POLITECNICO DI MILANO

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

Studio empirico sulla relazione tra influence e contenuto multimediale su Twitter

Relatore: **Prof.ssa Chiara Francalanci**
Correlatori: **Ing. Leonardo Bruni**
Dott.ssa Fiamma Petrovich

Tesi di laurea Specialistica di:
Marco Longhitano
Matr. 739450

Anno accademico 2011/2012

Indice

Indice delle figure

Sommario

Questo lavoro di Tesi si inserisce all'interno di un progetto riguardante la gestione della reputation online e sentiment analysis condotto dal Politecnico di Milano per il Comune di Milano. Propone uno studio empirico della relazione tra propagazione dei messaggi in Twitter ed il loro contenuto.

Obiettivo finale è capire come tale contenuto possa influenzare la diffusione del messaggio in rete, anche in rapporto al sentiment espresso.

Lo studio è stato effettuato sia da un punto di vista quantitativo, comparando i volumi di retweet di post contenenti video con post contenenti foto, sia da un punto di vista di dinamiche temporali, considerando non solo il numero di retweet ma anche gli istanti (espressi in secondi) intercorsi tra un post e i propri retweet.

In particolare, sono state formulate sei ipotesi relative alla viralità dei messaggi in relazione al loro contenuto multimediale e sono state validate su un dataset di circa 1,5 milioni di tweet relativi alle città di Milano, Berlino, Londra e Madrid postati nel mese di Luglio 2011. I risultati ottenuti supportano le ipotesi formulate e dimostrano come il contenuto multimediale assume un ruolo cruciale nel determinare la viralità di un messaggio.

1.0 Introduzione

«*The Web is more a social creation than a technical one*»

Tim Berners-Lee
inventore del World Wide Web

I social network hanno avuto un notevole impatto nella vita di tutti i giorni. Basti pensare all'acquisto di prodotti, alla ricerca di aiuto e risposte, tutte attività che è ormai consuetudine compiere online. Sono cambiate radicalmente le modalità di interazione tra utenti, la ricerca e la condivisione di informazioni. Il risultato è un'infinita mole di dati per nulla strutturata, che necessita di opportune operazioni di *data mining* per acquistare utilità ed estrapolare conoscenza.

Anche dal punto di vista delle aziende il risultato non cambia: immagine, dialogo con i potenziali clienti, riduzione dei costi... Risulta quindi necessario sviluppare una solida strategia in campo "social", con l'obiettivo di gestirne effetti e sfruttarne le opportunità, in maniera da utilizzarli proficuamente sotto diversi aspetti, *decision making* in primis.

Questo lavoro di Tesi, in aggiunta allo stage svolto presso la società di consulenza e marketing CommStrategy, si inserisce all'interno di un progetto di gestione della reputation online e sentiment analysis condotto dal Politecnico di Milano per il Comune di Milano. Propone uno studio empirico sul comportamento e la propagazione dei messaggi provenienti dall'ormai noto servizio di microblogging Twitter in relazione con il loro contenuto.

Obiettivo finale è quello di capire come tale contenuto possa influenzare la diffusione del messaggio in rete, anche in relazione al sentiment espresso. I tweet contenenti dei link, o più precisamente foto e video, provocano naturalmente reazioni diverse rispetto ai tweet con solo testo. Scopo di questa Tesi è dunque anche misurare queste reazioni, in maniera da capire l'efficacia di questi contenuti multimediali da un punto di vista di marketing e non solo.

La struttura della Tesi è la seguente:

- Il Capitolo 2 espone una panoramica della letteratura riguardo l'argomento trattato, introduce i concetti base della teoria dei grafi, le metriche più significative nello studio dei nodi, una breve descrizione dei principali social network ed un approfondimento su Twitter, un'introduzione al concetto di influence e si conclude con una rapida discussione sul gap esisistente in letteratura su quanto trattato.
- Il Capitolo 3 espone in maniera molto generale il progetto sviluppato dal Politecnico di Milano per il Comune, di cui questa Tesi è parte integrante.
- Nel Capitolo 4 vengono proposte le metodologie di analisi utilizzate in questo lavoro di Tesi, descrivendo il software e gli strumenti usati. Sono inoltre presentate le ipotesi che verranno poi dimostrate con dei risultati sperimentali nel capitolo successivo.
- Il Capitolo 5 riporta i dati sperimentali raccolti durante tutto il lavoro, utili a dimostrare le ipotesi, commentandone i vari test statistici effettuati per validare il risultato.
- Infine, il Capitolo 6 mostra una panoramica conclusiva, analizzando i risultati ottenuti e descrivendo eventuali possibilià di ulteriori sviluppi futuri.

2.0 Stato dell'arte

2.1. Introduzione

La comunicazione, nell'era del Web, si manifesta sotto svariate forme, dai siti ai blog, dai social network ai forum, ed il Web stesso è diventato un prezioso luogo d'incontro nel quale esprimere e soprattutto condividere le proprie opinioni.

Internet ha inoltre permesso una comunicazione bidirezionale tra le persone e le aziende, accorciando le distanze tra esse. Da un lato è possibile raggiungere un numero notevole di potenziali clienti rendendo noti i propri prodotti, dall'altro si possono esprimere liberamente i propri giudizi, portandoli a conoscenza dell'intera community, quindi anche delle stesse aziende.

Un esempio lampante può essere rappresentato da Twitter, il popolare social network che grazie all'utilizzo del meccanismo di *retweet*, permette ad una vasta rete di utenti di manifestare un eventuale proprio dissenso, ponendolo all'attenzione di una folta schiera di persone.

Altro esempio degno di nota, è quello di *eBay,* il famosissimo portale di e-commerce che ha inventato il sistema dei feedback, uno dei primi meccanismi di reputation apparsi su Internet. È un sistema facile da usare, che fornisce un minimo di garanzia anche nel caso di transazioni effettuate con sconosciuti [31]. Molti studi ne hanno dimostrato pregi e difetti, la sua incredibile diffusione ha dato l'opportunità di analizzare i comportamenti delle persone su vasta scala.

L'avvento del Web 2.0 ha dunque incrementato il potere dell'utente, che adesso ricopre un ruolo più attivo, grazie all'immediata e molto più semplice condivisione dei contenuti rispetto al passato. Un effetto amplificatore lo ha avuto l'esplosione di vari social network, forum e blog, che permettono a chiunque di condividere la propria user experience relativa ad un prodotto, facendo diventare il Web un luogo in cui è possibile valutare l'oggetto e non solo acquistarlo, come invece accadeva in principio.

In questo capitolo si tratterà dello stato dell'arte riguardo ai principali argomenti esposti, al fine di acquisire un quadro complessivo delle tecniche proposte e dunque ricavare un termine di paragone rispetto a quelle passate o attuali. Molte delle tematiche in questione richiederebbero maggiori approfondimenti vista la loro complessità, ma essendo altro lo scopo della Tesi, ci si limiterà a porre l'attenzione sugli aspetti maggiormente legati a questo lavoro.

In particolare, verrà dapprima fatta una rapida panoramica sulle reti sociali e ovviamente sui social network online, presentandone i principali soggetti (Twitter nello specifico). Poi si passerà ad un'attenta analisi di quella particolare categoria di persone chiamate *influencer*, determinanti dal punto di vista del marketing e non solo. Infine, si tratterà del gap attualmente esistente in letteratura, in merito ai principali temi della Tesi.

2.2. Le reti sociali

*«Una **rete sociale** (in inglese **social network**) consiste di un qualsiasi gruppo di persone connesse tra loro da diversi legami sociali, che vanno dalla conoscenza casuale, ai rapporti di lavoro, ai vincoli familiari» (Wikipedia).*

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

Quest'analisi permette sia di scomporre i comportamenti individuali sia, sfruttando il contesto sociale, di spiegare i comportamenti di gruppo.

Lo studio si basa su costrutti teorici tipici della sociologia e della matematica, come ad esempio la teoria dei grafi, che consentono di visualizzare ed evidenziare la struttura delle relazioni sociali.

Lo sviluppo dell'analisi delle reti sociali comincia dagli anni Trenta, in cui nacquero i primi approcci sistematici, volti a spiegare le cause di determinati fenomeni sociali. Grazie a Georg Simmel si diffuse un primo approccio sistematico. La sua teoria spiegava le cause dei fenomeni sociali e contribuì alla sociologia formale che può essere considerata la progenitrice dell'analisi delle reti sociali.

Fondamentale fu il contributo di Jacob Levi Moreno, il fondatore della sociometria, scienza che analizza le relazioni interpersonali. Fu infatti il primo, nel 1934, a descrivere matematicamente una rete sociale, mediante la teoria dei grafi, in cui i nodi rappresentano gli individui e gli archi le relazioni sociali. Grazie soprattutto all'introduzione dei collegamenti direzionali tra i nodi, si potevano spiegare pattern sociali con un grado di complessità più elevato rispetto ai risultati raggiunti sino a quel momento. Il collegamento tra nodi può essere indifferentemente direzionale o meno. Il primo caso è quello di Twitter, in cui un utente può essere legato ad un altro (un follower) ma senza alcuna relazione in senso opposto (ovvero non seguito a sua volta dal follower). Il secondo invece è quello di Facebook, dove sussiste una relazione bidirezionale tra due utenti (possono solo essere collegati oppure no), ossia un arco non orientato.

In seguito si formarono due filoni di ricerca separati: quello americano ad Harvard, l'altro britannico a Manchester. Il primo si occupava di tecniche per l'individuazione e lo studio di sottogruppi di persone all'interno di gruppi originari più ampi, con lo scopo di analizzare e comprendere i rapporti tra i sottogruppi stessi. Da qui, ha origine il cosiddetto *approccio sociocentrico,* in cui si preferisce analizzare la forma delle reti piuttosto che il loro contenuto. Si pensa che la forma delle relazioni sociali ne determini ampiamente il contenuto.

Tale metodo è stato illustrato brevemente in [6].

La seconda scuola di pensiero invece diede origine all'*approccio egocentrico,* più focalizzato sulle relazioni sottostanti agli individui, per poi generalizzarne le caratteristiche.

Del gruppo di Harvard faceva parte Milgram, che nel 1967 formalizzò la teoria dei *six degrees of separation* [23], in cui sosteneva e tentava di dimostrare la tesi del "mondo piccolo". Secondo lo studioso, ogni persona può essere collegata ad un'altra attraverso una rete di conoscenze che non supera i cinque intermediari. L'esempio empirico fornito fu quello della popolazione statunitense, in cui secondo Milgram, tutti gli individui erano connessi tra loro grazie (in media) a sei legami di conoscenza. Per provarlo, lo scienziato selezionò casualmente un gruppo di persone residenti nel Midwest e chiese loro di mandare un pacchetto ad una persona sconosciuta del Massachusetts, ossia a migliaia di chilometri di distanza. Conoscendo soltanto il

nome, la zona di residenza e l'occupazione del destinatario, ognuno degli individui selezionati mandò il pacco ad una persona conosciuta che, a proprio giudizio avesse le maggiori possibilità di conoscere il destinatario, così via fino alla consegna del pacco. Il risultato fu stupefacente, perché dopo soli cinque passaggi (in media) il pacchetto raggiunse la meta finale.

Quest'approccio fu poi ripreso da Watts, che confermò il risultato di Milgram grazie ad un esperimento in cui utilizzò un'email al posto del pacchetto da recapitare [33].

Un recente studio [18] ha rivelato che su Twitter, i sei gradi di separazione coprono il 98% delle relazioni sociali. Mentre Microsoft, analizzando i log di alcune conversazioni del suo *MSN*, ha ricavato una media pari a 6.6 collegamenti [22].

Attualmente la teoria delle reti è utilizzata per studiare l'influenza della struttura della rete sociale su un gruppo di lavoro e la gestione della conoscenza al loro interno. La gestione della conoscenza (*knowledge management*) comprende un insieme di tecniche volte ad identificare, creare, rappresentare e ridistribuire la conoscenza all'interno delle organizzazioni. Due sono i principali filoni di studio: uno riguardante l'applicazione di sistemi di supporto e l'altro relativo alle comunità. La prima soluzione studia la codifica e la distribuzione della conoscenza tramite l'*information and communication technology*, come i database e internet; invece la seconda, sostiene lo scambio di conoscenza tra persone con interessi e obiettivi comuni.

Da queste ricerche sono emersi risultati interessanti, come il fatto che chiedere consigli ai propri parigrado sia un importante canale che permette uno scambio di conoscenze molto specifiche. A tal proposito, alcuni degli studi in questo contesto si stanno focalizzando sulla relazione tra i benefici del *knowledge management* che derivano dalla struttura delle reti sociali corrispondenti, come illustrato in [25].

2.2.1. Cenni sulla teoria dei Grafi

In generale una rete sociale è composta da un insieme di attori (punti, o nodi, o agenti) che hanno relazioni gli uni con gli altri. Dato un insieme finito U di elementi:

$$U = \{X_1, X_2, \ldots, X_n\}$$

e un numero finito di relazioni R:

$$R_t \subseteq U \times U$$

Con

$$t = 1, 2, \ldots, r$$

si definisce rete sociale *N*, la *n*-upla composta dall'insieme finito di elementi U e da *(n−1)* relazioni tra di essi:

$$N = (U, R_1, R_2, \ldots, R_r)$$

Le relazioni rappresentano un legame qualunque tra due entità, ad esempio l'amicizia tra due persone e possono godere delle seguenti proprietà:

1. *riflessiva*: $\forall X \in U : XRX$
2. *irriflessiva*: $\forall X \in U : \neg XRX$
3. *simmetrica*: $\forall X,Y \in U : (XRY \Rightarrow YRX)$
4. *asimmetrica*: $\forall X,Y \in U : \neg(XRY \wedge YRX)$
5. *antisimmetrica*: $\forall X, Y \in U : (XRY \wedge Y RX \Rightarrow X = Y)$
6. *transitiva:* $\forall X, Y, Z \in U : (XRY \wedge Y RZ \Rightarrow XRZ)$
7. *intransitiva:* $\forall X, Y, Z \in U : (XRY \wedge Y RZ \Rightarrow \neg XRZ)$

Inoltre una relazione è detta di *ordinamento parziale*, se gode delle seguenti proprietà: riflessiva, antisimmetrica e transitiva.

Tale relazione è usata per descrivere i rapporti gerarchici all'interno di una organizzazione.

Mentre si dice che una relazione è di *equivalenza* se risulta riflessiva, simmetrica e transitiva. È utile nel modellare il caso in cui tutti gli elementi della rete sociale hanno lo stesso ruolo.

Una rete definita tramite una relazione R può essere rappresentata in modi diversi:

a) Tramite la matrice binaria R = $[r_{ij}]$ di dimensioni $n \times n$, detta anche *matrice di adiacenza*, dove:

$$r_{ij} = \begin{cases} 1 & se \quad X_i R X_j \\ 0 & altrimenti \end{cases}$$

nel caso di archi pesati, il valore di r_{ij} può essere il numero reale che indica il grado di legame R tra X_i e X_j.

b) Tramite la lista dei vicini, ovvero ciascun nodo possiede una lista dei nodi a lui prossimi.

c) Tramite un grafo $G = (V,L)$, dove V è l'insieme dei vertici (le unità della rete) e $L = \bigcup_{i=1}^{r} L_i$ rappresenta l'insieme degli archi che indicano le relazioni.

Inoltre, in base alla tipologia di connessione tra i nodi di un grafo, è possibile distinguere:

- *rete non direzionata*: la relazione *R* non è simmetrica, si verifica quando tutti gli archi non hanno una direzione specifica, come per esempio nella relazione tra due utenti di Facebook ;

- *rete direzionata*: la relazione *R* è simmetrica, ossia tutti gli archi sono orientati, come ad esempio nella relazione tra due utenti Twitter;

- *rete mista*: nella stessa rete si possono avere sia archi direzionati che non direzionati, per esempio quando si rappresentano due o più relazioni, come quella di matrimonio e paternità;

- *rete a due modi*: è formata da due insiemi di unità $U = U_a \cup U_e$ e una relazione *R* che connette i due insiemi, per esempio l'appartenenza di una persona ad una associazione.

2.2.2. Metriche di analisi delle reti sociali

L'obiettivo di questo paragrafo è descrivere le metriche per l'analisi delle reti sociali maggiormente utilizzate in letteratura.

- **Dimensione della rete:** ottenuta contando il numero di nodi presenti e definita come:

$$SIZE\ (N) = |U|$$

Parametro semplice ma allo stesso tempo fondamentale per lo studio delle reti.

- **Metriche di coesione:**
  - La *densità* di una rete è definita come il rapporto tra il numero di legami presenti effettivamente nella rete e quello massimo possibile, cioè:

$$DENSITY\ (N) = \frac{2l}{n(n-1)}$$

Parametro che permette di analizzare alcuni fenomeni come la velocità di diffusione dell'informazione tra i nodi o il grado di legame tra essi. Una rete molto densa permette di accedere in modo più efficiente all'informazione, grazie all'alto numero di legami tra i suoi nodi che quindi possono raggiungere più facilmente gli altri membri della rete.

  - Una variante della densità è la *sparsità*, definita come:

$$SPARSITY\ (N) = 1 - \frac{2l}{n(n-1)}$$

Questa metrica può essere utilizzata come indicatore nel processo di costruzione di un'intera rete sociale partendo da un suo sottoinsieme [20], utile nel caso in cui non vi siano dati sufficienti per studiare un determinato fenomeno.

Un attore a_j si dice raggiungibile da un attore a_i se esiste almeno un percorso p_{ij} che porta da a_i a a_j. Se le relazioni non sono simmetriche e quindi gli archi sono direzionati, è possibile che l'attore A sia raggiungibile dall'attore B, ma l'attore B non sia raggiungibile dall'attore A.

Un altro indicatore utilizzato per valutare il grado di coesione della rete è la *raggiungibilità*, utile per individuare se ci sono attori che non sono collegati con altri e quindi se ci sono esempi di sotto-gruppi e divisioni all'interno della rete.

2.2.3. Metriche di posizionamento dell'utente

È possibile utilizzare differenti metriche per misurare l'effettiva collocazione di un utente all'interno di un network [24]. Il posizionamento è una delle caratteristiche chiave che permette di definire l'importanza e dunque l'influenza, di un nodo in una rete.

Di seguito verranno descritte esclusivamente le metriche utilizzabili in un grafo costituito da nodi ed archi e in cui sono presenti legami direzionali.

Le relazioni di questo tipo tra utenti permettono di definire due posizioni principali:

- *Prestigio;*
- *Centralità.*

Per quanto concerne il *prestigio*, possiamo dire che un membro appartenente ad un network può essere considerato di prestigio quando sono presenti numerosi legami uscenti dagli altri utenti e diretti al nodo in questione.

Tra le varie metriche che permettono una misurazione del prestigio riportiamo:

- *Indegree centrality:* si basa sul numero di connessioni in ingresso ad un nodo, limitandosi a valutare esclusivamente il primo livello di vicinanza. Un utente sarà considerato tanto più interessante, quanto più verrà nominato dagli altri membri del network. La formula per il calcolo di questa metrica è:

$$IDC(x) = \frac{i(x)}{m-1}$$

  dove:

  - $i(x)$ indica il numero di membri della community adiacenti all'utente x;
  - m rappresenta il numero totale di membri.

- *Proximity prestige:* esprime la vicinanza dei membri ad un nodo x. Si basa sulla distanza geodesica, $d(x, y_i)$, ossia la distanza di tutti gli utenti y_i rispetto ad un utente x. La sua formula è la seguente:

$$PP(x) = \frac{\frac{p(x)}{m-1}}{\frac{1}{p(x)}\sum_{i=1}^{p(x)} d(x, y_i)} = \frac{p(x)^2}{(m-1)\sum_{i=1}^{p(x)} d(x, y_i)}$$

dove:

p(x) – indica il numero dei membri y_i appartenenti alla rete e in grado di raggiungere l'utente x;

m – rappresenta il numero totale di membri della rete.

Invece, per quel che riguarda le metriche di *centralità,* bisogna dire che esse permettono di individuare gli utenti che sono coinvolti in relazioni particolarmente rilevanti con gli altri utenti della community.

Si deve inoltre precisare che, questo tipo di misura si effettua su dei grafici in cui non importa se l'utente è destinatario o fonte dell'informazione.

Le metriche analizzate per la centralità sono le seguenti:

a) *Outdegree centrality:* misura il numero di archi che vanno dal nodo *x* verso gli altri nodi. Si calcola come:

$$ODC(x) = \frac{o(x)}{m-1}$$

con $o(x)$che indica il numero di nodi adiacenti a x (solo primo ordine di vicinanza) mentre *m* rappresenta il numero totale dei membri all'interno del network. Maggiori saranno le comunicazioni con gli altri utenti, maggiore sarà ODC;

b) *Eccentricity centrality*: indica il nodo più centrale della rete, ossia quello che ha distanza minima con tutti gli altri nodi del network:

$$EC(x) = \frac{1}{\max\{d(x, y) : y \in M\}}$$

$d(x, y)$ è la lunghezza del path più corto che collega *x* e *y*, invece M è l'insieme di tutti i nodi della rete.

c) *Closeness centrality*: al contrario della *proximity prestige*, rappresenta la vicinanza di un utente rispetto agli altri nodi della rete. Un utente viene

considerato “centrale” quando è in grado di raggiungere velocemente gli altri nodi, dunque chi ha tale valore elevato sarà anche indicato come un buon propagatore di informazioni. Si calcola come:

$$CC = \frac{m-1}{\sum_{y \neq x, y \in M} c(x,y)}$$

con $c(x,y)$ funzione che descrive la distanza tra i nodi *x* e *y* ed *M* che è il solito set dei membri del network.

d) *Betweenness centrality*: misura la centralità di un utente in base alla struttura della rete. Quindi un alto valore di tale metrica è un buon indice di come l’utente possa diffondere rapidamente le informazioni lungo la rete. È infatti calcolata come:

$$BC = \frac{\sum_{i \neq x \neq j;\, j \in M} b_{ij}(x)}{b_{ij}(x)}$$

dove $b_{ij}(x)$ indica appunto il numero di percorsi più brevi che vanno da *i* a *j* e passanti per *x*, mentre b_{ij} il numero di percorsi più brevi da *i* a *j*. *M* è il set di tutti gli utenti presenti nella rete.

Dopo aver elencato le metriche e le rispettive formule, vengono di seguito presentati in una tabella (Tabella 1), i vantaggi e gli svantaggi dell’utilizzo di ciascuna misura:

Metrica	**Vantaggi**	**Svantaggi**
IDC	• Semplice da calcolare; • Esaustiva in molti casi.	• Misura solo locale, considera solo conessioni di primo ordine; • Elevato numero di duplicati.
PP	• Prende in esame l’intera topologia della rete;	• Complessa ed inefficiente per reti di grandi dimensioni • Nei grafi disconnessi, viene assegnato 0 ad ogni nodo.
ODC	• Calcolo semplice da effettuare; • Esaustiva in molte applicazioni.	• Elevato numero di duplicati; • Considera solo connessioni primo livello.

EC	• Metrica globale; • Considera l'intera topologia della rete.	• Complessa ed inefficiente per reti di grandi dimensioni • Nei grafi disconnessi, viene assegnato 0 ad ogni nodo.
CC	• Metrica globale; • Considera l'intera topologia della rete.	• Complessa ed inefficiente per reti di grandi dimensioni • Nei grafi disconnessi, viene assegnato 0 ad ogni nodo.
BC	• Metrica globale; • Considera l'intera topologia della rete.	• Complessa ed inefficiente per reti di grandi dimensioni • Nei grafi disconnessi, viene assegnato 0 ad ogni nodo.

Tabella 1 - Riassunto metriche di posizionamento utente

2.2.4. Classificazione degli utenti

Di seguito verranno definiti degli archetipi di utente che differiscono in base ai fattori di dimensione e qualità della rete. Ognuno di essi è in grado di diffondere diversi tipi di messaggio a vari gruppi di utenti. In particolare, La categorizzazione utilizzata da Forrester [15] definisce le seguenti tipologie di utenti:

- **Fonte (source)**: Indica il nodo da cui parte l'informazione. Questa tipologia di individuo potrebbe anche avere un numero di collegamenti inferiore ad altri influencer, ma possiede un elevato livello di autorità su una determinata area tematica. Gli altri utenti del network spesso scoprono le informazioni dal nodo fonte prima di diffonderlo all'interno delle loro rispettive reti.

- **Ragno (spider)**: L'utente identificato come spider, è in grado di raggiungere una larga scala di individui grazie ad un elevata quantità di connessioni. Alcune delle quali sono a loro volta appartenenti ad una rete di collegamenti

molto estesa e contribuiscono ad una diffusione veloce ed ampia dei messaggi. Attraverso il suo robusto network sociale, questo tipo di influencer assume il ruolo di catalizzatore nella propagazione virale dell'informazione.

- **Sole (sun):** rappresenta un utente che ha il più elevato numero di legami diretti di primo ordine, ma proprio a causa di questa immensa quantità di connessioni, la robustezza relativa del suo network è tendenzialmente bassa.

La figura seguente (Figura 1) mostra una rappresentazione grafica di quanto detto sopra:

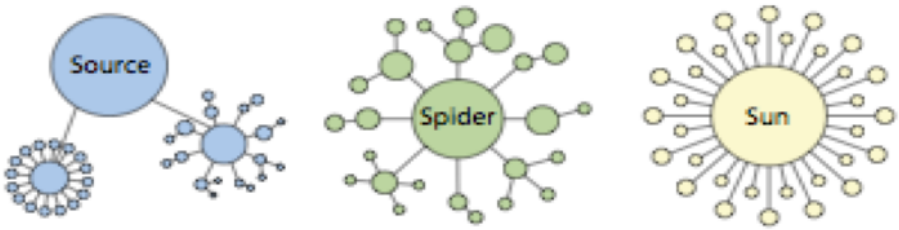

Figura 1 - Rappresentazione degli archetipi utente.

2.3. WOM ed eWOM

Negli anni '50 iniziarono i primi studi riguardanti il flusso delle informazioni nelle reti sociali. Cominciò a farsi strada la teoria secondo cui questo flusso di conoscenza ed opinioni venisse generato dai mass media, per poi passare agli opinion leaders ed infine da questi alle componenti meno attive della popolazione, attraverso una comunicazione definita a "due passi" (two-step communication flow), in cui il Word-Of-Mouth (passaparola) assumeva un ruolo fondamentale [14]. Proprio quest'ultimo, nella sua forma tradizionale (offline), è stato dimostrato come riesca ad essere un elemento primario e decisivo nelle rete sociali nel definire le decisioni di acquisto dei consumatori [27]. Inoltre ha un effetto diretto sulla diffusione dell'innovazione e di nuovi prodotti [34].

Nella tabella seguente (Tabella 1) è presente un breve riassunto delle motivazioni, identificate dalla letteratura, che spingono alle comunicazioni con il metodo del passaparola [11].

Autori	Motivazioni	Descrizione
Dichter	*Product involvement*	Il consumatore è coinvolto fortemente con il prodotto; la tensione causata dall'esperienza di consumo viene ridotta consigliando il prodotto ad altri individui.
	Self-involvement	Il prodotto rappresenta un mezzo attraverso il quale chi produce il messaggio può gratificare certi bisogni emozionali.
	Other involvement	L'attività di passaparola risponde al bisogno di dare qualcosa al ricevente del messaggio (ascoltatore).
	Message involvement	Si riferisce alla discussione che è stimolata dall'advertising o dalle pubbliche relazioni.
Engel, Blackwell & Miniard	*Involvement*	Il livello di interesse e coinvolgimento rispetto ad un topic serve a

		stimolare la discussione.
	Self-enhancement	Esprimere opinioni e raccomandazioni permette all'individuo di guadagnare attenzione, dare l'impressione di possedere informazione di valore e affermare una certa superiorità.
	Concern for others	Il desiderio di aiutare un amico o un parente nel fare una migliore decisione di acquisto.
	Message intrigue	Gradimento derivante dalla trattazione dell'advertising o dall'appeal della vendita.
	Dissonance reduction	Riduzione dei dubbi seguendo una decisione di acquisto condivisa da più individui.
Sundaram, Mitra & Webster	*Altruism (positive WOM)*	L'atto del fare qualcosa per gli altri senza ricevere alcuna forma di guadagno in cambio.
	Product involvement	Interesse personale nel prodotto, eccitazione derivante dal possesso e

		dall'utilizzo.
	Self-enhancement	Promozione della propria immagine con gli altri individui mostrandosi un compratore intelligente.
	Altruism (negative WOM)	Prevenire gli altri utenti da una cattiva esperienza di consumo del prodotto.
	Anxiety reduction	Ridurre rabbia, frustrazione e ansia.
	Vengeance	Cercare ri contro la compagnia associata all'esperienza di consumo negativa.
	Advice Seeking	Ottenere consigli su come risolvere problemi.

Tabella 2 - Le motivazioni del passaparola individuate in letteratura

Durante i primi anni del 2000, grazie all'esplosione dei social network come fenomeno di massa, l'online word-of-mouth, o eWOM (ossia electronic word-of-mouth), è diventato uno degli ambiti di ricerca principali all'interno della cosiddetta comunicazione *computer-mediated*, in particolare nel contesto delle comunicazione consumer to consumer [36].

L'eWOM è definito come «una qualsiasi affermazione positiva o negativa fatta da un consumatore potenziale, attuale o passato riguardo a un prodotto od un

brand, la quale è resa accessibile ad un moltitudine di persone ed istituzioni attraverso il Web» [11].

È internet a rendere più interessante l'eWom, grazie alle sue caratteristiche intrinseche. A differenza del contesto tradizionale, in cui l'interazione avviene attraverso il linguaggio parlato e vis-à-vis, l'eWOM si basa sulla trasmissione di opinioni ed esperienze in forma scritta.

È possibile individuare diverse ragioni per cui tali opinioni possano essere considerate dotate di un potere di diffusione e di capacità di influenza superiore rispetto al parlato:

- bidirezionalità dell'interazione: i social media permettono dei feedback ed uno scambio di opinioni in real time tra la fonte e il ricevente del messaggio [7];
- le comunicazioni online abilitano la possibilità precedentemente sconosciuta di connettere gli individui sia in modo sincrono (e.g. via istant messaging) che asincrono (e.g. via email) [35];
- velocità, convenienza e audience del messaggio, assenza della pressione del faccia a faccia: con uno sforzo oggettivamente inferiore abilitato dalle tecnologie Web, la scala di utenti raggiungibile è estesa in modo non paragonabile a quella con cui è possibile entrare in contatto in una comunicazione con un qualsiasi media tradizionale [26].

Proprio per questi motivi i servizi di social networking sono stati riconosciuti come un'importante fonte di informazione e scambio di opinioni in grado di influenzare in modo riconoscibile l'adozione e l'utilizzo di prodotti e servizi [35].

Il pericolo però, quando si affrontano queste tematiche è quello di far confusione tra diversi termini, a volte utilizzati come sinonimi. Di seguito i tre principali riguardanti la diffusione di un contenuto:

- *viral*: si intende il meccanismo di propagazione del messaggio, che come un virus colpisce gli individui che inglobano in sé questo batterio per poi contagiare, proprio come può avvenire per un raffreddore, i più prossimi e stretti parenti, conoscenti o amici. La prossimità, non necessariamente fisica o familiare ma anche emotiva, intellettuale o "tribale" in quanto membri di

una comunità o network, è un elemento indispensabile per l'attivazione del passaparola;

- *buzz*: questa espressione onomatopeica indica l'effetto "sonoro" che si verifica quando si sparge la voce e tutti cominciano a parlare di un argomento. Rappresenta la gente che racconta e conversa con il proprio "vicino" su qualcosa di particolarmente interessante;
- *word-of-mouth*: come già descritto in precedenza, è tradizionalmente definito come il processo di trasporto dell'informazione da persona a persona, rappresenta il mezzo legato al messaggio che si propaga tramite la parola, una comunicazione, verbale o telematica, che si trasmette da un individuo all'altro.

È ormai evidente che il passaparola sia indicato come la forma di comunicazione che più di ogni altra influenza la decisione di acquisto. Sostanzialmente ciò è dovuto al fatto che il passaparola riduce il rischio di scegliere un prodotto non conforme alle aspettative.

La difficoltà vera e propria, soprattutto per le aziende, è quella di attribuire una valutazione (numerica se possibile) a questo fenomeno, rendendolo in qualche modo misurabile.

McKinsey ha recentemente proposto una metrica utilizzabile per misurare gli effetti del passaparola, chiamata word-of-mouth *equity* [4]. Questo indicatore rappresenta l'impatto medio sulle vendite di un messaggio proveniente dal brand, moltiplicato dal numero di messaggi prodotti dal passaparola. L'obiettivo è di porre la giusta attenzione sull'impatto e sul volume di questi messaggi, in modo da permettere all'analista del marketing di valutarne in modo efficace l'effetto sulle vendite.

L'elemento chiave che determina l'impatto del word-of-mouth è il contenuto del messaggio, che deve riguardare la descrizione delle peculiarità primarie del prodotto/servizio affinché abbia un effetto significativo su altri potenziali clienti. Altro elemento critico è l'autorevolezza dell'individuo che diffonde il messaggio, poiché chi lo riceve deve avere fiducia nel mittente.

Concludendo, è necessario esaminare anche il contesto in cui il passaparola viene generato, al fine di capire le potenzialità di diffusione.

In generale, se la rete è ridotta e i nodi sono in un rapporto di elevata fiducia reciproca, il contenuto avrà una *reach* minore, ma allo stesso tempo un impatto maggiore se confrontato alla circolazione all'interno di una rete vasta e dispersiva. Questo fenomeno è in parte dovuto al legame rilevante che sussiste tra gli individui che godono di maggiore considerazione e fiducia all'interno di una community.

2.4. Online Social Network

Un social network, come sottolineato in precedenza, consiste in un gruppo di persone tenute insieme da diversi tipi di legami, che vanno dai rapporti di lavoro a quelli familiari, o semplicemente casuali.

Dunque la rete sociale storicamente è, in primo luogo, una rete fisica. Con l'avvento del Web, si è trasformata in un luogo d'incontro virtuale molto popolare, che permette agli individui 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.

Dunque, le principali attività all'interno di un social network, sono quelle di creare contenuti (video, messaggi, immagini, ecc..) e quelle relative al consumo degli stessi generati da altri utenti.

Ogni tipo di scambio di risorse è considerato uno scambio di relazione sociale, ovvero un legame. Il grado di forza o debolezza di tale legame, è principalmente dovuto alla frequenza, al numero o al tipo di risorse condivise [21].

Sauer e Coward [32] 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 2: 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.

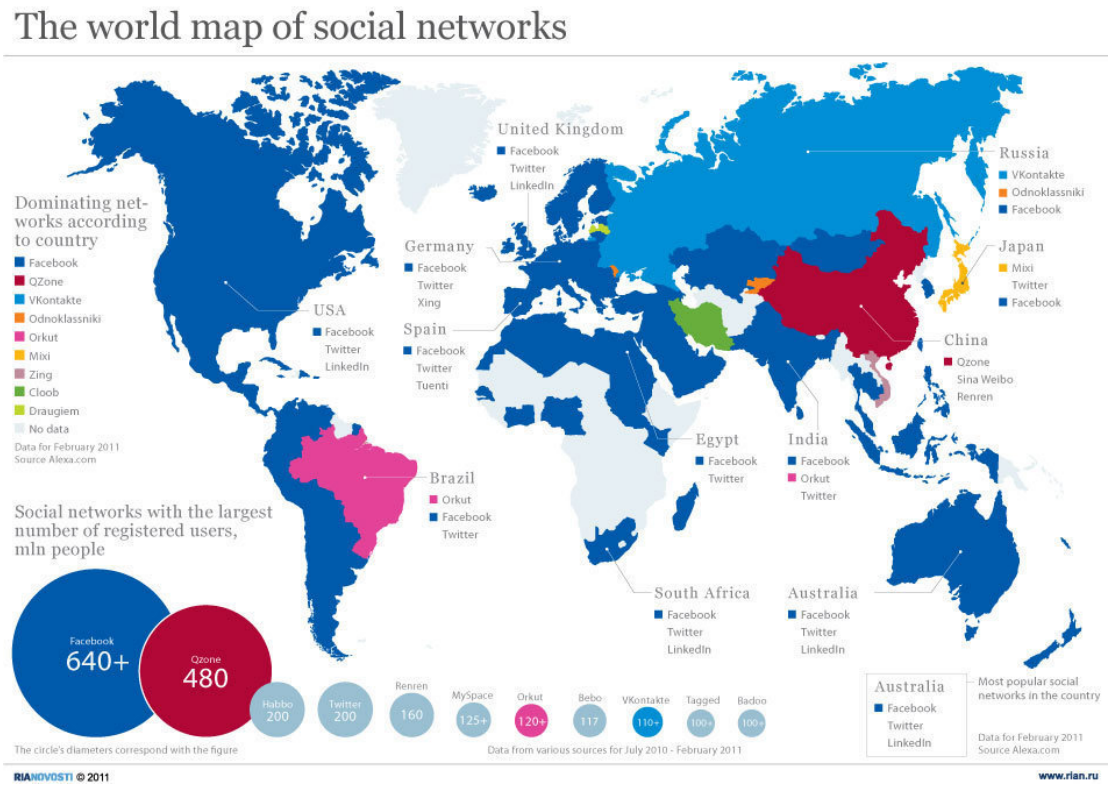

Figura 3: Mappa di utilizzo dei social network nel mondo, Febbraio 2011

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 [17].

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.

Come è facilmente intuibile, i numeri dei social media sono impressionanti. Proprio a questo proposito, è stata effettuata una ricerca da parte di ECHO Creative dalla quale si evince che il 75% della popolazione utilizza quotidianamente i social media, che le generazioni X e Y considerano obsoleto l'utilizzo delle e-mail, negli Usa l'80% delle aziende utilizza LinkedIn per il reclutamento del personale, Youtube è il

secondo motore di ricerca al mondo con un miliardo di visualizzazioni al giorno, il 34% dei blogger posta ogni giorno opinioni e recensioni su prodotti e marchi, il 25% dei risultati di ricerche web sui marchi più celebri riguarda contenuti generati dagli utenti, il 90% dei consumatori si fida dei pareri dei blogger online (Qualman, 2011). Semplice capire quanto sia sentita dalle aziende la necessità di cavalcare il fenomeno, cercando di capirne gli aspetti fondamentali per sfruttarne le elevate potenzialità, ad esempio dal punto di vista del marketing. Tutto ciò con l'obiettivo di avere la massima efficienza nelle campagne di marketing cioè ridurre i costi di contatto e pubblicizzazione del prodotto ed incrementare la propria *brand reputation* e naturalmente le vendite.

2.4.1. Twitter ed il micro-blogging

Il micro-blogging è una forma di comunicazione emergente che consente agli utenti di pubblicare dei brevi messaggi, che possono essere inoltrati su differenti canali. È dunque una pubblicazione costante di piccoli contenuti in Rete, sotto forma di messaggi di testo (normalmente 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 un articolo su un comune blog ad esempio. 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", in cui ogni utente può decidere chi vuol "seguire" per ricevere da esso tweet senza bisogno di alcuna autorizzazione, oppure può essere a sua volta seguito. Un *twitterer* i cui status sono seguiti è definito "amico", mentre uno che sta seguendo è chiamato "*follower*".

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 soprattutto negli Stati Uniti, grazie alla semplicità d'uso della piattaforma, rimasta ancora oggi il punto di forza del servizio.

L'utente crea una propria pagina web, su cui inviare gli aggiornamenti tramite il sito stesso, via SMS, con programmi di messaggistica istantanea, e-mail, 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 (followers). È anche possibile limitare la visibilità dei propri messaggi oppure renderli visibili a chiunque.

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

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.

2.5. 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 *one-to-one* 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 su Twitter, sono rappresentati con la relativa distribuzione percentuale nella Figura seguente (Figura 2) [16].

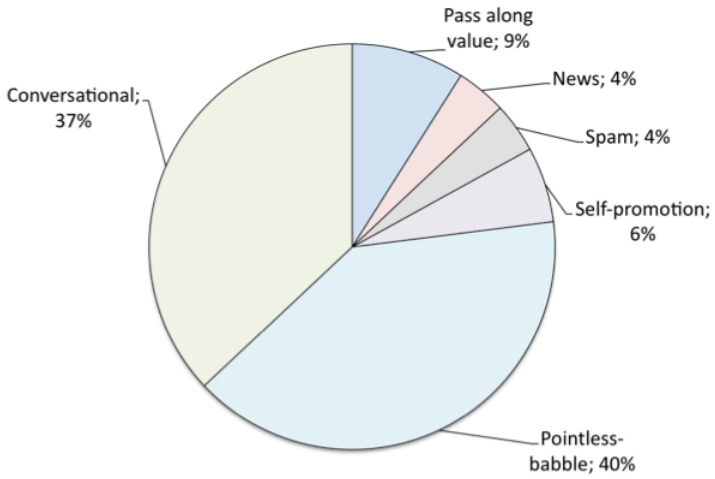

Figura 2 - Distribuzione dei contenuti per tipologia

Come si evince dal grafico (Figura 2), 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. Come è facile intuire, questo va a discapito dell'attendibilità.

Ciò è dovuto al fatto che non esiste alcun tipo di revisione dei contenuti, cosa che non avrebbe alcun senso in questo tipo di siti.

2.6. Gli influencer nelle reti sociali

Le principali teorie della comunicazione affermano che, una minoranza di individui, cosiddetti *influencers*, è dotata di un'elevata capacità di persuasione nei

confronti degli altri nelle decisioni di scelta [28]. Attraverso l'individuazione di uno specifico e ben definito gruppo di singoli, identificati come influenti, è possibile raggiungere una reazione a catena su larga scala grazie al passaparola, che dal punto di vista del marketing ha un'elevata efficienza e un costo molto basso. Questa tecnica è nota come *influencer outreach,* rappresenta un approccio strategico che prevede di coinvolgere gli influencer rendendoli promotori attivi, invece di insistere su una massa indistinta. Risulta quindi fondamentale sia andare a scegliere con cura i personaggi da influenzare, sia allo stesso tempo cercare di essere proprio gli influencer degli influencer. Così facendo sarebbe possibile sfruttare la viralità del marketing, ottenendo con limitati costi di investimento, alti ritorni in termini di visibilità.

Come ulteriore prova di quanto fin qui detto, si può citare uno studio recente che afferma che il 90% dei tweet è creato dal 10% degli utenti (B. Heil, 2009).

Negli ultimi anni, è cresciuta sempre di più l'attenzione di ricercatori e addetti al marketing su come la diffusione di un'informazione o di un nuovo prodotto possa essere massimizzata coinvolgendo proprio delle persone "chiave". Tali persone, evidenziano una combinazione di particolari attributi, sia personali come credibilità, esperienza, o entusiasmo, sia di network come connettività (inteso come il numero di connessioni) o centralità, che permettono loro di influenzare un vasto numero di altri individui, anche indirettamente.

Questa categoria di persone nello specifico del Web, è stata definita *e-fluentials* nel 1998 da Burson-Masteller e Roper Starch Worldwide. I due studiosi hanno indicato come *e-fluentials* quegli individui che si contraddistinguono per una significativa opinion leadership e che soprattutto utilizzano internet per diffondere le proprie idee.

È doveroso rilevare che, gli utenti presenti nelle reti sociali non sono tutti uguali, o meglio partecipano in maniera diversa alla creazione dell'informazione, in termini di frequenza, volume e qualità di contenuti.

Diventa quindi fondamentale riconoscere questo tipo di individui che influenzano l'attività degli altri, soprattutto per la corretta e più semplice gestione del network.

Tuttavia alcuni studi recenti, limitano il ruolo degli influencer all'interno di un network, indicando invece altre chiavi, quali:

- le relazioni interpersonali che intercorrono tra utenti ordinari [37];
- la prontezza e la predisposizione di una società nell'adottare un'innovazione [8].

Questa vision, dal punto di vista del marketing, porta ad intraprendere azioni di *collaborative filtering*, ovvero una classe di strumenti e meccanismi che consentono il recupero di informazioni predittive riguardo agli interessi di un dato insieme di utenti su larga scala (Wikipedia). L'assunzione fondamentale dietro il concetto di collaborative filtering è che ogni singolo utente che ha mostrato un certo insieme di preferenze continuerà a mostrarle in futuro.

Dunque, è evidente che non esista dimostrazione dell'entità dell'impatto degli influencer sulle reti sociali, né una causa che possa essere legata alla maggiore o minore distribuzione di uno specifico contenuto rispetto ad un altro.

2.6.1. Gli influencer su Twitter

Nel word-of-mouth la diffusione delle informazioni, è guidata in maniera sproporzionata da un piccolo numero di influencer, ma 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 pubblicato 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 sia un buon indicatore. Si è tuttavia osservato che [12]:

- Il 72.4% degli utenti segue più dell'80% dei propri followers;
- 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. 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 [19]. 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 [10], 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 [5]. Partendo da un database molto ampio composto da circa 6 milioni di utenti e prendendo come elementi determinanti 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

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

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, tenendo 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 [29], 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 ritrasmettano attivamente, dunque è necessario che sia in grado di superare la predisposizione passiva delle sue connessioni primarie.

Infine Weng et al [12] confrontando il numero dei follower e il page-rank con un page-rank modificato in relazione al topic, arrivano anche loro a dire 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 [12].

2.7. Literature Gap

Gli studi citati nel paragrafo precedente, tengono in considerazioni molteplici fattori, ma nessuno di essi mette al centro dell'analisi il contenuto della conversazione. Dunque le considerazioni fin qui fatte sembrerebbero cadere nel momento in cui ci si basa anche sul contenuto di un tweet, come ad esempio quando è presente un video o un'immagine. Si è notato infatti che i contenuti "interessanti" sono anche quelli che generano in cascata un effetto maggiore ed ovviamente certi argomenti vengono diffusi maggiormente e più rapidamente rispetto ad altri. Questo è proprio l'obiettivo di questa Tesi, che tenta di dimostrare tale intuizione.

Un primo tentativo di relazionare contenuto di un tweet e la sua propagazione, si può trovare in [9]. Sono stati scelti dei soggetti provenienti dal servizio *Mechanical*

Turk di Amazon[2], per classificare il contenuto di un campione formato da 1.000 url. La scelta degli "umani" al posto di un tool è dovuta alla scarsa capacità di classificare un contenuto come interessante o meno da parte delle "macchine". Questo scelta però porta anche ad avere un numero ridotto di campioni. Tramite un'opportuna filtrazione, volta a scartare spam e siti non in lingua inglese, si è arrivati a selezionarne 100 da sottoporre alle persone reclutate, i quali dovevano esprimere un giudizio su quanto veniva presentato, oltre che una categorizzazione del contenuto. Le figure seguenti descrivono i risultati trovati dalla ricerca in questione. In particolare la Figura 3 mostra come certi tipi di URL, come ad esempio quelli associati a contenuti multimediali e social network, tendono ad essere propagati più di altri.

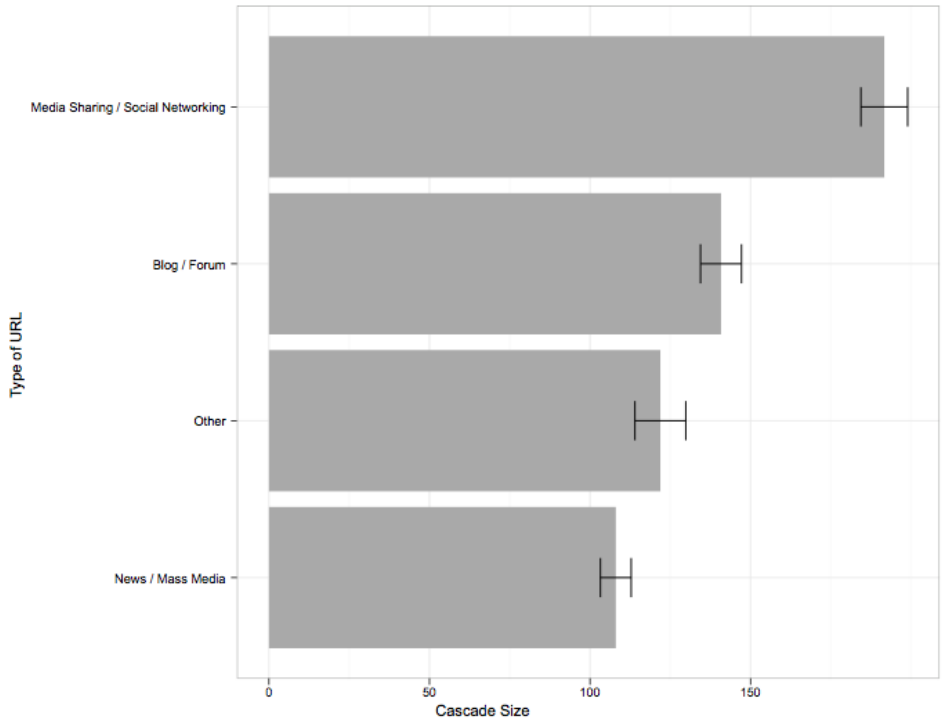

Figura 3 - Tipologia contenuti URL

[2] https://www.mturk.com/mturk/welcome

Così come alcuni tipi di contenuti abbiano più seguito rispetto ad altri, come si vede in Figura 4.

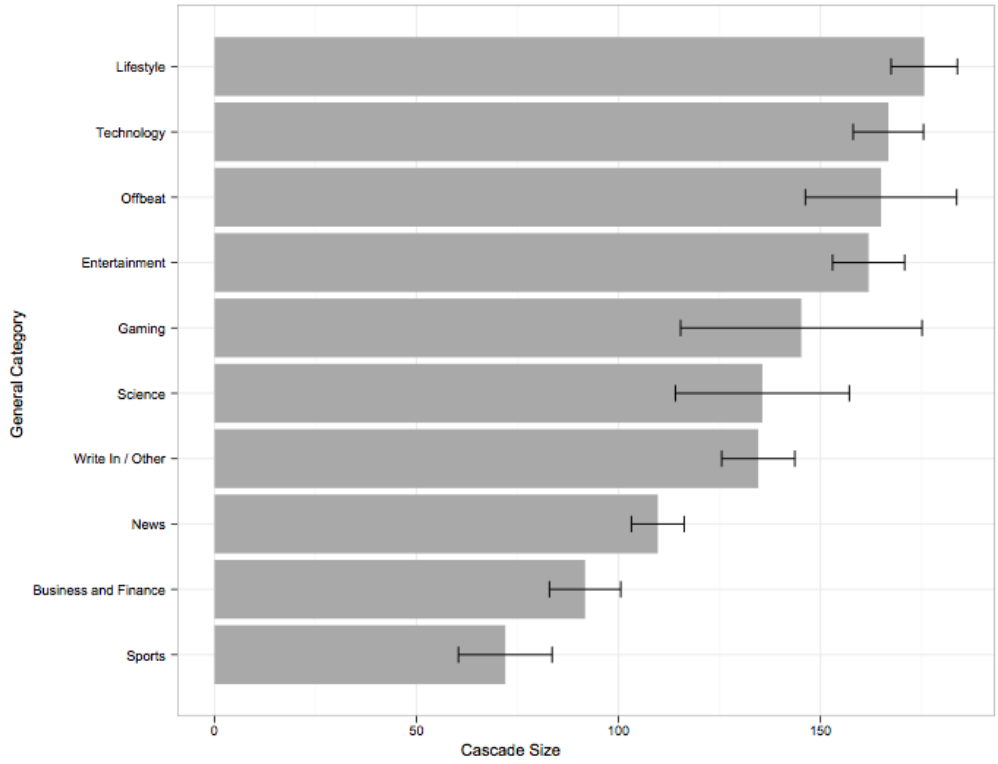

Figura 4 - Relazione contenuti/propagazione

Infine sempre la stessa ricerca, pone l'attenzione su come un messaggio giudicato più interessante possa generare in media più seguìto, rispetto ad uno meno interessante, come in Figura 5. Inoltre anche il sentiment espresso ne risente positivamente.

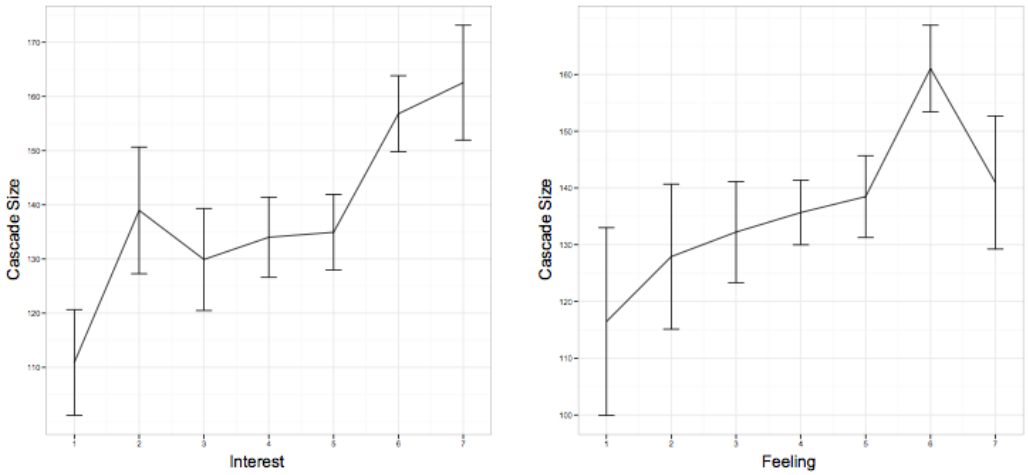

Figura 5 - Relazione interesse e sentiment rispetto alla propagazione

Questo studio, così come altri, ha il limite della pochezza del numero di campioni e del livello di dettaglio in cui si è arrivati nell'analisi dei post, al contrario dei risultati presentati in questa Tesi.

Anche dal punto di vista del SEO, i social media ed in particolare Twitter stanno ritagliandosi un ruolo importante. Google e Bing hanno recentemente affermato che vengono usati dai loro rispettivi motori di ricerca anche i post sui vari social network come fattore di posizionamento dei siti web, ossia i link condivisi tramite Facebook e Twitter hanno un impatto diretto sul posizionamento nelle SERP[3].

Mentre Bing dichiara che followers e followees possono contribuire ad aumentare leggermente il peso nei risultati di ricerca, Google aggiunge anche l'autorità degli autori come fattore decisivo per il posizionamento[4]. Entrambi però concordano come il contenuto del link "social" ne determini il posizionamento.

Proprio per questo motivo diventa importante capire le dinamiche di propagazione dei messaggi su questi nuovi mezzi di comunicazione. Questa Tesi offre uno spunto a tal proposito, focalizzandosi maggiormente su Twitter e su come il contenuto dei tweet possa influenzare tali dinamiche.

[3] Search Engine Results Page
[4] http://searchengineland.com/what-social-signals-do-google-bing-really-count-55389, 2010

3.0 Il Progetto

3.1. Introduzione

In questo capitolo verrà presentato il progetto realizzato dal Politecnico di Milano, in collaborazione con il Comune di Milano e CommStrategy, in cui questa Tesi si inserisce. Si tratta dello sviluppo di uno strumento informatico per la sentiment analysis nei social media, il cui nome è WISPO. Si è focalizzata l'attenzione sulla città di Milano dal punto di vista turistico, in relazione con tre importanti capitali europee come Londra, Berlino e Madrid.

Le analisi sono state effettuate sulla base di dati estrapolati da quattro social media, come Twitter e Facebook, e anche come quelli legati allo specifico dominio del turismo, come Trip Advisor e Lonely Planet.

Come si è potuto notare fin qui, uno degli aspetti curati maggiormente in questa Tesi è quello del marketing nella sua accezione più ampia di analisi di un brand, dei suoi punti di forza e di debolezza, al fine di promuoverne l'immagine percepita dai clienti. Il brand può essere di varia natura, come città, prodotti, servizi, aziende o anche persone. Nel caso affrontato da questa Tesi, il brand è la città di Milano.

Nel paragrafo seguente si discuterà riguardo la sentiment analysis, mentre in seguito si passerà ad una descrizione dettagliata del progetto che comprende questa Tesi, evidenziando gli obiettivi, le funzionalità e gli elementi che lo rendono innovativo.

3.2. Sentiment Analysis

Il contesto in cui si inserisce la sentiment analysis è quello dell'elaborazione del linguaggio naturale. Lo scopo è di identificare e classificare informazioni di tipo

soggettivo (fornite dagli utenti) presenti all'interno di documenti testuali di vario tipo (dagli articoli di riviste o editoriali, a dei semplici commenti, valutazioni di prodotti etc.).

Ci sono tanti modi per ricavare un giudizio da un testo presente sul Web, ad esempio tramite il traffico generato, ossia dal volume di dati che il contenuto genera. Per molte ragioni - tra le quali altruismo ed un particolare attaccamento al prodotto o in generale al brand di qualunque natura esso sia - è l'utente stesso a recensire spontaneamente un oggetto.

Quindi, un classico lavoro di sentiment analysis potrebbe essere quello di classificare la polarità di un testo come positiva, negativa o neutra in base alle informazioni soggettive (opinioni) lasciate dall'autore del testo.

Risulta subito evidente la grande utilità di questo tipo di analisi, che permette di avere molteplici vantaggi, tra cui:

- Avere un feedback riguardo le opinioni dei clienti, quindi una misura della loro soddisfazione;
- Individuare i cosiddetti influencer, della cui importanza si è ampiamente discusso nel capitolo precedente;
- Recuperare una serie di informazioni fondamentali in merito al contesto e il mercato in cui si opera;
- Avere un quadro preciso della presenza sui media online.

Tutto ciò ha aperto nuove prospettive ed, allo stesso tempo, ha creato nuovi problemi tecnici su come interpretare nella maniera corretta questa enorme mole di dati.

Bisogna infatti dire che, l'analisi del sentiment non è immune da difficoltà. Basti pensare all'elevato grado d'intelligenza che necessita un algoritmo utile a questo scopo. Non è facile avere piena comprensione di un testo, semplicemente elaborandone le singole parole. Si deve quindi capire l'argomento e contestualizzare la frase. Ciò impone un'elevata conoscenza linguistica, per tradurre efficientemente testi poco comprensibili in altri più chiari, al fine di ricavarne dati consistenti.

Peraltro bisogna essere capaci di distinguere le fonti affidabili e riconoscere un giusto peso a tutte le opinioni riscontrate.

La sentiment analysis rappresenta dunque una nuova frontiera nel mondo del marketing, ma come si è discusso in precedenza, anche questo tipo di analisi è purtroppo affetta da qualche criticità, soprattutto dal punto di vista tecnico. Infatti, software troppo restrittivi possono far perdere dati rilevanti, mentre altri troppo permissivi possono catturare molto spam.

Si tratta di un task molto complesso, per cui è necessario:

- Individuare l'argomento della discussione
- Raccogliere le possibili opinioni multiple presenti in una frase
- Fare i conti con post ironici o sarcastici che potrebbero essere male interpretati da un software.

È chiaro che individuare una certa categoria di persone, che possano influenzarne potenzialmente molte altre, può portare ad avere campagne di marketing sempre più mirate ed efficaci. L'obiettivo è quello di rendere sempre più affezionati i clienti già presenti e di attrarne sempre di nuovi, magari sottraendoli alla concorrenza.

Inoltre, saper sfruttare questa gigantesca mole di informazioni per conoscere la *Web reputation* di un certo brand, permetterebbe alle aziende di abbandonare i tradizionali strumenti per il rilevamento della customer satisfaction basati su questionari. Tali strumenti presentano una serie di svantaggi:

- costo elevato;
- lentezza nel rilevare situazioni di crisi;
- feedback non continuo sulle opinioni dei clienti;
- campione statistico abbastanza limitato quindi meno significativo.

Per quanto riguarda l'adozione di questo tipo di tecniche, la situazione delle aziende italiane è critica in quanto molte di esse o non sono propense a questo tipo

di ricerche o non danno la giusta importanza. In generale, vi è un palese ritardo riguardo l'adozione di tali misure, rispetto ad aziende americane, molto più spinte in questo settore.

Il principale motivo di dubbio da parte dei manager delle aziende, è rappresentato dal data quality. Vi è ancora un limite tecnico per cui non è possibile garantire dei risultati certi, né d'altro canto sono presenti procedure completamente automatiche.

Non si può però negare che con il Web 2.0 gli utenti non sono più semplici spettatori, ma interagiscono con i siti Web e sono essi stessi fornitori di contenuti di varia natura.

I numeri lo confermano e sono impressionanti. Da una recente indagine sulle 500 principali aziende americane[5], si evince che tutte le aziende nelle prime posizioni hanno un account Twitter, così come quasi tutte le altre aziende nelle posizioni inferiori della classifica. Tutti i 173 account Twitter analizzati erano attivi con retweet e repliche negli ultimi trenta giorni, mostrando un'interazione persistente con gli altri utenti.

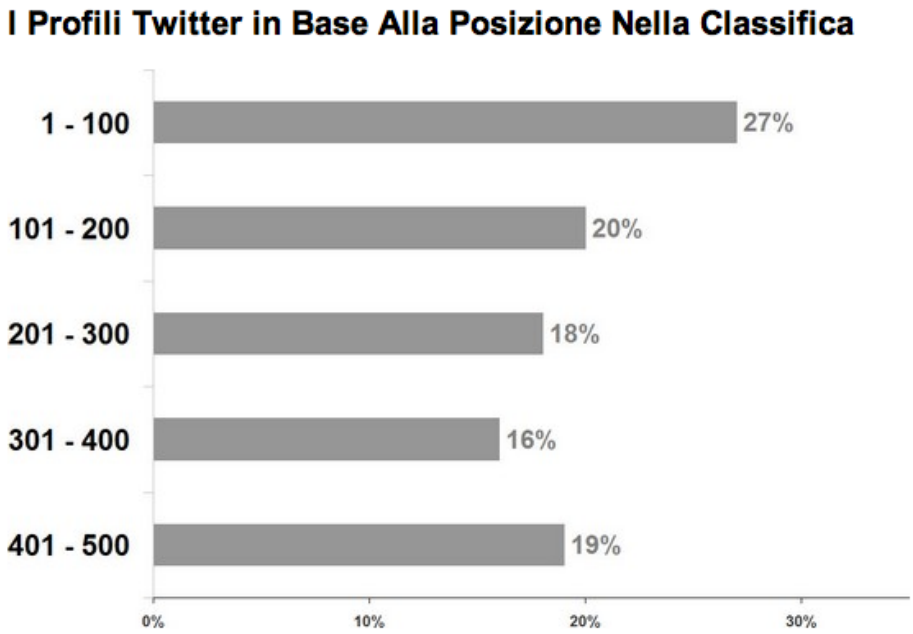

Figura 6 - Profili Twitter in base alla posizione in classifica Fortune

[5] Indagine curata da Nora Ganim Barnes ed Eric Mattson, sulle aziende presenti per l'anno 2009 nella classifica Fortune 500.

Per concludere con i dati, basti pensare che i contenuti pubblicati ogni mese su Facebook sono 30 miliardi, mentre Twitter ha visto nel 2010 la pubblicazione di 25 miliardi di tweet[6].

Questi numeri dimostrano le grandi potenzialità che oggi Internet mette a disposizione. Molti dei contenuti condivisi dagli utenti rappresentano le loro opinioni, preferenze, esperienze riguardo un qualche tema, con l'ulteriore vantaggio di essere sempre costantemente aggiornate.

Sfruttare queste informazioni per saggiare la Web reputation di un certo brand permetterebbe di:

- Migliorare il loro approccio comunicativo
- consolidare le conoscenze interne
- migliorare il marketing e di conseguenza aumentare le vendite
- garantirsi una crescita ed una sostenibilità a lungo termine.

3.3. Il Progetto

Il campo dei Social Network rappresenta ormai una sfida per la Business e Marketing Intelligence. Il Laboratorio sulla Reputazione Digitale del Politecnico di Milano in collaborazione con CommStrategy ha sviluppato una metodologia e un tool informatico per l'analisi semi-automatica della reputazione online di un city brand, chiamato WISPO. L'obiettivo è di confrontare i dati raccolti con quelli relativi ai principali competitor, in questo caso altre tre importanti capitali europee: Londra, Berlino e Madrid. Tutto ciò, con lo scopo di trarne indicazioni per appropriate operazioni di marketing territoriale, visto che l'aspetto su cui è focalizzato il progetto è prettamente turistico.

L'innovatività dello strumento risiede nella sua capacità di categorizzare e valutare il sentiment di contenuti prelevati da social media quali Twitter,

[6] Indagine realizzata da Pingdom, http://royal.pingdom.com/2011/01/12/Internet-2010-in-numbers/

LonelyPlanet e TripAdvisor, che sono basati su un linguaggio e una sintassi poco standard e difficilmente analizzabili con l'approccio semantico tradizionale. L'analisi delle informazioni raccolte ha portato un consistente miglioramento nelle campagne marketing del Comune, con l'intento di promuovere il turismo e le iniziative sul territorio. Con questi dati, è dunque possibile sapere le zone più apprezzate e quelle meno gradite della città, consentendo di intervenire in maniera puntuale.

L'approccio utilizzato si basa sull'identificazione di schemi di linguaggio parlato, ciò risulta utile anche nel caso di testi brevi come nel micro-blogging.

I vantaggi di questo approccio sono molteplici:

- Permette di monitorare in tempo quasi reale volumi elevati di messaggi su temi rilevanti.
- È basato su un modello solido che può essere utilizzato in modo continuativo nel tempo, magari adattandolo a future esigenze.
- È versatile ed efficace nella visualizzazione dei risultati, anche provenienti da diverse fonti.

Non è sufficiente formulare un giudizio globale sulla città, ma vi è la necessità di differenziare il sentiment in base ad una serie di categorie predefinite. Queste vanno a formare il cosiddetto modello di attrattività, ispirato al modello di Anholt [30], il quale tiene conto di cinque fattori competitivi che caratterizzano il vissuto e l'attrattività di un centro urbano: places, pulse, people, presence e prerequisites.

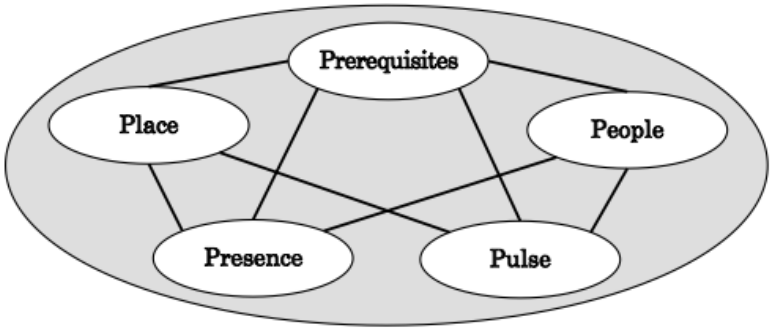

Figura 7 - Metodologia per l'analisi del passaparola online

In base a quanto detto, sono state selezionate nove diverse categorie per il modello:

- *Arts&Culture*
- *Food&Drink*
- *Life&Entertainment*
- *Services&Transport*
- *Fashion&Shopping*
- *Night&Music*
- *Events&Sport*
- *Weather&Environmental*
- *Tickets*

Il tool permette l'analisi del sentiment secondo tre coordinate fondamentali: le categorie del modello, le città e il tempo. Per migliorare la qualità e la significatività dei dati stessi, è stato necessario applicare questo tipo di categorizzazione, in modo da raggruppare le conversazioni in base all'argomento principale di discussione. Si è scelta una suddivisione per tag basati su un modello adatto al contesto turistico, che consente di elaborare minuziosamente i singoli post. Il confronto tra questi dati con quelli di altre città europee, come Londra (benchmark assoluto per Twitter), Berlino e Madrid (città molto simile a Milano sotto vari aspetti), ha permesso un'utile analisi comparativa molto dettagliata.

3.4. Architettura software

L'architettura del software in questione è composta da diversi moduli:

- Crawler

- User interface
- Data quality
- Sentiment Analysis

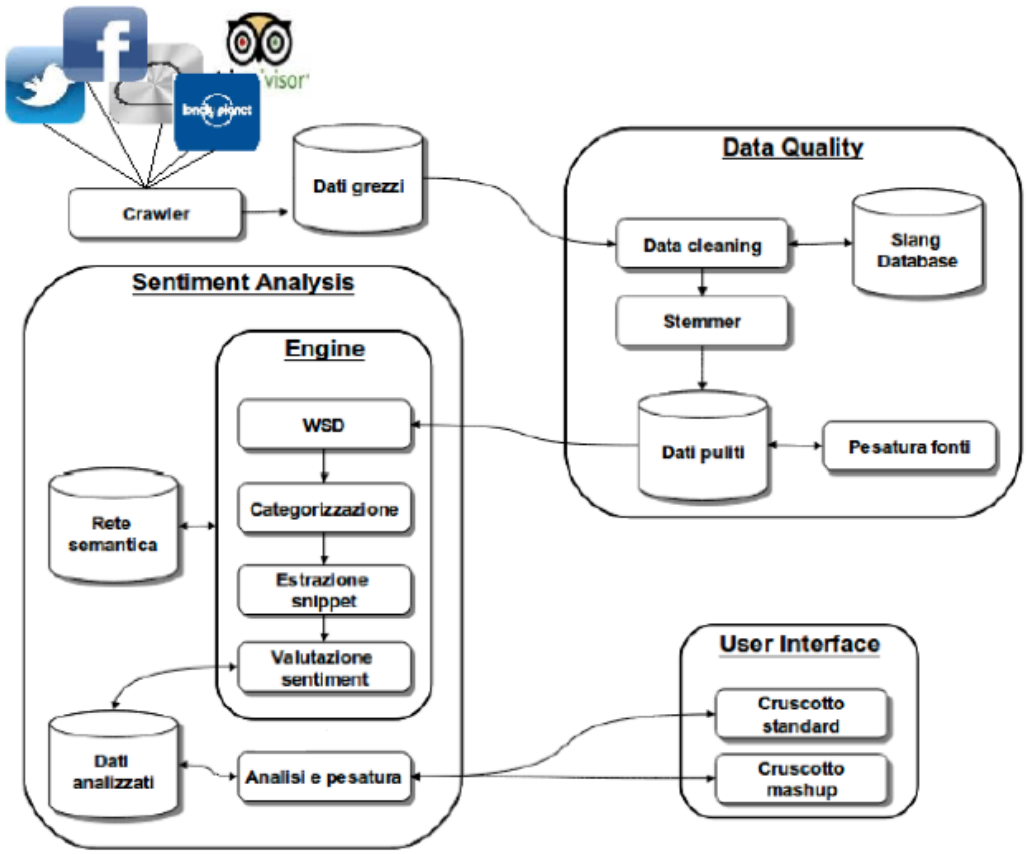

Figura 8 - Architettura dello strumento prototipale di Web Reputation

Il database viene popolato dal crawler, che si occupa appunto di prelevare i dati grezzi dalla rete ed immagazzinarli in una banca dati per una futura elaborazione. Il compito di pulire questi dati spetta invece al modulo di data quality, che tramite dei sottomoduli esegue un'analisi sintattica delle frasi e ne verifica la correttezza. Uno dei problemi da affrontare in questa fase delicata, è quello dello *slang*, ovvero una parola o espressione che non è propria né del lessico standard del linguaggio parlato né del linguaggio tecnico. Si fa quindi uso di un dizionario di slang, utile nell'individuazione, quando possibile, di un termine comune equivalente. Ove non possibile invece, vengono fissate delle regole per aiutare a decidere della corretta interpretazione.

Nella fase di cleaning vengono inoltre recuperate le frasi che contengono errori di ortografia e che potrebbero creare problemi al parser, ad esempio quelle contenenti emoticon o url. Il modulo di pesatura delle fonti invece, fornisce le metriche per valutare la reputazione in base alle sorgenti di informazione. Questi dati, ripuliti e pesati, verranno dati in pasto al modulo di sentiment analysis che è poi il cuore del progetto, di conseguenza anche la parte più complessa, poiché deve eseguire la disambiguazione. La stima del sentiment di una frase passa attraverso la valutazione di avverbi, aggettivi, sostantivi o verbi chiave che esprimono emozione (positiva o negativa). Anche questo valore deve essere pesato in base alle fonti. È in questa fase che vengono individuati gli influencer.

Infine vi è un interfaccia chiamata DashMash (modulo di user interface) che fornisce all'utente una serie di funzionalità utili ad elaborare dati, attraverso grafici di varia natura, permettendo inoltre di effettuare operazioni di roll-up e drill-down o usare mappe come quelle del tagcloud.[7]

3.4.1. Il Crawler

È possibile estrarre dei dati da una fonte Web, sostanzialmente attraverso due metodi:

- Parser Html;
- API.

Il primo metodo fa uso del parsing, ovvero un processo atto ad analizzare uno stream continuo di dati in input, pagine web in questo caso, al fine di determinarne la struttura grammaticale grazie ad una precisa grammatica formale di riferimento.

Dunque un crawler html, analizza la struttura di una singola pagina web, passando di pagina in pagina semplicemente tramite i collegamenti presenti in essa. Insomma, un comportamento molto simile ad un utente che naviga le pagine.

[7] http://it.wikipedia.org/wiki/Tag_cloud

Il parser estrae i dati e li memorizza in un apposito database, con cui è predisposto per la connessione. È facile intuire che questo tipo di operazione è molto dispendiosa in termini di tempo, nonostante sia un metodo semplice ed efficace.

Un metodo alternativo consiste nello sfruttare le API (Application Programming Interface) per ricavare dati da una data piattaforma Web. Con API si indica "ogni insieme di procedure disponibili al programmatore o in generale ad un utente, di solito raggruppate a formare un set di strumenti specifici per l'espletamento di un determinato compito all'interno di un certo programma" (Wikipedia). In questo caso, si tratta di funzionalità messe a disposizione dal social network, che permettono di gestire i dati in maniera semplice, purtroppo però con alcune limitazioni (ad esempio il numero di post visualizzabili). Gli esempi più noti ed utilizzati di questo tipo di interfacce, sono quelli di Facebook e Twitter.

3.4.2. Data Quality

Quando si affronta il problema della valutazione della qualità dei dati in ambito sentiment analysis, bisogna tener conto del fatto che i risultati di questo tipo di analisi devono essere presi con la necessaria cautela. Infatti, la correttezza o meno della stima della polarità di un documento non è dimostrabile in modo oggettivo poiché questo lavoro è dipendente da un certo livello di soggettività. Questa considerazione vale per l'uomo e a maggior ragione per le macchine.

Detto questo, il modulo di data quality di WISPO prevede tre sottomoduli:

- Data cleaning
- Pesatura delle fonti
- Stemmer

Per descrivere l'attività di cleaning bisogna prima fare un'assunzione: i testi prelevati da fonti autorevoli ed ufficiali, come ad esempio giornali o riviste online, sono considerate affidabili ovvero sintatticamente e grammaticalmente corretti. Di

conseguenza sono anche più facili da interpretare. Quando invece, i testi provengono da fonti meno formali, come nel caso di Twitter, allora è altamente probabile trovare degli errori o parti poco decifrabili, a causa di slang e forme espressive tipiche di questo contesto.

Proprio per ciò, bisogna eliminare le parti inutili e rendere questi testi facilmente interpretabili da un software al fine di aumentare le performance a livello di sentiment analysis.

Per quanto riguarda l'attività di pesatura delle fonti, è necessario porre l'accento sull'importanza della scelta delle metriche in questo contesto. Sostanzialmente quello che accade con i motori di ricerca, che basano la loro efficacia sulla classificazione dei siti Web. Ne è un esempio Google (Brin&Page,1998), col suo algoritmo chiamato PageRank, chiave del suo successo.

Nel caso della social media, si è resa necessaria l'introduzione di un nuovo metodo di ranking, che tenga conto delle caratteristiche eterogenee di blog e community (Carminati, 2010). Tempo, partecipazione e traffico sono le tre categorie che caratterizzano le metriche per le fonti web:

- Tempo medio speso dagli utenti sul sito, rappresenta un buon indice di gradimento. Usato spesso in combinazione con il buonce rate, ossia il tasso di abbandono entro un determinato periodo temporale.
- Partecipazione individuata tramite il livello di vivacità e dinamicità di un sito Web, calcolata ad esempio come media di post per ogni discussione, numero giornaliero di discussioni aperte, ecc.
- Volumi di traffico per un sito, valutato in base al numero di pagine visualizzate, numero di visitatori giornalieri, o parametri simili.

3.4.2.1. Stemmer

Questo componente si occupa di sostituire aggettivi, verbi e sostantivi con la loro forma base, ad esempio i verbi vengono tradotti in forma infinita. Questo allo scopo di semplificare le analisi successive effettuate per l'analisi del sentiment.

3.4.3. L'interfaccia mashup

I punti di forza dell'interfaccia WISPO basata su mashup, sono flessibilità e facilità di utilizzo. Le funzionalità sono invece indirizzate verso l'analisi dei volumi delle categorie, effettuato rispetto ad un intervallo di tempo definibile dall'utente, in modo da elaborare un'analisi consistente dei trend.

Tag cloud, percentuale divisa per categorie rispetto al totale dei post, andamento del sentiment nel tempo, sono tutti esempi delle possibili combinazioni eseguibili tramite interfaccia.

Si possono creare liste di influencer, divisi per categoria e con granularità temporale, a cui si aggiunge la possibilità di rilevare il *klout score*, la geolocalizzazione dei post e il link al profilo sul rispettivo social network.

Tutto questo conferisce all'interfaccia grafica di WISPO un elevato grado di innovatività. L'utilizzo di tecnologie di mashup invece, consente di includere dinamicamente contenuti e formati di rappresentazione provenienti da varie fonti. Il punto di forza di questo tipo di tecniche sta proprio nel saper creare nuovo valore dai servizi integrati, nella misura in cui li combinano in un modo innovativo, fornendo funzionalità che prima non erano presenti [13].

La figura seguente (Figura 9) mostra un classico esempio di utilizzo dell'interfaccia mashup implementata per il tool di sentiment analysis.

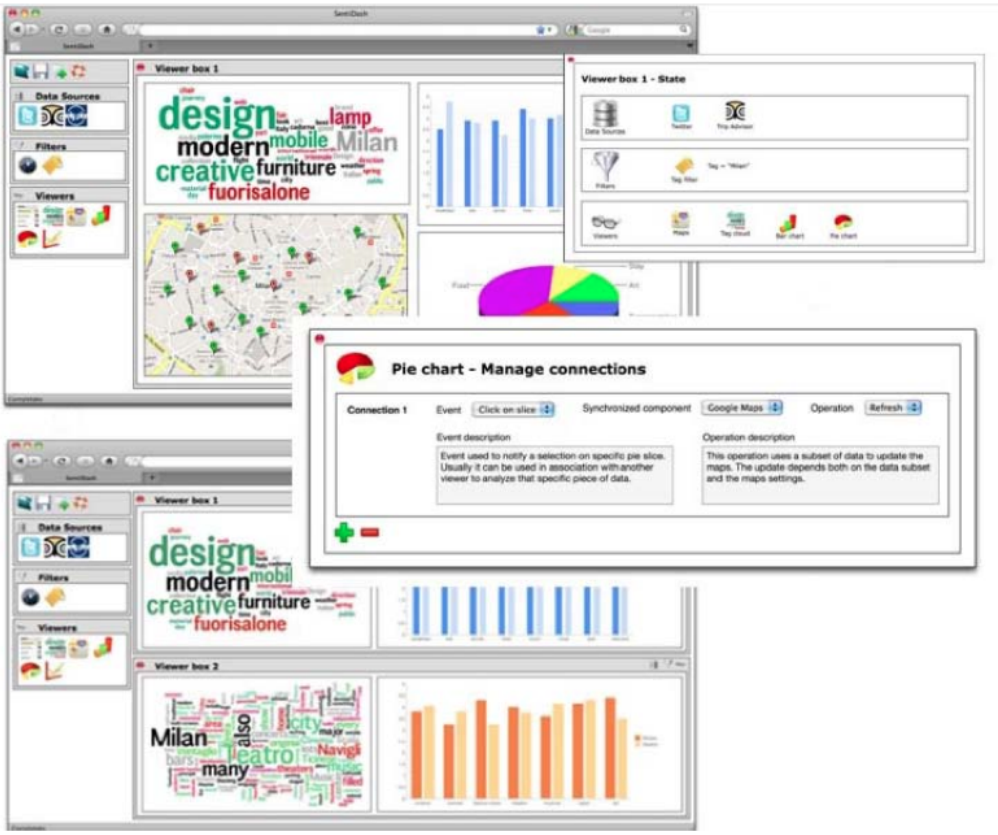

Figura 9 - Interfaccia mashup di WISPO

I moduli presenti all'interno di questo tool grafico, sono stati pensati in maniera da renderne facile l'interazione, dunque ne viene fuori uno strumento sicuramente user-friendly, come è possibile capire dalle seguenti immagini.

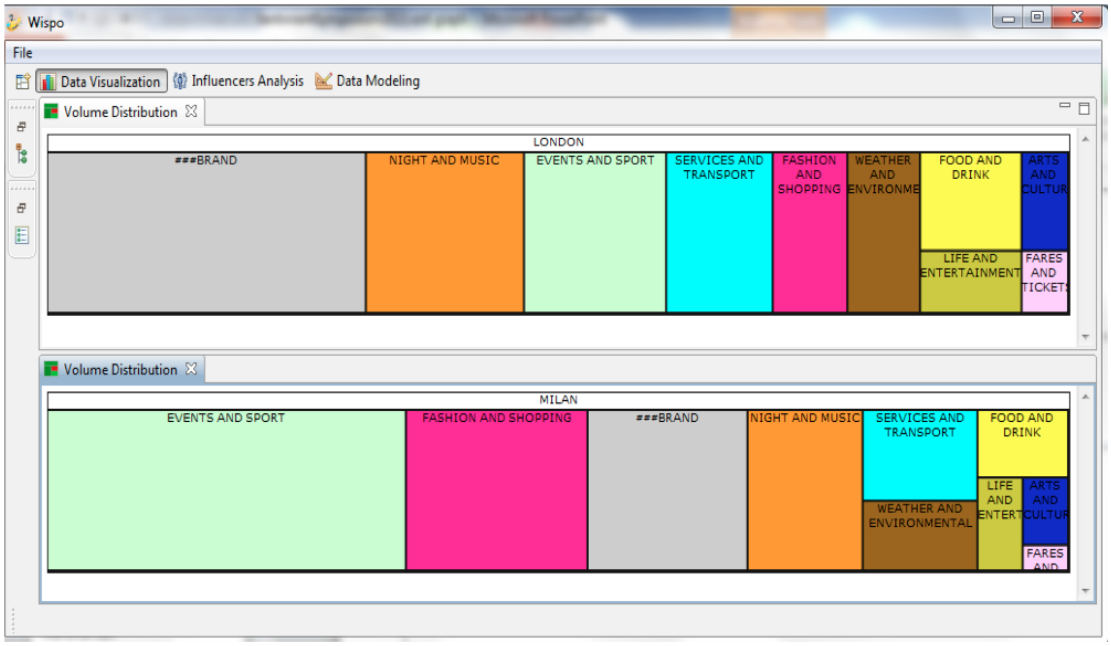

Figura 10 - Esempio di Volume Distribution (Milano Vs Londra)

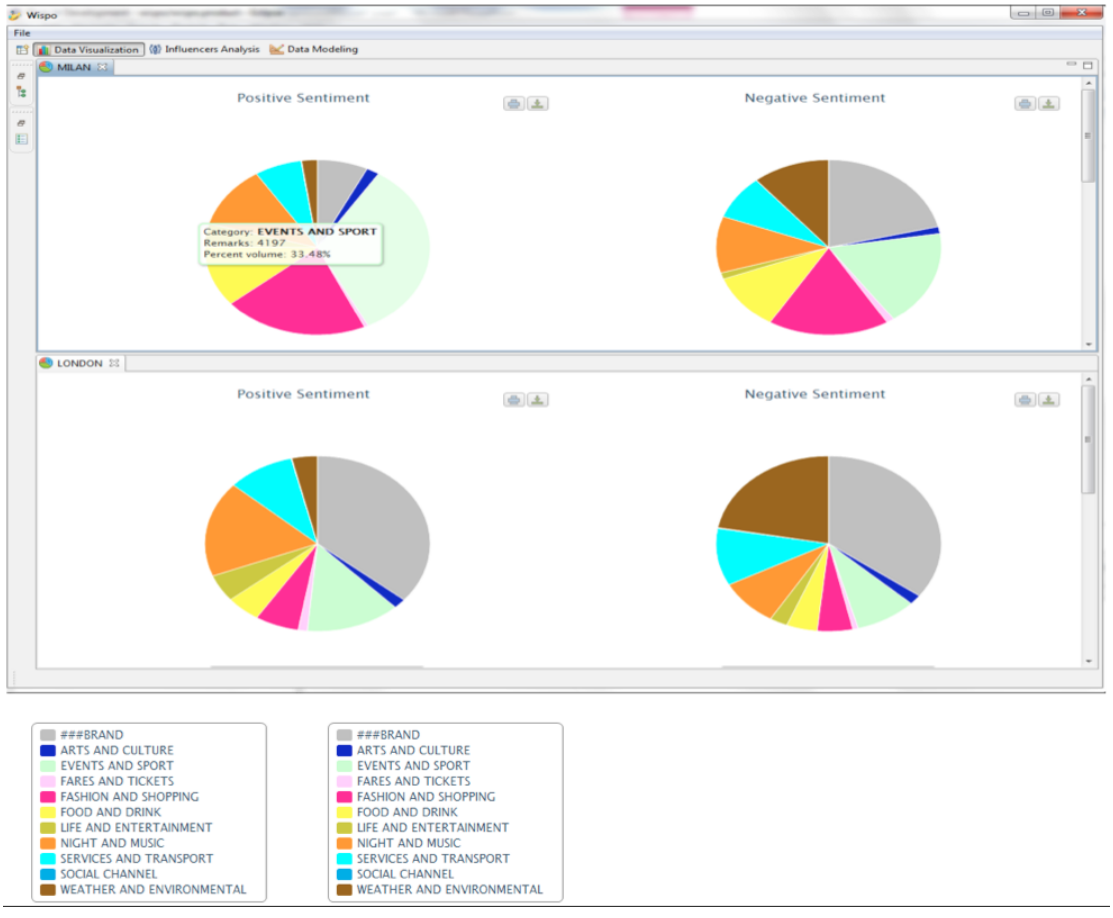

Figura 11 - Esempio di Polarity Pie (Milano Vs Londra)

3.5. Analisi degli Influencer

Il passaparola online è diventato un vero e proprio *mass media* attraverso cui veicolare valori, tendenze e stile. Le grandi aziende hanno l'esigenza di essere presenti su blog, social network e community online con l'obiettivo di fidelizzare la clientela e creare con essa un legame affettivo. Diventa importante capire chi, tra la moltitudine di utenti sul Web, ha la capacità di influenzare le opinioni della gente, questi sono detti appunto opinion leader o influencer.

A tal proposito, i tool di sentiment analysis offrono diverse funzionalità di ricerca di influencer rispetto ad un topic. Anche WISPO prevede quattro differenti tipi di visualizzazioni:

- Influencer List, una tabella che mostra la classifica degli influencer
- Influencer Map, il mashup tra la Influencer List e una mappa, in modo da visualizzare la location di ogni influencer
- Post List, l'elenco dei post pubblicati da un determinato utente ordinati rispetto alla data di pubblicazione
- Post Map, il mashup tra la Post List e una mappa, in modo da mostrare per un dato utente la posizione di ciascun post.

3.6. Analisi dei trend

WISPO annovera, tra le altre funzioni, la possibilità di analizzare il flusso temporale delle opinioni degli utenti, suddividendole anche per categoria.

Permette cioè di capire le variazioni che esse hanno, in termine non solo di volume ma anche di sentiment.

Quest'attività è nota come analisi dei trend ed ha come prerequisito l'individuazione dell'argomento di discussione relativo al singolo post.

L'analisi dei trend consiste nel tentativo di capire in tempo reale i principali topic di discussione che coinvolgono gli utenti sul Web. L'obiettivo è quello di cogliere informazioni utili sul brand osservando con che intensità e con che reattività

si condividono le diverse opinioni ed esperienze del mondo reale. Si possono anche percepire rapidamente eventuali situazioni critiche legate al brand, che gli utenti possono manifestare con le loro lamentele.

Va inoltre sottolineato quanto sia fondamentale carpire in fretta questo tipo di informazione, al fine di prendere subito le adeguate misure correttive (eventualmente sfruttando lo stesso canale Web), con l'obiettivo di ridurre al minimo il disagio vissuto dagli utenti.

4.0 Metodologia: analisi info multimediali

4.1. Introduzione

Questo lavoro di Tesi fa parte di un progetto (di cui si è discusso ampiamente nel Capitolo 3), realizzato con la collaborazione del Comune di Milano, ai fini dell'analisi della reputazione Web del brand della città, rispetto a Londra, tenuta in considerazione come benchmark, Berlino e Madrid, quest'ultima selezionata per la sua forte somiglianza, in termini di caratteristiche strutturali, al capoluogo lombardo.

L'analisi della reputazione sul Web ha come scopo quello di definire una metodologia e degli strumenti necessari al fine di valutare il grado di importanza dei nodi all'interno di social network. Tali analisi consentono di catalogare un insieme di utenti più influenti, ovvero gli *influencer*, in relazione ai diversi domini di appartenenza.

L'obiettivo di questo lavoro di Tesi è di analizzare le informazioni multimediali (come foto o video) contenute nei tweet relativi alle quattro città in questione, per capire come la presenza di un contenuto multimediale possa modificare la capacità di influenza sugli utenti.

La reputazione sul Web applicata alle aziende è un elemento in divenire, che sta cambiando col tempo, in quanto cambiano i modi e gli strumenti per manifestare la propria opinione.

"La reputazione online è un riflesso del prestigio, stima di una persona o brand online" (Wikipedia). Quindi, a differenza del concetto tradizionale di "marchio", che può essere gestito e generato attraverso la pubblicità, la reputazione non è sotto il controllo assoluto della persona o dell'organizzazione, ed è dunque più difficile da amministrare. Ma il concetto di reputazione online o *brand reputation*, non è altro che una conversazione tra clienti dove si esplicitano i propri punti di vista e in sempre più casi le aziende interagiscono. Questo particolare modo di comunicare ha anche un risvolto economico, perché avere una buona reputazione significa avere

buon profitto per l'azienda oggetto della comunicazione. Normalmente questo tipo di interazione avviene attraverso meccanismi quali forum, blog o social network.

Il costo inferiore di questo approccio lo mette in una posizione di vantaggio rispetto a sondaggi e analisi di mercato tradizionali. Inoltre, essendo un'analisi in cui i diretti interessati non sanno di essere sotto osservazione, si riusciranno ad estrapolare opinioni sincere che riflettono il vero stato della percezione dell'azienda da parte del mondo esterno.

È però necessaria una grande quantità di contenuti con relativo sentiment, al fine di ricavarne il feedback di quegli utenti in grado di influenzare altri potenziali clienti. Tuttavia, sono ancora necessari degli sforzi, soprattutto dal punto di vista tecnologico, per automatizzare l'intero processo.

4.2. Architettura 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 appositamente per questo progetto ha un struttura abbastanza semplice. I componenti principali sono tre, tutti sviluppati in Java che permettono sia di interrogare il database prelevando i tweet d'interesse, sia di raggruppare questi post e effettuare il conteggio necessario ai fini del computo delle statistiche presentate in seguito.

Dunque i tre moduli che compongono il tool sono:

- *Database.java:* questo modulo ha lo scopo di interfacciarsi con il database, che in questo caso ha un motore MySQL Server. Gestisce e avvia le singole query, oltre ovviamente a fornire i risultati secondo un formato ben definito (una lista ordinata) ai due moduli successivi.

- *Main.java*: ha come input la lista di tweet ordinati in uscita al primo modulo, quindi tweet pronti per essere analizzati e raggruppati. Si preoccupa di

scrivere su un file ogni tweet elaborato, in modo da renderlo disponibile per le successive analisi. Genera dunque in uscita, un file *CSV* (Comma-Separated Values)[8] contenente tutti i tweet catalogati secondo numero di retweet, numero di snippet con sentiment positivo, numero di snippet con sentiment negativo e altro all'occorrenza. Per snippet si intende in questo caso una porzione del post (dunque una frase o parte di essa), tendenzialmente di senso compiuto, individuata dal motore semantico ed a cui esso assegna il sentiment espresso. È infatti possibile che un tweet abbia più snippet, in quanto presenta diverse frasi all'interno, ognuna considerata significativa grazie al sentiment che esprime.

- *MainTempo.java*: anche questo componente ha come input la struttura dati ordinata proveniente dal primo modulo. La differenza con il modulo precendente sta tutta nel tipo di analisi effettuata sui tweet trattati. In questo caso si tratta di post di cui bisogna effettuare un'analisi temporale. Anch'esso produrrà in output un file *CSV*, che classifica tutti i tweet elaborati in termini di retweet ma presentanso informazioni sul tempo trascorso dal primo post "sorgente".

4.2.1. 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.

[8] http://it.wikipedia.org/wiki/Comma-separated_values

- *Retweet:* è il modo più immediato per manifestare interesse per un tweet, che appunto essendo ritwittato ai propri follower, acquista un certo grado di influenza. Recentemente il meccanismo di retweet è cambiato, quello considerato in questa Tesi è il concetto più ortodosso introdotto dal nuovo Twitter: viene considerato retweet un post che non viene modificato (non viene aggiunto alcun carattere al tweet originale) e che è stato ritwittato seguendo la procedura automatica imposta da Twitter.

  È comunque possibile ritwittare un post tramite la sintassi *"RT @username:"* che però è una procedura manuale molto diffusa ma non riconosciuta come ufficiale da Twitter, dunque può essere considerato un retweet non ufficiale seppur valido. Si è quindi deciso di procedere analizzando solo i retweet classici, effettuati mediante procedura automatica (nient'altro che usando il tasto retweet). Twitter mette a disposizione dei suoi utenti delle API tramite le quali è possibile interfacciarsi col servizio di microblogging attraverso applicazioni di terze parti. Tra le funzioni disponibili è presente anche il retweet, che risulta comunque effettuato in maniera automatica solo se conforme alle condizioni precedenti.

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.

4.2.2. Metodo di ricerca

Una volta chiarita la scelta effettuata riguardo l'interpretazione del retweet, si può procedere alla descrizione della metodologia di ricerca usata per classificare i tweet.

Al fine di formare i gruppi sopracitati si è proceduto alla sperimentazione di varie tecniche sia lato database (via MySql) sia tramite software dedicato per la ricerca (scritto in Java).

Mediante l'utilizzo della clausola MySql "*like"*, si è provveduto a fare una prima scrematura dei post in quanto tale comando, permette di impostare un confronto esatto del testo passatogli dalla query, obbligando ad una corrispondenza esatta fra la stringa all'interno del messaggio e la stringa passatagli come query. Tale metodo assicura l'individuazione dei soli tweet "ben formati" ossia del tipo "*RT @username: testo originale"* (proprio il risultato del retweet automatico accennato sopra).

In un secondo momento, entra in gioco il modulo Java che si occupa di perfezionare la ricerca e contare i tweet appartenenti ad un gruppo, oltre che a produrre in output i dati aggregati. L'aggregazione di tweet in gruppi è la fase cruciale di tutto il processo, in quanto bisogna selezionare i post aventi testo identico (ovvero quello che Twitter considera retweet) ed unirli incrementando un contatore. Così facendo, al termine dell'analisi si avrà un conteggio totale dei retweet per ogni tweet analizzato.

4.2.3. I dati in output

I file CSV in uscita dai moduli Java sono due, uno utilizzato per l'analisi classica e l'altro per quella temporale. L'unica differenza tra i due sta nel campo *timestamp*, che nel secondo caso indica la differenza in secondi trascorsi dal primo tweet.

Nello specifico, sono presenti diverse informazioni riguardanti i tweet classificati, in particolare:

- Id del tweet;
- Timestamp del tweet se si tratta del tweet sorgente (o distanza in secondi rispetto al primo post quando si tratta di analisi temporale);
- Testo del tweet;
- Numero di retweet presenti nel gruppo;

- Numero di snippet presenti nel gruppo a cui è stato assegnato un valore di sentiment neutro (/);
- Numero di snippet presenti nel gruppo a cui è stato assegnato un valore di sentiment negativo (-);
- Numero di snippet presenti nel gruppo a cui è stato assegnato un valore di sentiment positivo (+).

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

4.3. Ipotesi

In letteratura sono presenti diverse ricerche riguardo lo studio delle dinamiche di diffusione dell'informazione e dei messaggi all'interno di Twitter, che incentrano la propria analisi sulla propagazione del messaggio e sulla maggiore o minore passività dei membri della rete sociale come elementi chiave.

L'assunto su cui si basano questi studi è che l'opinion leadership di un *twitterer* può essere paragonata a quella di una comune pagina Web. Dunque ne viene fuori che l'autorità di un nodo è tanto elevata, quanto più lo è la somma delle autorità dei suoi follower.

Recentemente, sono state introdotte nuove metriche che si adattano meglio ai meccanismi di Twitter, allo scopo di identificare i nodi particolarmente significativi all'interno della rete. Il fine ultimo è capire attraverso quali nodi un messaggio possa essere propagato in maniera più efficiente ed efficace.

Per tali nodi, chiamati influencer, sono state analizzate varie metriche che hanno permesso di andare ad analizzare determinati pattern comportamentali, ma di questo si è ampiamente discusso nel paragrafo 2.5.

Oltre a chi comunica il messaggio, bisogna però anche studiare il contenuto del messaggio.

I social networks sono ormai considerati anche in letteratura alla pari dei mezzi di informazione tradizionali [3]. In questi ultimi, si può facilmente notare come le notizie che esprimono un sentiment negativo (ne è un esempio la cronaca nera) abbiano un seguito e un clamore maggiore rispetto alle notizie a cui viene attribuito un livello di sentiment positivo [3].

Basta guardare le prime pagine dei giornali per capire che le notizie negative sono anche quelle più enfatizzate ed in generale la politica delle principali fonti di informazione è quella di catturare l'attenzione con questo tipo di notizie.

Dal punto di vista del contenuto invece, foto e video ricoprono sicuramente un ruolo fondamentale. Come è possibile notare da quest'infografica di Crowd Science[9] di fine 2011, video e soprattutto foto occupano le prime posizioni per quanto concerne i contenuti che ottengono maggiori *like*[10] su Facebook.

[9] http://crowdscience.com/
[10] http://en.wikipedia.org/wiki/Like_button

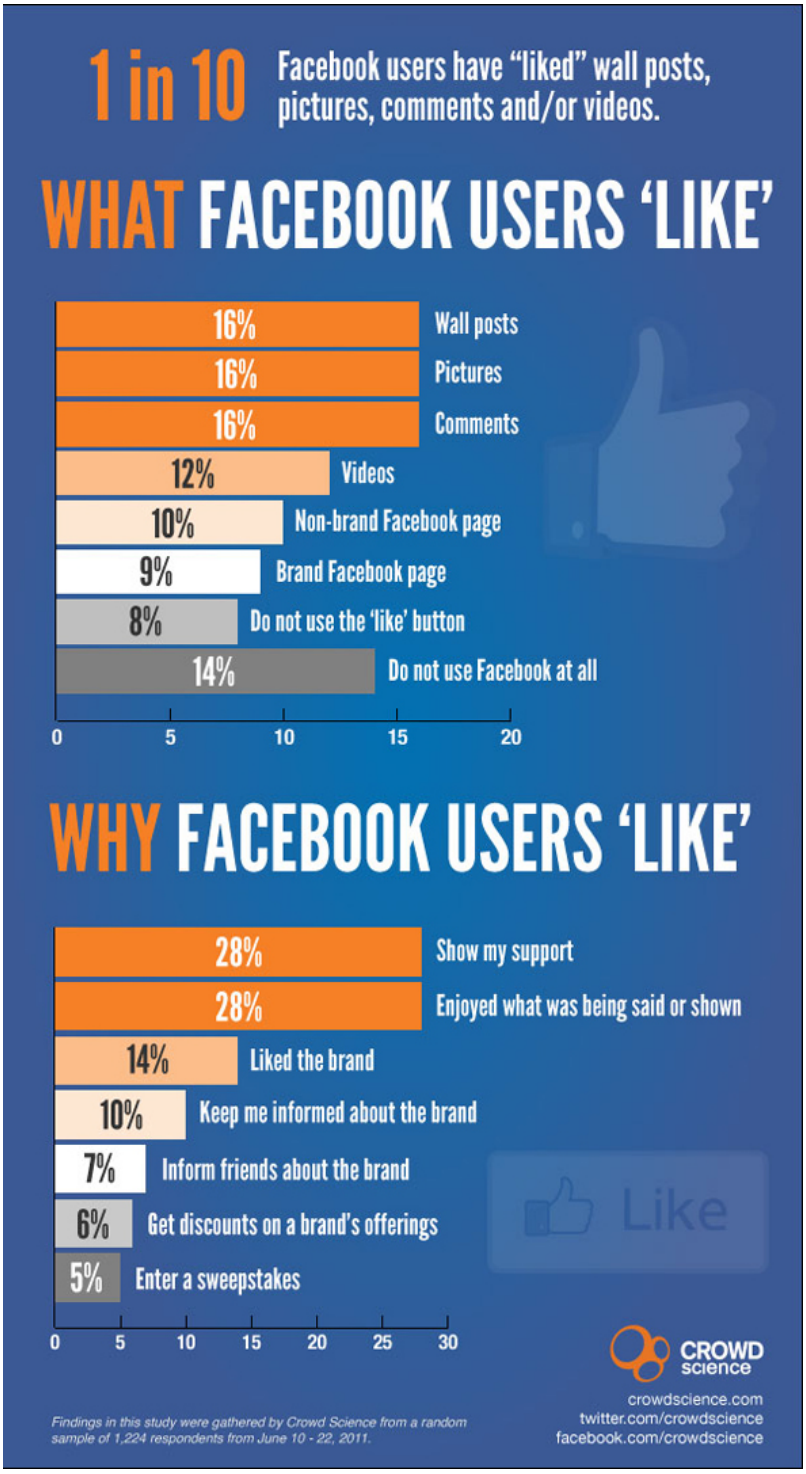

Figura 12 - Infografica Facebook Like

In questo ambito la letteratura manca di una trattazione adeguata per quanto riguarda Twitter, al contrario di Facebook dove i primi studi cominciano a far

emergere i gusti degli utenti in relazione ai contenuti condivisi e soprattutto i motivi per cui certi *status* suscitino più interesse di altri.

Da qui lo spunto per l'idea che sta alla base di questo lavoro di Tesi, ossia fare un'analisi su come i contenuti di un tweet possano influenzare la reach e il sentiment espresso. In particolare:

- **Ipotesi 1 (H1):** i post contenenti foto o video suscitano più interesse, dunque sono più ritwittati, dei post che non contengono alcun tipo di link.

Sembra plausibile credere che la presenza di una foto, o ancora di un video, all'interno di uno status di Twitter, possa modificare la potenziale influenza che il tweet stesso abbia nei confronti degli utenti.

Basti pensare alle immagini e a come, negli ultimi anni, il loro ruolo nell'informazione ha sicuramente preso il sopravvento, anche nella carta stampata, cosa che ha scaturito un dibattito perfino sulla natura stessa dell'informazione.

Quando si percepisce un oggetto, quello che rimane in mente altro non è che l'immagine dello stesso. A questa immagine si collegano concetti, sensazioni, pensieri, un vero e proprio "bagaglio emozionale" che viene rievocato alla sola vista dell'immagine.

Dunque è facile intuire come, anche nel mondo dei social media ed in particolare di Twitter, sia più efficace un tweet contenente una foto rispetto ad uno "piatto", ovvero privo di alcun contenuto multimediale a supporto del concetto espresso.

Quindi l'analisi è stata ulteriormente dettagliata in foto e video, mettendoli a confronto.

- **Ipotesi 2 (H2):** i tweet contenenti foto vengono maggiormente ritwittati rispetto ai video.

A differenza dei video, le immagini stimolano una comunicazione efficace anche a livello temporale (guardare un'immagine richiede meno tempo che guardare

un video), ma soprattutto necessitano di meno impegno da parte del fruitore del contenuto (guardare un video implica un grado di attenzione maggiore).

Inoltre i video sono da un punto di vista prettamente tecnico più "pesanti", al contrario delle immagini che in generale, perfino al variare del formato sono sempre visibili su qualsiasi dispositivo.

- **Ipotesi 3 (H3):** i video esprimono più sentiment rispetto alle immagini.

L'obiettivo di questa ultima ipotesi è dimostrare che i video, proprio per la loro natura e per le caratteristiche descritte sopra, riescano a coinvolgere emotivamente l'utente in maniera sicuramente diversa e forse più profonda, rispetto ad un'immagine.

Certamente, il fatto che un video possa essere ricondotto ad un insieme di foto in movimento, quindi un contenuto dinamico, influisce sul grado di appeal che esercita su chi lo sta guardando, basti pensare ai suoni o agli effetti che è possibile associare alle immagini.

Analogamente le stesse ipotesi esposte fin qui, si possono elaborare da un punto di vista temporale, ossia analizzare la relazione che sussiste tra il contenuto di un tweet e la sua propagazione nel tempo.

- **Ipotesi 4 (H4):** i tweet che presentano al loro interno immagini o filmati, sono ritwittati più velocemente rispetto a quelli che non hanno alcun link.

La presenza o meno del link, come visto in precedenza, influisce non solo sulla probabilità di retweet, ma è determinante anche nella velocità con cui viene propagato il messaggio.

Stesso comportamento sembrano avere i tweet che hanno al loro interno un'immagine, messi a confronto con quelli che hanno un video. Tale concetto è alla base della seguente ipotesi.

- **Ipotesi 5 (H5):** i post contenenti un'immagine, sono più velocemente ritwittati rispetto ai tweet che includono video.

Si può inoltre considerare il sentiment suscitato da tali tweet per formulare l'ultima ipotesi.

- **Ipotesi 6 (H6):** tra i tweet che esprimono sentiment (ossia composti da almeno uno snippet positivo o negativo) quelli che contengono una foto sono ritwittati più velocemente rispetto ai post contenenti video.

4.4. Dataset

Per questo lavoro di Tesi è stato utilizzato un ampio dataset, contenente i tweet del mese di Luglio 2011 relativi alle città di Milano, Madrid, Berlino e Londra. Questi post, rigorosamente in lingua inglese, sono stati ottenuti filtrando semanticamente rispetto alle parole chiavi "Milan", "Berlin", "Madrid", "London".

Sono stati raccolti i dati sui tweet e sui retweet tramite le Twitter API. Si è posta l'attenzione sul dominio del turismo, collezionando tutti i post contenenti appunto le parole chiave sopra citate e usando un motore di ricerca semantico per filtrare la ricerca e valutarne il sentiment [2]. È stato scelto come riferimento il modello di Anholt [30] e sono state create differenti categorie per il dominio in uso, che vanno dai servizi ai trasporti, all'arte e alla cultura, ecc…

Il dataset utilizzato per la verifica delle ipotesi descritte in seguito, risulta essere composto da un totale di circa 1.5 milioni di tweet, di cui solo i retweet sono circa 270 mila. Tramite alcune tecniche descritte nel Capitolo 4 si sono suddivisi i retweet in circa 110 mila gruppi, ciascun gruppo formato da tutti i retweet generati nel periodo di riferimento ossia Luglio 2011, facente capo allo stesso tweet "sorgente".

La base di dati cosi ottenuta ha una dimensione media di gruppo pari a 2. In particolare, il gruppo di dimensioni maggiori è costituito da 8.515 retweet mentre quello di dimensioni minori è composto da un singolo tweet.

La stessa suddivisione in gruppi è stata adottata per i tweet contenenti link, individuati tramite la stringa "http://" o "https://". Questo tipo di post, sono stati

ulteriormente suddivisi in tweet che includono foto e video, poi usati per verificare le ipotesi presentate nei paragrafi successivi.

I link al cui interno si trovano immagini sono stati individuati grazie alle applicazioni (principalmente di tipo mobile) usate per caricarle su Twitter, che saranno brevemente descritte in seguito. Sono stati presi in considerazione i quattro servizi per la condivisione immagini più diffusi, ovvero: Flickr, Instagram, TwitPic, Yfrog.

Sono stati catalogati circa 17.000 tweet che includono immagini, in 7.343 casi si tratta di retweet ed è stato possibile dividerli in 1.627 gruppi (ossia 1.627 tweet unici presenti all'interno del dataset).

Anche per quanto riguarda i tweet contenti video è stata adottata la stessa tecnica, tenendo in considerazione le seguenti quattro applicazioni: TwitVid, TwitCam, Youtube, Vimeo. Sono stati individuati circa 10.000 di questo tipo di post, raggruppati in 586 gruppi (tweet unici), per un totale di 1.011 retweet.

Ognuno di questi servizi, utilizza un link proprio, facilmente riconoscibile, che permette di distinguerli e di conoscere il contenuto foto o video a seconda dell'applicazione in questione (es. http://instagr.am per le foto o http://youtu.be per i video).

Infine sono stati individuati 189.104 retweet che non contengono alcun collegamento ipertestuale. Anch'essi raggruppati in base ai retweet formano 74.295 tweet unici.

4.4.1. Le Applicazioni

Per catalogare i tweet, è stato necessario analizzare i link generati dalle diverse applicazioni prese in considerazione. La scelta è ricaduta su quelle più diffuse e principalmente multipiattaforma o web-based. Sono state selezionate quattro applicazioni per quanto riguarda le foto e quattro per i video.

Di seguito una breve panoramica, partendo da quelle relative alle foto.

Flickr

Flickr è una piattaforma che permette agli iscritti di condividere fotografie personali con chiunque abbia accesso a Internet, in perfetto stile web 2.0.

Il sito, di proprietà del gruppo Yahoo!, può contare su circa sette milioni di utenti.

Inizialmente è nato come strumento per ospitare le proprie immagini da pubblicare su altri siti, ha avuto grande successo grazie al fenomeno dei blog. In seguito il suo utilizzo è cambiato parecchio, si è evoluto. Fino a diventare esso stesso una comunità virtuale grazie ai gruppi tematici ed ai forum. Oggi viene principalmente utilizzato per raccogliere le foto della propria vita in un'unica bacheca virtuale e rimanere aggiornati su quella dei propri conoscenti ed amici. Questo anche grazie alla crescita tecnologica ed al largo numero di strumenti fotografici attualmente in circolazione.

Instagram

Applicazione gratuita di condivisione foto che permette agli utenti di scattare foto, applicare filtri e poi condividere tutto su numerosi servizi di social networking.

La diffusione sempre più massiccia degli iPhone e dell'uso dell'internet mobile, il lancio di applicazioni per telefonini Android e Blackberry, il miglioramento delle funzioni dell'applicazione, sono tra i principali fattori dell'enorme successo ottenuto da Instagram.

TwitPic

TwitPic è un sito web che consente agli utenti di inviare immagini a Twitter e ad altri social media. È anche usato dai giornalisti per caricare e distribuire le immagini in tempo reale durante un evento.

Può essere usato in maniera indipendente da Twitter (come accade per Flickr), ma è proprio il legame stretto con tale social network uno dei suoi punti di forza. Infatti le stesse applicazioni Twitter ufficiali per Android e iOs permettono di caricare foto tramite TwitPic di default.

Yfrog

Si tratta di un servizio di hosting per le immagini, fornito da ImageShack[11] a cui si ispira il nome (Yellow Frog è il logo di ImageShack). È stato principalmente concepito per permettere agli utenti di condividere su Twitter, foto e video tramite link.

Anch'esso è presente nell'applicazione ufficiale Twitter per iPhone.

Per quanto concerne i video, si è optato per le quattro applicazioni fra le più utilizzate dagli utenti, descritte di seguito.

Youtube

Youtube è il più diffuso e famoso sito per la condivisione video. Permette di condividere video sui principali social network, tra cui Twitter, per cui include già un abbreviatore url.

TwitCam

TwitCam presenta una leggera differenza rispetto agli altri servizi fin qui citati. Si tratta infatti di una piattaforma per il live streaming, che permette agli utenti Twitter di realizzare il proprio stream video in diretta. Ai fruitori di questo contenuto live multimediale, possono partecipare attivamente a questo tipo di trasmissione, attraverso una chat associata ad ogni stream.

TwitVid

Si definisce come *open social video network*, non limitandosi ad essere una classica piattaforma video. Possiede infatti molte delle caratteristiche di un comune social network, offrendo all'utente la possibilità di creare o far parte di una vera e propria community, unita da interessi comuni.

Vimeo

[11] http://en.wikipedia.org/wiki/ImageShack

Il nome Vimeo descrive perfettamente il significato stesso del progetto. Si tratta infatti dell'anagramma di *movie*, ossia filmato, oltre ad avere *me* al centro, per indicare che i contenuti video sono a cura degli utenti stessi del servizio. È uno dei principali e agguerriti concorrenti di Youtube.

5.0 Risultati sperimentali

5.1. Introduzione

In questo capitolo verranno presentati i dati ottenuti dalle analisi descritte in precedenza.

In particolare, nel paragrafo successivo, saranno commentati i risultati del test statistico condotto in merito alle diverse analisi svolte su tweet con presenza di foto e video confrontati con quelli che non contengono alcun link multimediale. Successivamente si discuterà delle differenze trovate, a livello statistico, tra i tweet contenenti immagini contro quelli contenenti video. Infine, nell'ultima analisi, foto e video saranno nuovamente comparati tra loro aggiungendo però il sentiment come termine di paragone.

Nel paragrafo 4.2, i risultati esposti saranno caratterizzati dal tempo come fattore del confronto, tentando di dimostrare come il contenuto influisca anche da un punto di vista temporale.

5.2. Analisi

Per effettuare i test statistici si è scelto di utilizzare il *Test di Student*[12], in quanto è stato ritenuto il più indicato per confrontare le medie di due campioni indipendenti.

Risulta però necessario rispettare le seguenti ipotesi:

[12] http://it.wikipedia.org/wiki/Distribuzione_t_di_Student

- la distribuzione dei dati deve essere una distribuzione *Normale*, o la si può approssimare tale se il campione è abbastanza ampio (il caso trattato in questa Tesi);
- le osservazioni devono essere raccolte in modo indipendente.

È stato condotto un *t-test* per campioni indipendenti, per confrontare innanzitutto la differenza tra le medie di retweet di post contenenti immagini e video contrapposti a quelli che non contengono alcun tipo di link. Dalla seguente (Figura 13), si evincono le differenti medie tra i due tipi di campioni.

Statistiche di gruppo

Tipo	N	Media	Deviazione std.	Errore std. Media
No Link	74295	1,55	34,937	,128
Foto e Video	2213	2,77	26,390	,561

Figura 13 - Statistiche di gruppo: Foto e Video Vs No Link

Si può notare ancora più chiaramente (Figura 14) che, basandosi su un livello di significatività basso (*p-value* = 0.033), la differenza delle medie di retweet per i post con foto/video (Media = 2,77, Deviazione Standard = 26,39) rispetto ai post senza (Media = 1,55, Deviazione Standard = 34,94) risulta netta e non dovuta ad un fattore casuale.

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
#RT	Assumi varianze uguali	7,145	,008	-1,642	76506	,101	-1,230	,749	-2,698	,238
	Non assumere varianze uguali			-2,137	2448,782	,033	-1,230	,575	-2,358	-,101

Figura 14 - Statistiche T-Student Foto e Video Vs No Link

Questo significa sostanzialmente che la presenza di un contenuto multimediale all'interno di un tweet garantisce maggior probabilità di retweet.

Un'evidente differenza la possiamo inoltre costatare confrontando i post con immagini rispetto ai post con filmati. È dunque andando a dettagliare l'analisi che si trovano ulteriori importanti diversità nella media tra i vari tipi di campioni.

In particolare, nella Figura 15 si osserva tale divario in maniera inequivocabile, permettendo di affermare che nei dati riscontrati, le foto sono ritwittate in media 5 volte di più che un video.

Statistiche di gruppo

	Foto	N	Media	Deviazione std.	Errore std. Media
0	Foto	1627	3,51	30,698	,761
	Video	586	,73	2,889	,119

Figura 15 - Statistiche di gruppo: Foto Vs Video

Questi dati si basano su un campione di più di 2000 tweet, divisi tra 1627 immagini e 586 filmati. Tutti filtrati attraverso i link delle principali applicazioni video e foto, come descritto in precedenza.

Dalla Figura 16 ricaviamo il livello di significatività, ovvero la percentuale di errore nel considerare casuale tale diversità tra medie, 0 in questo caso.

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
0	Assumi varianze uguali	13,572	,000	2,196	2210	,028	2,790	1,271	,298	5,282
	Non assumere varianze uguali			3,620	1702,950	,000	2,790	,771	1,278	4,302

Figura 16 - Statistiche T-Student Foto Vs Video

Dunque, esiste una forte evidenza empirica (*p-value* pari a 0%) a favore del fatto che vi sia una differenza (non dovuta al caso) tra le medie dei due campioni. Segue che in media, le foto sono più oggetto di retweet rispetto ai video.

Nel caso del sentiment invece l'analisi è stata effettuata considerando, per ogni tweet, la somma del numero di snippet negativi più quelli positivi. Del risultato ottenuto si è poi fatta la media e i risultati sono quelli mostrati di seguito nella Figura 17.

Questo significa che i video in media esprimono un sentiment maggiore rispetto alle più statiche immagini.

Statistiche di gruppo

	Video	N	Media	Deviazione std.	Errore std. Media
0	Foto	1627	,33	3,043	,075
	Video	586	,58	2,223	,092

Figura 17 - Statistiche di gruppo: Foto Vs Video con Sentiment

Dalla Figura 18 si nota il basso grado di significatività (pari a 0,40) che valida ulteriormente il risultato ottenuto.

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
0	Assumi varianze uguali	8,652	,003	-1,781	2211	,075	-,244	,137	-,514	,025
	Non assumere varianze uguali			-2,057	1409,729	,040	-,244	,119	-,478	-,011

Figura 18 - Statistiche T-Student Foto Vs Video con Sentiment

5.3. Analisi Temporale

Lo stesso dataset di campioni precedenti sono stati usati per un'accurata analisi temporale, prendendo in esame in particolare, gli istanti (espressi in secondi) trascorsi tra il primo tweet "sorgente" ed ognuno dei successivi retweet.

Nel caso dei post contenenti foto o video, sempre messi in relazione con quelli senza alcun tipo di link, si evince subito una differenza tra le medie temporali di retweet, anche piuttosto netta, riportata nella figura seguente (Figura 19).

Statistiche di gruppo

Tipo	N	Media	Deviazione std.	Errore std. Media
No Link	114809	412,58	206,274	,609
Foto e Video	6141	283,18	169,649	2,165

Figura 19 - Statistiche di gruppo: Foto e Video Vs No Link (analisi temporale)

Su circa 114.809 campioni infatti, la media di retweet del primo tipo di post (senza link) è di 413 secondi, ovvero un retweet entro poco meno di 7 minuti. Invece la media di retweet in secondi degli status che presentano almeno un link multimediale è di 283 secondi, ovvero poco meno di 5 minuti.

Questo significa che un tweet contenente una foto o un video viene ritwittato più velocemente rispetto ad uno che non contiene all'interno alcun collegamento ipertestuale.

La Figura 20 aiuta a chiarire meglio le idee e fornisce un dettaglio dei parametri statistici utili a supportare la tesi. Da notare anche in questo caso il livello di significatività pari a 0, che rafforza il risultato ottenuto.

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
#RT	Assumi varianze uguali	1591,084	,000	48,292	120948	,000	129,395	2,679	124,143	134,646
	Non assumere varianze uguali			57,539	7147,067	,000	129,395	2,249	124,986	133,803

Figura 20 - Statistiche T-Student Foto e Video Vs No Link (analisi temporale)

Si evince quindi che vi è assoluta evidenza empirica a favore del fatto che le medie siano diverse e questa diversità non è imputabile a fattori casuali, proprio come dimostra il *test di Student*.

Nessuno stravolgimento rispetto alle analisi precedenti per quanto riguarda il confronto delle medie tra tweet che includono immagini e quelli che includono filmati.

Tali medie risultano infatti nettamente diverse e in questo caso trattandosi della media di retweet, ovvero degli istanti di tempo trascorsi tra un retweet ed un altro, indicano che le immagini sono ritwittate più velocemente rispetto ai video. Come si diceva in precedenza, ciò può essere dovuto al fatto che i video hanno bisogno di più tempo per essere visionati rispetto ad un'immagine.

Si è eseguito il solito *t-test* per campioni indipendenti, volto al confronto tra la media di retweet di post contenenti foto contro la media di quelli contenenti video ed i risultati sono esposti nella Figura 21.

Statistiche di gruppo

	Tipo	N	Media	Deviazione std.	Errore std. Media
#RT	Foto	5716	274,59	166,682	2,205
	Video	425	398,78	167,058	8,104

Figura 21 - Statistiche di gruppo: Foto Vs Video (analisi temporale)

Le statistiche riportate di seguito validano ulteriormente questo concetto:

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
#RT	Assumi varianze uguali	2,277	,131	-14,817	6139	,000	-124,194	8,382	-140,625	-107,763
	Non assumere varianze uguali			-14,788	488,892	,000	-124,194	8,398	-140,695	-107,693

Figura 22 - Statistiche T-Student Foto Vs Video (analisi temporale)

La significatività del test di Student a 2 code indica che i risultati sono attendibili, ovvero che le medie differiscono in maniera significativa e tale differenza non è dovuta a fattori casuali.

Anche prendendo in considerazione i tweet che hanno almeno uno snippet (positivo o negativo), risulta comunque evidente che le foto vengono ritwittate più rapidamente rispetto ai filmati (Figura 23).

Statistiche di gruppo

	Tipo	N	Media	Deviazione std.	Errore std. Media
#rt	Foto	234	332,19	212,080	13,864
	Video	36	429,89	197,001	32,834

Figura 23 - Statistiche di gruppo: Foto Vs Video con sentiment (analisi temporale)

I risultati riportati nella Figura 24 sottolineano il grado di validità del test, con un *p-value* pari a 0,01 che permette di rifiutare l'ipotesi che la differenza tra le medie dei due campioni sia casuale.

Test per campioni indipendenti

		Test di Levene di uguaglianza delle varianze		Test t di uguaglianza delle medie						
									Intervallo di confidenza per la differenza al 95%	
		F	Sig.	t	df	Sig. (2-code)	Differenza fra medie	Differenza errore standard	Inferiore	Superiore
#rt	Assumi varianze uguali	1,490	,223	-2,596	268	,010	-97,697	37,627	-171,778	-23,615
	Non assumere varianze uguali			-2,741	48,363	,009	-97,697	35,641	-169,343	-26,050

Figura 24 - Statistiche T-Student Foto Vs Video con sentiment (analisi temporale)

6.0 Conclusioni

Questo capitolo ha lo scopo di riassumere e commentare i risultati ottenuti nel Capitolo 5. Saranno inoltre illustrati i possibili sviluppi futuri di questo lavoro di Tesi.

Nel corso della Tesi si è riusciti a verificare empiricamente che il contenuto dei post su Twitter è decisivo ai fini della sua propagazione. Ad esempio, includere foto o video nei tweet influisce sull'impatto che il messaggio avrà sugli altri utenti, sia in termini di retweet che in termini di sentiment espresso.

Gli utenti comuni, al pari degli *influencer*, non sono semplici utilizzatori dei social network ma recitano un ruolo attivo nella diffusione dei contenuti.

Le caratteristiche intrinseche del processo di espressione e condivisione delle opinioni su questi nuovi media focalizzano l'attenzione sulla qualità dei contenuti. Grazie inoltre alla natura "social" del mezzo di comunicazione, l'informazione trasmessa risulta poco soggetta a manipolazione.

Ponendo l'attenzione in particolare sui contenuti multimediali presenti nei tweet, un importante risultato raggiunto evidenzia l'impatto positivo che hanno sulla viralità dei tweet. Si è ampiamente dimostrato inoltre, come la presenza di foto e video sia determinante rispetto alla divulgazione di un contenuto.

Il primo stadio dell'analisi è stato incentrato sul raggruppamento dei tweet, catalogandoli uno per uno in base al numero di retweet ed effettuando anche un conteggio circa il sentiment (numero snippet negativi, positivi e neutri). Questo è stato possibile grazie a tre moduli software ad-hoc sviluppati in Java. Poi con l'ausilio di opportuni strumenti statistici, sono stati effettuati dei test, in particolare si è scelto di usare il test di Student per validare i risultati già previsti nelle ipotesi e citati sopra.

Le analisi sono state ulteriormente dettagliate, sia da un punto di vista qualitativo, ad esempio comparando post contenenti video con post contenenti foto, sia da un punto di vista quantitativo, considerando non solo il numero di retweet ma anche gli istanti (espressi in secondi) intercorsi tra un retweet ed un altro. Anche il

sentiment espresso dai singoli tweet ha avuto un ruolo determinante nelle varie elaborazioni, per cui si è potuto constatare come i video siano più efficaci da questo punto di vista rispetto alle immagini.

Tutto ciò ha portato diversi punti di riflessione, tra cui il fatto che le immagini abbiano un effetto più immediato rispetto ai filmati, sia in termini di tempo di retweet che di quantità di retweet stessi. D'altro canto però, i filmati in media esprimono più sentiment rispetto alle immagini.

Una possibile chiave di lettura di questa differenza è che le immagini al contrario dei video, stimolano una comunicazione più immediata anche a livello temporale (guardare un'immagine richiede meno tempo che guardare un video). Inoltre una foto necessita di meno impegno da parte del fruitore del contenuto (guardare un video implica un grado di attenzione maggiore).

Per quanto riguarda i possibili sviluppi futuri, si sta già pensando all'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.

Inoltre, gli studi futuri allargheranno i confini della ricerca a nuovi indici per la misurazione delle dinamiche di diffusione di un contenuto, come ad esempio il livello di penetrazione di un'informazione.

Anche il concetto di influencer sarà soggetto a future revisioni, applicando le nuove metriche ad un campione di utenti selezionati fra i vari social network.

7.0 Bibliografia

[1] B.Heil, M. P. (2009). New Twitter Research: Men Follow Men and Nobody Tweets. Harvard Business

[2] Barbagallo, D. & Cappiello, C. & Francalanci, C. & Matera, M. (2011). Semantic sentiment analyses based on the reputation of Web information sources. Applied Semantic Web Technologies, Sugumaran, V. and Gulla J. A. (eds), Taylor & Francis.

[3] Barbagallo, D., Bruni, L., Francalanci, C., Giacomazzi, P. (2012). An empirical study on the relationship between sentiment and influence in the tourism domain. In Proceedings of the 19th eTourism community conference (ENTER2012).

[4] Bughin, J., Doogan, J., & Vetvik, O. J. (2010 April). A new way to measure word-of- mouth marketing. McKinsey Quarterly .

[5] 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.

[6] Chung, K. K., Hossain, L., & Davies, J. (2005). Exploring sociocentric and egocentrici approaches for social network analysis. In Proceedings of International Conference on Knowledge Management, (pp. 1-8). Wellington, New Zealand, Asia Pacific.

[7] Dellarocas, C. (2003 October). The digitization of the word-of-mouth: Promise and challenges of online feedback mechanisms. Management Science 49 , 49 (10), pp. 1407- 1424.

[8] Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 57-66

[9] E. Bakshy, J. Hofman, W. Mason, and D. J. Watts (2011). Everyon has an influencer: Quantifying influence on twitter. In Proceedings of WSDM.

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

[11] Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer opinion platforms: What motivates consumers to articulate themselves on the internet? Journal of Interactive Marketing , 18 (1), 38-52.

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

[13] J. Yu, B. Benatallah, F. Casati, F. Daniel (2008). Understanding mashup development. IEEE Internet Computing, 12(5):44–52.

[14] Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. The Public Opinion Quarterly , 21 (1), 61-78.

[15] Katz, J. M. (2009 December). Defining influence as a strategic marketing metric. Forrester Research Inc.

[16] 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.

[17] 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.

[18] Lardinois, F. (2010 29-April). On Twitter, it's just five degrees of separation. From Read Write Web: http://rww.to/9YpuJ0

[19] 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.

[20] Makrehchi, M. (2006). Learning social networks from Web documents using support vector classifier. Proceedings of IEEE/WIC/ACM International Conference on web intelligence, (p. 88-94).

[21] Marsden, P. V. and Campbell K. E., (1984). Measuring Tie Strength. Social Forces, 63: 482-501.

[22] Microsoft. (2007). Worldwide Buzz: Planetary ‐ Scale Views on an Instant ‐ Messaging Network.

[23] Milgram, S. (1967). The small world problem. Psichology Today , 1 (1), 60-67.

[24] Musial, K., Kazienko, P., & Brodka, P. (2009). User position measures in social networks. In Proceedings of the 3rd Workshop On Social Network Mining and Analysis (pp. 1-9). Paris, France: ACM.

[25] Ohira, Ohsugi, Ohoka, & Matsumoto. (2005). D-sns: a knowledge exchange mechanism using social network density among mega-community users.

[26] Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing of electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. Journal of Advertising Research , 44 (4), pp. 333-348.

[27] Richins, M., & Root-Shaffer, T. (1988). The role of involvement and opinion leadership in consumer word-of-mouth: an implicit model made explicit. Advances in Consumer Reasearch , 15, 32-36.

[28] Rogers, E. (1962). Diffusion of innovation. New York, NY, USA: Free Press.

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

[30] S Anholt (2009). Places: Identity, image and reputation. Palgrave Macmillan.

[31] S. Ba and P. Pavlou (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. MIS Quarterly, 26(3): 243–268.

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

[33] Sheridan, D., Muhamad, R., & Watt, D. (2003). An Experimental Study of Search in Global Social Networks. Science, 301 (5634), p. 827-829.

[34] Strang, D., & Soule, S. (1998). Diffusion in organization and social movements: From hybrid corn to poison pills. Annual Review of Sociology , 24 (1), pp. 265-290.

[35] Subramani, M. R., & Rajagopalan, B. (2003 December). Knowledge-sharing and influence in online social networks via viral marketing. Commun. ACM , 46 (12), pp. 300-307.

[36] Sun, T., Youn, S., Wu, G., & Kuntaraporn, M. (2006). Online word-of-mouth (or mouse): An exploration of its antecedents and consequences. Journal of Computer-Mediated Communication , 11 (4).

[37] Watts, D., & Dodds, P. (2007). Influentials, networks and public opinion formation. Journal of Consumer Research , 34 (4), 441-458.