesse.

Il primo vero contributo sul concetto di influencer verrà dato, a fine anni Settanta, da Freeman [7], il quale concentrò l'attenzione sull'identificazione di nodi rilevanti in una rete e sulle metriche per misurarne la rilevanza.

Gli influencer formano legami e connessioni forti con individui affini all'interno di una community e fortificano queste relazioni attraverso interazioni molto significative e di valore.

La capacità di "influenzare" di ogni influencer dipende da vari fattori come:

- Il seguito dell'utente sul social network (es n° di friends, follower);
- Il proprio status;
- L'autorevolezza all'interno della rete;
- La dimensione e la fedeltà della loro audience.

Le tipologie di influencer con cui interagiamo tutti i giorni possono essere racchiuse in queste tre seguenti categorie [22]:

- **Social broadcaster**: persone con un reach molto ampio, che possono non avere una competenza focalizzata su un brand o un argomento. Un'estensione del concetto di "VIP";
- **Mass influencer**: persone che hanno un reach ampio, un'alta affinità con il proprio target e che spesso sono specializzate su uno o pochi temi molto rilevanti per la community con cui interagiscono;
- **Potential influencer**: chiunque possa influenzare il proprio network di persone, anche molto piccolo.

Allo stesso modo, i pilastri cardine dell'influenza sono sostanzialmente tre [8]:

- **Reach** che esprime la capacità di far "viaggiare" un'informazione grazie al proprio grafo sociale o community estesa; dipende a sua volta dalla popolarità (che esprime quanto sei amato, stimato e seguito dalla persone), dalla prossimità (è infatti evidente come la location possa incidere quando si vuole raggiungere un effetto localizzato) e dalla buona volontà (che fa aumentare l'apprezzamento e la possibilità di una collaborazione o azione);
- **Relevance** e il grafo di interessi, dato che la perizia di un esperto su un determinato argomento è il collante delle communità di interesse (focus specifico); tale pilastro dipende dall'autorevolezza (livello di expertise riconosciuto su un argomento), dalla fiducia (parametro difficile da descrivere e misurare: affidabilità, forza, veridicità, onestà, ecc..) e dall'affinità (naturale gradimento per qualcosa o qualcuno);
- **Resonance** che misura la durata, il tasso e il livello di interattività attorno a un contenuto o conversazione; dipende dalla frequenza (quanto spesso un oggetto social appare nelle conversazioni), dal periodo (periodo temporale in cui l'oggetto rimane visibile) e dall'ampiezza (livello di engagement, coinvolgimento all'interno del network).

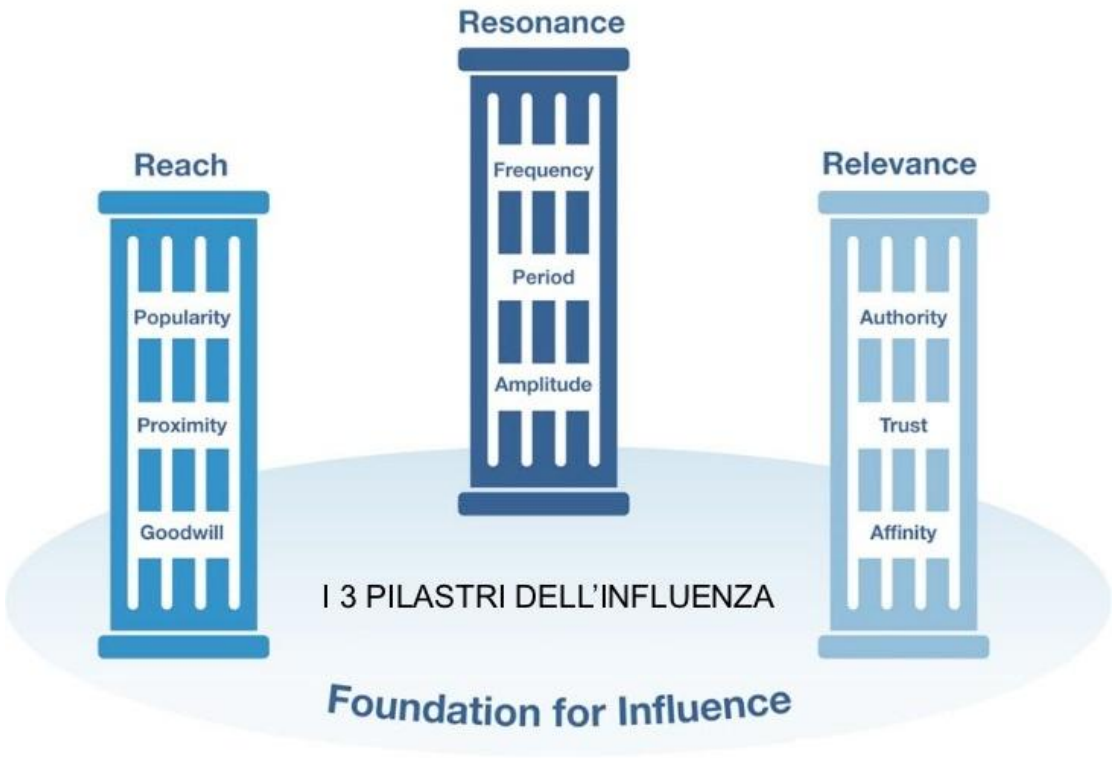

Figura 5 - I pilastri dell'influenza

Al fine di avere un elevata influenza, non è importante tanto la dimensione del network in cui si opera ma risulta fondamentale essere connessi a persone autorevoli, che condividono interessi comuni e influenti su temi specifici.

2.4.1. Gli influencer su Twitter

Nel passaparola, il vero motore che porta alla diffusione delle informazioni sono gli influencer.

La rete in cui si opera è, in generale, difficile da osservare accuratamente; Twitter invece rappresenta un perfetto laboratorio per lo studio del processo di diffusione. Permette di ricostruire l'intero percorso di una notizia, attraverso un semplice crawling del corrispondente grafico dei follower.

Chiaramente questo tipo di individui, sono capaci di influenzare un gran numero di persone differenti, ma esercitano anche differenti tipi di influenza su essi. Ad esempio un commento su un prodotto osannato da una celebrità, avrà un'influenza diversa rispetto a quello fatto da un amico o conoscente o addirittura da un esperto.

Su Twitter non si fa differenza tra i tipi di utente, ma si forza la comunicazione verso un'unica modalità, ovvero via tweet ai propri follower.

Come è facile intuire, l'influenza dei twitterer è parzialmente dovuta al numero di follower, ma ancora non è stato dimostrato che questo sia un buon indicatore. A riprova di quanto affermato, recentemente uno dei creatori di Twitter Evan Williams in una conferenza a Manhattan ha affermato come il numero dei follower non risulta più un buon indicatore riguardante l'importanza e l'autorevolezza di un utente poiché, secondo lo stesso fondatore, il dato a cui bisogna dare più importanza è senza dubbio il numero dei retweet.

<<Il numero dei follower non è sufficiente per rappresentare il peso di una persona, *penso che ci sia una cosa più interessante del numero di follower: ovvero il numero di retweet. La metrica perfetta è sapere quante persone vedono i tuoi tweet>>* (Williams, Manhattan).

Si è tuttavia osservato che [9]:

- Il 72.4% degli utenti segue più dell'80% dei propri follower;
- L'80.5% degli utenti conta di un 80% dei loro amici che li seguono a loro volta.

Ciò sembra essere dovuto a due ragioni. Innanzitutto, potrebbe esserci casualità nella scelta di chi seguire, e chi viene seguito ricambia la "cortesia" diventando follower a sua volta [23]. Oppure potrebbe succedere l'esatto contrario, ossia la relazione tra follower è proprio individuata dagli interessi comuni. In altre parole, un twitterer segue un amico proprio perché condivide gli stessi interessi. Questo fenomeno è chiamato *homophily*, ed è stato riscontrato in diversi social network [10]. La causa di questa reciprocità ha implicazioni rilevanti.

Tuttavia il ranking degli utenti più influenti dipende dal tipo di misura che si adotta. Ad esempio, Kwak et al [11], hanno confrontato tre differenti misurazioni di influenza, numero di follower, page-rank e numero di retweet, notando proprio una certa discrepanza nelle diverse misurazioni.

Uno degli studi più interessanti e discussi in questo ambito, è quello di Cha, Haddadi, Benvenuto, Gummadi, The Million Follower Fallacy [12]. Partendo da un database molto ampio composto da circa 6 milioni di utenti e prendendo come elementi determinanti per l'autorità di ciascuno gli indicatori di indegree (numero di follower), retweet e mention, ne è stata analizzata la correlazione attraverso l'indice di Spearman[1]. Secondo tale analisi, considerando gli utenti appartenenti al primo e al decimo percentile dell'intero set, si è rilevato un notevole valore di correlazione tra i retweet e le mention. Questo legame statistico risulta invece non significativo effettuando la misurazione con l'indegree, portando quindi alla conclusione che la popolarità di un utente ha una scarsa incidenza sull'attenzione e sulle reazioni che è in grado di generare negli altri individui, ovvero la influence che potenzialmente esercita. Lo studio ha inoltre analizzato le dinamiche dell'opinion leadership al variare dei topic e del tempo, sempre tenendo in considerazione retweet e mention. Secondo la ricerca, un ristretto gruppo definito di top influentials sarebbe in grado di mantenere una rilevante autorità su una varietà di argomenti, giungendo infine alla conclusione che la influence non viene guadagnata in modo spontaneo o casuale, ma attraverso uno sforzo mirato che implica anche un coinvolgimento personale.

Un altro studio nello stesso contesto, è stato svolto dal Web Ecology Project (Leavitt, 2009), gruppo di ricerca di Boston, Massachusetts. Basandosi sui contenuti e le risposte generate da un set di 12 utenti popolari, appartenenti a tre cluster definiti a priori come celebrity, news outlet e social media analyst, vengono categorizzate le azioni secondo il contenuto e la conversazione per comprendere come differenti tipologie di utenti e i relativi follower interagiscano in modo differente. In questa ricerca si distinguono due categorie di risposte, *conversation-related* date dalla somma di *reply* e *mention*, e *content-related* date dall'utilizzo dei retweet. Pesandole opportunamente sia con il numero di follower che con l'attività registrata durante il periodo di analisi, risultano evidenti forti discrepanze tra i ranking con i valori assoluti da quelli pesati.

Altre ricerche si sono invece focalizzate su un altro ambito, quello della dinamica di diffusione dell'informazione e dei messaggi all'interno di Twitter,

[1] http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

tenenendo conto della propagazione del messaggio e della maggiore o minore passività dei membri della rete sociale come elementi determinanti. Queste teorie si basano sul fatto che l'opinion leadership di un utente twitter può essere paragonata con quella di una pagina Web, analizzandola mediante gli stessi fattori con cui si giudica un sito. L'autorità di un nodo risulta dunque proporzionale a quella dei suoi follower, quindi più saranno autorevoli i follower più lo sarà l'utente in questione. Questa analogia rende possibile l'uso del PageRank (Brin & Page, 1998), lo stesso utilizzato da Google per indicizzare le pagine Web, o di algoritmi simili con delle varianti, come strumento per effettuare le misurazioni.

Hp Labs [13], ha effettuato un'analisi su un ampio set formato da 22 milioni di tweet contenenti la stringa "http" (i.e. dei Web link). Grazie alla creazione di un algoritmo chiamato IP, ha valutato la propagazione dell'informazione nella rete in termini di riproposizione da parte degli utenti. Questo algoritmo assegna a ciascun utente un *influence score* e un *passivity score*. Quest'ultimo può essere definito come la tendenza a visionare i tweet altrui senza però condividerli con gli altri utenti, insomma un indice di quando si tenda ad essere influenzati. Si è arrivati alla conclusione che il legame tra popolarità e influence è più debole di quanto si creda, inoltre sulla capacità di influenza incidono in modo determinante sia la quantità, sia soprattutto la qualità dell'audience. Risulta chiaro che un tweet vedrà ovviamente una maggiore *reach* se gli altri utenti non ne effettuano esclusivamente un consumo passivo ma lo ritrasmettono attivamente, dunque è necessario che esso sia in grado di superare la predisposizione passiva delle sue connessioni primarie.

Infine Weng et al [9] confrontando il numero dei follower e il page-rank con un page-rank modificato in relazione al topic, arrivano anche loro ad affermare che il ranking dipende proprio dalla misura effettuata.

Sempre questi studi più recenti, tendono a sottolineare il fatto che gli individui che sono stati influenzati nel passato e che hanno molti follower, sono a loro volta più indotti ad essere influenti nel futuro, ovviamente in media [9].

2.5. Essere influenti: cosa è davvero importante?

Da una analisi effettuata [24] e poi pubblicata da SocialFlow, società di servizi di social analytics, riguardante l'audience di sei importanti fonti d'informazione internazionali presenti su Twitter («Al-Jazeera», «BBC News», «CNN», «The Economist», «Fox News» e «The New York Times»), è emerso come non sia la quantità dei follower a fare la differenza e a generare quindi il maggior numero di click e retweet.

I retweet servono a creare consapevolezza e credibilità in un brand, a creare notorietà e fama ad una marca e misurano in qualche modo la fiducia nella fonte d'informazione; i click, al di là dell'aspetto speculativo di generare traffico al sito, indicano il livello di coinvolgimento e di adesione tra i temi proposti e la partecipazione da parte dell'audience, ovvero del pubblico di riferimento.

Dall'analisi emerge come la capacità di attrarre l'interesse delle persone e di avere la loro fiducia siano gli attributi e i requisiti necessari per avere successo in un social network come Twitter e, non dipendono dal numero dei propri follower.

Il grafico di sintesi dei risultati sotto riportato indica chiaramente come non siano «CNN» e «The New York Times», le due fonti ad avere il maggior numero di follower tra quelle esaminate, ad avere né il maggior numero di click verso il proprio sito web né il maggior numero di retweet.

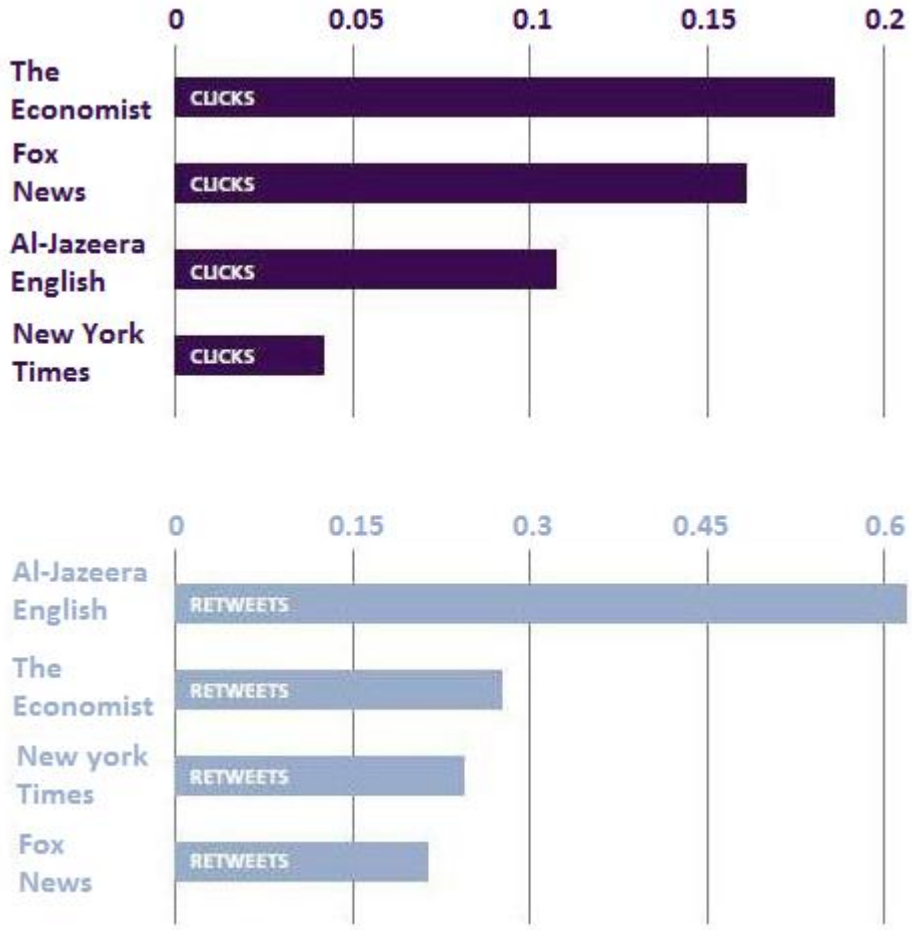

Figura 6 - Click e retweet per Fonte d'informazione

Da ulteriori analisi [14] è emerso come alcuni post vengano condivisi più di altri, senza essere "rimbalzati" dagli utenti della rete, perché possiedono cinque caratteristiche che li rendono particolarmente retwittabili come:

- Contengono un link; risulta infatti come il 56,69% dei retweet include un link. Questa fatto si può facilmente spiegare dalla necessità da parte degli utenti di esprimere maggiori informazioni (rispetto a quelle contenute nei 140 caratteri) e di condividere i contenuti da tutto il web;

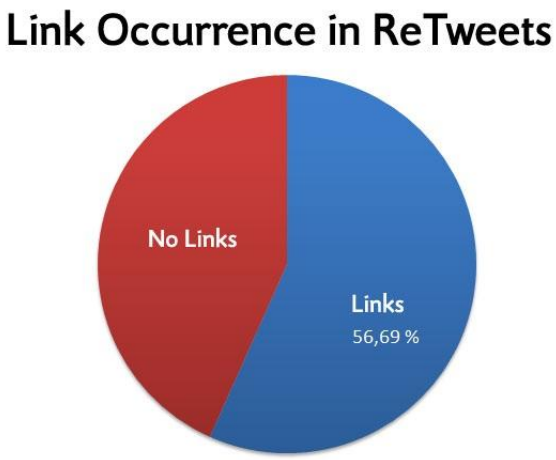

Figura 7 - % di retweet contenenti link

- Contengono parole più lunghe (con più sillabe); sembra infatti che le parole più lunghe e complesse attirino maggiormente l'attenzione altrui e non risultino invece ostacoli al retweet. Post ragionati e articolati come un pensiero "rilevante", possono trovare la strada del retweet in maniera molto più facile;

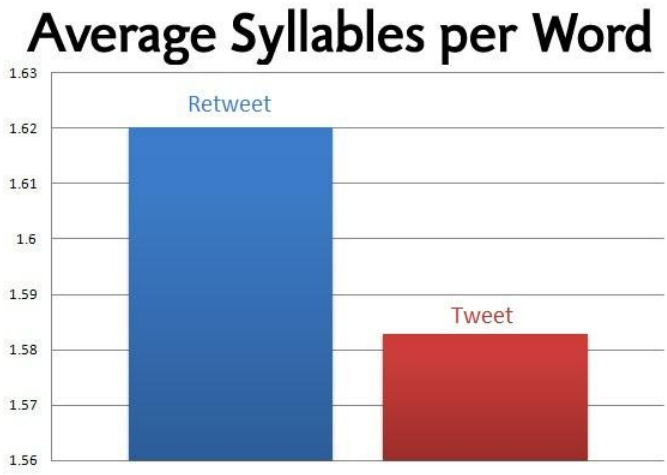

Figura 8 - Sillabe medie per parola: tweet VS retweet

- Risultano più complessi, "brillanti" e, per essere capiti, richiedono un livello di istruzione elevato; riflessioni argute, pungenti sono la norma per aspettarsi un buon numero di retweet;
- Contengono parole meno comuni; è opportuno che i post possiedano non solo parole lunghe, ma anche termini poco usati nelle conversazioni comuni;
- Sono meno autoreferenziali; sembra infatti che la possibilità di retwittare qualche contenuto sia meglio accettata dai follower quando il contenuto stesso è slegato dall'autore originale.

3.0 Il Progetto: raccolta, analisi dei dati ed ipotesi di ricerca

3.1. Introduzione

Il web rappresenta ad oggi una fonte inestimabile di informazione che, se opportunamente elaborata, può essere il punto di forza di molte aziende.

Il lavoro realizzato si pone diversi obiettivi. Si è cercato di costruire un tool in grado di, in maniera del tutto automatizzata, lavorare grandi quantità di dati provenienti da tutti i tweet raccolti dal Politecnico di Milano, per individuare le dinamiche e le caratteristiche, o almeno osservare se ne esistono, che permettono all'informazione di diffondersi nel web.

Le aziende, quando parlano di marketing aziendale, danno sempre più importanza ai dati reperibili su internet poiché sono fonti ricche e sempre aggiornate di informazioni facilmente reperibili e che presentano determinati vantaggi come:

- Costi minori rispetto, per esempio, ai classici questionari;
- Le opioni risultano essere più sincere, ognuno di noi non si sente sotto esame ed analisi quando esprime una opinione in un social network;
- Non vi sono tutti i suggerimenti impliciti che si riscontrano nella maggior parte dei questionari.

In assenza di strumenti concreti che permettano tali analisi automatiche, l'obiettivo di questo progetto è quello di realizzare un tool in grado di analizzare, attraverso studi empirici, alcune caratteristiche dei social network, in particolare di Twitter; si vuole infatti cercare di capire tutte le varie ragioni e comportamenti che permettono ad alcuni utenti di avere una elevata influence.

3.2. Architettura generale dello strumento utile alla raccolta,"pulizia" ed archiviazione dei dati

Per svolgere questo lavoro di tesi è stato utilizzato un database contenente tweet raccolti da Twitter dal Politecnico di Milano attraverso uno strumento chiamato WISPO [25].

L'architettura, vedi Figura 9, è composta principalmente da quattro moduli:

- Modulo di crawling;
- Modulo di data quality;
- Modulo di sentiment analysis;
- User inferface.

Il crawler è l'elemento dedicato alla raccolta di dati grezzi. Si tratta sostanzialmente di un agente software in grado di effettuare ricerche per keyword o per concetti semantici su differenti tipologie di fonti, come social network e forum.

Nello specifico sono presenti differenti tipi di crawler, ognuno specifico per ogni fonte trattata.

Nel modulo di data quality sono presenti tutti quei meccanismi per il cleaning e lo stemming dei dati crawlati precedentemente. Scopo di questi sottomoduli è quello di eseguire una corretta analisi sintattica dei messaggi.

Nello specifico :

- Il modulo di cleaning cerca di recuperare quei messaggi che contengono errori di ortografia o slang specifici di una determinata lingua;
- Il modulo di stemming consente la riduzione in forma flessa di una parola alla sua forma radice;
- Il modulo di pesatura fornisce metriche per misurare la reputazione dei dati secondo le varie fonti di informazione.

Il modulo di sentiment analysis ha il compito di eseguire la disambiguazione sulle parole ed assegnare un livello di sentiment positivo, negativo o neutro, attraverso la valutazione di aggettivi, verbi, sostantivi ed avverbi.

Il modulo di user interface, infine, fornisce l'interfaccia utente, chiamata DashMash, a tutto lo strumento attraverso l'utilizzo di mashup, come tagcloud, mappe e grafici, i quali consentono all'utente finale di personalizzare ed estendere facilmente l'insieme di strumenti di analisi disponibili in maniera facile ed intuitiva.

Tramite tale interfaccia è anche possibile effettuare operazioni di roll-up e drill-down utili per avere una visione generale a livelli di granularità differenti.

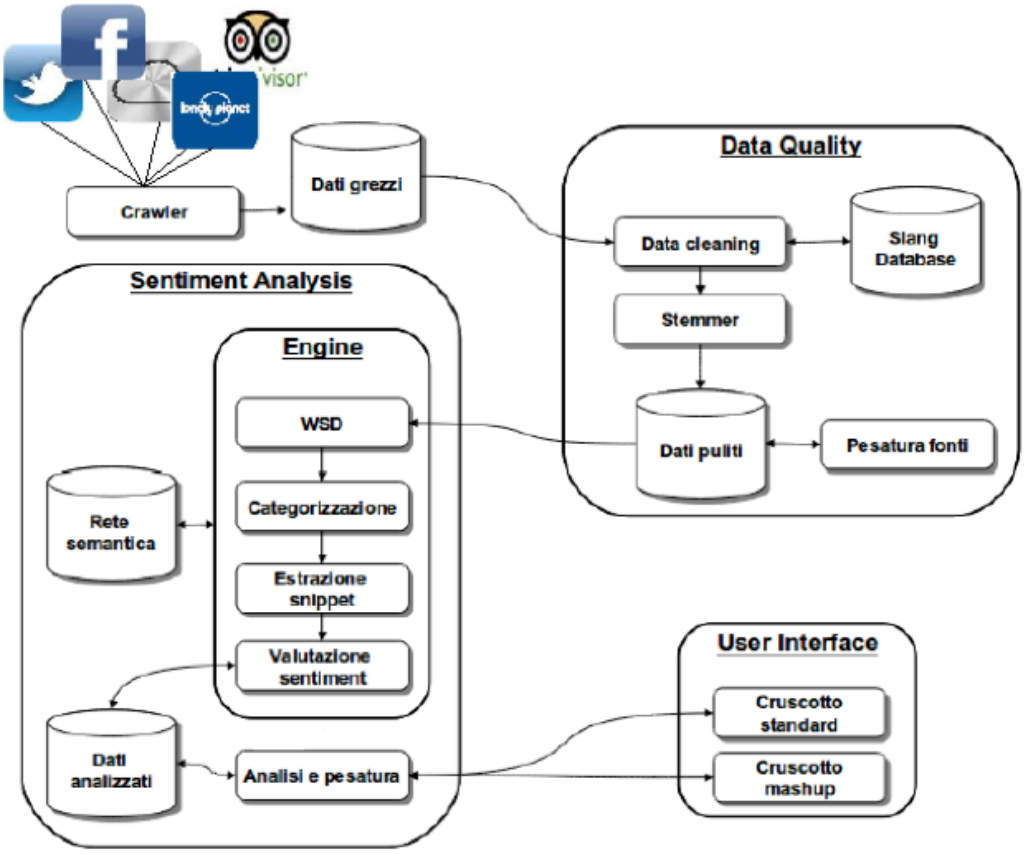

Figura 9 - Architettura generale dello strumento per la raccolta, "pulizia" ed archiviazione dei dati

3.3. Dataset

Per questo lavoro di tesi sono stati utilizzati due ampi dataset di cui viene data una spiegazione di seguito.

3.3.1. Dataset A

Il dataset A è composto da tutti i post inglesi crawlati avente come dominio il turismo riguardante la città di Milano (keyword di selezione "Milan") nel periodo che intercorre tra il mese di luglio e quello di dicembre 2011.

Il dataset risulta essere composto da un totale di 1.378.659 tweet, di cui 353.458 retweet.

Le analisi che sono state effettuate hanno riguardato non solo i post in lingua inglese ma anche quelli in lingua italiana; pertanto un ulteriore dataset simile conterrà tutti i post italiani crawlati negli stessi mesi con keyword di selezione "Milano" (1.617.853 Tweeet, di cui 201.445 retweet).

Tali dataset sono stati utilizzati per effettuare l'analisi del sentiment (H1) e delle 10 diverse categorie sui tweet (H2), (ipotesi specificate nel paragrafo 3.7); l'indagine è stata eseguita in maniera distinta sui post italiani, così come su quelli inglesi ed, in ultimo, sul dataset composto da tutti i tweet di entrambi gli idiomi.

3.3.1. Dataset B

Il dataset B è composto da tutti i post inglesi crawlati avente come dominio il turismo riguardante la città di Milano (keyword di selezione "Milan") nel periodo che intercorre tra il mese di ottobre e quello di dicembre 2011.

Il dataset risulta essere composto da un totale di 751.871 tweet, di cui 193.125 retweet.

Anche in questo caso, le analisi che sono state effettuate hanno riguardato non solo i post in lingua inglese ma anche quelli in lingua italiana; pertanto un ulteriore

dataset simile conterrà tutti i post italiani crawlati negli stessi mesi con keyword di selezione "Milano" (856.541 tweet, di cui 126.088 retweet).

Tali dataset sono stati utilizzati per effettuare l'analisi che studia le modalità con cui i post con sentiment negativo, positivo o di qualsiasi sentiment vengono retwettati da parte dei follower e non follower (H3) (specificate nel paragrafo 3.7) .

3.4. Architettura generale del progetto di tesi

In questo paragrafo verranno descritti i metodi e le tecnologie utilizzate per questo lavoro di tesi.

L'architettura del tool sviluppato in Java, appositamente per questo progetto, è costituito dai seguenti moduli ognuno dotato di specifiche caratteristiche:

- *Database.java:* questo modulo ha lo scopo di interfacciarsi con il database in esame, che in questo caso ha un motore MySQL Server, permettendo il prelievo dei tweet di interesse. Gestisce e avvia le singole query, oltre ovviamente a fornire i risultati secondo un formato ben definito ai moduli successivi.

  Questo modulo viene, in ultimo, utilizzato anche per eseguire il salvataggio dei risultati ottenuti in appositi database realizzati in precedenza;

- *Main.java*: Permette all'utente che utilizza il tool di inserire dei parametri che consentono una selezione sui post da utilizzare nella fase di analisi vera e propria; questi parametri sono l'arco temporale di interesse ed eventuali parole che si vuole compaiano necessariamente all'interno del testo dei post.

  Ha quindi come input una lista di tweet "selezionati" forniti dal primo modulo e che sono pronti per essere analizzati e raggruppati. Genera dunque in uscita, o meglio provvede al riempimento di, un ulteriore database contenente tutte le coppie tweet/retweet che sono state identificate durante

la fase di elaborazione del programma stesso e che serviranno successivamente per l'esecuzione di ulteriori analisi quantitative;

- *TwitterFollower.java*: questo modulo ha come input la struttura dati che è stata riempita nella fase di elaborazione del secondo modulo.
  Per ogni tupla fornita in input, ovvero per ogni corrispondenza tweet/retweet identificata, il modulo provvederà a verificare se l'autore che ha effettuato il post "originale" risulta essere follower dell'autore che ha effettuato il retweet, e viceversa; questo tipo di analisi viene effettuata tramite le API messe a disposizione da Twitter.
  Anch'esso provvederà al riempimento di un ulteriore database contenente tutte le informazioni già precedentemente raccolte, arricchite a loro volta di queste caratteristiche aggiuntive che saranno utili per successive analisi.

3.5. Analisi di un tweet

Twitter prevede principalmente due tipi di interazione tra gli utenti:

- *Menzioni:* danno la possibilità di rispondere direttamente ad uno o più utenti, che siano follower o meno. È forse uno dei metodi più usati e anche più intuitivi, basta infatti utilizzare la sintassi "*@username*" all'inizio del tweet per recapitarlo al diretto interessato. Può anche essere usato come citazione, se presente all'interno (e non all'inizio) del corpo del messaggio;

- *Retweet:* è il modo più immediato per manifestare interesse per un tweet, che appunto essendo retweettato ai propri follower, acquista un certo grado di influenza. Recentemente il meccanismo di retweet è cambiato; in questa tesi non viene considerato solo il concetto più ortodosso introdotto dal nuovo Twitter dove con il termine retweet, si intende un post il cui testo non ha subito variazioni (non viene aggiunto alcun carattere al tweet originale) e

che è stato retweettato seguendo la procedura automatica imposta da Twitter.

Infatti, essendo possibile retwittare un post anche tramite la sintassi *"RT @username:"*, procedura manuale molto diffusa ma non riconosciuta come ufficiale da Twitter, prenderemo in considerazione per l'analisi anche tali tipi di retweet che considereremo a tutti gli effetti validi.

Un discorso analogo viene fatto per tutti quei retweet che sono stati realizzati citando l'utente "originale" ovvero scrivendo "via @username" nel tweet copiato.

Bisogna precisare una differenza sostanziale tra retweet e mention. Mentre il primo ha come caratteristica principale quella di aumentare la reach del tweet sorgente, puntando alla diffusione del contenuto, il secondo è più orientato all'interazione tra gli utenti, dunque sulla conversazione.

È doveroso ricordare che tutti i tipi di messaggi non possono superare il limite massimo di 140 caratteri, pena il troncamento del messaggio stesso.

3.5.1. Metodo di individuazione di un retweet

Una volta chiarita la scelta effettuata riguardo l'interpretazione di un retweet, si può procedere alla descrizione della metodologia di ricerca usata per classificare i vari post presi in esame.

Al fine di determinare e distinguere i post considerati originali dai retweet e, trovare una eventuale corrispondenza tra tweet e retweet analizzati, si è proceduto alla sperimentazione di varie tecniche sia lato database (via MySql) sia tramite software dedicato per la ricerca.

Mediante il metodo String.LastIndexOf (String) si è provveduto a fare una prima scrematura dei tweet ricevuti in input (post selezionati di un determinato periodo e con parole chiave di interesse), osservando se nel loro rispettivo testo si trovasse una occorrenza delle stringhe "RT@", "RT @" o "via@".

Tale metodo assicura l'individuazione e quindi la separazione di tutti quei post che non risultano essere "originali" bensì retweet di altri.

In un secondo momento si è provveduto a perfezionare la ricerca e l'analisi. Per identificare la corrispondenza tra un post "originale" ed un suo retweet, qualora questa esista, si è deciso di utilizzare le espressioni regolari (ER) sul testo di quest'ultimi; tramite le ER è stato infatti possibile non solo trovare la parte di testo di ogni singolo post che lo identifica appunto come un retweet ma, anche la sua totale eliminazione.

Modifica che è stata necessaria per confrontare poi tale contenuto con quello di tutti gli altri post "originali".

Così facendo è stato possibile, al termine dell'analisi, avere un conteggio totale dei retweet per ogni tweet analizzato.

Di seguito un esempio per esemplificare quanto fatto:

il testo del retweet seguente "RT @username: sono un retweet" è stato identificato e successivamente trasformato tramite le espressioni regolari nel seguente "sono un retweet", per poi essere confrontato con il testo di tutti i post "originali", al fine di trovare una possibile correlazione.

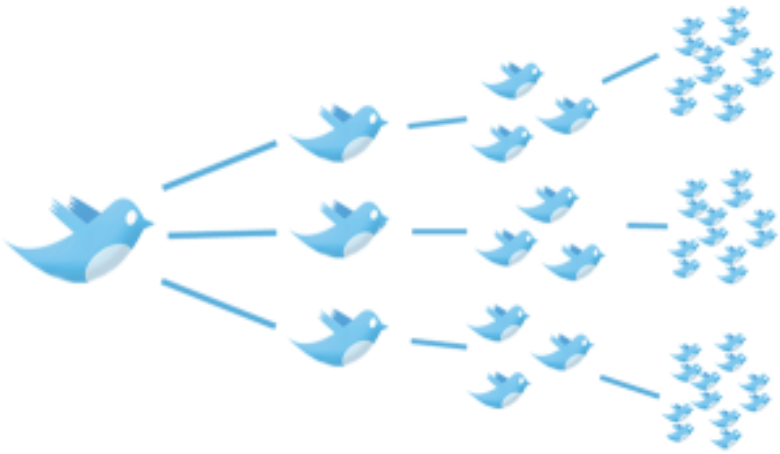

Figura 10 - Forza del retweet

3.6. I dati in output

Le informazioni in uscita dai moduli Java che andranno a riempire il database opportuno sono delle "tuple" contenenti tutte le caratteristiche che riguardano, essenzialmente, un tweet ed il rispettivo retweet.

Un ulteriore database conterrà oltre a queste informazioni, degli attributi aggiuntivi che si riferiscono alle relazioni di amicizia che esistono tra i vari autori che hanno realizzato il tweet ed il retweet in questione; informazioni necessarie per testare l'ipotesi H3 (paragrafo 3,7).

Nello specifico, sono presenti diverse informazioni come:

- Id del tweet;
- Id del retweet;
- Autore del tweet;
- Autore del retweet;
- Data di pubblicazione del tweet;
- Data di pubblicazione del retweet;
- Testo del tweet;
- Tipo del retweet (un retweet, per come viene inteso, può essere eseguito da un autore differente da colui che ha postato il tweet originale o, dallo stesso autore).

Informazioni aggiuntive, che sono presenti nel secondo database sono le seguenti:

- Follower (che permette di sapere se l'utente che ha effettuato il retweet segue l'utente che ha postato in origine il tweet);
- Following (che permette di sapere se l'utente che ha postato il tweet in origine segue l'utente che ha effettuato il retweet).

3.6.1. Dati aggregati

Grazie ai dati in output memorizzati riguardanti le occorrenze trovate di tweet e retweet e i rispettivi rapporti di amicizia tra gli utenti che li hanno realizzati, e grazie alle informazioni salvate in un ulteriore database contenenti le caratteristiche, come il sentiment e le categorie di appartenenza, di tutti i post raccolti dal Politecnico di Milano, è stato poi possibile procedere tramite studiate query alla raccolta di informazioni aggregate relative ad ogni singolo tweet "originale"; così facendo è stato creato un ulteriore dataset costituito dalle seguenti informazioni:

- Id del tweet "originale";
- Numero totale di retweet individuati;
- Numero di retweet effettuati da utenti follower;
- Numero di retweet effettuati da persone che sono seguite dall'utente che ha realizzato inizialmente il post;
- Numero di retweet che sono combinazione delle precedenti due colonne; ovvero numero di retweet in cui l'utente che ha realizzato il post "originale" è seguito e segue a sua volta l'utente che ha effettuato il retweet del post stesso;
- Numero di retweet che sono stati effettuati da utenti che non hanno nessun tipo di relazione con l'utente che ha effettuato il post; ovvero da account che non sono né follower né following dell'utente che ha postato il tweet "originale";
- Numero di post presenti nel gruppo a cui è stato assegnato un valore di sentiment positivo (+);
- Numero di post presenti nel gruppo a cui è stato assegnato un valore di sentiment negativo (-);
- Numero di post presenti nel gruppo a cui è stato assegnato un valore di sentiment neutro (/);
- Numero di post presenti nel gruppo facenti parte della categoria relativa ad una categoria.

Questa ultima informazione sarà presente rispettivamente per tutte le 10 categorie con la quale un post può essere caratterizzato; quest'ultime vengono elencate di seguito, nello stesso ordine con cui sono memorizzate nel dataset, *EVENTS AND SPORT, LIFE AND ENTERTAINMENT, WEATHER AND ENVIRONMENTAL, FOOD AND DRINK, FASHION AND SHOPPING, NIGHT AND MUSIC, ARTS AND CULTURE, FARES AND TICKETS, SERVICES AND TRASPORT* ed in ultimo *###BRAND*.

Grazie a queste informazioni e ad opportuni strumenti per il calcolo statistico, è possibile verificare le ipotesi presentate di seguito.

3.7. Ipotesi

Lo scopo sostanziale di questo lavoro è quello di osservare se nei social network, e nello specifico in Twitter, il contenuto di un determinato tweet svolge un ruolo chiave nel determinare l'influenza dell'informazione stessa; è stata infatti svolta un'analisi che cerca di osservare e capire come i contenuti ed il sentiment di un post possano, o meno, condizionare la diffusione del messaggio.

Verranno presentate 3 ipotesi differenti che cercheranno di sostenere e definire quanto detto in precedenza.

- **Ipotesi 1 (H1):** le notizie con sentiment negativo vengono propagate maggiormente rispetto a quelle con sentiment positivo.

Si vuole quindi osservare se il numero medio di retweet dei post con sentiment negativo sia maggiore del numero medio di retweet dei post a cui è stato attribuito un sentiment positivo.

Sempre più nei media tradizionali, televisioni e giornali, veniamo a contatto con notizie negative che sembrano avere un maggior riscontro e diffusione di quelle positive; l'idea di base è quello di osservare se anche nei media alternativi come i social network, ed in particolar modo in Twitter, questo trend viene riprodotto nel medesimo modo.

Questa diffusione di notizie infelici può dipendere da vari fattori e, una spiegazione psicologica, può essere il fatto che le persone tendono a memorizzare maggiormente gli argomenti negativi rispetto a quelli positivi; da qui la ovvia conseguenza di come, anche nelle televisioni, questo andamento venga mantenuto.

In maniera del tutto analogo delle ulteriori spiegazioni sul verificarsi di questo fenomeno possono essere ricercate in due motivi distinti; il primo è la curiosità innata degli uomini, spinti dalla necessità di sapere; il secondo è il bisogno di ogni individuo di distogliere l'attenzione dalle brutte esperienze personali di vita che risultano nulle se paragonate a quello che ascoltiamo nei tg di tutti i giorni.

In letteratura [16] è stato studiato come gli utenti dei social network tendono statisticamente ad autopromuoversi, generando quindi un numero maggiore di messaggi con sentiment positivo. In modo del tutto speculare, come già ampiamente discusso, nei media tradizionali è invece più facile ascoltare notizie negative [17] che trovano un maggior riscontro ed interesse nelle persone rispetto alle notizie a cui viene attribuito un sentiment positivo.

Alla base di quanto detto, si vuole quindi osservare, attraverso l'analisi e lo studio di questa ipotesi se anche nei media online le notizie con sentiment negativo abbiano la stessa maggiore diffusione che hanno nei media tradizionali.

- **Ipotesi 2 (H2):** il numero di retweet di un determinato post è strettamente correlato all'argomento che tratta il tweet stesso.

Come nella vita normale, anche nei social media alcune categorie di argomenti hanno un seguito maggiore rispetto ad altre poiché riescono ad attirare maggiormente l'attenzione del "pubblico" online.

Nel dicembre 2011, è stato pubblicato sul blog ufficiale di Facebook il *Memology 2011* [18], ossia gli argomenti più condivisi, menzionati e commentati a livello globale nel 2011 sul social network in questione.

Figura 11 - Facebook Memology 2011

E' estremamente importante, soprattutto in questi grandi social network dove vengono postati circa 30 miliardi di contenuti mensili [26], sapere di cosa si parla; proprio per questo motivo l'analisi effettuata tenta di scoprire quali siano gli argomenti che vengono maggiormente trattati.

E' opportuno specificare che, lo scopo di questa analisi non vuole essere quello di determinare quale categoria viene maggiormente ritrovata nei post ma, si vuole determinare quale argomento riscuote il maggior numero di retweet.

- **Ipotesi 3 (H3):** i tweet vengono propagati maggiormente da parte di utenti che seguono l'autore del post rispetto a chi non lo segue.

Questa terza ipotesi è stata analizzata poiché in Twitter, a differenza di altri social network come Facebook, il meccanismo di "amicizia" è diretta (non biunivoca); una conseguenza di quanto riportato è il fatto che un utente decide di seguire solo chi vuole, senza nessuna autorizzazione da confermare, e, allo stesso modo, un utente viene seguito solo da chi è interessato a ciò che esso scrive.

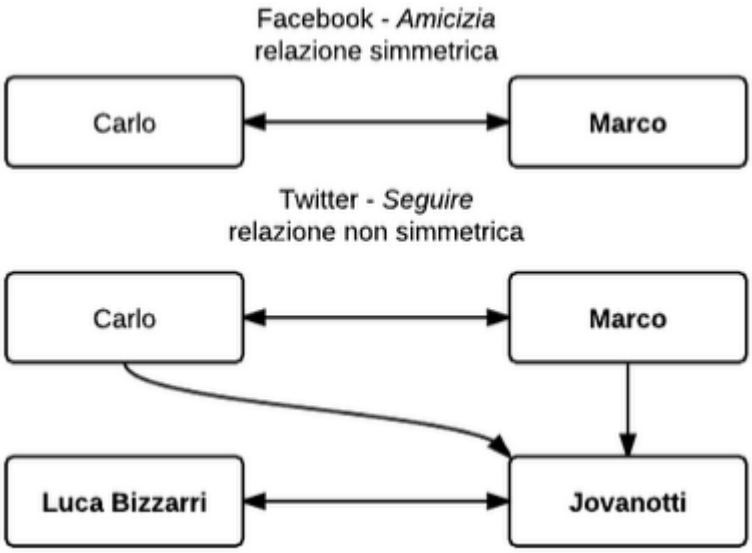

Figura 12 - Relazioni di amicizia: Twitter VS Facebook

Tale struttura induce a pensare che gli utenti di Twitter siano portati a seguire soltanto coloro che, ai propri occhi, pubblicano con alta frequenza contenuti interessanti e quindi degni di essere retwettati.

I contenuti riproposti sono quindi legati, o strettamente connessi, agli argomenti di discussione che gli account seguiti creano e, dal momento in cui un utente riceve le notifiche sulla propria bacheca in modalità push da parte di tutti gli account che segue, la possibilità di retwettare i tweet ricevuti in questa modalità risulta presumibilmente superiore rispetto a quelli ricevuti in modalità pull, ovvero tramite la ricerca indipendente di parole chiave.

Normalmente condividiamo le idee delle persone che ci piacciono, che decidiamo di seguire e di renderle, in qualche modo, partecipi della nostra vita.

Come nei rapporti di amicizia reali, ognuno di noi è portato a relazionarsi maggiormente con qualcuno rispetto a qualcun altro e, il motivo di tutto questo, può dipendere dalle più svariate ragioni: interessi comuni, sensazioni a pelle per citarne alcune.

Si vuole osservare se, anche nei social network, le persone sono portate a stringere determinati legami, dettati prevalentemente da interessi comuni, attraverso il quale la condivisione delle proprie "idee" diventa quasi automatica.

Intuitivamente risulta del tutto normale il fatto che un utente, in Twitter, condivida un tweet di un altro utente che segue e, che quindi, ha osservato nella propria home page; risulterebbe invece anomalo retwettare un post di un utente che non si segue e, che proprio per questo, è stato ricercato con l'apposito form di ricerca grazie al quale è possibile accedere a tutto il contenuto presente in Twitter rispetto al solo contenuto generato dagli account amici.

3.8. Identificazione dei follower

E' lecito aspettarsi che i messaggi che vengono retwettati o comunque inoltrati da un utente, siano post prodotti dall'account di una persona follower e che quindi, seguendoci, visualizza sulla propria home page i tweet da noi realizzati.

Un ulteriore modo per ricercare topic in generale è quello di fare uso dell'apposita form di ricerca, così da accedere al contenuto di tutto il social network piuttosto che esclusivamente al contenuto generato dagli account amici.

Facendo uso delle API messe a disposizione da Twitter si è andati ad indagare se fosse vero il comportamento citato in precedenza utilizzando le informazioni presenti nel dataset B.

3.9. API di Twitter

Le API di Twitter, come riporta la documentazione ufficiale [19] sono al momento composte da due differenti tipologie, due appartenenti al tipo REST e una al tipo streaming.

Per quanto riguarda i metodi REST si hanno:

- REST API methods che consentono l'accesso ai dati di Twitter. Tale categoria è a sua volta suddivisa in: Timeline Methods, Status Methods, User Methods, List Methods, List Members Methods, List Subscribers Methods, Direct Message Methods, Friendship Methods, Social Graph

Methods, Account Methods, Favorite Methods, Notification Methods, Block Methods, Spam Reporting Methods, Saved Searches Methods, Trends Methods, Geo methods, Help Methods ed in ultimo OAuth Methods che racchiude il nuovo metodo di autenticazione basato su token introdotto di recente;

- Search API che permettono l'interazione con il search di Twitter.

Per quanto riguarda le streaming API forniscono l'accesso ad un elevato volume di tweet in near-realtime di un sottoinsieme di dati pubblici e protetti.

In generale le API di Twitter al momento supportano e restituiscono i dati sotto forma di diversi formati:

- XML, acronimo di extensible markup language, metalinguaggio utilizzato per creare nuovi linguaggi, atti a descrivere documenti strutturati e con il quale è possibile definire dei tag propri a seconda delle esigenze;
- JSON, acronimo di javascript object notation, è un altro formato adatto per lo scambio di dati in applicazioni client-server. A differenza dell'XML non è un linguaggio di marcatura, ma un formato di interscambio di dati;
- RSS, acronimo di really simple syndication. Conosciuto come uno dei più diffusi formati di distribuzione di contenuti web, è basato su XML, con cui condivide le caratteristiche di semplicità, estensibilità e flessibilità;
- ATOM, l'atom syndication format è un formato di documento basato su XML per la sottoscrizione di contenuti web, come blog o testate giornalistiche.

La comunità di Twitter ha inoltre messo a disposizione differenti librerie per linguaggi di programmazione diversi come C, C++, .Net, Java, PHP, Python e altri.

Nello specifico è stata utilizzata la libreria java twitter4j[2].

[2] http://twitter4j.org/en/index.html

E' stato quindi progettato un modulo che, preso in input due utenti, colui che ha effettuato il post e colui che ha effettuato il retweet dello stesso, verifica se i due sono tra loro follower.

3.9.1. Problemi relativi alle API di Twitter

Negli ultimi anni Twitter è riuscito ad imporsi grazie anche alle sue API, con le quali è stato possibile realizzare e far "nascere" una infinità di applicazioni di terzi che utilizzano tali interfacce, così in poco tempo il traffico generato dalle API ha abbondantemente superato quello della tradizionale interfaccia web.

Tutto questo ha portato al verificarsi di un serio problema che è stato risolto dagli sviluppatori del social network imponendo delle restrizioni sull'uso delle Twitter REST API; infatti è noto che, quest'ultime, sono soggette ai seguenti limiti:

- 150 richieste non autenticate ogni ora (basate sull'indirizzo IP dal quale proviene la richiesta);
- 350 richieste autenticate all'ora (basate sull'identificativo dell'utente che fa la richiesta).

Una ulteriore limitazione è posta nel numero massimo di tweet, 1500, che possono essere restituiti con una richiesta.

La documentazione che riguarda le Twitter Search API [15] specifica inoltre che la ricerca non dà accesso all'indice completo di tutti i tweet ma solo di quelli recenti, fino a 6-9 giorni prima, e che non si possono usare le Search API per trovare tweet più vecchi di una settimana.

In ultimo, ma non meno importante, i tweet reperiti attraverso questa API non garantiscono la "completezza" (la documentazione parla infatti di focus sulla rilevanza) e alcuni tweet potrebbero non essere restituiti per raggiunti limiti di richieste, perché l'utente che ha generato il tweet ha un basso ranking o, infine, semplicemente perché, a causa della limitatezza delle risorse, non tutti i tweet possono essere indicizzati in Twitter Search.

Per ovviare a questo problema, come suggerisce la documentazione di Twitter, conviene usare le Streaming API le quali restituiscono i tweet in tempo reale.

4.0 Verifica di ipotesi

4.1. Introduzione

In questo capitolo verranno presentate le analisi effettuate, ed i risultati ottenuti, che sono state eseguite per verificare le ipotesi introdotte nel Capitolo 3.

In particolare, nel paragrafo successivo, si cercherà di dare una risposta all'ipotesi H1 analizzando i tweet presenti nel dataset A per verificare se la propagazione di un messaggio può dipendere dal tipo di sentiment espresso.

Nel paragrafo 4.3 verrà esaminata l'ipotesi H2; si cercherà, anche in questo caso, di osservare se, dai risultati ottenuti, è possibile stabilire qualcosa sulle dinamiche con cui avviene la propagazione di un tweet in funzione della categoria di appartenenza del post stesso.

Infine, nel paragrafo 4.4 si cercherà di rispondere all'ipotesi H3 secondo cui gli utenti che ripropongono un messaggio, siano principalmente amici dell'utente che ha generato il post stesso.

4.2. Propagazione dei messaggi in base al sentiment

In questo paragrafo verrà verificata l'ipotesi H1 introdotta nel capitolo 3:

> **Ipotesi 1 (H1):** le notizie con sentiment negativo vengono propagate maggiormente rispetto a quelle con sentiment positivo.

Di seguito verranno descritte le analisi compiute per verificare tale ipotesi e verranno illustrati i risultati che si sono ottenuti.

4.2.1. Analisi dei dati aggregati

Dai dati aggregati, filtrati in base al sentiment positivo o negativo, sono state calcolate le statistiche descrittive ed è stato applicato, ad entrambi i dataset, il test di Kolmogorov-Smirnov al fine di verificarne la normalità. Il test di Kolmogorov-Smirnov è un test non parametrico che verifica la forma delle distribuzioni campionarie.

Dai risultati ottenuti per entrambi i dataset (sentiment positivo e sentiment negativo), il test di di Kolmogorov-Smirnov risulta avere livello di significatività pari allo 0,000 , ossia viene confutata l'ipotesi nulla secondo la quale il campione in esame risulta gaussiano, e statistica test che risulta rispettivamente di 74,000 per il dataset con sentiment positivo e di 31,593 per il dataset con sentiment negativo.

Si riportano le frequenze con la quale vengono retwettati i tweet, suddivisi per sentiment, che sono stati presi in esame.

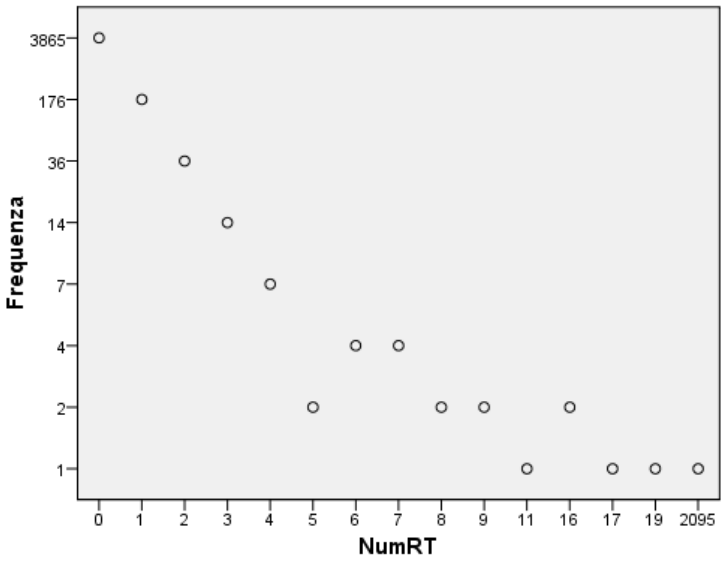

Figura 13 - Grafico a dispersione delle frequenze del numero di retweet (tweet inglese, sentiment negativo)

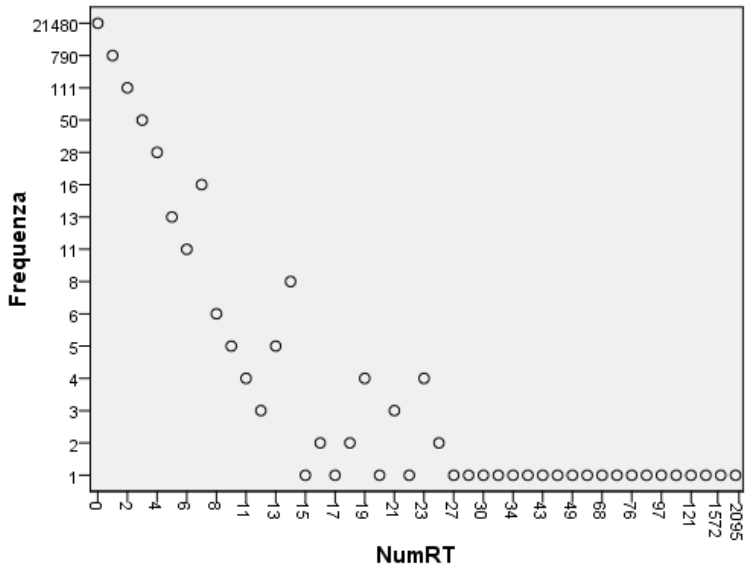

Figura 14 - Grafico a dispersione delle frequenze del numero di retweet (tweet inglese, sentiment positivo)

Si riportano inoltre le statistiche descrittive di entrambe le distribuzioni nella tabella seguente.

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Negativi	4118	4118	0	2095	0,63	32,654
Positivi	22571	22571	0	2095	0,37	20,005

Tabella 1 - Statistiche descrittive di entrambi i campioni (tweet inglese)

Dato che il test di KS applicato ai due dataset di dati aggregati suddivisi per sentiment non ha identificato i campioni come distribuzioni normali è stato possibile, per verificare l'ipotesi H1, eseguire il test di Mann-Whitney sui dati aggregati contenenti i tweet con entrambi i sentiment positivo e negativo.

Il test di Mann-Whitney è un test non parametrico che permette di verificare se due campioni statistici provengono dalla stessa popolazione, attraverso il confronto delle medie dei due gruppi dei campioni indipendenti; i campioni indipendenti, in tale tipo di test, non devono avere una distribuzione di tipo

gaussiana e, per i campioni in esame, quanto detto è già stato verificato attraverso il test di KS realizzato precedentemente.

Il test di Mann-whitney realizzato sui due campioni risulta avere significatività pari allo 0,000 , ossia viene confutata l'ipotesi nulla secondo la quale i due campione in esame provengono dalla stessa popolazione, e statistica test standardizzata che risulta pari a Z = -3,547.

Dai risultati ottenuti con il test di Mann-Whitney, è possibile osservare come i volumi assoluti di retweet con sentiment positivo siano estremamente superiori rispetto a quelli con sentiment negativo (22.571 contro i 4.118) e questo è dovuto al fenomeno dell'auto promozione presente nei social network, come già anticipato nel capitolo 3.6 .

Ciò nonostante, osservando le medie restituite dal test è possibile dedurre, a conferma di quanto ipotizzato nell'ipotesi H1, come i tweet con sentiment negativo siano maggiormente retwettati di quelli con sentiment positivo.

Viene riportato il grafico a dispersione delle frequenze dei tweet inglesi non suddivisi per sentiment.

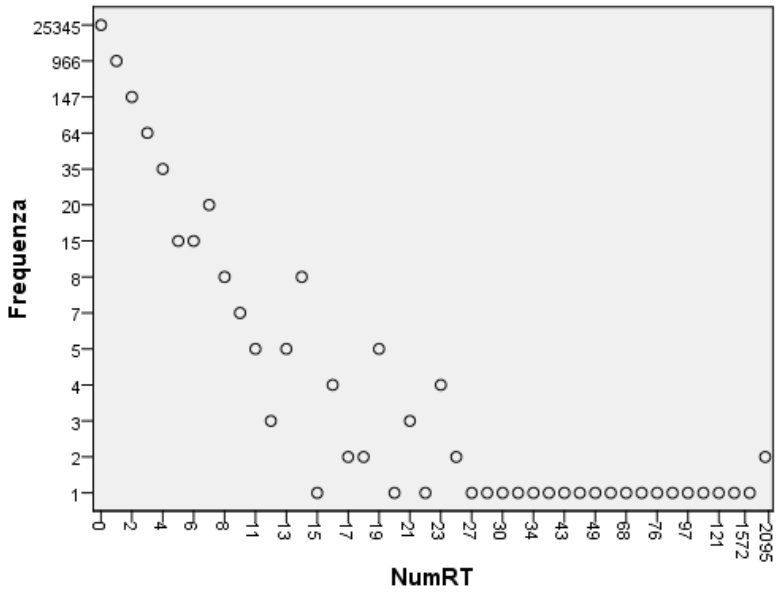

Figura 15 - Grafico a dispersione delle frequenze del numero di retweet (tweet inglese)

Queste analisi sono state effettuate, nel medesimo modo, anche con il dataset di tweet italiani; sono stati ottenuti risultati del tutto analoghi che hanno contribuito a confermare maggiormente quanto già ipotizzato.

Di seguito vengono riportati tali risultati.

Il test di di Kolmogorov-Smirnov applicato alle due nuove distribuzioni di tweet italiani suddivisi per sentiment, risulta avere per entrambi livello di significatività pari allo 0,000 , ossia viene confutata l'ipotesi nulla secondo la quale il campione in esame risulta gaussiano, e statistica test che risulta 58,518 per il dataset con sentiment positivo e 36,165 per il dataset con sentiment negativo.

Si riportano le statistiche descrittive e i grafici a dispersione delle frequenze di entrambe le distribuzioni.

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Negativi	7044	7044	0	128	0,20	2,358
Positivi	14528	14528	0	65	0,15	1,366

Tabella 2 - Statistiche descrittive di entrambi i campioni (tweet italiano)

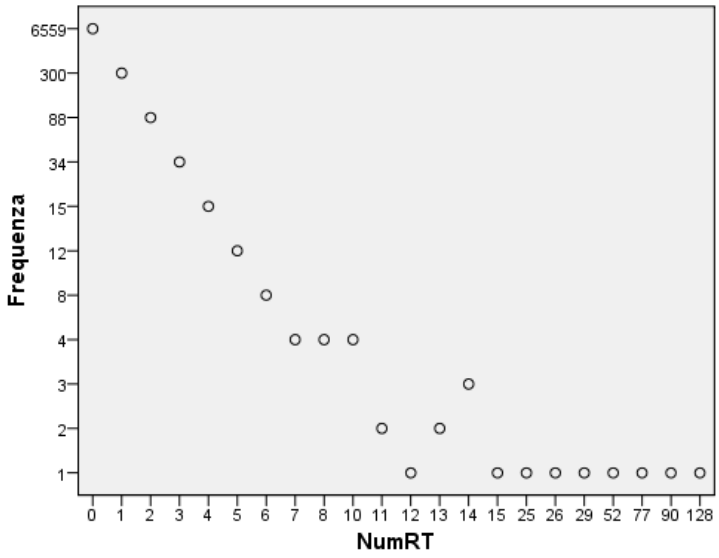

Figura 16 - Grafico a dispersione delle frequenze del numero di retweet (tweet italiano, sentiment negativo)

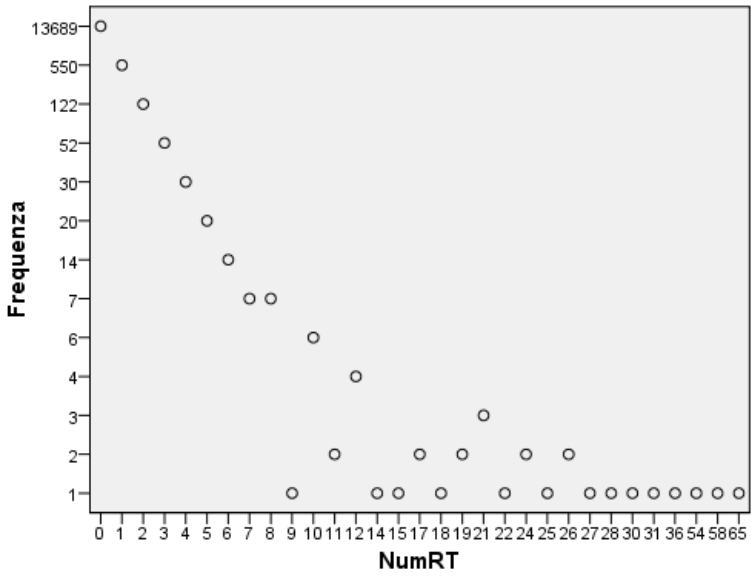

Figura 17 - Grafico a dispersione delle frequenze del numero di retweet (tweet italiano, sentiment positivo)

Dato che il test di KS applicato ai due dataset di dati aggregati suddivisi per sentiment non ha identificato i campioni come distribuzioni normali, è stato possibile, per verificare l'ipotesi H1, eseguire il test di Mann-Whitney sui dati aggregati contenenti i tweet con entrambi i sentiment positivo e negativo.

Il test di Mann-whitney realizzato sui due campioni risulta avere significatività pari allo 0,001 , ossia viene confutata l'ipotesi nulla secondo la quale i due campione in esame provengono dalla stessa popolazione, e statistica test standardizzata che risulta pari a Z = -3,217.

Dai risultati ottenuti con il test di Mann-Whitney, è possibile osservare come, anche in questo caso, i volumi assoluti di retweet con sentiment positivo siano estremamente superiori rispetto a quelli con sentiment negativo (14.528 contro i 7.044) come conseguenza del fenomeno dell'auto promozione presente nei social network.

Allo stesso modo, osservando le medie restituite dal test è possibile dedurre quanto già peraltro confermato con le analisi precedenti, ovvero come i tweet con sentiment negativo siano maggiormente retwettati di quelli con sentiment positivo.

Viene riportato, infine, il grafico a dispersione delle frequenze dei tweet italiani non suddivisi per sentiment.

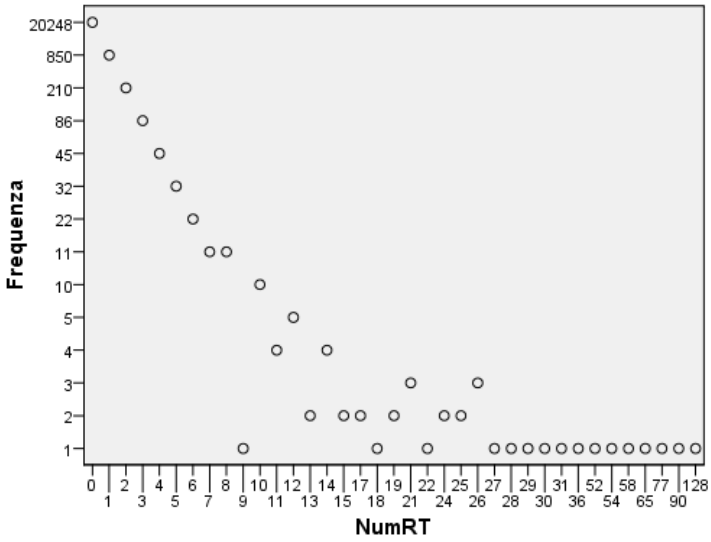

Figura 18 - Grafico a dispersione delle frequenze del numero di retweet (tweet italiano)

4.3. Propagazione dei messaggi in base alle categorie di appartenenza del tweet

In questo paragrafo verrà verificata l'ipotesi H2 introdotta nel capitolo 3:

> **Ipotesi 2 (H2):** il numero di retweet di un determinato post è strettamente correlato all'argomento che tratta il tweet stesso.

Di seguito verranno descritte le analisi compiute per verificare tale ipotesi e verranno illustrati i risultati che si sono ottenuti.

4.3.1. Analisi dei dati aggregati

Come già anticipato, ad ogni tweet analizzato è stata associata una determinata categoria, la quale gli viene accostata in funzione dell'argomento che esso tratta; è opportuno specificare che un determinato tweet può far parte di più categorie differenti (le possibili categorie identificate dal Politecnico di Milano sono infatti 10 (come specificato in 3.6.1)).

Il dataset A, contenente i post inglesi, utilizzato per le analisi è costituito da un totale di 1378659 tweet di cui 395475 retweet.

Dai dati aggregati ottenuti dal dataset in esame, sono state apportate delle suddivisioni che hanno riguardato appunto la categoria di appartenenza dei tweet; sono quindi stati creati i seguenti nuovi dataset, su cui poi sono state effettuate le analisi vere e proprie:

- retweet inglesi appartenenti alla categoria 1 (D1);
- retweet inglesi appartenenti alla categoria 2 (D2);
- retweet inglesi appartenenti alla categoria 3 (D3);
- retweet inglesi appartenenti alla categoria 4 (D4);
- retweet inglesi appartenenti alla categoria 5 (D5);
- retweet inglesi appartenenti alla categoria 6 (D6);
- retweet inglesi appartenenti alla categoria 7 (D7);
- retweet inglesi appartenenti alla categoria 8 (D8);
- retweet inglesi appartenenti alla categoria 9 (D9);
- retweet inglesi appartenenti alla categoria 10 (D10).

Ogni distribuzione è stata dapprima ripulita di tutti quei valori che sono risultati numericamente distanti dal resto dei dati raccolti, ovvero sono stati eliminati i cosiddetti outlier; questo ha permesso che le statistiche ed i test realizzati non restituissero risultati fuorvianti.

Come valore limite si è deciso di tenere tutti quei tweet che hanno ottenuto non più di 500 retweet; questa scelta ha comportato la rimozione di 9 post ovvero:

- 1 tweet appartenente alla categoria 1, *Events and Sport*;

- 3 tweet appartenenti alla categoria 2, *Life and Entertainment*;
- 1 tweet appartenente alla categoria 5, *Fashion*;
- Ed infine, 4 tweet appartenenti alla categoria 10, *Brand*.

Per ogni distribuzione, attraverso il test di Kolmogorov-Smirnov, è stata osservata l'eventuale normalità; vengono riportati in una tabella i risultati del test e le statistiche descrittive di ogni distribuzione ottenuta.

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
D1	83452	83452	0	352	0,22	6,228
D2	22717	22717	0	234	0,43	19,946
D3	3884	3884	0	149	0,19	2,982
D4	14762	14762	0	221	0,20	3,276
D5	58976	58976	0	431	0,19	4,915
D6	19526	19526	0	149	0,15	2,014
D7	5319	5319	0	149	0,18	2,509
D8	2794	2794	0	25	0,07	0,627
D9	28018	28018	0	234	0,15	2,166
D10	156027	156027	0	431	0,21	8,971

Tabella 3 - Statistiche descrittive delle 10 distribuzioni sulle categorie (tweet inglese)

Distribuzione	Statistica Test	Significatività	Risultato
D1	140,441	0,000	Rifiuta l'ipotesi nulla
D2	74,063	0,000	Rifiuta l'ipotesi nulla
D3	29,616	0,000	Rifiuta l'ipotesi nulla
D4	57,776	0,000	Rifiuta l'ipotesi nulla
D5	117,773	0,000	Rifiuta l'ipotesi nulla
D6	66,407	0,000	Rifiuta l'ipotesi nulla
D7	34,362	0,000	Rifiuta l'ipotesi nulla
D8	26,757	0,000	Rifiuta l'ipotesi nulla
D9	79,507	0,000	Rifiuta l'ipotesi nulla

D10	193,747	0,000	Rifiuta l'ipotesi nulla

Tabella 4 - Risultati KS delle 10 distribuzioni sulle categorie (tweet inglese)

Dato che i campioni non sono risultati approssimabili a distribuzioni gaussiane, è stato applicato un test di Kruskal-Wallis; questo tipo di test è un metodo non parametrico utilizzato per verificare l'uguaglianza delle mediane di diversi gruppi; serve per controllare se tali gruppi provengono da una stessa popolazione o, da popolazioni con uguale mediana.

Il test di Kruskal-Wallis realizzato sui dieci campioni risulta avere significatività pari allo 0,000 , ossia viene confutata l'ipotesi nulla secondo la quale i vari campioni in esame provengono dalla stessa popolazione, e statistica test che risulta pari a 377,219.

Dato che i campioni non provengono da una stessa distribuzione, è possibile affermare che gli argomenti trattati in un determinato post influiscono sulla diffusione che quest'ultimo ha; viene quindi validata l'ipotesi di partenza.

In particolare, dalle medie restituite dal test, emerge come i tweet appartenenti alla categoria 2, *Life and Entertainment*, siano maggiormente retwettati degli altri.

Un discorso del tutto analogo viene effettuato sul dataset A contenente i post realizzati in lingua italiana; quest'ultimo è costituito da 1617853 tweet di cui 731590 retweet.

Anche in questo caso sono stati realizzati i 10 diversi dataset, ognuno contenente i tweet appartenenti alle diverse categorie, le cui distribuzioni, dopo essere state ripuliti dagli outlier, sono state esaminate con un test di Kolmogorov-Smirnov per accertarne l'eventuale normalità.

La ripulitura dagli outlier, usando come valore limite lo stesso scelto nelle analisi effettuate in precedenza, ha comportato la rimozione di un singolo post facente parte della categoria 10, *Brand*.

Di seguito vengono dettagliati i risultati ottenuti dal test e le statistiche descrittive di ogni campione.

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
D1	54535	54535	0	70	0,12	0,977
D2	24938	24938	0	62	0,14	1,024
D3	8262	8262	0	90	0,11	1,316
D4	29728	29728	0	273	0,12	2,160
D5	92904	92904	0	273	0,08	1,355
D6	23399	23399	0	182	0,20	2,053
D7	5839	5839	0	43	0,15	1,038
D8	8577	8577	0	38	0,10	0,855
D9	93213	93213	0	65	0,05	0,632
D10	390195	390195	0	338	0,14	2,859

Tabella 5 - Statistiche descrittive delle 10 distribuzioni sulle categorie (tweet italiano)

Distribuzione	Statistica Test	Significatività	Risultato
D1	114,380	0,000	Rifiuta l'ipotesi nulla
D2	77,143	0,000	Rifiuta l'ipotesi nulla
D3	44,343	0,000	Rifiuta l'ipotesi nulla
D4	82,416	0,000	Rifiuta l'ipotesi nulla
D5	149,161	0,000	Rifiuta l'ipotesi nulla
D6	70,457	0,000	Rifiuta l'ipotesi nulla
D7	37,362	0,000	Rifiuta l'ipotesi nulla
D8	46,307	0,000	Rifiuta l'ipotesi nulla
D9	155,329	0,000	Rifiuta l'ipotesi nulla
D10	300,476	0,000	Rifiuta l'ipotesi nulla

Tabella 6 - Risultati KS delle 10 distribuzioni sulle categorie (tweet italiano)

Dato che i campioni non sono risultati approssimabili a distribuzioni gaussiane, è stato applicato ad essi un test di Kruskal-Wallis; tale test realizzato sui dieci

campioni risulta avere significatività pari allo 0,000 , ossia viene confutata l'ipotesi nulla secondo la quale i vari campioni in esame provengono dalla stessa popolazione, e statistica test che risulta pari a 2853,032.

Dato che i campioni non provengono da una stessa distribuzione, anche in questo caso, è possibile affermare che gli argomenti trattati in un determinato post influiscono sulla diffusione che quest'ultimo ha; viene quindi ancora validata l'ipotesi di partenza.

In particolare, in questa circostanza, dalle medie restituite dal test, emerge come i tweet appartenenti alla categoria 6, *Night and Music*, siano maggiormente retwettati degli altri.

4.4. Propagazione dei messaggi in base al tipo di rapporto di amicizia

In questo paragrafo verrà verificata l'ipotesi H3 introdotta nel capitolo 3:

> **Ipotesi 3 (H3):** i tweet vengono propagati maggiormente da parte di utenti che seguono l'autore del post rispetto a chi non lo segue.

Di seguito verranno descritte le analisi compiute per verificare tale ipotesi e verranno illustrati i risultati che si sono ottenuti.

4.4.1. Analisi dei dati aggregati

Dalle analisi effettuate sul dataset B contenente i post inglesi, utilizzato per verificare appunto tale ipotesi, sono emersi i seguenti risultati dove, su un totale di 36.186 retweet trovati risulta che:

- 17.441 retweet sono stati effettuati da utenti follower dell'account che ha prodotto il messaggio (48% del totale);

- 18.444 retweet risultano essere stati realizzati da utenti che, al tempo dell'analisi, non sono risultati essere né follower né following dell'account di cui hanno retweettato il messaggio (51% del totale);
- 11.729 retweet (non contati nei 36.186 totali) sono invece stati eseguiti da utenti che, al momento dell'analisi, risultano essere disabilitati/bannati da Twitter.

Analisi del tutto replicate sono state effettuate anche sul dataset B contenente i post italiani dove su un totale di 37.581 retweet trovati sono emersi i risultati seguenti:

- 17.859 retweet sono stati effettuati da utenti follower dell'account che ha prodotto il messaggio (47,5% del totale);
- 19.231 retweet risultano essere stati realizzati da utenti che, al tempo dell'analisi, non sono risultati essere né follower né following dell'account di cui hanno retweettato il messaggio (51% del totale);
- 11.181 retweet (non contati nei 37.581 totali) sono invece stati eseguiti da utenti che, al momento dell'analisi, risultano essere disabilitati/bannati da Twitter.

Successivamente dai dati aggregati ottenuti dai dataset in esame, sono state apportate delle suddivisioni che hanno riguardato il sentiment; si sono così creati i seguenti nuovi dataset, su cui poi sono state effettuate le analisi vere e proprie:

- retweet inglesi con sentiment negativo (D1);
- retweet inglesi con sentiment positivo (D2);
- retweet italiani con sentiment negativo (D3);
- retweet italiani con sentiment positivo (D4).

Dal dataset D1 sono state estrapolate le distribuzioni dei retweet effettuati rispettivamente da utenti follower, following, mutui ed utenti non follower.

Con il termine mutui si vuole intendere tutti quei retweet effettuati da utenti che sono allo stesso tempo follower e following dell'utente originale.

Per ogni distribuzione, attraverso il test di Kolmogorov-Smirnov, è stata osservata l'eventuale normalità; vengono riportati in una tabella i risultati del test e le statistiche descrittive di ogni distribuzione ottenuta dal dataset D1.

DATASET D1

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Follower	157	157	0	1013	7,34	80,792
Following	157	157	0	35	0,54	2,818
Mutui	157	157	0	35	0,50	2,816
Non Follower	157	157	0	1082	7,90	86,279

Tabella 7 - Statistiche descrittive dei 4 campioni (tweet inglese, sentiment negativo)

Distribuzione	Statistica Test	Significatività	Risultato
Follower	6,029	0,000	Rifiuta l'ipotesi nulla
Following	5,322	0,000	Rifiuta l'ipotesi nulla
Mutui	5,377	0,000	Rifiuta l'ipotesi nulla
Non Follower	6,273	0,000	Rifiuta l'ipotesi nulla

Tabella 8 - Risultati KS dei 4 campioni (tweet inglese, sentiment negativo)

Queste analisi oltre che per il dataset D1 sono state effettuate per ogni dataset in esame (D2, D3, D4) ottenendo così i seguenti risultati (elencati nello stesso ordine con cui sono stati citati):

DATASET D2

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Follower	721	721	0	1013	4,83	54,148
Following	721	721	0	35	0,48	1,475
Mutui	721	721	0	35	0,43	1,460

Non Follower	721	721	0	1082	5,06	57,250

Tabella 9 - Statistiche descrittive dei 4 campioni (tweet inglese, sentiment positivo)

Distribuzione	Statistica Test	Significatività	Risultato
Follower	12,472	0,000	Rifiuta l'ipotesi nulla
Following	10,029	0,000	Rifiuta l'ipotesi nulla
Mutui	10,316	0,000	Rifiuta l'ipotesi nulla
Non Follower	12,480	0,000	Rifiuta l'ipotesi nulla

Tabella 10 - Risultati KS dei 4 campioni (tweet inglese, sentiment positivo)

DATASET D3

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Follower	266	266	0	36	1,14	2,818
Following	266	266	0	10	0,47	0,887
Mutui	266	266	0	9	0,42	0,821
Non Follower	266	266	0	127	1,69	8,520

Tabella 11 - Statistiche descrittive dei 4 campioni (tweet italiano, sentiment negativo)

Distribuzione	Statistica Test	Significatività	Risultato
Follower	5,839	0,000	Rifiuta l'ipotesi nulla
Following	5,648	0,000	Rifiuta l'ipotesi nulla
Mutui	6,113	0,000	Rifiuta l'ipotesi nulla
Non Follower	6,874	0,000	Rifiuta l'ipotesi nulla

Tabella 12 - Risultati KS dei 4 campioni (tweet italiano, sentiment negativo)

DATASET D4

Distribuzione	N	N validi	Minimo	Massimo	Media	Deviazione std.
Follower	530	530	0	43	1,38	3,462
Following	530	530	0	4	0,50	0,694
Mutui	530	530	0	4	0,46	0,667
Non Follower	530	530	0	23	1,26	2,734

Tabella 13 - Statistiche descrittive dei 4 campioni (tweet italiano, sentiment positivo)

Distribuzione	Statistica Test	Significatività	Risultato
Follower	8,056	0,000	Rifiuta l'ipotesi nulla
Following	8,136	0,000	Rifiuta l'ipotesi nulla
Mutui	8,564	0,000	Rifiuta l'ipotesi nulla
Non Follower	8,686	0,000	Rifiuta l'ipotesi nulla

Tabella 14 - Risultati KS dei 4 campioni (tweet italiano, sentiment positivo)

Dato che i campioni non sono risultati approssimabili a distribuzioni gaussiane, è stato applicato, per ogni dataset trovato, un test Mann-Whitney per osservare se i campioni analizzati, di volta in volta, possano essere assimilati ad una stessa distribuzione; i confronti che sono stati effettuati riguardano separatamente:

- La distribuzione dei follower con quella dei non follower;
- La distribuzione dei follower con quella dei mutui;
- La distribuzione dei non follower con quella dei mutui;
- Ed, in ultimo, la distribuzione dei following con quella dei non follower.

Il test applicato alle distribuzioni sopra elencate ha evidenziato, per i quattro dataset in esame, i risultati riassunti nelle tabelle sottostanti.

DATASET D1

Distribuzione	Statistica Test Standardizzata	Significatività	Risultato
Follower vs Non Follower	-3,303	0,001	Rifiuta l'ipotesi nulla
Follower vs Mutui	-4,382	0,000	Rifiuta l'ipotesi nulla
Non Follower vs Mutui	-8,044	0,000	Rifiuta l'ipotesi nulla
Following vs Non Follower	-7,594	0,000	Rifiuta l'ipotesi nulla

Tabella 15 - Risultati MW (tweet inglese, sentiment negativo)

DATASET D2

Distribuzione	Statistica Test Standardizzata	Significatività	Risultato
Follower vs Non Follower	-1,841	0,066	Mantieni l'ipotesi nulla
Follower vs Mutui	-8,803	0,000	Rifiuta l'ipotesi nulla
Non Follower vs Mutui	-10,809	0,000	Rifiuta l'ipotesi nulla
Following vs Non Follower	-9,502	0,000	Rifiuta l'ipotesi nulla

Tabella 16 - Risultati MW (tweet inglese, sentiment positivo)

DATASET D3

Distribuzione	Statistica Test Standardizzata	Significatività	Risultato
Follower vs Non Follower	-0,362	0,717	Mantieni l'ipotesi nulla
Follower vs Mutui	-6,325	0,000	Rifiuta l'ipotesi nulla
Non Follower vs Mutui	-6,390	0,000	Rifiuta l'ipotesi nulla
Following vs Non Follower	-5,553	0,000	Rifiuta l'ipotesi nulla

Tabella 17 - Risultati MW (tweet italiano, sentiment negativo)

DATASET D4

Distribuzione	Statistica Test Standardizzata	Significatività	Risultato
Follower vs Non Follower	2,187	0,029	Rifiuta l'ipotesi nulla
Follower vs Mutui	-8,576	0,000	Rifiuta l'ipotesi nulla
Non Follower vs Mutui	-6,237	0,000	Rifiuta l'ipotesi nulla
Following vs Non Follower	-5,349	0,000	Rifiuta l'ipotesi nulla

Tabella 18 - Risultati MW (tweet italiano, sentiment positivo)

Come si può evincere dai risultati ottenuti, i campioni confrontati tra di loro hanno evidenziato la non appartenenza ad una stessa distribuzione; in particolare osservando le medie emerge come, per tutti i dataset in esame, risulta che:

- I tweet vengono maggiormente retwettati da utenti follower piuttosto che da utenti mutui;
- I tweet vengono maggiormente retweettati da utenti non follower piuttosto che da utenti mutui;
- I tweet vengono maggiormente retweettati da utenti non follower piuttosto che da utenti following.

Un discorso differente occorre farlo per i risultati restituiti dal test MW sul confronto tra i campioni follower ed non follower; è infatti evidente dagli esiti del test come i due campioni possano quasi essere assimilati ad una stessa distribuzione.

Emerge inoltre come, in media, i tweet vengano maggiormente retwettati da utenti non follower piuttosto che da utenti follower; occorre specificare che questa superiorità risulta comunque minima; questi risultati vanno quindi a contrastare l'ipotesi fatta.

5.0 Conclusioni

5.1. Introduzione

Questo capitolo ha lo scopo di riassumere e commentare i risultati ottenuti nel Capitolo 4.

Saranno inoltre illustrati i possibili sviluppi futuri di questo lavoro di tesi.

5.2. Conclusioni

In questo lavoro di tesi sono state dimostrate per via empirica le ipotesi introdotte nel capitolo 3.

I risultati ottenuti hanno evidenziato determinate caratteristiche interessanti; è infatti emerso come esistano delle peculiarità comuni tra i media tradizionali ed i social media.

Per quanto riguarda l'ipotesi sullo studio del sentiment è stato eseguito un test per l'identificazione delle distribuzioni dei campioni considerati. In particolare si è svolto un test di Kolmogorov-Smirnov il quale ha concluso che detti campioni non siano delle distribuzioni gaussiane. E' stato quindi applicato un test di Mann-Whitney sulle medie il quale ha rivelato come, nonostante i volumi assoluti dei tweet con sentiment positivo siano di gran lunga superiori a quelli con sentiment negativo (bias di positività), sono le notizie negative ad avere un maggiore coefficiente di propagazione. Tale risultato conferma quindi l'ipotesi H1.

Per quanto concerne lo studio sull'ipotesi riguardante le modalità di diffusione dei tweet in base agli argomenti che trattano, dopo aver osservato attraverso il test di Kolmogorv-Smirnov la non normalità dei campioni in esame, ottenuti dalla suddivisione del dataset di partenza in base alle dieci diverse categorie, è stato applicato ad essi un test di Kruskall-Wallis che ha evidenziato come le diverse

distribuzioni non siano assimilabili ad una stessa popolazione; questo risultato conferma quanto ipotizzato, ovvero che il contenuto dell'informazione di un determinato post influisce sulle modalità di diffusione di quest'ultimo.

Inoltre è emerso come i tweet scritti in lingua inglese che parlano di vita ed intrattenimento generale, inteso come divertimento, siano maggiormente retweettati; per quanto concerne i tweet italiani invece risulta che l'argomento che riscuote un maggior numero di retweet riguarda la vita notturna e la musica.

Nello studio dell'ultima ipotesi introdotto, quella riguardante la diffusione dei tweet confrontata con i diversi rapporti di amicizia di coloro che li hanno retwettati, è stata compiuta una analisi empirica sul dataset B. Sfruttando le API messe a disposizione da Twitter, è stata fatta una ricerca esaustiva su tutti gli account presenti nel campione al fine di verificare se i retweet del messaggio generato fossero fatti da nodi follower o meno.

I risultati hanno mostrato come solo circa il 48% dei retweet fossero fatti da follower del nodo generatore del messaggio mentre l'altro 51% fosse fatto da account completamente non coinvolti in nessun tipo di rapporto di amicizia con il nodo sorgente.

Tramite questo risultato è possibile dedurre come Twitter sia un mezzo fortemente utilizzato per la ricerca attiva di contenuti e, come gli account non si limitino ad agire passivamente retwettando e interessandosi solo ai contenuti proposti dagli account seguiti.

Tali considerazioni portano quindi a confutare l'ipotesi fatta.

5.3. Sviluppi Futuri

Un possibile sviluppo futuro riguarda l’estensione delle analisi svolte in questo lavoro di tesi, a più Social Network, Facebook in primis. Così facendo il campione risulterebbe molto più ampio, aprendo la strada a possibili nuovi spunti di riflessione e allo studio di nuove dinamiche di diffusione dei messaggi.

Un altro sviluppo interessante sarebbe quello di verificare ipotesi simili a quelle affrontate in questo lavoro di tesi considerando invece dei tweet e delle parole, i flussi generati dagli hashtag.

Un ultimo sviluppo proposto riguarda il fatto di saper esaminare e determinare, a monte di tutte le analisi effettuate, i cosiddetti tweet spam; in questo lavoro, peraltro, sono già state omesse dalle indagini tutte quelle coppie di tweet e retweet che sono risultati essere stati effettuati dallo stesso autore e che, dopo opportune analisi manuali, si sono dimostrati post pubblicitari del tutto inutili al fine della verifica delle ipotesi.

6.0 Bibliografia

[1] https://www.google.com/adplanner/?hl=it#siteSearch/

[2] http://socialnetwork.toweb.co/influencer-social-media-business/
http://www.mrassociati.it/perche-twitter-e-importante-per-chi-comunica/

[3] Enciclopedia Deagostini, Scienze umane - Sociologia - Generalità - *rete sociale*.

[4] Sauer, W. J. and Coward, R. T. (1985). *Social Support Networks and the Care of the Eldery.* NY: Springer Publishing Company.

[5] Lampe, C., Ellison, N., & Steinfeld, C. (2006). *A familiar Face(book): Profile elements as signals in an online social network*. In Proceedings of SIGCHI Conference on Human Factors in Computing Systems (pp. 435-444). San Jose, CA, USA: ACM.

[6] Kelly, R. (2009 August). Twitter Study. From PearAnalytics Blog: http://bit.ly/a9c8iE Krishnamurthy, B., Gill, P., & Arlitt, M. (2008). *A few chirps on Twitter*. In Proceedings of the 1st Workshop on Online Social Networks (pp. 19-24). Seattle, USA: ACM.

[7] Linton C. Freeman (1978/79). Social Network. *Centrality in Social Networks: Conceptual Clarification.*

[8] Brian Solis. (21 marzo 2012). *The Rise of Digital Influence: A "How-To" Guide for Businesses.*

[9] J. Weng, E. P. Lim, J. Jiang, and Q. He (2010). *Twitterrank: finding topic-sensitive influential twitterers*. pages 261–270. ACM.

[10] M. McPherson, L. Smith-Lovin and J.M. Cook (2001). *Birds of a feather: Homophily in social network*. Annal Review of Sociology, 27(1): 415-444.

[11] H. Kwak, C. Lee, H. Park, and S. Moon (2010). *What is twitter, a social network or a news media?*, 591 - 600. ACM.

[12] Cha, M., Haddadi, H., Benvenuto, F., & Gummadi, K. P. (2010*). Measuring user influence in Twitter: The million follower fallacy*. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (pp. 10-17). Association for the advancement of artificial intelligence.

[13] Romero, D. M., Asur, S., Galuba, W., & Huberman, B. A. (2010*). Influence and passivity in social media* . ACM.

[14] Statistiche sul comportamento e sulle dinamiche di retweet effettuate da Dan Zarrella (viral marketing scientist) (luglio 2009).

[15] https://dev.twitter.com/docs/using-search/

[16] Sobel, K. & Jansen, B. J. & Zhang, M. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* 60 (11), pp. 2169-2188.

[17] Mass Media e Controllo di Massa: "L'Agenda delle Cattive Notizie"; http://newapocalypse.altervista.org/blog/2012/09/03/mass-media-e-controllo-di-massa-lagenda-delle-cattive-notizie/

[18] Facebook Memology 2011, Blog ufficiale di Facebook. https://blog.facebook.com/blog.php?post=10150391956652131/

[19] https://dev.twitter.com/docs/

[20] La mappa dei social network nel mondo - giugno 2012 http://vincos.it/2012/06/11/la-mappa-dei-social-network-nel-mondo-giugno-2012/

[21] D. Barbagallo, L. Bruni, C. Francalanci, P. Giacomazzi, ENTER2012, *An empirical study on the relationship between Twitter sentiment and influence in the tourism domani*.

[22] http://www.olmr.it/2012/06/come-si-misura-la-social-influence/

[23] http://www.marketingo.it/following/

[24] http://it.paperblog.com/analisi-dell-audience-delle-fonti-d-informazione-su-twitter-1338442/

[25] PhD Thesis Barbagallo.

[26] http://www.ninjamarketing.it/2011/12/01/facebook-twitter-linkedin-e-youtube-luniverso-dei-social-media-in-un-video/

[27] Morise G, Massaro D. (2007) *Ricerca e percorsi di Analisi Dati con SPSS*. Pearson Education.

[28] Norusis MJ. (2008) *SPSS 16.0 advanced Statistical Procedures Companion*. Milano: McGraw-Hill.

[29] Norusis MJ. (2008) *SPSS 16.0 guide to data analysis*. Milano: McGraw-Hill.

[30] Emanuel Rosen. (2002) *The Anatomy of Buzz: How to Create Word of Mouth Marketing*.