POLITECNICO DI MILANO

Corso di Laurea Specialistica in Ingegneria Gestionale

Dipartimento di Ingegneria dei Sistemi

"SOCIAL MEDIA INTELLIGENCE: L'ANALISI DELLA INFLUENCE NEL MICROBLOGGING"

Tesi di Laurea Specialistica di:

Cesare D'ADDA

(matr. 735644)

Relatore:

Prof.ssa Chiara FRANCALANCI

Correlatori:

Dott.ssa Fiamma PETROVICH

Ing. Donato BARBAGALLO

Anno accademico 2009-2010

Per mio papà.

Sommario

Il Web 2.0 e il social networking hanno radicalmente modificato il modo in cui gli utenti interagiscono, ricercano e condividono informazione. L'utilizzo dell'immenso bacino di dati non strutturati provenienti in tempo reale dai social media come fonte di data mining, processo complesso e tecnologicamente abilitato dalle cosiddette piattaforme di ascolto, è un mezzo sempre più diffuso ed impiegato da aziende e istituzioni pubbliche, finalizzato ad analisi d'intelligence e all'acquisizione di informazione di valore. Aspetti rilevanti e di grande interesse, soprattutto per il marketing, sono le modalità con le quali gli individui generano e scambiano, attraverso il passaparola online, i.e. eWOM, opinioni ed esperienze su brand, prodotti e servizi. Gli studi sulla circolazione dei messaggi all'interno delle reti sociali e tematiche come l'opinion leadership, la influence e la viralità dei contenuti, i quali risalgono già i primi anni '50, sono diventati più che mai attuali proprio grazie alla significativa diffusione dei servizi 2.0., tra i quali il microblogging.

Il lavoro di tesi, accompagnato da uno stage svolto presso la società di consulenza e ricerca di marketing CommStrategy, si inserisce all'interno di un progetto di gestione della reputation online e sentiment analysis condotto dal Politecnico per il Comune di Milano. Il periodo di lavoro ha permesso una profonda comprensione, grazie anche all'utilizzo prolungato di alcuni tool, di quello che è la *social media analysis*, delle caratteristiche e della complessità che la contraddistinguono. L'obiettivo primario della tesi è quello di individuare delle linee guida finalizzate alla realizzazione e all'implementazione all'interno di un tool prototipo di monitoraggio, di un modulo specificatamente dedicato alla *influencer analysis* in Twitter, la più conosciuta piattaforma di microblogging.

Ringraziamenti

Desidero ringraziare la Prof.ssa Chiara Francalanci per avermi dato l'opportunità di lavorare a questo progetto, per aver creduto nella mia capacità e per il tempo dedicato alla mia tesi.

Un sincero ringraziamento all'Ing. Donato Barbagallo per avermi affiancato costantemente durante la stesura del lavoro, per l'aiuto e la pazienza mostrata.

Ringrazio sentitamente il Dott. Paolo Barbesino e la Dott.ssa Fiamma Petrovich di CommStrategy per la splendida esperienza professionale vissuta durante il periodo di stage, per la fiducia sempre mostrata nel mio lavoro e per le preziose conoscenze trasmesse.

La mia gratitudine ai compagni di corso di questi anni e agli amici di sempre, che hanno condiviso con me esperienze ed avventure universitarie e nella vita.

Il grazie più grande va ai miei genitori e alla mia famiglia, per il costante sostegno ed aiuto che mi hanno sempre dato durante questi anni di studi e perché è soprattutto merito loro se sono arrivato sin qui.

Cesare

Politecnico di Milano
Ottobre 2010

Indice

Sommario ..... **II**

Ringraziamenti ..... **III**

Indice ..... **IV**

Elenco delle figure ..... **VII**

Elenco delle tabelle ..... **IX**

Capitolo 1 - Introduzione **1**

Capitolo 2 - Stato dell'arte **3**

2.1 Introduzione 3

2.2 Le reti sociali 4

2.2.1 Cenni teorici sulle reti sociali 7

2.2.2 Metriche per l'analisi delle reti sociali 8

2.3 I social network online 10

2.3.1 Il microblogging e Twitter 12

2.4 Il WOM e i Social Media: nozioni introduttive 13

2.4.1 Dalla brand image alla brand reputation 17

2.4.2 La misurazione del WOM 17

2.4.3 La influence e l'influencer outreach 19

2.4.4 Il microblogging come strumento per il Web marketing 21

2.5 Gli influencer e le rete sociali 23

2.5.1 Caratteristiche del network 24

2.5.2 Metriche di posizionamento dell'utente 26

2.5.3 Tassonomia degli influencer 31

2.6 Twitter e la influence: la letteratura 34

Capitolo 3 - Twitter: Il profilo sociale 38

3.1 Introduzione 38

3.2 Informazioni dei profili e funzionalità 39

3.2.1 Mobilità e applicazioni di terze parti 41

3.3 Le conversazioni in Twitter 42

3.3.1 Contenuto vs conversazione 43

3.3.2 Il tagging dei contenuti 44

3.3.3 Le tipologie di contenuti presenti nel microblogging 45

3.3.4 Volatilità, attualità e credibilità dei contenuti 46

3.4 Caratteristiche del network 48

3.5 Demografiche 50

3.5.1 Distribuzione geografica 52

3.6 Tassonomia degli utenti 54

Capitolo 4 - Il processo di analisi e la tecnologia 57

4.1 Introduzione 57

4.2 Il processo di analisi 59

4.2.1 Definizione dei tre processi 60

4.3 Componenti dei tool 61

4.4 L'offerta del mercato 67

4.4.1 Esempi di soluzioni integrate 68

4.4.2 Esempi di freemium tool 72

4.5 L'estrazione dei dati da Twitter 75

4.5.1 Il crawling 75

4.5.2 Le API di Twitter 77

4.5.3 Il database e i diagrammi E/R 80

4.5.4 La Data visualization 82

Capitolo 5 - Analisi e risultati 95

5.1 Introduzione 95

5.2 Il data set di utenti 96

5.3 I paramentri della influence..... 98

5.4 Analisi per categorie..... 106

5.5 Conclusioni..... 111

Capitolo 6 - Conclusioni e sviluppi futuri 117

Bibliografia 121

Elenco delle figure

Figura 1 - Il social media prism (Solis, Introducing the conversation prism, 2008)..... 11
Figura 2 - Le componenti del net promoter score 18
Figura 3 - Le componenti del word-of-mouth equity..... 19
Figura 4 - Componenti generali del modello di branding 23
Figura 5 - Rappresentazione degli archetipi sun, spider e surce (Forrester)..... 32
Figura 6 - Esempio di profilo utente in Twitter 40
Figura 7 – Metodi di input e output in Twitter..... 41
Figura 8 – Distribuzione dei contenuti per tipologia..... 46
Figura 9 - Posizionamento di Twitter rispetto alla volatilità e credibilità dei contenuti... 48
Figura 10 - Grafo delle relazioni tra gli utenti di Twitter..... 48
Figura 11 - Distribuzione degli utenti per fascia di età e reach di Twitter..... 51
Figura 12 – Mappa delle principali nazioni per traffico sul sito 52
Figura 13 - Parti costituenti il monitoraggio dei social media 58
Figura 14 - Componenti del processo di social media analysis 61
Figura 15 - Architettura di un tool di benchmark..... 64
Figura 16 - Interfaccia utente di Radian6..... 66
Figura 17 - Schermata del modulo per l'analisi degli utenti di Radian6 69
Figura 18 - Schermata del tool Map di Sysomos 71
Figura 19 - Schermata di Klout..... 72
Figura 20 - Schermata di Twitalyzer..... 75
Figura 21 - Interazione client-server in Twitter 80
Figura 22 - Diagramma ER di Twitter 81
Figura 23 – Schermata del tool Revisit..... 83
Figura 24 – Schermata del tool Streamdin..... 84
Figura 25 - I cinque fattori competitivi secondo il modello di Anholt 85
Figura 26 – Sezione dell'alberatura con alcune label 87
Figura 27 - Architettura del tool del progetto 90
Figura 28 - Interfaccia prototipo del tool 92
Figura 29 - Grafico log-log di distribuzione dei follower..... 100

Figura 30 - Grafico log-log di distribuzione delle mention 101
Figura 31 - Grafico log-log di distribuzione dei retweet..... 101
Figura 32 - Grafico lognormale Q-Q di distribuzione delle mention..... 103
Figura 33 - Grafico lognormale Q-Q di distribuzione dei retweet..... 103
Figura 34 - Matrice di confronto del sentiment influencer/rest of us..... 115

Elenco delle tabelle

Tabella I - Le motivazioni del passaparola individuate in letteratura 14
Tabella II - Le caratteristiche del network 26
Tabella III - Riassunto delle metriche di posizionamento dell'utente 31
Tabella IV - Caratteristiche degli archetipi source, spider e sun (Forrester) 33
Tabella V - Le tipologie di influencer secondo Gartner 34
Tabella VI - Utilizzo dell'hashtag in un set di messaggi 45
Tabella VII – Valutazione utente in base al rapporto indegree/outdegree 49
Tabella VIII - Dimensione dell'audience 51
Tabella IX - Principali nazioni per contributo all'attività e al numero di utenti 53
Tabella X - Principali città per attività al minuto 53
Tabella XI - Principali città per contributo all'attività e al numero di utenti 54
Tabella XII - Categorizzaione degli utenti per modalità di utilizzo 55
Tabella XIII - Categorizzazione per tipologia di account 56
Tabella XIV - Indicatori presi in considerazione per le metriche di Klout 73
Tabella XV - Caratteristiche riassuntive delle metodologie di crawling 77
Tabella XVI - Estratto di esempio del data set 97
Tabella XVII - Statistiche descrittive del data set 98
Tabella XVIII - Metriche per la misura della influence in Twitter 99
Tabella XIX - Indici di correlazione di Spearman per i cluster A e B 104
Tabella XX - Statistiche descrittive per i cluster A e B 105
Tabella XXI - t-test per l'uguaglianza media dei due cluster 105
Tabella XXII – Analisi ANOVA per il confronto delle medie per le diverse tipologie di account 108
Tabella XXIII - Comaparazioni multiple con il metodo della correzione di Bonferroni 109

Capitolo 1

Introduzione

"The web is more a social creation than a technical one. I designed it for a social effect - to help people work together - and not as a technical toy. The ultimate goal of the Web is to support and improve our weblike existence in the world. We clump into families, associations, and companies. We develop trust across the miles and distrust around the corner."

Tim Barners-Lee

Il Web 2.0 e il social networking hanno largamente e profondamente influenzato le modalità con le quali gli utenti ricercano informazione e interagiscono tra di loro online. L'immensa mole di dati non strutturati e in real time provenienti da questi servizi richiede però tecnologie adeguate che abilitino un processo estremamente complesso che va dall'estrazione fino allo riutilizzo di questa informazione, ripresentata attraverso dashboard interattive e/o report che impiegano metriche e indicatori sintetici efficaci. Aziende e istituzioni pubbliche stanno sempre più comprendendo l'imprescindibile necessità di integrare questa tipologia di analisi all'interno della loro business intelligence, con la possibilità poi di servirsene finalizzando azioni o strategie significativamente differenti tra di loro, e.g. ricerche di marketing, brand management & reputation, customer relationship management.

Di grande rilevanza specialmente per le azioni marketing, è il processo di interazione tra utenti grazie al quale avviene lo scambio di opinioni ed esperienze, i.e. il passaparola, diffusamente considerato come la primaria fonte di informazione sulle decisioni di acquisto dal 20% al 50% dei casi. Il suo potere di influenzare le scelte, elevato soprattutto quando l'acquisto di un particolare prodotto avviene per la prima volta o nel caso si tratti di un bene molto costoso, fattori questi che inducono a una ricerca più approfondita di pareri altrui, ha reso il word-of-mouth non più un atto di comunicazione che avviene in modalità privata e *one-to-one*, ma, ed è proprio qui che si rileva il suo aspetto di maggiore potenzialità, che opera su una base assolutamente *one-to-many*, con i social network che abilitano la condivisione su larga scala di recensioni o esperienze postate online da singoli individui.

Diverse piattaforme 2.0 visti i significativi bacini di utenza di cui sono dotate, e.g. il microblogging e nel caso specifico Twitter per le sue caratteristiche intrinseche, vengono da un lato sottoposte a studi ed analisi focalizzati sulla diffusione dei contenuti e della opinion leadership all'interno dei rispettivi network, dall'altro sempre più coinvolte ed utilizzate dai brand come mezzo di comunicazione, promozione ed engagement dei consumatori finali.

Il lavoro si propone uno studio e una comprensione delle teorie e delle innovative tecnologie esistenti volte proprio all'estrazione e all'analisi di informazioni provenienti dai siti Web 2.0, con il preciso scopo finale di individuare delle linee guida finalizzate alla realizzazione e all'implementazione all'interno di un tool prototipo di monitoraggio, di un modulo specificatamente dedicato alla *influencer analysis* in Twitter, la più conosciuta piattaforma di microblogging.

La tesi è strutturata come segue. Il Capitolo 2 presenta una completa e piuttosto eterogenea panoramica sulle tematiche trattate, soffermandosi in particolare sulla storia dell'analisi delle reti sociali, fornendo alcune metriche fondamentali per valutazione di queste ultime; le principali teorie e nozioni fondamentali riguardanti il passaparola, sia offline che online; ed infine le teorie sulla opinion

leadership e la diffusione dell'informazione all'interno di un network, supportate da alcune metriche di posizionamento degli utenti e possibili tassonomie per gli utenti che fanno parte della rete sociale.

Il Capitolo 3 è interamente dedicato a Twitter, la piattaforma di microblogging studiata ed impiegata nell'analisi. Lo scopo del Capitolo è trattarne e evidenziarne gli aspetti e le caratteristiche principali.

Il Capitolo 4 descrive in modo dettagliato tutte le fasi del processo molto complesso di *social network analysis*, definibile anche con il nome di social media intelligence. Saranno inoltre indicati degli esempi tratti dalla practice attuale e descritti nello specifico alcuni elementi tecnologici riguardanti Twitter e l'estrazione dei dati.

Il Capitolo 5 spiega, partendo dalla selezione del data set, le analisi svolte e i risultati ottenuti.

Capitolo 2

Stato dell'arte

2.1 Introduzione

L'analisi delle reti sociali è un ambito di ricerca nato ad inizio secolo scorso e da sempre utilizzato per spiegare fenomeni legati a differenti ambiti di studio, quali la sociologia, la psicologia e l'economia. Con la comparsa di social network online, blog e forum e soprattutto in seguito all'esplosione del fenomeno Web 2.0, gli studi riguardanti le community online si sono intensificati, da un lato con l'obiettivo di utilizzare i dati di queste strutture chiuse per l'analisi di alcuni pattern e comportamenti sociali, dall'altro, visto il crescente utilizzo del Web come canale di marketing e comunicazione, per valutare l'opportunità e l'impatto delle campagne online.

Il capitolo è strutturato come segue: la Sezione 2.2 fornisce una panoramica della storia e delle metriche impiegabili per la valutazione delle reti sociali; nella Sezione 2.3 è presente, oltre alla definizione di social network online, una breve introduzione sul microblogging; la Sezione 2.4 è dedicata alla trattazione di tutti gli aspetti riguardanti il fenomeno del passaparola all'interno delle reti sociali, con il passaggio dalla forma tradizionale a quella online, il tema della influence e la possibilità d'impiego del microblogging come strumento virale; la Sezione 2.5 definisce il ruolo degli influencer all'interno di un network, introducendo alcune metriche utilizzabili per la misura del posizionamento e delle possibili tassonomie; infine nella Sezione 2.6 viene presentata brevemente la letteratura riguardante gli studi già effettuati sulla influence all'interno di Twitter.

2.2 Le reti sociali

Gli studi sulle reti sociali si sono guadagnati dei riconoscimenti significativi negli ultimi anni sia in termini di avanzamenti teorici che di metodologie, andando ad impattare notevolmente su vari domini come il capitale sociale, la gestione della conoscenza e le teorie organizzative. Basate su costrutti teorici della sociologia e su quelli matematici propri della teoria dei grafi, complice anche la recente evoluzione di hardware e software, l'analisi delle reti sociali offre una metodologia che unifica la visualizzazione e l'investigazione delle strutture e delle relazioni sociali. Se da un lato, per alcuni problemi, un sondaggio generale di carattere sociale può aiutare nello studio delle proprietà individuali in prima approssimazione, l'analisi delle reti sociali incorpora il contesto sociale per spiegare sia i comportamenti sociali che di gruppo, in questo modo le relazioni tra gli attori diventano di primaria importanza, mettendo in seconda posizione le caratteristiche proprie del singolo individuo.

Lo sviluppo del campo delle reti sociali ha inizio negli anni Trenta tramite gli studi di diversi gruppi di ricerca al tempo al lavoro in modo indipendente. Grazie a Simmel cominciò ad emergere un approccio sistematico, costruendo una teoria che spiegava le cause dei fenomeni sociali e contribuì alla sociologia formale che può quindi essere considerata la progenitrice dell'analisi delle reti sociali.

Nel 1934 Moreno fu il primo a operazionalizzare una rete sociale creando un sistema per rappresentarla come una combinazione di nodi e collegamenti tra essi.

Successivamente Cartwright e Harary (Harary, Norman, & Cartwright, 1965; Cartwright, 1959) continuarono il lavoro di Moreno cominciando ad applicare i concetti della teoria dei grafi a quelli che erano allora detti "sociogrammi". Grazie soprattutto all'introduzione dei collegamenti direzionali tra i nodi, nei loro studi furono capaci di spiegare pattern sociali di complessità molto maggiore rispetto ai risultati che si erano raggiunti sino a quel momento.

Alla fine degli anni Trenta si formarono due scuole separate: una, americana, costituita da un gruppo di lavoro presso l'università di Harvard, l'altra, britannica, dagli antropologi dell'università di Manchester.

La prima si focalizzava soprattutto su tecniche per individuare e studiare sottogruppi di persone all'interno di gruppi originari più ampi, con lo scopo di analizzare e comprendere i rapporti tra i sottogruppi stessi. Da questa scuola, come illustrato brevemente in (Chung, Hossain, & Davies, 2005), si generò il cosiddetto *approccio sociocentrico*, che concerne lo studio quantitativo di relazioni tra le persone all'interno di un determinato gruppo, con l'obiettivo di misurare dei pattern strutturali tra le interazioni e di come questi pattern riescano a spiegare i risultati.

La seconda scuola, invece, si concentrò maggiormente sullo studio delle comunità dando origine all'*approccio egocentrico*. I ricercatori di questa scuola analizzarono le reti di relazioni sottostanti agli individui piuttosto che focalizzarsi sull'intera società, con l'obiettivo finale di generalizzare le caratteristiche trovate sulla rete personale all'interno del gruppo.

Successivamente, i lavori di Barres (Barres, 1954), Granovetter (Granovetter, The strenght of weak ties, 1973; Granovetter, The strenght of weak ties: a network theory revisited, 1983) e Milgram (Milgram, 1967) formalizzarono la teoria dell'analisi delle reti sociali tra gli anni 50' e 70'.

In particolare, Barres ha dato origine alla nozione di rete sociale di cui successivamente Granovetter sottolineò l'importanza come mezzo di diffusione

della conoscenza e dell'informazione attraverso i suoi lavori sui legami deboli tra individui.

Milgram introdusse il famoso concetto dei *six degrees of separation* (Milgram, 1967) con cui tentava di dimostrare l'idea di quello che definì "mondo piccolo" (*small world phenomenom*). Mediante un approccio empirico supportò la sua tesi mostrando che tutte le persone negli Stati Uniti sembravano essere connesse tra di loro in media da sei legami di conoscenza con qualsiasi altro connazionale attraverso la presentazione di circa sei persone.

Contemporaneamente, importanti lavori indipendenti si svilupparono presso l'università della California (Irvine) soprattutto intorno a Freeman, che si concentrò sulla definizione dei nodi importanti all'interno di una rete e delle relative metriche (Freeman, 1979).

Oggi la teoria delle reti sociali è largamente utilizzata per studiare l'influenza della struttura della rete sociale in un'organizzazione o in un gruppo di lavoro e la gestione della conoscenza al loro interno. La gestione della conoscenza (*knowledge management*) comprende un insieme di pratiche volte ad identificare, creare, rappresentare e ridistribuire la conoscenza all'interno delle organizzazioni.

La maggior parte degli studi si è focalizzata su due principali pratiche per gestire la conoscenza in maniera formale: l'applicazione di sistemi di supporto e le comunità. La prima soluzione punta sulla codifica e la distribuzione della conoscenza tramite l'*information and communication technology*, come i database e internet; la seconda, invece, promuove lo scambio di conoscenza tra persone con interessi e obiettivi comuni. Queste ricerche hanno dimostrato che chiedere consigli ai propri parigrado è un importante canale che favorisce quotidianamente uno scambio di conoscenza molto specifica. Per questo motivo una parte della ricerca in quest'ambito si sta focalizzando sulla relazione tra i benefici del *knowledge management* che derivano dalla struttura delle reti sociali corrispondenti, come illustrato in (Ohira, Ohsugi, Ohoka, & Matsumoto, 2005; Ohira, Nakakoji, & Matsumoto, D-sns: a knowledge exchange mechanism using social network density among mega-community users, 2006).

2.2.1 Cenni teorici sulle reti sociali

L'idea alla base delle reti sociali è molto semplice: una rete è un insieme di attori (o 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 essi:

$$N = (U, R_{1,} R_{2,\ldots,} R_r)$$

Le relazioni possono rappresentare qualunque tipo di legame: per esempio il legame padre e figlio oppure l'appartenenza a uno stesso progetto da parte di due sviluppatori o i rapporti di amicizia tra le persone.

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

a. tramite la matrice binaria $R = [r_{ij}]_{nxn}$, detta anche matrice di adiacenza, dove:

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

se gli archi sono pesati il valore di r_{ij} può essere il numero reale che indica la "forza" del legame *R* tra X_i e X_j.

b. tramite la lista dei vicini, specificando per ciascun nodo la lista degli altri nodi a cui è relazionato;

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

E' possibile elencare le tipologie di reti sociali più utilizzate in letteratura:

a. *rete non direzionata*: la relazione R è simmetrica, ovvero tutti gli archi non hanno una direzione specifica, come per esempio nella relazione matrimonio tra due persone;

b. *rete direzionata*: la relazione R è simmetrica, ovvero tutti gli archi sono orientati, come ad esempio nella relazione di paternità padre e figlio;

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

d. *rete a due modi*: è formata da due insiemi di unità $U = U_a \cup U_e$, spesso definiti in letteratura come attori U_a ed eventi U_e e una relazione R che connette i due insiemi, per esempio l'appartenenza di una persona ad una associazione.

Le proprietà e le relative misure sulle reti si basano su due differenti livelli di analisi: quelle relative alla rete considerata nella sua interezza come relazioni tra insiemi di attori e quelle relative al singolo individuo (o elemento).

2.2.2 Metriche per l'analisi delle reti sociali

In questa sezione sono descritte le metriche per l'analisi delle reti sociali maggiormente utilizzate in letteratura. Verranno qui illustrate inizialmente solo quelle che si occupano di descrivere le caratteristiche della rete nella sue

interezza, mentre nella Sezione 2.5.2 quelle pertinenti l'analisi del posizionamento dei singoli nodi.

Dimensione della rete

La *dimensione della rete*, ottenuta semplicemente contando il numero di nodi presenti e quindi definita come:

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

E' un parametro tanto semplice quanto importante da tenere in considerazione nello studio di un problema basato su reti sociali. Basti considerare, per esempio, un insieme di studenti che frequentano un dato corso: se il numero di studenti è pari a qualche decina si ha un'alta probabilità che questi si conoscano tutti tra di loro e quindi possano maggiormente condividere conoscenza.

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, ovvero:

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

Questo parametro permette di analizzare alcuni fenomeni come ad esempio la velocità con quale l'informazione si può diffondere tra i nodi o quanto gli attori in gioco siano legati tra di loro e quindi condividano il capitale sociale e i suoi vincoli. Una rete molto densa, infatti, presenta un alto numero di legami tra i suoi attori che, di conseguenza, possono raggiungere più facilmente ciascun altro membro della rete e quindi accedere in modo più efficiente all'informazione.

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

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

Questa metrica può essere utilizzata come indicatore durante il processo di costruzione di un'intera rete sociale partendo da un suo sottoinsieme (Makrehchi, 2006), utile nel caso in cui non vi siano dati sufficientemente ampi 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. Bisogna sottolineare come, se le relazioni non sono simmetriche e quindi gli archi sono direzionati, è possibile che se l'attore *A* sia raggiungibile dall'attore *B*, 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.

2.3 I social network online

Un social network, termine con cui è comunemente indicata una rete sociale online, è definibile come un servizio Web-based che permette agli individui di (Ellison & Boyd, 2007):

a. creare un profilo pubblico o semi-pubblico in un sistema limitato;

b. articolare una lista di altri utenti con il quale condividere una connessione;

c. visualizzare e navigare tra le liste di connessioni degli altri utenti attraverso il sistema.

L'elemento principale di ciascun sito di social networking è quindi la collezione di un insieme di profili appartenenti agli utenti registrati, dove questi possono pubblicare dell'informazione che intendono condividere con gli altri individui della rete sociale. Gli utenti sono coinvolti per lo più in due differenti tipologie di attività: la creazione di contenuti attraverso l'editing dei loro profili (e.g. caricamento di foto, musica, messaggi, scrivere post su un blog etc.) ed il consumo di contenuto generato dagli altri nodi del network (e.g. lettura dei blog, guardare le foto, scaricare musica etc.) (Trusov, Bodapati, & Ucklin, 209).

Lo sviluppo delle reti sociali online, come blog, wiki e piattaforme di social networking ha dimostrato l'abilità di generare comunità online a crescita molto rapida, dove gli utenti comunicano, condividono informazione e si mantengono in contatto, molto spesso anche senza conoscersi direttamente gli uni con gli altri. A differenza del cosiddetto Web 1.0, che era unicamente e largamente organizzato intorno al contenuto, il Web 2.0, espressione introdotta per la prima volta nel 2004 da Tim O'Reilly (O'Reilly, 2005), è organizzato avendo come punto di riferimento gli utenti.

Sin dalla loro nascita, piattaforme come Facebook[1], Myspace[2], Flickr[3] e Twitter hanno attratto milioni di utenti, la maggior parte dei quali ha integrato questi servizi all'interno delle proprie attività giornaliere, rendendo il social networking la prima attività, ed in continua crescita per peso relativo, rispetto al totale del tempo speso sul Web (Nielsen Company, 2009; Nielsen Company, 2010) .

Figura 1 - Il social media prism (Solis, Introducing the conversation prism, 2008)

[1] http://www.facebook.com
[2] http://www.myspace.com
[3] http://www.flickr.com

Le piattaforme di social sono delle fonti molto ricche di dati riguardanti i comportamentali naturalistici degli individui. I dati riguardanti i singoli profili e/o i collegamenti possono essere raccolti attraverso l'impiego di tecniche di estrazione automatizzate oppure attraverso data set forniti dalle stesse società proprietarie, consentendo lo svolgimento di analisi delle reti volte ad esplorare su larga scala diversi pattern riguardanti l'utilizzo (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006), la natura e la struttura dei legami tra nodi (Lampe, Ellison, & Steinfeld, 2006; Kumar, Novak, & Tomkins, 2006) e ulteriori metriche significative (Hogan, 2008).

2.3.1 Il microblogging e Twitter

Il microblogging[4] è definibile come un mezzo di comunicazione che utilizza lo stile ed il formato del blogging, differendo dal blog tradizionale per un contenuto tipicamente e significativamente ridotto; un post infatti, può essere costituito unicamente da una frase, da un immagine o da un video *embedded*.

Twitter[5], nato a San Francisco da una start-up chiamata Obvious e lanciato nell'Ottobre 2006, è la più diffusa e conosciuta piattaforma di microblogging, con più di 145 milioni di utenti registrati e 180 milioni di visitatori unici al mese (Huffington Post, 2010; Techcrunch, 2010). In grado di guadagnare una discreta popolarità fin dal suo lancio (Milstein, Chowdhury, Hochmuth, Lorica, & Magoulas, 2008), e' stato protagonista assoluto della celebre conferenza SXSW '07 (*South by Southwest*) (Douglas, 2007), durante il quale, oltre a triplicare i volumi di traffico giornaliero, il suo nome ha cominciato a circolare attivamente all'interno della community del Web, con un forte incremento del tasso di crescita, tutt'ora assolutamente significativo con la registrazione di circa 300.000 nuovi account al giorno.

Twitter combina elementi tipici di altri servizi di social networking (Ellison & Boyd, 2007) e dei blog (Marlow, 2005), seppur con alcune indicative differenze. La connessione tra gli utenti avviene attraverso lo scambio di brevi ma frequenti

[4] http://en.wikipedia.org/wiki/microblogging
[5] http://www.twitter.com/

messaggi, noti come *tweets* (letteralmente "cinguettii"), che costituiscono sostanzialmente dei post in formato testuale con una lunghezza massima di 140 caratteri. Quest'ultima caratteristica, che potrebbe sembrare una forte limitazione, ne ha rappresentato invece uno dei principali punti di forza nella competizione con gli altri servizi.

Twitter è stato in grado di attrarre un forte interesse anche da parte dei brand per il suo potenziale utilizzo come mezzo per il marketing virale, per pubblicizzare prodotti e fornire informazioni e supporto agli utenti interessati. La piattaforma inoltre, è pesantemente utilizzata dalla grande maggioranza delle fonti di news, le quali la impiegano come mezzo di diffusione degli aggiornamenti, successivamente filtrati e commentati dalla community di iscritti.

2.4 Il WOM e i Social Media: nozioni introduttive

Come detto, gli studi riguardanti il flusso dell'informazione e delle idee all'interno delle reti sociali offline, i.e. le società, cominciarono già negli anni '50, in cui venne avanzata ed analizzata l'ipotesi che questo flusso di conoscenza ed opinioni procedesse partendo dai mass media, arrivando direttamente agli *opinion leaders* e poi da questi alle componenti meno attive della popolazione, attraverso una comunicazione definita a "due passi" (*two-step communication flow)*, in cui il passaparola (*word-of-mouth*) assumeva un ruolo cruciale (Katz E. , 1957).

E' stato dimostrato come quest'ultimo nella sua forma tradizionale, quindi offline, sia un elemento primario e decisivo nelle rete sociali nel definire le decisioni di acquisto dei consumatori (Richins & Root-Shaffer, 1988), oltre ad esser dotato di un effetto diretto sulla diffusione dell'innovazione e di nuovi prodotti (Strang & Soule, 1998).

Nella tabella seguente è presente un breve riassunto delle motivazioni, identificate dalla letteratura, che spingono alle comunicazioni con il metodo del passaparola (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004).

Autori	Motivazioni	Descrizione
Dichter (Dichter, 1966)	*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 (Engel & Blackwell, 1994)	*Involvement*	Il livello di interesse e coinvolgimento rispetto ad un topic serve a stimolare la discussione.
	Self-enhancement	L'esprimere opinioni e raccomandazioni permette all'individuo di guadagnare attenzione, dare l'impressione di possedere informazione di valore e asserire una certa superiorità.
	Concern for others	Un genuino desiderio di aiutare un amico o un parente nel fare una migliore decisione di acquisto.
	Message intrigue	Piacere 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 (Sundaram & Mitra, 1998)	*Altruism (positive WOM)*	L'atto del fare qualcosa per gli altri senza il ricevimento di 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 come 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 un ripago contro la compagnia associata all'esperienza di consumo negativa.
	Advice Seeking	Ottenere consigli su come risolvere problemi.

Tabella I - Le motivazioni del passaparola individuate in letteratura

L'esponenziale crescita dei social network online durante i primi anni del 2000, tale da renderli un fenomeno assolutamente di massa, ha reso *l'online word-of-mouth*, o *eWOM* (i.e *electronic word-of-mouth*), uno degli ambiti di ricerca principali all'interno della comunicazione cosiddetta *computer-mediated*, in modo particolare nel contesto delle comunicazione *consumer to consumer* (Sun, Youn, Wu, & Kuntaraporn, 2006).

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" (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004).

Il valore e l'interesse diffusosi per il passaparola online è dovuto in modo particolare alle caratteristiche intrinseche che possiede il mezzo di propagazione in questione, e cioè Internet. A differenza del contesto tradizionale, in cui è l'interazione avviene attraverso il linguaggio parlato e faccia a faccia (Wellman, Salaff, Dimitrova, Garton, Giulia, & Haythornthwaite, 1996; Bickart & Schindler, 2001), l'*eWOM* si basa sulla trasmissione di opinioni ed esperienze in forma scritta, le quali sono da considerarsi dotate di un potere di diffusione e relativa influenza significativamente superiore per le seguenti ragioni:

a. 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 con una comunicazione con qualsiasi media tradizionale (Phelps, Lewis, Mobilio, Perry, & Raman, 2004);

b. 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) (Subramani & Rajagopalan, 2003);

c. bidirezionalità dell'interazione: i social media permettono feedback ed uno scambio di opinioni in real time tra la fonte e il ricevente del messaggio (Dellarocas, 2003).

Per questa serie di motivazioni, i servizi di social networking sono stati progressivamente riconosciuti come un'importante fonte di informazione e scambio di opinioni in grado anch'essa di influenzare in modo riconoscibile l'adozione e l'utilizzo di prodotti e servizi (Subramani & Rajagopalan, 2003).

Molto diffuse quando si parla di queste tematiche sono espressioni come *viral (marketing)*, *buzz online* e come detto, *WOM* e *eWOM*, sui quali, benchè sinonimi, è utile fare alcune precisazioni e chiarimenti (Guya, 2006).

- *viral* – per viral si intende il meccanismo di propagazione del messaggio, appunto perché come un virus attecchisce, colpisce gli individui che inglobano in se 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 questa o quella 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 che lo ha particolarmente colpito, soddisfatto o interessato;

- *word-of-mouth* – è tradizionalmente definito come il processo di trasporto dell'informazione da persona a persona (Richins & Root-Shaffer, 1988), 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.

2.4.1 Dalla brand image alla brand reputation

Da sempre le marche hanno cercato di costruirsi, attraverso gli strumenti del marketing, dall'advertising alle PR, un'immagine. La *brand image* può essere considerata come "la marca che parla di sé", mentre, al contrario, la *reputation* è il risultato più o meno diretto di un processo di creazione collettiva della percezione del brand (Cova, Giordano, & Pallera, 2007). Quest'ultimo "è costituito dall'insieme dei discorsi tenuti su di essa dalla totalità dei soggetti (individuali e collettivi) coinvolti nella sua generazione" (Semprini, 2006). Le persone che parlano della marca, che esprimono opinioni e giudizi su di essa, sono quindi responsabili di una buona o cattiva reputazione della stessa. Quest'ultima è inoltre molto più legata alla sostanza dei comportamenti dell'azienda, rispetto all'apparenza delle dichiarazioni di chi ne gestisce le attività di comunicazione. E se da un lato non risulta particolarmente difficile aumentare nel breve periodo la propria visibilità in termini di notorietà, la reputazione non può essere manipolata o condizionata facilmente, in quanto la fiducia delle persone non può essere acquistata come l'immagine.

Il concetto di *brand reputation* riconosce quindi il crescente potere di accesso all'informazioni delle persone e la situazione di aumentata trasparenza in cui si trovano le aziende al tempo del Web. Più che alla semplice opinione positiva che le persone hanno sulla qualità dei prodotti o dei servizi, la reputazione ha a che fare con la capacità di entusiasmare gli animi e mobilitare le persone, in definitiva dipende dalla capacità di creare veri e proprio sostenitori. In questo senso è legata intrinsecamente alla raccomandazione e al passaparola che è in grado di stimolare.

2.4.2 La misurazione del WOM

Il valore del word-of-mouth per chi compra un prodotto è indiscutibile: sistematicamente il passaparola viene indicato come la forma di comunicazione che più di ogni altra influenza la decisione di acquisto. La logica che sta dietro a questo è molto semplice: il passaparola riduce il rischio. La difficoltà vera e propria, soprattutto per le aziende, è quella di attribuire una valutazione ed un preciso un valore numerico a questo fenomeno, rendendolo in qualche modo

misurabile. Una possibile soluzione del problema consiste nell'utilizzare le ricerche per studiare i "tassi di raccomandazione" (*recomendation rates*) relativi a prodotti e servizi per poi metterli in relazione con l'aumento delle vendite. Il *net promoter score* (Reichheld, 2003), misurato con un valore percentuale (da 1 a 100) e definito dalla differenza della quota-parte di *promotori* rispetto a quella di *detrattori*, rappresenta la probabilità che gli acquirenti di un prodotto o servizio consiglino questi ultimi agli amici e sarebbe in grado di permettere eventuali previsioni e stime sulle vendite.

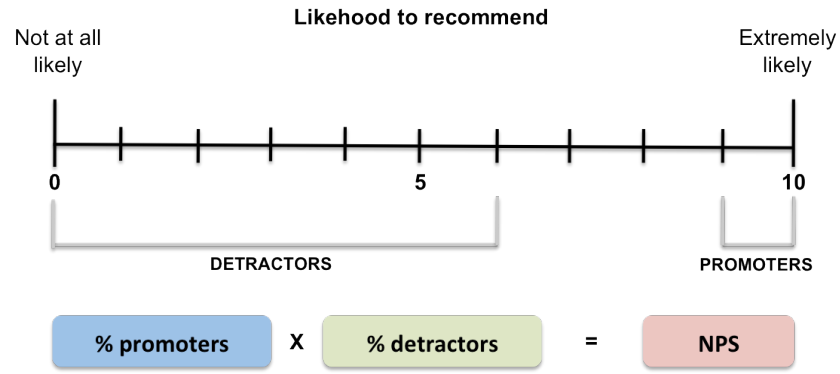


Figura 2 - Le componenti del net promoter score

McKinsey ha recentemente proposto un'altra metrica impiegabile per misurare gli effetti del passaparola, indicata con il nome di *word-of-mouth equity* (Bughin, Doogan, & Vetvik, 2010). Questo indicatore rappresenta l'impatto medio sulle vendite di un messaggio proveniente dal brand, moltiplicato dal numero di messaggi prodotti dal passaparola. Ponendo l'attenzione sull'impatto, così come sul volume, di questi messaggi, consente all'analista del marketing di valutarne in modo accurato l'effetto sulle vendite, sulla quota di mercato, sulle campagne individuali, così come sulle aziende nel loro complesso.

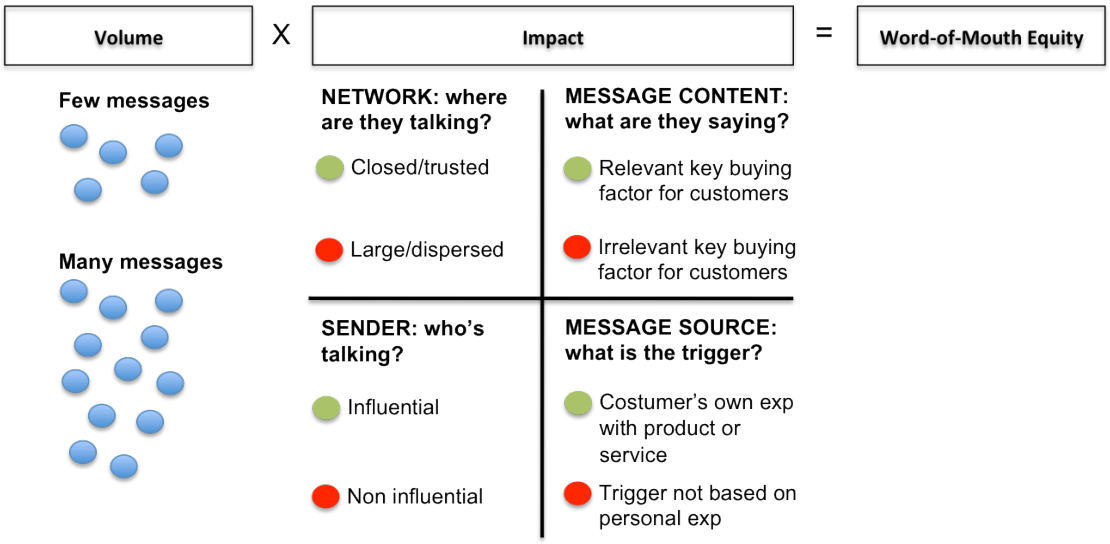


Figura 3 - Le componenti del word-of-mouth equity

Il primo elemento chiave che determina l'impatto del word-of-mouth è il contenuto effettivo del messaggio, quest'ultimo dovrebbe infatti essere indirizzato a caratteristiche primarie del prodotto/servizio perché possa avere un effetto significativo su un altro utente. La seconda componente critica è l'identità dell'individuo che crea il messaggio, in quanto chi lo riceve deve avere fiducia nel mittente e credere che costui conosca realmente il prodotto/servizio in questione. Infine è necessario esaminare anche il contesto in cui il passaparola circola, in quanto determinante per il potere di diffusione del messaggio stesso. Tipicamente se la rete è stretta e i nodi sono in un rapporto di elevata fiducia reciproca, il contenuto avrà una *reach* minore, ma un impatto superiore se confrontato alla circolazione all'interno di una rete molto dispersiva. Questo fatto è in parte spiegato da un legame molto spesso rilevante che sussiste tra gli individui che all'interno di un network teniamo maggiormente in considerazione e quelli in cui riponiamo una maggior fiducia.

2.4.3 La influence e l'influencer outreach

Il concetto di influence e le sue implicazioni sono state a lungo soggetto di studio nei campi della sociologia, della comunicazione, del marketing e dell'economia

(Rogers, 1962; Lazarsfeld & Katz, 1955). Questa infatti possiede un ruolo determinante all'interno di come agisce il business e di come funziona la società, comprovato in alcuni ambiti come la diffusione dei trend nel fashion (Gladwell, 2002) e le scelte di voto (Keller & Berry, 2003).

Lo studio delle modalità di diffusione attraverso cui la influence agisce sugli individui, può essere d'aiuto per capire al meglio perché certi trend o innovazioni sono adottati in modo più veloce rispetto ad altre e come è possibile aiutare gli uomini del marketing a creare campagne più efficaci. Tuttavia, l'individuazione di pattern significativi è risultata essere parecchio difficoltosa, questo innanzitutto perché le analisi non conducono immediatamente ad aspetti quantitativi e inoltre per il fatto che componenti essenziali come le scelte umane e i meccanismi di azione delle società non riescono ad essere efficacemente riprodotti all'interno di uno studio di ricerca. Per questo motivo la letteratura esistente è giunta a conclusioni ed affermazioni talvolta radicalmente differenti.

Le tradizionali teorie della comunicazione affermano che una minoranza di individui, noti come *influentials* (o *influencers*), è dotata di un'elevata capacità di persuasione nei confronti degli altri nelle decisioni di scelta (Rogers, 1962). Queste teorie affermano che, attraverso l'individuazione di uno specifico e ben definito gruppo di singoli identificati come influenti, è possibile raggiungere, proprio grazie al word-of-mouth, una reazione a catena su larga scala, contraddistinta lato marketing da un'elevata efficienza e da un costo molto basso (Lazarsfeld & Katz, 1955). Questa tecnica, nota anche come *influencer outreach*, non è considerabile tanto uno strumento di per sé quanto un approccio strategico al *targeting*[6], dove invece che rivolgersi direttamente ed indistintamente alla massa, il focus si sposta sugli *infuentials*, individuandoli e tentando poi di coinvolgerli con lo scopo ben preciso di trasformarli in promotori attivi.

Nell'anno 1998 Burson-Masteller e la Roper Starch Worldwide coniarono il termine *e-fluentials*, impiegato per descrivere tutti quegli individui, nel caso specifico del Web, utenti, contraddistinti da una significativa opinion leadership e che utilizzano Internet come mezzo di diffusione delle proprie opinioni ed idee.

[6] http://en.wikipedia.org/wiki/Target_market

Al contrario, alcuni studi recenti tendono a limitare pesantemente il ruolo assunto dagli opinion leader all'interno di un network, indicando altri come i reali fattori chiave:

a. le relazioni interpersonali che intercorrono tra utenti ordinari (Watts & Dodds, 2007);

b. la prontezza e la predisposizione di una società nell'adottare un'innovazione (Domingos & Richardson, 2001).

Questa visione, che come detto riduce l'importanza attribuita al potenziale di persuasione di un ridotto gruppo, dal lato dell'azione di marketing conduce nella maggioranza dei casi a strategie d'azione che prevedono quello che è meglio conosciuto come *collaborative filtering*[7].

Come dovrebbe risultare chiaro da quanto finora detto, è evidente che non esistano regole uniche, provate e condivise che spieghino come la influence impatti effettivamente la circolazione virale di un messaggio all'interno di una rete sociale e come questa sia variabile rispetto a differenti topic e al tempo, ne tantomeno le cause che determinano la maggiore o minore viralità di uno specifico contenuto piuttosto che di un altro.

2.4.4 Il microblogging come strumento per il Web marketing

Il microblogging è diffusamente ritenuto come un nuovo e possibile mezzo impiegabile per i programmi Web marketing (Jansen, Zhang, & Sobel, 2009).

Uno dei paradigmi applicabili per lo studio in area commerciale dei moderni servizi di social networking, contraddistinti da una connettività pressoché costante, è noto come *attention economy* (Beck & Davenport, 2001), i.e. economia dell'attenzione, teoria secondo cui i brand sono coinvolti in una competizione continua per guadagnare l'attenzione dei potenziali consumatori.

Il microblogging è considerabile come una nuova forma di comunicazione nella quale gli utenti possono, attraverso l'utilizzo di post molto brevi, parlare e

[7] http://en.wikipedia.org/wiki/Collaborative_filtering

descrivere contenuti relativi ai loro interessi ed esprimere opinioni e raccomandazioni. Uno degli elementi che fa sì che questo strumento abbia un potenziale effetto così diretto e rilevante nel passaparola online è senza dubbio la facilità di aggiornamento e condivisione effettuabile ovunque e praticamente in ogni momento (e.g. mentre si è al computer, mentre si guida o si prende un caffè), diretto a praticamente chiunque sia connesso (i.e. via Web, mobile) e su una scala di utenza raggiungibile mai vista in passato.

La caratteristica principale del microblogging è la brevità intrinseca del servizio che obbliga l'utente a rinunciare ad espressioni di pensiero troppo lunghe e articolate ed è esattamente questo "micro" aspetto che rende Twitter e servizi simili totalmente differenti da altri mezzi adatti all'*eWOM*, includendo blog, altri social network e portali prettamente dedicati allo scambio e alla consultazione di recensioni. La lunghezza standard di un messaggio, i.e. un massimo consentito di 140 caratteri, è approssimativamente quella tipica di titoli e sottotitoli dei giornali, la quale rende la piattaforma estremamente agevole sia per la produzione che per il consumo dei contenuti. Il messaggio è inoltre asincrono[8] e non invasivo, poiché ciascun utente può scegliere se ricevere o meno aggiornamenti da un altro user. I post sono archiviabili, nel senso che sono permanenti e ricercabili attraverso dei *search* sia via Web, che via mobile.

Un altro aspetto rilevante e di notevole interesse è quello legato al microblogging come fonte da cui estrarre dati utili per effettuare analisi di sentiment, vista l'evidente attrattività del servizio come risorsa di opinioni create da consumatori capaci di fornire una possibilità di comprensione per le reazioni effettive nei confronti di un prodotto, un brand, un evento o un esperienza.

Di seguito nella Figura 4 è rappresentato il modello generale di *branding* (Esch, Langner, & Schmitt, 2006), rivisto e riletto alla luce dei ragionevoli effetti di viralità causati dal microblogging.

[8] In realtà è sia asincrono che sincrono. In questo caso nel testo si vuole evidenziare il fatto che non ha necessità di avere una risposta immediata, come accade nell'IM, ma c'è anche la possibilità di rispondere come accadrebbe per una e-mail.

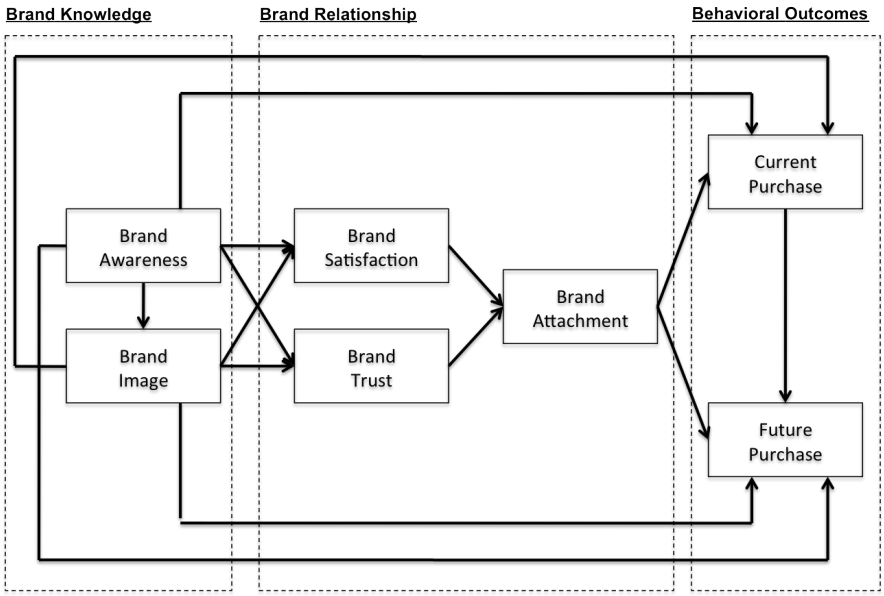


Figura 4 - Componenti generali del modello di branding

Le componenti di *brand image* e *brand awareness*, le quali formano la *brand knowledge*, sono le aree primarie su cui l'*electronic word-of-mouth*, attivato grazie ai post su Twitter, potrebbe avere un effetto diretto e rimarcabile. Il monitoraggio e la gestione del proprio account e quindi dei contenuti e delle risposte avrebbe un impatto sulla *brand relationship,* fase in cui i consumatori possono essere coinvolti in un'interazione con i brand, allo stesso modo in cui si svolgono le relazioni personali tra individui. Infine, data la diretta associazione tra le comunicazioni in passaparola e la *brand satisfaction*, fase in cui avviene anche la decisione di acquisto, si sarebbe indotti a ritenere come necessario per i brand manager prendere quanto meno in considerazione la possibilità di operare attivamente su un network come Twitter.

2.5 Gli influencer e le rete sociali

Una dimensione accessoria al WOM è data, come già introdotto nella Sezione 2.4.3, dall'analisi dei potenziali influencer e del processo di opinion leadership. Secondo alcuni studi, viste le caratteristiche del Web, che non solo facilita la condivisione di un messaggio, ma d'altra parte ne semplifica decisamente il

consumo stesso, sarebbe quindi necessario all'interno della valutazione del passaparola tenere in considerazione anche il processo effettivo di ricerca, i.e. *opinion seeking* (Flynn, Goldsmith, & Eastman, 1996).

L'insieme dei contenuti digitali presenti su un sito di social networking, o più in generale 2.0, il quale costituisce l'elemento trainante la vitalità e l'attrattività di una specifica piattaforma, è prodotto in maniera pressoché totale dagli utenti. All'interno di una rete sociale gli utenti non sono tuttavia uguali tra loro e non sono tutti responsabili allo stesso modo del complesso di informazione prodotta, ma differiscono in termini di frequenza, volume, tipologia e qualità del contenuto creato e consumato.

Da un punto di vista di gestione del network diventa quindi fondamentale capire e individuare quel ristretto gruppo di individui che influenzano anche l'attività degli altri, così come definire la struttura complessiva della rete, dato che il processo di ricerca e più in generale i pattern e le modalità di trasmissione del messaggio incidono in modo diretto facilitando la diffusione dell'informazione in un processo di comunicazione interpersonale (Reynolds & Darden, 1971).

2.5.1 Caratteristiche del network

Per definire e valutare al meglio la influence è fondamentale considerare due indicatori riguardanti il network che si sta analizzando:

a. *dimensione del network* – definibile come il numero di utenti un individuo può raggiungere attraverso connessioni primarie o estese (metrica nella Sezione 2.2.2);

b. *qualità del network* – indica il livello di familiarità e di fiducia tra l'individuo e le sue connessioni.

Vediamoli ora in dettaglio.

Dimensione del network

Gli elementi che è necessario comprendere e misurare riguardano sostanzialmente la portata e il volume dell'attività: quanto ben connesso è un individuo agli altri

appartenenti al network e quanto le opinioni espresse da questo possono propagarsi attraverso network estesi. Ci sono due fattori principali che contribuiscono alla dimensione globale della rete:

- *numero di connessioni primarie* – le connessioni dirette di un utente all'interno di un network rappresentano l'audience su cui può esercitare la sua influenza;

- *numero di connessioni estese* – indica la quantità di contatti non connessi direttamente, ma che l'utente può raggiungere traverso le sue connessioni primarie.

Qualità del network

Il valore intrinseco di una rete è strettamente correlabile alla qualità delle connessioni presenti. Quest'ultima varia in base a:

- *autorità percepita dell'influencer* – l'autorità di un influencer deriva da un passato di informazioni accreditate ed accurate;

- *robustezza della relazione tra l'influencer e il suo network* – forti legami tra l'individuo e le sue connessioni danno come risultato dei legami di alta qualità. Generalmente le relazioni più robuste sono quelle sviluppate offline, come rapporti familiari o di amicizia, ma le connessioni online possono essere altrettanto personali e fidate. La influence ha la possibilità di agire in entrambe le direzioni: molti legami sui social network riescono a continuare offline attraverso i *meetup* e quelli nati offline raggiungono ulteriori dimensioni attraverso i social media;

- *rilevanza dell'informazione che l'influencer condivide* – la qualità della influence è dipende fortemente dal contesto dell'informazione. Il messaggio dell'influencer deve essere legato agli interessi dei suoi contatti per essere percepito come rilevante.

Di seguito la Tabella II riassume quanto appena descritto fornendo degli esempi:

Fattori influence	Compomenti influence	Definzioni	Esempi
Dimensione del network	*Numero di connessioni primarie*	Quantità di contatti diretti un'influencer ha all'interno del network.	Twitter *follower*, Facebook *friends.*
	Numero di connessioni estese	Quantità di contatti le connessioni di un influencer hanno all'interno del network.	Facebook *friends of friends*, Twitter *retweet* su altri network.
Qualità del network	*Autorità percepita di un influencer*	Livello di credibilità riconosciuta di un influencer.	Esperti in un determinato settore (e.g. blogger di una industry).
	Robustezza delle relazioni	La probabilità che le connessioni seguano un consiglio/opinione dell'influencer.	Amici stretti, famiglia e colleghi sono considerati legami forti.
	Rilevanza dell'informazione	Quanto l'informazione è legata all'interesse dell'audience dell'influencer.	L'informazione rientra tra le aree di *expertise* dell'influencer.

Tabella II - Le caratteristiche del network

2.5.2 Metriche di posizionamento dell'utente

L'analisi delle reti offre differenti metriche utilizzabili per misurare l'effettiva collocazione di utente all'interno di un network di appartenenza (Musial, Kazienko, & Brodka, 2009). Il posizionamento è una delle caratteristiche chiave che permette di definire l'importanza, quindi in un certo qual modo l'influenza, di un nodo in una rete. Verranno di seguito prese in considerazione 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:

- *posizione di prestigio*;
- *posizione di centralità*.

Prestigio

Un membro appartenente ad un network può essere considerato di prestigio nel caso in cui è presente un elevato numero di legami uscenti dagli altri utenti e diretti al primo. Tra le varie metriche che permettono una misurazione del prestigio riportiamo:

a. *indegree centrality;*
b. *proximity prestige.*

a. Indegree Centrality

L'*indegree centrality* si basa sul numero di connessioni in ingresso ad un nodo, prendendo quindi in considerazione il numero di nodi che sono adiacenti ad un particolare utente della community. In altre parole quest'ultimo sarà considerabile tanto più interessante quanto più verrà nominato dagli altri membri del network.
Questa la formula:

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

dove:

i(x) – indica il numero di membri della community adiacenti all'utente *x*, considerando solo il primo livello di prossimità;
m – rappresenta il numero totale di membri all'interno del network.

Come è immediatamente osservabile, questa metrica è di tipo locale, limitandosi a considerare esclusivamente il primo livello di vicinanza.

b. Proximity Prestige

Questa metrica riflette la vicinanza di tutti i membri della community al nodo *x.* La misura si basa sulla distanza geodesica, indicata con $d(x,y_i)$, la quale indica la distanza di tutti gli utenti y_i dal soggetto *x.* La formula della *proximity prestige* 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 di tutti i membri y_i appartenenti al network ed in grado di raggiungere l'utente *x*, in quanto esiste un *path* di collegamento;
m – rappresenta il numero totale di membri all'interno del network.

Centralità

Le misura relative alla centralità permettono di individuare gli utenti che sono estensivamente coinvolti in relazioni con gli altri utenti della community. Tendenzialmente queste metriche sono applicate in grafici non diretti, cioè in cui non è rilevante il fatto che l'utente sia destinatario o fonte dell'informazione.

Le metriche prese in considerazione per la centralità sono le seguenti:

a. *outdegree centrality;*
b. *eccentricity centrality;*
c. *closeness centrality;*
d. *betweenness centrality*.

Andiamo ad analizzarle nel dettaglio come fatto per le precedenti.

a. Outdegree centrality

Misura il numero di archi che vanno dal nodo *x* verso altri nodi. La formula per il calcolo è la seguente:

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

dove:

o(x) – indica il numero di utenti adiacenti ad *x*, considerando solo la vicinanza di primo ordine;

m – rappresenta il numero totale dei membri all'interno del network.

Sostanzialmente più gli utenti comunicano con un elevato numero di individui, più otterranno un valore di ODC elevato.

b. Eccentricity centrality

L'*eccentricity* indica il nodo più centrale della rete come quello che minimizza la distanza da tutti gli altri nodi del network. Di seguito la formula:

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

dove:

d(x,y) – indica la lunghezza del *path* più corto che collega *x* e *y*;
M – rappresenta il set del totale dei membri all'interno del network.

c. Closeness Centrality

La *closeness centrality*, in contrasto con la *proximity prestige*, esprime la vicinanza di un utente rispetto a tutti gli altri facenti parte della rete. L'idea di base è che un nodo occupa una posizione centrale nel momento in cui può raggiungere in modo veloce gli altri utenti della rete. Questa metrica misura quindi la qualità della posizione all'interno della community. Un utente con alti valori di CC sarà quindi considerabile un buon propagatore di informazioni ed opinioni. Viene calcolata nel modo seguente:

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

dove:

c(x,y) – è una funzione che descrive la distanza tra i nodi *x* e *y* (e.g. max, min, mean);
M – rappresenta il set del totale dei membri all'interno del network.

d. Betweenness Centrality

Questa metrica misura la centralità di un utente sulla base di una particolare strutturazione della rete. Gli utenti con un elevato valore di *betweenness centrality* sono fondamentali per la diffusione dell'informazione all'interno del network. La metrica è calcolata dividendo il numero di percorsi più brevi che vanno da *y* a *z* rispetto al numero di quello che passano attraverso *x*:

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

dove:

$b_{ij}(x)$ – indica il numero di percorsi più corti che vanno da *i* a *j* e passanti per *x*;
b_{ij} – numero di percorsi più brevi che vanno da *i* a *j*;
M – rappresenta il set del totale dei membri all'interno del network.

Vantaggi e svantaggi

Nella Tabella III vengono evidenziati in modo chiaro e sintetico i vantaggi e gli svantaggi di ciascuna delle metriche analizzate:

Nome	Vantaggi	Svantaggi
IDC	• Semplice da calcolare; • Sufficientemente informativo per molte applicazioni.	• Elevato numero di duplicati; • Misura locale – prende in considerazione solo le connessioni di primo ordine.
PP	• Misura globale; • Considera la topologia totale della rete.	• I grafi disconnessi assegnano valore 0 ad ogni nodo; • Molto complessa e inefficiente per le grandi reti.
ODC	• Semplice da calcolare; • Sufficientemente informativo per molte applicazioni.	• Alto numero di duplicati; • Misura locale – prende in considerazione solo le connessione di primo livello.
EC	• Misura globale; • Considera la topologia totale della rete.	• I grafi disconnessi assegnano valore 0 ad ogni nodo; • Molto complessa e inefficiente per le grandi reti.
CC	• Misura globale; • Considera la topologia totale della rete.	• I grafi disconnessi assegnano valore 0 ad ogni nodo; • Molto complessa e inefficiente per le grandi reti.
BC	• Misura globale; • Considera la topologia totale della rete.	• I grafi disconnessi assegnano valore 0 ad ogni nodo; • Molto complessa e inefficiente per le grandi reti.

Tabella III - Riassunto delle metriche di posizionamento dell'utente

2.5.3 Tassonomia degli influencer

E' possibile definire degli archetipi di utente che differiscono in base ai fattori di dimensione e qualità della rete. Ciascuno di questi è in grado di diffondere diversi tipi di messaggio a vari gruppi di individui.

La categorizzazione utilizzata da Forrester (Katz J. M., 2009) definisce le seguenti tipologie di utenti:

La fonte (*the source*)

Spesso rappresenta il nodo da cui parte l'informazione, questa tipologia di individuo può normalmente avere un numero di contatti inferiore ad altri influencer, ma possiede un elevato livello di autorità su una determinata area tematica. Gli altri utenti del network spesso scoprono informazione dal nodo fonte prima di diffonderlo all'interno delle loro rispettive reti.

Il ragno (*the spider*)

L'utente identificato come *spider* è in grado di raggiungere una larga scala di individui grazie ad un elevato numero di connessioni, alcune delle quali sono a loro volta normalmente 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.

Il sole (*the sun*)

Questo archetipo di utente ha, tra le tipologie viste, 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.

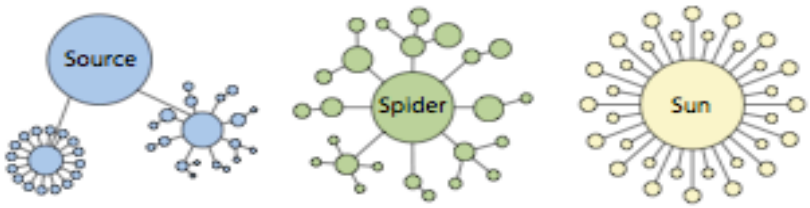


Figura 5 - Rappresentazione degli archetipi sun, spider e surce (Forrester)

Nella Tabella IV sono riassunte le caratteristiche principali dei tre tipi di *influencer* individuati da Forrester:

Influencer	***The source***	***The spider***	***The sun***
Definizione	Spesso rappresenta l'origine dell'informazione, è connesso agli altri influencers.	Socialmente connesso attraverso una rete estesa più ampia.	Connesso ad un network molto ampio grazie a connessioni di primo livello.
Dimensione del network	Da piccolo a medio.	Da medio a largo.	Molto largo.
Qualità del network	Molto forte.	Da medio a forte.	Da debole a medio.

Tabella IV - Caratteristiche degli archetipi source, spider e sun (Forrester)

Gartner propone invece una differente suddivisione (Ingelbrecht, Patrick, & Foong, June 2010), mutuata dai testi di Gladwell (Gladwell, 2002) e Clippinger (Clippinger, 2007), che descrive i seguenti ruoli potenzialmente presenti all'interno di un social network:

Tipologia	Descrizione
Connector	La funziona primaria di un *connector* è quella di collegare differenti gruppi di persone. Possono essere a loro volti suddivisi in due categorie: (1) *heavy connector*, i quali hanno un circolo di legami molto forti con famiglia e amici; (2) *light connector*, presenti in un range più ampio per diversità di gruppi, ma inevitabilmente con legami più deboli e meno frequenti.
Salesman	I *salesman* (i.e. venditore) hanno ampie connessioni sociali, ma la loro caratteristica distintiva è la capacità di persuadere e influenzare altre persone (e.g. nell'acquisto di certi prodotti o nel comportarsi in un certo modo). Questo ruolo non è da intendersi come collegato ad un'attività commerciale, ma dipende da un'abilità personale.
Seeker	I *seeker* si connettono con altri utenti al fine di trovare informazioni e *skills* di cui hanno bisogno per condurre la loro vita quotidiana.
Maven	La caratteristica principale dei *maven* è lo scambio di informazione e conoscenza. Sono utenti esperti in particolare aree e sono ricercati dagli altri per ricevere consigli. A differenza dei *salesman*, non hanno l'obiettivo primario di persuadere, ma utilizzano e acquisiscono informazione per i proprio interessi.
Self-sufficient	Questa tipologia di utente preferisce trovare in modo autonomo l'informazione di cui necessita per soddisfare i proprio bisogni. Tendenzialmente non pongono molta attenzione alle raccomandazioni altrui e sono considerati un target market difficile da raggiungere perché molto poco sensibile ai messaggi virali e agli effetti su banda larga.
Unclassified	La categoria rappresenta all'incirca i due-terzi dell'intera popolazione e non rientra in alcuno dei cinque casi descritti. Spesso però, utenti che non esibiscono chiaramente caratteristiche specifiche di una categoria, possono essere in grado di assumere differenti ruoli a seconda della rete sociale contestuale.

Tabella V - Le tipologie di influencer secondo Gartner

2.6 Twitter e la influence: la letteratura

Nonostante la breve età del servizio di microblogging, ma vista la sua forte attrattività sotto vari punti di vista, diversi gruppi di ricerca si sono impegnati a valutare quali siano i parametri rilevanti da misurare e tenere in considerazione

all'interno di Twitter per valutare l'autorità degli utenti e ad analizzare la diffusione dell'informazione e dei contenuti all'interno del suo network.

Uno degli studi più interessanti e discussi, anche grazie all'enorme risonanza che ha avuto nel Web, è quello di Cha, Haddadi, Benvenuto, Gummadi, dal provocatorio titolo *The Million Follower Fallacy* (Cha, Haddadi, Benvenuto, & Gummadi, 2010). Partendo da un data set molto ampio composto da circa 6 milioni di utenti e prendendo come elementi determinanti l'autorità di ciascuno gli indicatori di *indegree* (i.e. numero di *follower*), *retweet* e *mention*[9], ne è stata analizzata la correlazione attraverso l'indice di Spearman[10]. Secondo le analisi, considerando specificatamente gli utenti appartenenti al primo e al decimo percentile dell'intero set, si rileva un considerevole valore di correlazione tra i *retweet* e le *mention*, legame statistico che è invece non significativo effettuando la misurazione per ciascuno di quest'ultimi con l'*indegree*, giungendo quindi alla conclusione che la popolarità di un utente ha una scarsa incidenza sull'attenzione e sulle reazioni che è in grado di generare negli altri individui, i.e. la influence potenziale che potenzialmente esercita.

Sempre tenendo in considerazione *retweet* e *mention* lo studio ha esplorato le dinamiche della opinion leadership al variare dei topic e del tempo. Secondo la ricerca, un ristretto gruppo definito di *top influentials* sarebbe in grado di mantenere una significativa autorità su una varietà di argomenti, asserendo infine che la influence tra gli "utilizzatori ordinari" non viene guadagnata in modo spontaneo o accidentale, ma attraverso uno sforzo focalizzato che comporta un coinvolgimento personale.

Sugli stessi elementi si è concentrato anche uno studio svolto dal Web Ecology Project (Leavitt, 2009), gruppo di ricerca di Boston, Massachusetts, il quale basandosi sui contenuti e le risposte generate da un set di 12 utenti molto popolari, appartenenti a tre cluster definiti a priori – *celebrity, news outlet, social media analyst* – categorizza le azioni in contenuto e conversazione per comprendere come differenti tipologie di utenti e i relativi *follower* interagiscano in modo

[9] Spiegati in dettaglio e con esempi nella Sezione 3.2
[10] http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

diverso. Distinguendo risposte *conversation-related*, date dalla somma di *reply* e *mention,* e *content-related*, date dall'utilizzo dei *retweet,* e pesando opportunamente queste sia con il numero di *follower* che con l'attività registrata durante il periodo di analisi, vengono messe in luce forti discrepanze tra i ranking con i valori assoluti da quelli pesati, con delle variazioni anche per quanto riguarda la differenziazione dei messaggi in *conversazione* e *contenuto*. Le *celebrity*, dotate del seguito più ampio, sono in grado di produrre significativi volumi di risposte con uno sforzo (i.e. attività) molto basso; i *social media analyst,* in cima ai ranking se si valutano contenuti diffusi e conversazioni generate pesando gli indicatori per il numero di *follower*, ma a seguito di uno sforzo molto elevato; i *news outlet*, con la migliore capacità di avere i proprio contenuti "spinti" dagli altri utenti.

Altre ricerche si sono concentrate maggiormente su un ambito ben specifico, la dinamica di diffusione dell'informazione e dei messaggi all'interno del network Twitter, considerando la propagazione attuale del messaggio e la maggiore o minore passività dei membri della rete sociale come elementi determinanti. L'assunto su cui si basano questi studi è il fatto che l'opinion leadership di un *twitterer* può essere confrontata con quella di una pagina Web: l'autorità di un nodo è tanto elevata, tanto più lo è la somma di quella dei suoi *follower*. Questa similarità motiva l'uso del PageRank (Brin & Page, 1998), lo stesso utilizzato da Google per indicizzare le pagine Web, o di algoritmi con delle varianti, come strumento per effettuare le misurazioni.

Una delle metodologie proposte implica l'individuazione degli influencer attraverso una misura chiamata *effective readers*, basando l'analisi primariamente sulla struttura delle connessioni del network e sull'ordine dell'adozione dell'informazione (Lee, Kwak, Park, & Moon, 2010). Crawlando all'incirca 41 milioni di utenti, per un totale di 223 milioni e 4.262 *trending topics* ed esplorando i pattern di diffusione dei contenuti e prendendo in considerazione il rank di adozione, è emerso come gli utenti con molti *follower* non rappresentino sempre i migliori diffusori. La quantità cumulata di *potential readers*, i.e. lettori potenziali, aumenta rapidamente nei primi stadi e la crescita rallenta mano a mano

al passare del tempo. Questo comportamento dimostra come l'informazione si diffonda per lo più immediatamente dopo la sua introduzione. Gli influencer sono stati calcolati in base al numero di *effective readers*, i.e. lettori effettivi, che, a differenza del *potential*, è definito come colui che è stato esposto per la prima volta al topic dall'utente che ha postato il messaggio. Gli stessi autori in un altro studio, hanno valutato la correlazione tra i ranking formati basandosi su parametri diversi (Kwak, Lee, Park, & Moon, 2010), tra cui lo stesso PageRank.

Hp Labs infine (Romero, Asur, Galuba, & Huberman, 2010), attraverso l'analisi di un ampio set formato da 22 milioni di *tweet* contenenti la stringa *http* (i.e. dei Web link) e la creazione di un algoritmo rinominato *IP*, il quale assegna a ciascun utente un *influence score* e un *passivity score,* dove quest'ultimo può essere definito come la tendenza a visionare i *tweet* altrui senza però condividerli con il resto del network, elemento che rappresenta in un certo qual modo una barriera ad essere influenzati, ha valutato la propagazione dell'informazione nella rete in termini di riproposizione da parte degli utenti e consumo. Le conclusioni raggiunte sostengono che il legame tra popolarità e influence è più debole di quanto ci si possa aspettare e su quest'ultima incidono in modo determinante sia la quantità ma soprattutto la qualità dell'audience. Al singolo utente, il cui contenuto vedrà ovviamente una maggiore *reach* se gli altri individui non ne effettuano esclusivamente un consumo passivo ma lo ritrasmettano attivamente, non è sufficiente attirare l'attenzione altrui, i.e. essere popolare, ma è necessario sia in grado di superare la predisposizione passiva di base delle sue connessioni primarie.

Capitolo 3

Twitter: Il profilo sociale

3.1 Introduzione

All'interno della definizione di microblogging è possibile individuare una serie di diverse piattaforme 2.0, tra cui Ping.fm[11], Jaiku[12], Tumblr[13] e Twitter. Nessuna come quest'ultima è stata però in grado di raggiungere un successo e una diffusione così rilevante, tale da incentivare brand, varie *celebrity*, testate giornalistiche e magazine ad entrare far parte della community. Soprattutto grazie al suo potenziale di diffusione di contenuti in real-time, Twitter è diventato una sorta di sismografo umano, in grado di misurare e trasmettere il polso non solo del Web, ma anche di eventi locali e mondiali.

[11] http://ping.fm/
[12] http://www.jaiku.com/
[13] http://www.tumblr.com/

Lo scopo del capitolo è quello di delineare al meglio le caratteristiche principali di Twitter e renderne l'utilizzo il più chiaro possibile. La Sezione 3.2 spiega brevemente le informazioni accessibili dal profilo di ciascun utente e i metodi di *input/output* per la creazione e la condivisione di contenuti; la Sezione 3.3 tratta le funzionalità principali, le modalità di comunicazione tra utenti e i contenuti presenti; nella Sezione 3.4 vengono descritte le caratteristiche strutturali del network; la Sezione 3.5 presenta alcune informazioni demografiche e di distribuzione geografica relative all'utilizzo del servizio; infine la Sezione 3.6 illustra alcune delle possibili tassonomie applicabili agli utenti.

3.2 Informazioni dei profili e funzionalità

Il profilo di ciascun utente Twitter è raggiungibile via Web accedendo alla indirizzo *www.twitter.com/username*, dove *username* è associato in modo univoco ad un determinato user. Il layout è composto da un form principale nel quale viene inserito l'aggiornamento del proprio stato, rispondendo alla semplice domanda "*what's happening?"*, e dalle seguenti informazioni:

- *descrizione del profilo* – tra le informazioni dell'utente compaiono il campo *name*, il quale non corrisponde necessariamente allo *username* (e.g. *name "*Pete Cashmore", *username* "Mashable"); la *location*, che indica il luogo di provenienza; il campo *Web*, utilizzato per inserire eventuali siti Web dell'utente (e.g. blog, pagina di un altro social network); e la *bio*, spazio in cui è possibile inserire una breve descrizione;

- *follower (indegree)* – indica il numero di utenti iscritti al network che hanno deciso di sottoscrivere i contenuti di un determinato profilo (e.g. Figura 6, un totale di 2.001.523 utenti visualizzano all'interno della propria timeline i contenuti condivisi da *@mashable*). Rappresenta il numero di connessioni in ingresso al nodo;

- *following (outdegree)* – indica il numero di profili a cui un utente ha deciso di sottoscriversi, visualizzandone i contenuti sulla timeline. Rappresenta il numero di connessioni in uscita dal nodo;

- *listed* – è possibile organizzare i contatti in liste personalizzabili consentendo un filtraggio del flusso di aggiornamenti (e.g. per area tematica, per area geografica etc.), il valore indica la quantità di liste in cui un utente è stato incluso;

- *tweet* – rappresenta il numero complessivo di messaggi generati dal profilo dal momento della sua iscrizione al servizio.

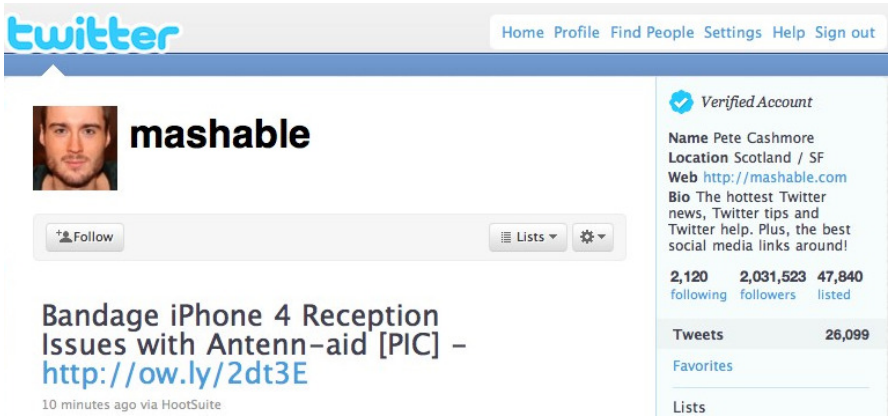


Figura 6 - Esempio di profilo utente in Twitter

Utilizzando il servizio, ciascun utente registrato ha a disposizione, oltre alla pagina contenente il proprio profilo, una timeline all'interno della quale scorrono in real-time gli aggiornamenti postati dagli account a cui questo ha deciso di sottoscriversi, i.e. *following*. Allo stesso modo i contenuti da lui generati saranno accessibili da tutti coloro che avranno scelto di effettuarne la sottoscrizione, i.e. *follower*.

3.2.1 Mobilità e applicazioni di terze parti

Uno dei fattori distintivi del servizio è la sua abilità di trasmettere dati agli utenti interessati attraverso molteplici canali di comunicazione (Krishnamurthy, Gill, & Arlitt, 2008). I messaggi di Twitter possono essere ricevuti come messaggi di testo (i.e. sms) su dispositivi cellulari, piuttosto che attraverso l'applicazione di Facebook o di altri piattaforme attraverso la connessione degli account, o ancora via e-mail, tramite un feed RSS o un instant messenger (e.g. Jabber, Google Talk etc.).

La Figura 7 di seguito mostra i potenziali sistemi di *input/output* attraverso i quali mandare e ricevere i messaggi. Gli utenti possono decidere se mantenere i contenuti che hanno generato come pubblici, caso in cui i messaggi appariranno in ordine cronologico inverso sulla timeline pubblica della pagina di Twitter e sulla pagina personale dell'utente, oppure come privati, dove solo gli utenti che hanno deciso di sottoscriversi ai contenuti di quell'utente avranno la possibilità di visualizzarli.

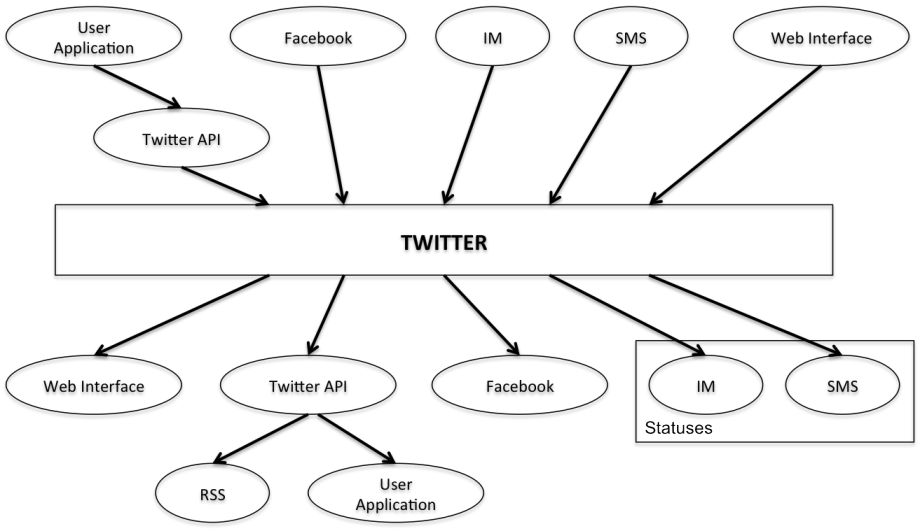


Figura 7 – Metodi di input e output in Twitter

Le ultime statistiche ufficiali (Techcrunch, 2010) evidenziano la forte caratterizzazione mobile del servizio, secondo cui il 46%, costantemente in crescita, dell'utenza avrebbe compiuto almeno un accesso al sito non via

www.twitter.com, ma tramite sito mobile, che rappresenta la seconda interfaccia più impiegata con il 14%, o un'applicazione mobile di terze parti, le quali occupano i primi posti per utilizzo del totale composto da più di 300mila *apps* al momento disponibili, di tipo sia mobile che desktop, le quali sottopongono le API[14] di Twitter a circa 3 miliardi di chiamate al giorno.

3.3 Le conversazioni in Twitter

Vediamo ora quali sono le modalità di comunicazione che permettono lo scambio di messaggi e l'interazione tra utenti all'interno del network:

a. *retweet* – consiste nel riproporre un messaggio generato da un altro utente. E' tipicamente identificato dalla sintassi "*RT @username*", ma può comparire anche come "*via @username*" oppure "*from @username*". La differenza dipende sostanzialmente dall'applicazione utilizzata al momento dell'azione, alcune di terze parti hanno come sintassi di default il *via* o il *from.* Esempi:

   ***RT @mashable** bandage iPhone 4 reception issues with antenn-a http://bit.ly/cd9iNm #accessories #antenna-aid #antennagate*

   *Foursquare in Talks with Major Search Engines [REPORT] (**via @mashable**) http://bit.ly/cFkEMO*

b. *mention* (o *reference*): è possibile rispondere ai post altrui oppure citare un altro utente in un messaggio utilizzando la sintassi *"@username*", la quale può essere posta all'inizio del messaggio, così come in un suo punto casuale. Vediamo la distinzione:

   i. *reply*: quando *"@username*" viene inserito nella parte iniziale del messaggio, quest'ultimo è rivolto direttamente all'utente

[14] http://en.wikipedia.org/wiki/Application_programming_interface descritte nel dettaglio nella Sezione 4.5.2

"*username"* ed è visibile esclusivamente dal mittente, dal ricevitore e da tutti gli individui che questi hanno comune come *following*. Esempio:

***@mashable** your site is not loading*

ii. *mention*: la sintassi *@username* può essere in qualsiasi posizione del messaggio. In questo caso, il *tweet* sarà sia inviato direttamente a *"@username"*, ma sarà contemporaneamente un contenuto *one-to-many*, cioè visibile da tutti gli utenti. Esempio:

*seems like **@mashable**'s site is down*

3.3.1 Contenuto vs conversazione

L'utilizzo diffuso della sintassi "*@username*" per indirizzare il messaggio, tipica *dell'Internet Relay Chat* (IRC), è considerata come una forma di *addressitivity,* la quale può essere impiegata, come detto, per inviare messaggi diretti o per riferirsi a un altro utente attirandone l'attenzione (i.e. *reply* vs *mention*). Data la struttura di Twitter, che verrà analizzata meglio nella Sezione 3.4, le conversazioni sono disperse attraverso un network di utenti interconnessi, più che essere limitate all'interno di spazi o gruppi ristretti e lo *stream* fornito dalla piattaforma permette agli individui di essere marginalmente informati senza partecipare attivamente.

Honeycutt e Herring hanno analizzato le funzioni di utilizzo del segno *"@"* e le caratteristiche delle conversazioni (Honeycutt & Herring, 2009), ottenendo come risultato che circa il 91% ha lo scopo di indirizzare il messaggio direttamente verso un altro utente, supportando la tesi che la sintassi *@username* abbia un impiego soprattutto conversazionale.

Intuitivamente si potrebbe pensare che intercorra una differenza di utilizzo tra le comunicazioni che adottano le due funzionalità, *retweet* e *mention*, considerando quest'ultimo, come esclusivamente dedicato all'interazione uno a uno, mentre il primo, con la principale caratteristica di poter aumentare la *reach* di

un messaggio, con un impiego maggiormente focalizzato alla diffusione di contenuto.

Secondo uno studio di Boyd, Golder e Lotan invece, quando la comunicazione è distribuita su un network non coeso come Twitter, nel quale i ricevitori di ciascun messaggio cambiano a seconda di chi lo condivide, vengono a mancare delle strutturale conversazionali prestabilite, facendo sì che la partecipazione non si articoli in uno scambio ordinato di comunicazioni, ma ci sia un libero adattamento ad una molteplicità di contesti conversazionali. Per questo motivo anche il *retweet* favorisce la conversazione e la sua funzione non è limitata alla sola diffusione di contenuto (Boyd, Golder, & Lotan, 2010), invitando altri utenti a partecipare anche senza indirizzare loro un messaggio. Nonostante le analisi rivelino che più di metà dei tweet contenenti la sintassi "*RT*" includano un link ad un altro sito Web, elemento che rafforza l'idea di propagazione di un messaggio, agevolandone l'esposizione ad un audience di dimensione maggiore, la riproposizione è comunque considerabile come una modalità di validazione del contenuto e di relazione con gli altri.

Queste considerazioni sono interessanti se relazionate ad alcune metriche presentate nella Sezione 4.4.2 e all'analisi svolte nel Capitolo 5.

3.3.2 Il tagging dei contenuti

I topic delle conversazioni sono indicati attraverso una sintassi che prevede la combinazione di un *hashtag* (#) e una keyword (e.g. *#android, #plastikman, #milan*), come evidenziato nell'esempio seguente:

Got Android 2.2 update for HTC Desire this morning via Meteor. ***#android***

I messaggi possono contenere al loro interno anche più di un *hashtag* e l'utilizzo ne favorisce la tracciabilità e la gestione dei contenuti medesimi e delle conversazioni, in quanto effettuando una query, ad esempio nel *search* di Twitter[15], contenente come keyword *"#topic*", è possibile visualizzare tutti i

[15] http://search.twitter.com/

messaggi che contengono quel particolare tag. Questa metodologia di etichettatura dei contenuti è del tutto similare al *tagging*[16], largamente impiegato per categorizzare i contenuti Web.

La Tabella VI mostra un esempio dell'utilizzo e dell'effettiva incidenza dei contenuti etichettati rispetto al totale all'interno di un argomento specifico. Quest'ultimo, preso come riferimento per il sample, è la London Book Fair svoltasi a Londra il 21-23 Aprile 2010 ed i post estratti appartengono al periodo 12-26 Aprile.

totale contenuti microblogging	4483	100%
contenenti keyword "*London book fair*" e "*London*" + "*book fair*"	1557	34,7%
contenenti hashtag *"#LBF", "#LBF10"* e *"#LBFDC"*	2926	65,3%

Tabella VI - Utilizzo dell'hashtag in un set di messaggi

Come dimostra il data set utilizzato, una percentuale abbondantemente superiore al 50% include all'interno del messaggio un *hashtag*, necessario a etichettare i contenuti e ad inserirli in uno flusso conversazionale ben preciso. Questa caratteristica è importante anche in relazione alle modalità con cui vengono estratte i dati dal servizio di microblogging, aspetto trattato nel Capitolo 4.

3.3.3 Le tipologie di contenuti presenti nel microblogging

Esistono siti focalizzati su determinate aree tematiche, quali possono essere ad esempio il turismo o il fashion, ed altri più generalisti, nei quali vengono presi in considerazione casuali e differenti tipologie di argomenti, lasciando all'utente la più totale libertà di scelta. Quest'ultimo è esattamente il caso del microblogging e di Twitter, dove da un *tweet* all'altro il topic del discorso può essere totalmente diverso.

Come detto, un fattore importante in un qualsiasi network riguarda le possibili relazioni e modalità in cui avvengono le comunicazioni tra gli utenti. Twitter si fonda su un concetto di comunicazione *one-to-one*, una persona pone una

[16] http://en.wikipedia.org/wiki/Tag

domanda, un'altra risponde, e così via, formando una catena di messaggi brevi. Se si vuole fare un'altra domanda sullo stesso argomento è quindi necessario iniziare una nuova catena, in cui verranno postate risposte diverse da quelle della catena precedente.

E' comunque possibile differenziare le tipologie di contenuti principali presenti in Twitter, rappresentati con la relativa distribuzione percentuale nella Figura seguente (Kelly, 2009).

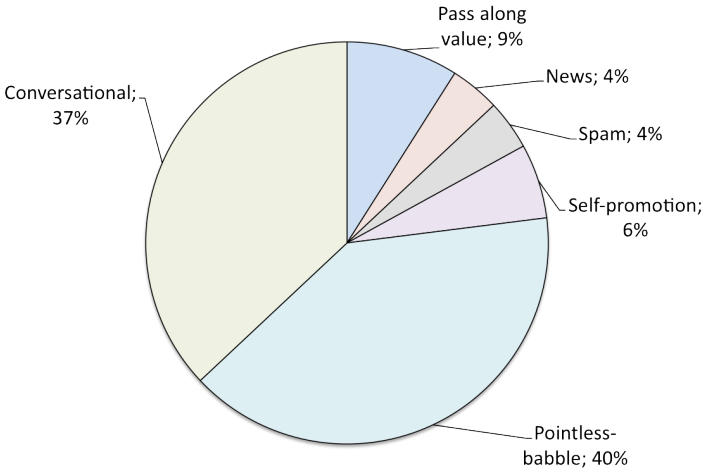


Figura 8 – Distribuzione dei contenuti per tipologia

Come si può vedere, poco più di un terzo è occupato dalla conversazioni tra coppie o gruppi di utenti, mentre i *retweet* rappresentano circa un decimo del totale (i.e. *pass along value*). La porzione etichettata come *pointless babble* contiene in realtà proprio quelle opinioni ed espressioni personali che sono di forte interesse per la *social media analysis*, il cui processo effettivo verrà trattato nel Capitolo 4.

3.3.4 Volatilità, attualità e credibilità dei contenuti

Ci sono dei fattori rilevanti che contribuiscono a definire meglio e contraddistinguono i contenuti presenti nel microblogging. Vediamoli nel seguito prendendo come confronto un servizio, come Wikipedia, che differisce notevolmente per caratteristiche:

- *volatilità* – indica contenuti freschi, addirittura in realtime, aggiornati ogni giorno, ogni ora e minuto, quindi altamente volatili. Differenti dai contenuti trovabili su Wikipedia, i quali rimangono online per anni;

- *attualità* – ulteriore variabile che si differenzia dalla volatilità perché un commento è attuale se parla di un argomento recente all'istante di tempo in cui viene pubblicato. Di conseguenza un post volatile può essere o non essere attuale;

- c*redibilità* – i dati sono accettati se considerati veri, reali e credibili da più persone.

Com'è possibile vedere dalla Figura 9, Twitter presenta un forte grado di volatilità e novità in real-time, in quanto i post sono perennemente aggiornati. Se attraverso il motore di ricerca, sia esso Google o il *search* di Twitter, si vuole scoprire cosa viene detto di un determinato evento o topic, vengono restituiti tutti i *tweet* più recenti fino a pochi secondi prima dell'invio della query.

D'altra parte però, questo social network ha una bassa attendibilità, in quanto non c'è, e sarebbe probabilmente impossibile da attuare, una verifica di contenuti postati dagli utenti, a differenza di Wikipedia, in cui il meccanismo *peer-to-peer* su cui si basa il servizio porta gli stessi utenti ad effettuare dei cambiamenti e/o delle correzioni in caso di necessità. In Figura 9 è indicato il posizionamento di Twitter e Wikipedia rispetto agli assi di *volatilità* e *credibilità*.

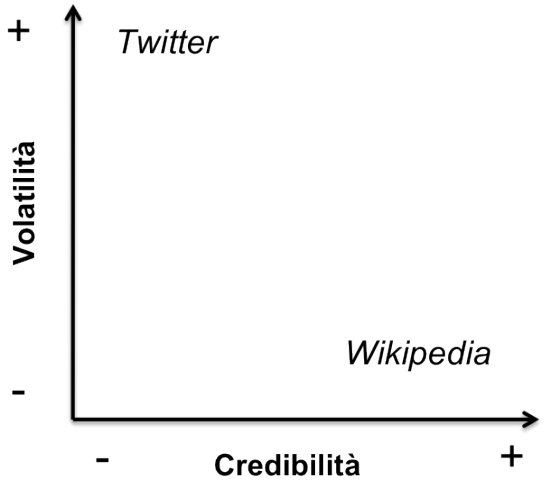


Figura 9 - Posizionamento di Twitter rispetto alla volatilità e credibilità dei contenuti

3.4 Caratteristiche del network

La semplicità della piattaforma permette di mappare gli utenti come nodi e seguire le relazioni come collegamenti, andando così a generare un grafo diretto utilizzabile per analizzare la rete sociale (Teutle, 2010).

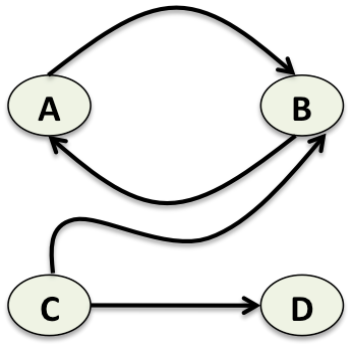


Figura 10 - Grafo delle relazioni tra gli utenti di Twitter

E' da notare che l'informazione fluisce nel verso opposto rispetto alla direzione del collegamento diretto indicata dalla freccia, dal momento che questa rappresenta la relazione di sottoscrizione ai contenuti di un altro utente.

Quest'ultimo aspetto è di notevole interesse in quanto, al momento, il numero di *follower* è considerato tra i parametri principali per valutare l'autorità di un

determinato utente. La relazione di *following* può essere intesa come non del tutto casuale, situazione che renderebbe la sottoscrizione agli aggiornamenti altrui, un indicatore debole e legato soprattutto a un concetto di similarità tra individui. In pratica un utente effettuerebbe la sottoscrizione ad un altro, solo nel caso in cui il primo sia effettivamente interessato ai contenuti prodotti dal secondo e questo ricambierà la relazione se gli ambiti d'interesse sono effettivamente condivisi. Questo fenomeno, noto con il nome di *omofilia* è stato studiato e confermato all'interno di diverse reti sociali (McPherson, Smith-Lovin, & Cook, 2001), così come in Twitter (Weng, Lim, Jiang, & He, 2010).

Date queste considerazioni potrebbe essere rilevante utilizzare come indicatore più che il numero assoluto di *follower*, il rapporto *indegree* (i.e. *follower*) su *outdegree* (i.e. *following*), valore che può essere utilizzato come ulteriore punto di vista per la valutazione dell'autorità, con la possibilità di identificare delle categorie prefissate di utenti (Teutle, 2010).

Valore rapporto (*follower*/*following*)	**Descrizione utente**
rapporto < 1	L'utente "segue" più utenti di quanti si siano sottoscritti lui, tipicamente ricerca e raccoglie informazioni. Se il rapporto è troppo basso, il nodo potrebbe essere un bot o uno spider che raccoglie informazione sui *trending topic*.
rapporto ≅ 1	Valori di *indegree* e *outdegree* simili sono tipici di una community.
rapporto > 1	Utente rispettato dalla community e che tendenzialmente condivide risorse apprezzate dagli altri.
rapporto > 10	Rappresenta nodi che hanno un largo impatto e che non hanno interesse nel sottoscriversi ad altri utenti.

Tabella VII – Valutazione utente in base al rapporto indegree/outdegree

Altre due caratteristiche distintive del network sono le *reciprocità* delle connessioni e i *gradi di separazione* tra gli utenti.

Per quanto riguarda il primo, Twitter mostra un livello molto basso di reciprocità tra i nodi, indicando che il rapporto tra due utenti è biunivoco (i.e. reciproco) solo per una piccola porzione della rete sociale, mentre nella maggioranza dei casi è assolutamente univoco, con studi che mostrano un valore pari a circa il 78% (Kwak, Lee, Park, & Moon, 2010), significativamente superiore a quanto rilevato per altri social network (Cha, Mislove, & Gummadi, A measurement-driven analysis of information propagation in the Flickr social network, 2009).

Il concetto dei *gradi di separazione*, come già citato nel Capitolo 2, è un elemento chiave nello studio e nella comprensione di una rete sociale[17]. Contrariamente a quanto si possa pensare vista la bassa reciprocità, è stato infatti provato che il 98% degli utenti rientra nei 6 gradi di connessione, con un valore medio per il network pari a 4,62 (Lardinois, 2010). Questo è un dato degno di nota che rileva come i legami tra utenti siano stretti in particolar modo con l'obiettivo di ricercare informazione, rendendo l'impiego di Twitter non limitato a quello tipico di un servizio di social networking.

3.5 Demografiche

Nel seguente paragrafo verranno forniti alcuni dati riguardanti le statistiche demografiche di Twitter. Coerentemente con le finalità del lavoro, secondo cui il criterio di selezione degli utenti del data set è stato la localizzazione, e al fine di disporre di un elevato livello di accuratezza, i dati, presi da CommScore[18], si riferiscono al traffico sul sito nel mese di Aprile 2010 di un audience esclusivamente collocata a Londra.

La Tabella VIII, sintetizza la dimensione del set, mostrando un incidenza del 13,2 % dei visitatori di Twitter sul totale degli utilizzatori di Internet.

[17] Milgram, S. (1967). The small world problem. *Psichology today , 2* (1), 60-67.
[18] www.commscore.com/

Target audience	***Total Internet (000)***	***Twitter.com (000)***
Audience based: London	10.388	1.372
% reach	-	13,2

Tabella VIII - Dimensione dell'audience

Il grafico seguente, Figura 11, mostra, sempre tenendo come riferimento sia l'utilizzo del Web che quello del microblogging, l'incidenza percentuale dei visitatori unici sul totale distribuiti per fasce di età e la *reach* dell'utenza del microblogging rispetto al Web.

I risultati confermano altre statistiche e infografiche facilmente reperibili, secondo le quali la fascia in cui il servizio è più diffuso è quella tra i 30 e i 45 anni.

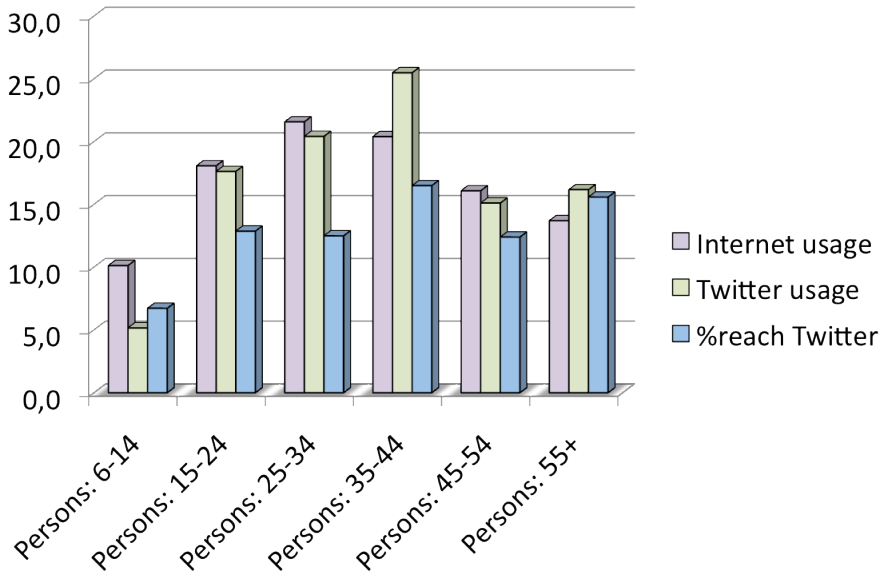


Figura 11 - Distribuzione degli utenti per fascia di età e reach di Twitter

3.5.1 Distribuzione geografica

Vediamo di seguito com'è la distribuzione degli utenti per utilizzo del servizio a livello di nazioni e di città del mondo. Il paragrafo assume un certo interesse in relazione al criterio con cui, come detto, è stato selezionato il data set per analisi.

La prima infografica in Figura 12 indica i paesi che registrano il maggior utilizzo del microblogging (Pals, 2010).

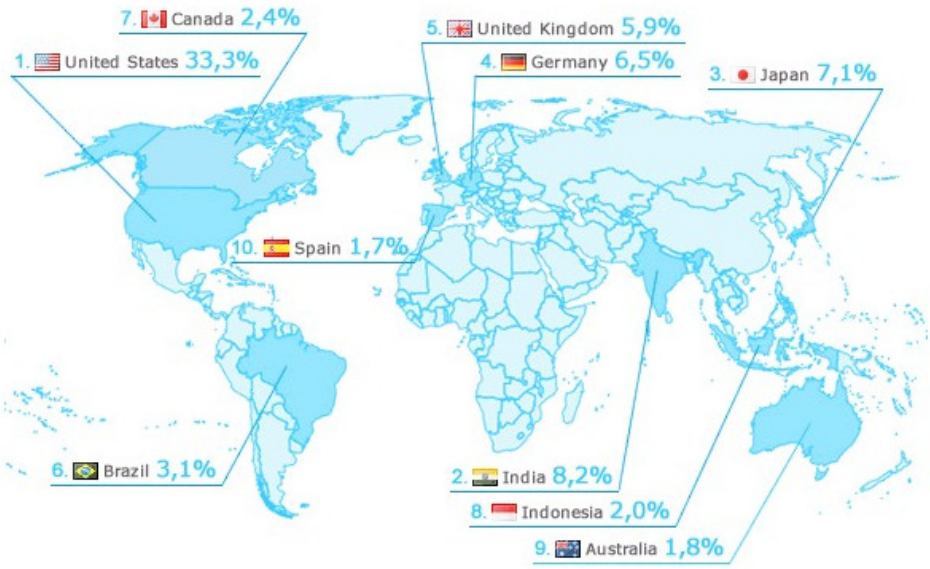


Figura 12 – Mappa delle principali nazioni per traffico sul sito

Le percentuali indicano il traffico sul sito, con all'incirca un terzo del totale proveniente dagli Stati Uniti, mentre per quanto riguarda il continente europeo le principali nazioni sono la Germania e al secondo posto l'UK.

Secondo invece un report pubblicato da Sysomos (Sysomos Inc., 2010), basato su un set comprensivo di 13 milioni di account attivi e su un insieme di *tweet* registrati nel periodo dal 16 Ottobre 2009 al 16 Dicembre 2009, i dati sono i seguenti (dati riportati per le prime cinque nazioni):

Country	% total tweets contributed	% of Twitter users
U.S.A	56.59%	50.88%
UK	8.09%	7.20%
Brazil	6.73%	8.79%
Canada	4.36%	4.35%
Australia	2.63%	2.39%

Tabella IX - Principali nazioni per contributo all'attività e al numero di utenti

I risultati, indicati nella Tabella IX, mostrano in questo caso il contributo relativo in termini di attività e utenti sul totale del set, con gli USA che occupano la prima posizione per entrambe i parametri, così come l'UK mostra i valori più significativi per quanto riguarda l'Europa. Secondo i trend è comunque registrabile una riduzione progressiva del contributo relativo proveniente dagli Stati Uniti, dovuto ovviamente alla crescita nell'utilizzo da parte di altri paesi.

Gli stessi dati sono di seguito presentati in relazione alle città. La tabella seguente (Branckaute, 2010), misura la media dei *tweet* al minuto registrati durante il periodo dal 10 al 14 Giugno 2010, utilizzando il tool Tweet-o-Meter[19].

City	Tweet per minute
Los Angeles (USA)	1244
New York (USA)	1176
San Francisco (USA)	479
Tokyo (JPN)	342
London (UK)	188

Tabella X - Principali città per attività al minuto

I risultati, da considerarsi necessariamente contingenti al breve periodo di misurazione, indicano come principali città al mondo per contributo New York e Los Angeles, entrambe negli USA, mentre per l'Europa troviamo in cima Parigi, subito seguita da Londra e Barcellona.

Come effettuato precedentemente confrontiamo con i dati proposti dal report di Sysomos (Sysomos Inc., 2010), presentati nella Tabella XI:

[19] http://www.casa.ucl.ac.uk/tom/

City	% total tweet contributed	% of Twitter users
New York	2.37%	1.44%
London	2.12%	2.08%
Los Angeles	2.10%	1.63%
Chicago	1.46%	1.20%
San Paolo	1.18%	1.47%

Tabella XI - Principali città per contributo all'attività e al numero di utenti

Quello che emerge complessivamente dalle statistiche è che UK e Londra appaiono come la nazione e la città europea che mostrano il livello maggiore di utilizzo di Twitter e proprio per la forte diffusione e gli eccellenti flussi di attività è la città stata selezionata come criterio di localizzazione per la selezione del data set finale.

3.6 Tassonomia degli utenti

A differenza di altri social media, come ad esempio Facebook, in cui le pagine utente sono distinte da quelle tradizionalmente impiegate dai brand, i.e. le cosiddette *fan page*, per quanto riguarda Twitter, al di là di una categorizzazione per area tematica fornita dal servizio stesso per suggerire nuovi utenti a cui sottoscriversi, non c'è nessuna differenziazione a priori tra i profili che permetta di effettuarne una qualche categorizzazione.

In quello che è una dei primi studi sull'utilizzo di Twitter (Java, Finn, Song, & Tseng, 2007), vengono identificati tre semplici categorie di utenti:

a. *information sources* – pubblicano notizie e tendono ad avere una larga base di *follower*, possono essere sia individui che account che postano messaggi in automatico;

b. *friends* – è un categoria varia che comprende la maggior parte dell'utenza, includendo a famiglia, colleghi di lavoro e sconosciuti (offline);

c. *information seekers* – sono utenti che tendono a generare pochi contenuti, ma che "seguono" altri regolarmente.

Ricercando nel Web è possibile individuare diverse proposte riguardanti una tassonomia per gli utenti, realizzate soprattutto in ambito di *social media analysis* e Web marketing&PR. Ovviamente è molto difficile definire una clusterizzazione rigorosa che si applichi perfettamente a qualsiasi utente e per questo che le categorie indicate, provenienti da fonti molto autorevoli in ambito Web, sono funzionali piuttosto ad una rappresentazione di stereotipi molto polarizzati che nel complesso riassumo molto bene i profili di utenza di Twitter.

La prima, creata dal guru di Internet Guy Kawasaky, differenzia gli user in sei principali tipologie, basate sostanzialmente sull'utilizzo effettivo che ciascuna di queste fa del microblogging (Kawasaky, 2009).

Tipologia	**Caratteristiche**	**Motivazioni**
The Newbie	Utente che si è registrato da poco al servizio e la cui maggioranza dei contenuti sono relativi al proprio *lifestreaming*. Questi utenti si evolvono in un'altra tipologia o abbandonano il servizio.	Curiosità di utilizzo.
The Brand	Il brand divide la sua attività tra l'utilizzo di Twitter come canale di marketing e l'engage degli altri utenti in modo da non sembrare stia usando il servizio esclusivamente come marketing tool.	Aumentare la *brand awareness*.
The Smore	Twitter è utilizzato come mezzo di self-promotion per ottenere qualcosa dagli altri utenti.	Guadagnare *follower* o opportunità di business.
The Bitch	Twitter è effettuato come mezzo di provocazione – i.e. un equivalente dei troll all'interno dei forum.	Generare reazioni rabbiose o *flame*.
The Maven	Questa tipologia di utente è un esperto all'interno del proprio campo (molto spesso web marketing e web design).	Diffondere i propri post ed essere riconosciuti come esperti.
The Mensch	Rappresentano una tipologia non molto diffusa, che partecipa poco alle conversazioni, ma interviene ogni qual volta qualcuno ha bisogno d'aiuto, sapendo o sapendo come trovare una risposta.	Aiutare gli altri utenti.

Tabella XII - Categorizzaione degli utenti per modalità di utilizzo

L'altra tassonomia è presa da un articolo apparso su Mashable (Deal, 2009), che ben definisce quali tipologie di account è possibile incontrare su Twitter, focalizzandosi quindi più sui profili in sé, che sull'utilizzo dello servizio.

Tipologia	**Caratteristiche**
Memes, games & activities	Sono legati a semplici giochi, attività (e.g. quiz) e *meme*.
Company, product or brand	Nel migliore dei casi ci sono brand che hanno capito l'importanza dell'engagement e dell'interazione con i clienti. In altri casi utilizzano Twitter come canale di diffusione di notizie riguardanti l'azienda e/o i prodotti.
Suspended accounts (spam)	Account di spam che molto spesso finiscono per violare i termini di servizio e per essere sospesi.
Guy in a suit, corporate back round, with more following than followers	Il loro avatar e la bio servono come business card con i quali utilizzano Twitter per trovare affari e guadagnare fiducia. Sono molto spesso CEO o fondatori di qualcosa e nelle descrizioni compaiono keywords come: startup, expert, marketing etc.
Default avatar:spam, n00b or something else	Rappresentano tendenzialmente n00b – i.e. nuovi utenti – o account di spam.
Web gurus & evangelists	Sono personaggi molto importanti e noti all'interno del mondo Web e dei Social Media e contraddistinti da un seguito di utenti molto ampio. Esempi sono Pete Cashmore, Guy Kawasaky, Kevin Rose etc.
Entertainers, atlete and otherwise famous people	Personaggi famosi e conosciuti attraverso i media tradizionali. Sono tra i top user di Twitter perché raccolgono in tempi molto brevi un ampio seguito di *follower*.
News sources	Fonti di informazioni che utilizzano Twitter come canale aggiuntivo ai tradizionali.
Characters, personalities and unusual entities	Personaggi famosi ma che non sono reali, provenienti da: film, serie tv ed altre fonti.
The rest of us	Persone comuni che utilizzano Twitter per i più svariati motivi.

Tabella XIII - Categorizzazione per tipologia di account

Capitolo 4

Il processo di analisi e la tecnologia

4.1 Introduzione

Le fonti Web 2.0, dai social network (e.g. *Facebook, Myspace, Twitter*), alle realtà virtuali (e.g. *SecondLife*) e alle comunità online (e.g. *YouTube, Wikipedia*), stanno sempre più rapidamente trasformando il modo in cui i consumatori ottengono le informazioni e formano le loro opinioni e giudizi su prodotti, marchi, servizi ed esperienze, offrendo l'opportunità alla aziende di integrarli nei propri sistemi di business intelligence ed in quelli marketing intelligence. Una delle sfide chiave nell'adozione di un tool di monitoraggio dei social media è senza dubbio il fatto che si tratti di piattaforme orizzontali che vanno a supportare una varietà di funzioni aziendali ciascuna con differenti necessità.

In particolare, come esemplifica la Figura 13, è possibile identificare tre diversi macroaspetti – *security, business intelligence* e *marketing* – le cui rispettive

sottocategorie ben rappresentano le finalità che possono richiedere l'impiego della *social media monitoring & analysis,* con obiettivi sia orientati all'attuazione di strategie offensive di crescita, che difensive e focalizzate sulla gestione del rischio.

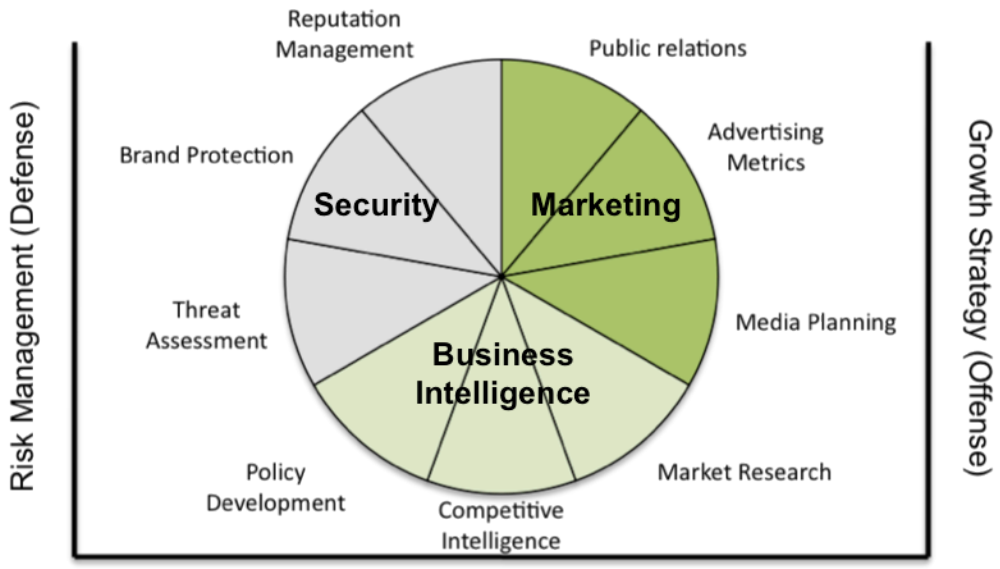


Figura 13 - Parti costituenti il monitoraggio dei social media

Se consideriamo ad esempio le ricerche di mercato, l'analisi dell'opinione presenta una serie di vantaggi rispetto alla metodologia classica con cui queste sono tradizionalmente svolte:

a. coloro che esprimono le opinioni lo fanno senza la consapevolezza di essere esaminati;

b. gli elementi di opinione e valore contenuti nei pari sui network online sono aperti e non predefiniti in un questionario;

c. coloro che esprimono le opinioni hanno una certa esperienza in relazione all'impresa o al brand e quindi la loro visione può fortemente influenzare il pubblico target.

Un altro aspetto importante e abilitato dalle enorme quantità di informazioni che arriva dalle reti sociali online è quello di *reputation management*, in cui un tracciamento continuo dei canali sociali permette un monitoraggio dell'opinione degli utenti, eventualmente del sentiment di queste opinioni, con la possibilità di generare degli alert nel caso in cui vengano individuate delle possibili crisi.

Per monitorare il *buzz online* al momento ci si può avvalere di diversi sistemi in grado analizzare l'opinione in rete su un brand, un'impresa e le sue attività. Con "opinione" s'intende quell'insieme di percezioni e relazioni creato dai pareri postati su Internet attraverso le varie piattaforme del Web 2.0, opinioni libere che come ampiamente spiegato, possono avere un effetto enorme, andando a influenzare direttamente il valore percepito di prodotti e servizi.

Le tecnologie impiegate in questo campo estraggono e processano le informazioni partendo dai dati non strutturati provenienti da diverse fonti Web (e.g. blog, forum, Twitter), andando a costituire delle vere e proprie "macchine del clima sociale", i cui elementi e procedimenti di lavoro sono descritti nei paragrafi seguenti.

Il capitolo è composto come segue: nella Sezione 4.2 saranno presentate le fasi che definiscono il processo di analisi seguito dalle piattaforme di ascolto; la Sezione 4.3 è dedicata alla descrizione delle componenti che costituiscono effettivamente i tool; la Sezione 4.4 tratta l'offerta di mercato attuale, con una breve disanima di alcuni prodotti; nella Sezione 4.5 sono analizzati in dettaglio alcuni aspetti tecnologici relativi all'estrazione dei dati da Twitter; infine la Sezione 4.6 spiega il progetto in ambito *social media analysis* attualmente in corso per il comune di Milano.

4.2 Il processo di analisi

Esistono una molteplicità di differenti servizi e tecnologie che permettono di approcciarsi alle informazione fortemente disgregate che provengono dai social media. Questi tool, generalmente etichettati con il nome di *listening platforms* o *social media analysis/monitoring tools*, sono tecnologie specializzate e dedicate

all'estrazione di dati dalle fonti online. Nonostante possano esserci rilevanti differenze, metodologicamente tutte si adattano ad un processo costituito principalmente da tre step:

- *data collection;*
- *data processing;*
- *delivery.*

4.2.1 Definizione dei tre processi

Data Collection

Consiste sostanzialmente nell'estrazione dei contenuti e dati dall'insieme di fonti *social* preselezionate. Elementi e sfide indicative da questo punto di vista sono l'ampiezza delle fonti e il relativo problema del *data-overload*, la riduzione dello spam e la possibilità di tracciamento dei dati in real-time.

Data Processing

La maggior parte delle piattaforme è costantemente impegnata in un progressivo miglioramento dei motori di analisi testuale, con l'obiettivo principale di perfezionare e incrementare metriche automatizzate relative alla definizione dei topic, alle informazioni demografiche, all'analisi del sentiment e all'individuazione degli influencer. Un altro aspetto, che riguarda soprattutto le necessità di grandi imprese che operano a livello internazionale, è la capacità di eseguire il *processing* su un'ampia varietà di lingue.

Come verrà spiegato a breve, questa fase è senza dubbio quello più complessa e delicata dal punto di vista delle soluzioni tecnologiche e dell'automazione.

Delivery

La modalità più diffusa di *delivery* è quella di una dashboard utilizzabile dall'utente finale. Quelle implementate da una buona parte dei tool più conosciuti incorporano molte funzionalità e, in alcuni casi, elevate possibilità di personalizzazione e un'ottima usabilità. Allo stato attuale ci troviamo però di

fronte a piattaforme molto semplici da utilizzare ma ben poco performanti per quanto riguarda la *collection* ed il *processing*, oppure molto potenti per questi'ultimi aspetti, ma con una limitata facilità di navigazione.

Alcuni vendor aggiungono all'offerta di delivery una reportistica periodica con eventuali analisi aggiuntive e/o veri e propri servizi di consulenza per sfruttare al meglio l'utilizzo delle piattaforme.

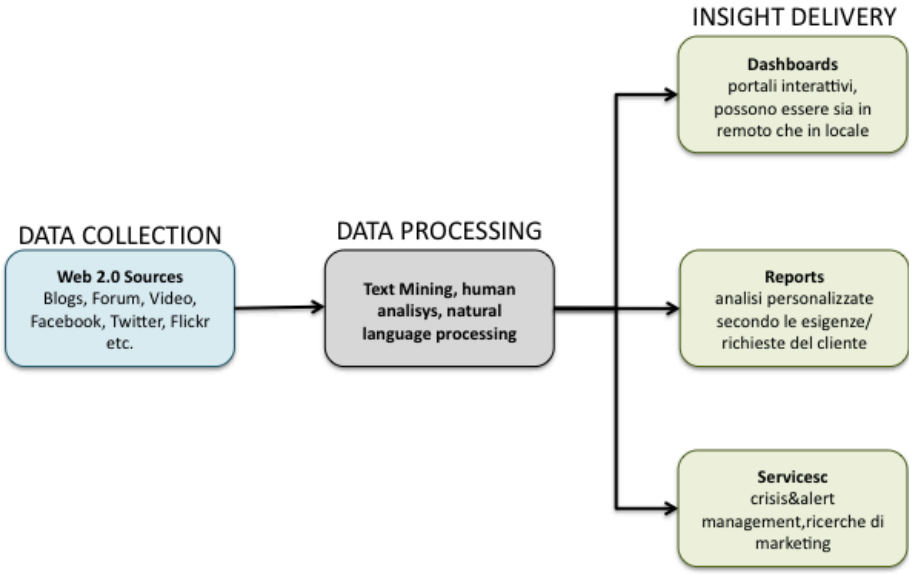


Figura 14 - Componenti del processo di social media analysis

4.3 Componenti dei tool

Vediamo ora quali sono gli elementi caratterizzanti e costituenti i tool, mostrandone la stretta relazione con i tre processi in precedenza definiti.

Web Crawler

Il crawler è l'elemento dedicato alla raccolta dei dati, collegato quindi al processo di *data collection*. Si tratta sostanzialmente di un agente software in grado di effettuare ricerche per keyword o per concetti semantici su differenti tipologie di fonte e mappare le connessioni esistenti tra servizi 2.0 ed individui in modo da tracciare ed analizzare l'andamento dei contenuti *user-generated*.

Le Sezioni 4.5.1 e 4.5.2 sono dedicate alla descrizione delle API di Twitter e di come viene effettuato il *crawling*.

Natural Language Processing (NLP)

Il motore che svolge la funzione di *natural language processing*, corrispondente al secondo dei tre passi del processo complessivo, ha il compito primario di filtrare le conversazioni per rilevanza o per qualsiasi altro indicatore/metrica significativo e può svolgere ulteriori funzioni accessorie, come per esempio l'estrazione e lo *scoring* del sentiment.

Innanzitutto è necessario effettuare una distinzione per quanto concerne la fase iniziale di definizione dei topic di interesse, la quale può essere *keyword-based*, situazione in cui l'utente definisce una serie di keyword in modo da filtrare il set complessivo dei dati, correndo fortemente il rischio di dati non corretti perché non disambiguati (e.g. "*apple"* come *apple inc.* e non come frutto) e dovendo in pratica effettuare manualmente una pulitura dei risultati, per esempio con relazioni tra keyword più stringenti; oppure per ricerca semantica (e.g. cerca "*apple"* come frutto e non come *apple inc*.), caso in cui si disponga di un motore automatizzato in grado di agire a livello semantico, o di un team di analisti che controllino i post manualmente, con lo scopo di ottenere un'analisi delle lingue naturali quantomeno comparabile a quanto fanno gli esseri umani, abbastanza simile da poter essere utilizzata in applicazioni che interagiscono con l'utente finale in forma linguistica.

L'NLP è sicuramente un ambito complesso e che esibisce un grado di interdisciplinarità molto alto, richiedendo un lavoro comune di competenze differenti tra loro: linguistica, informatica e psicologia.

Altri due aspetti che è importante citare, legati alla fase di *processing*, e parzialmente anche di *crawling*, sono le cosiddette *precision* e *recall*. Con la prima si intende la percentuale di documenti corretti restituiti, che sono rilevanti per la query effettuata. La *recall* è invece la percentuale di documenti restituiti in rapporto al totale dei documenti restituibili, metrica non semplicissima da

misurare in quanto bisogna essere a conoscenza del totale complessivo dei contenuti sarebbero teoricamente disponibili.

I seguenti possono essere presi in considerazione come fattori chiave e differenziali tra i vari tool:

- metriche di analisi proprietarie e livelli di analisi;
- accuratezza degli indicatori;
- determinazione del sentiment:
- capacità di operare con più lingue;
- integrazione della tecnologia con l'intelligenza umana.

Questo modulo incide notevolmente nella valutazione di tool competitor, con alcuni vendor che fanno precise scelte di investimento in tecnologia e propongono soluzione fortemente automatizzate, piuttosto di altre che hanno approcci più *analyst-intensive,* con una maggior rilevanza assunta dalla componente umana all'interno del flusso di processo.

La Figura 15 mostra l'architettura di un tool di benchmark, con la potenzialità di effettuare sia la parte di analisi semantica che quella del sentiment in modo automatizzato.

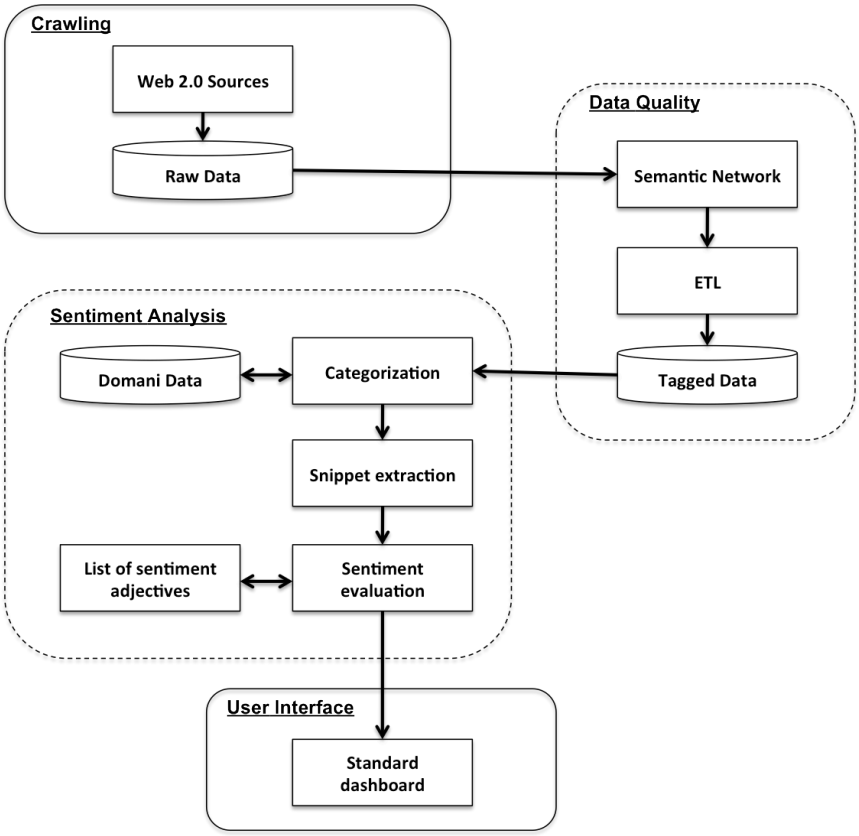


Figura 15 - Architettura di un tool di benchmark

Modalità di delivery: dashboard e report

La dashboard può essere costituita da un'interfaccia online utilizzabile in remoto, un'applicazione software *stand-alone* in locale, un cruscotto di dati personalizzabili o un'applicazione che interviene sul *workflow* con delle notifiche di alert in situazioni prestabilite. Gli elementi differenzianti le offerte sono rappresentati soprattutto da:

- modello di accesso;
- capacità di fornire dati ed informazioni in real-time;
- visualizzazione dei dati e *look-and-feel* dell'interfaccia utente.

In particolare l'interfaccia utente rappresenta forse l'elemento più determinante per l'esperienza d'uso finale. La modalità in cui questa è realizzata influisce

direttamente sull'utilizzo e sul modo in cui l'utente effettua le sue analisi. Una buona interfaccia dovrebbe garantire la giusta flessibilità e un'elevata autonomia all'utente finale, permettendogli di personalizzare al meglio le proprie analisi.

Alcune aziende offrono, come detto, servizi *analyst-intensive* di reportistica ad hoc e consulenza a integrazione della piattaforma, gestiti da un team di analisti e specialisti, i quali, oltre ad avere la possibilità di intervenire al fine di correggere o assistere le valutazioni della piattaforma tecnologica, forniscono prestazioni finalizzate a trarre maggiori benefici ed un miglior utilizzo dei dati a disposizione. I fattori chiave di valutazione sono:

- ampiezza dello staff di professionisti;
- dominio di esperienza e competenza.

La Figura 16 mostra la schermata di Radian6, descritto brevemente nella Sezione 4.4.1, che rappresenta al momento uno dei benchmark per quanto riguarda l'interfaccia e l'esperienza d'uso.

Figura 16 - Interfaccia utente di Radian6

4.4 L'offerta del mercato

La diversità di requisiti per le metriche dei social media e delle esigenze che distinguono differenti brand e industry ha condotto necessariamente ad una varietà di approcci tecnologici, ma non solo, che secondo Gartner (Frank, 2008) possono essere ristretti e schematizzati in cinque aree principali, illustrate nel seguito:

a. *freemium search tools* – la crescita e la diffusione del Web 2.0 hanno portato alla comparsa di una vasta gamma di servizi in grado di combinare capacità di base di ricerca con sistemi di alert e/o feed, realizzando così delle soluzioni *keyword-based* che possono essere sia free (generalmente supportate da advertising), che offerte in una modalità freemium che permette un livello di funzionalità di base e rende disponibili a pagamento i livelli di servizio più avanzato;

b. *self-service media monitors* – dal punto di vista delle funzionalità, questi tool normalmente forniscono fonti e funzioni analitiche addizionali rispetto ai freemium, disponibili attraverso una sottoscrizione a pagamento;

c. *tech-enabled agencies* – l'assunzione di un'agenzia specializzata rappresenta probabilmente il mezzo più comune e meno rischioso per integrare le capacità della *social media analysis* con le attività di marketing. Il problema principale di questo approccio, in cui sono sostanzialmente coinvolte agenzie di PR e marketing, è legato soprattutto a costi, tempistiche e alla possibile creazione di una dipendenza significativa dal lavoro dell'agenzia.

d. *analytic intelligence firms* – insieme con le agenzie, anche le aziende di ricerche media tradizionali hanno acquisito o sviluppato piattaforme e si sono specializzate in questi ambiti. Questa tipologia ha il vantaggio di essere in grado di integrare il *social media monitoring* con i tradizionali servizi di ricerca sui media.

e. *software as a service (SaaS) platforms* – questa è una categoria emergente di specialisti tecnologici che sta facendo concorrenza all'interno del mercato per soddisfare la crescente domanda di soluzioni complete. Queste piattaforme hanno tendenzialmente lo scopo di trasformare le pratiche di marketing, con lo svantaggio che il loro successo potrebbe essere largamente dipendente dalla disponibilità del cliente ad abbracciare e a investire risorse per l'innovazione del cambiamento.

Di seguito saranno forniti alcuni esempi di soluzioni appartenenti a diverse categorie, che per semplicità e brevità verranno differenziate in *soluzioni integrate* e *freemium tool*. Riguardo a questi ultimi l'interesse è focalizzato esclusivamente su quelli che si propongono di eseguire misurazioni della influence degli utenti di Twitter, con l'obiettivo di illustrarne le caratteristiche e le metriche impiegate; mentre con *soluzioni integrate* si intendono tutte le tipologie di servizi che sono stati inclusi nelle categorie identificate con le lettere *b,c,d* e *e.*

4.4.1 Esempi di soluzioni integrate

Radian6

Radian6[20] è un tool prodotto dall'omonima azienda canadese. Analizza un'ampia gamma di fonti (blog, forum, video, immagini, microblogging) ed è organizzato in profili, creati da un insieme di keyword funzionali alla definizione del topic di analisi (*topic profile*). La piattaforma, che consiste in una dashboard con accesso da browser, Figura 16, dispone di diverse funzionalità tra cui: andamento dei trend nel tempo, analisi di volumi assoluti segmentabili per svariate metriche (e.g. fonte, lingua), tag-cloud per i contenuti relativi al *topic profile* o ad un suo subset e l'analisi degli *influencer*.

Relativamente al topic impostato, il tool è infatti in grado di estrarre anche le informazioni sugli utenti che hanno contribuito alle conversazioni. All'interno dell'area di configurazione, sono presenti alcune impostazioni definite *influencer*

[20] http://www.radian6.com

EQ weightings, in cui vengono mostrati i parametri che andranno a definire il punteggio attribuito a ciascuno degli utenti che ha partecipato alla conversazione, parametri sui quali è possibile intervenire modificandone l'incidenza relativa. Nel caso di Twitter, gli elementi presi in considerazione dal tool per il calcolo del punteggio di rank sono:

- numero di post (*tweet*) creati riguardanti il topic*;*
- numero dei *following;*
- numero dei *follower;*
- conteggio degli aggiornamenti (*tweet*) totali postati dell'utente.

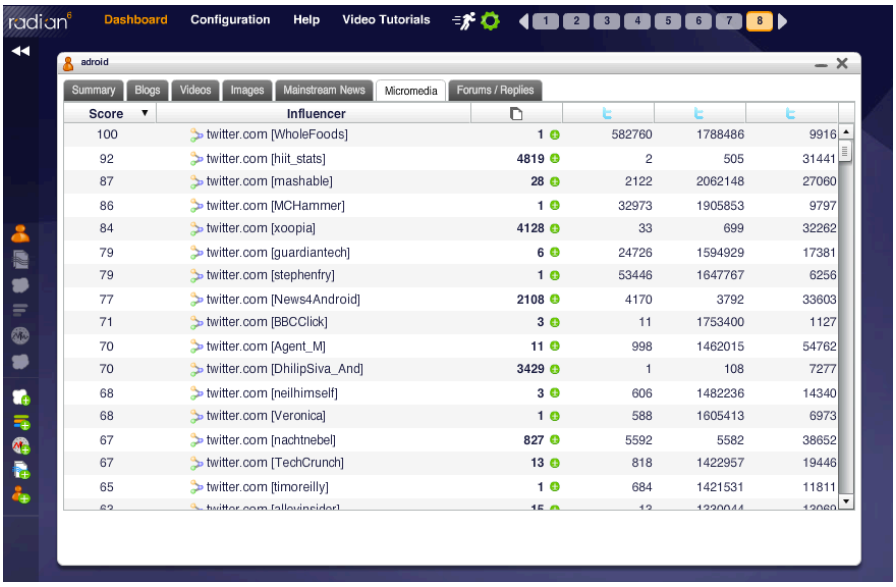


Figura 17 - Schermata del modulo per l'analisi degli utenti di Radian6

Nella Figura 17 è visibile il widget con il quale gestire l'analisi degli *influencer*. Una volta attivato è possibile, oltre alla modifica dell'ordinamento degli utenti secondo gli indicatori a disposizione, visualizzare per intero i post e verificare, attraverso l'opzione *social profile*, l'eventuale presenza di altri profili Web legati di un determinato utente.

Buzzmetrics

Buzzmetrics è l'offerta sviluppata e commercializzata da Nielsen[21], leader mondiale per le ricerche di mercato e la Web analytics su prodotti e servizi. Il prodotto si presta molto bene per eseguire analisi dei volumi e comparazioni tra competitor, dispone di una funzione di individuazione degli opinion leader e l'azienda fornisce il supporto di analisti per realizzare report integrativi. L'analisi semantica e una valutazione automatizzata delle opinioni sono parzialmente supportate.

TruCAST

Trucast[22] è una *product suite* per il tracciamento e l'analisi delle community online offerta da Visibile Technologies. L'orientamento principale del prodotto è costituito dall'analisi della reputation, ma vengono forniti anche dei tool di supporto per intervenire in caso di situazioni critiche. Per questa ragione la piattaforma è in grado di effettuare sia l'analisi degli influencer, sia una pesatura con punteggio delle fonti, così da poter pesare la reputation complessiva in base all'importanza relativa delle stesse.

Buzzlogic

Buzzlogic[23] è il prodotto per la Web analytics realizzato dall'omonima azienda. Si tratta di un tool molto orientato alle pratiche di marketing, che ha tra le sue funzionalità principali il posizionamento online delle iniziative di Web marketing, i.e. l'individuazione di siti Web strategici per campagnie di advertising. Fornisce l'analisi degli influencer e delle più importanti fonti online.

MAP - Sysomos

MAP (*media analysis platform*) è la soluzione principale offerta dalla canadese Sysomos[24], società recentemente divenuta sussidiaria di Marketwire, un'importante media company sempre con base in Canada. Il tool offerto è molto

[21] http://bit.ly/cqCDem (shortened link della pagina dedicata al tool)
[22] http://www.trucast.net
[23] http://www.buzzlogic.com
[24] http://sysomos.com

completo e dispone sia di un motore semantico per la distillazione dei contenuti, sia dell'analisi del sentiment in modo automatizzato. Oltre alla possibilità di compiere analisi dei trend e comparazioni tra competitor ed estrarre le informazioni geo-demografiche, uno dei moduli della piattaforma, l'*influencer search*, è dedicato all'individuazione e all'engagement degli influencer, con l'obiettivo di definirne l'autorità e la rilevanza sulla reputation del brand/prodotto in analisi.

Figura 18 - Schermata del tool Map di Sysomos

Blogmeter

Blogmeter[25] è un'azienda italiana risultato della collaborazione di un team di specialisti operanti nell'Internet intelligence, nella media research e nel *natural language processing*. L'offerta è una soluzione integrata che comprende oltre all'accesso ad un tool, una reportistica periodica ad-hoc. Il servizio, che è parzialmente automatizzato per quanto riguarda il processing dei dati estratti, è decisamente *analyst-intensive* sia con un intervento in quest'ultima fase, sia per l'analisi del sentiment, effettuabile su richiesta del cliente. E' disponibile tra le funzionalità anche l'individuazione degli opinion leader e dei gruppi di influenza.

[25] http://www.blogmeter.it

4.4.2 Esempi di freemium tool

Klout

Klout[26] è un servizio Web-based e gratuito che permette di misurare l'autorità di utente Twitter. I dati estratti riguardano la composizione del network dell'individuo, i contenuti creati e come gli altri user interagiscono con questo contenuto. La influence è definita attraverso un indicatore sintetico, il *Klout score*, che può variare su una scala da uno a cento. Il punteggio è principalmente determinato da tre aspetti, descritti di seguito, e tiene conto dei dati appartenenti ad un range di tempo pari ai trenta giorni precedenti la "richiesta di valutazione".

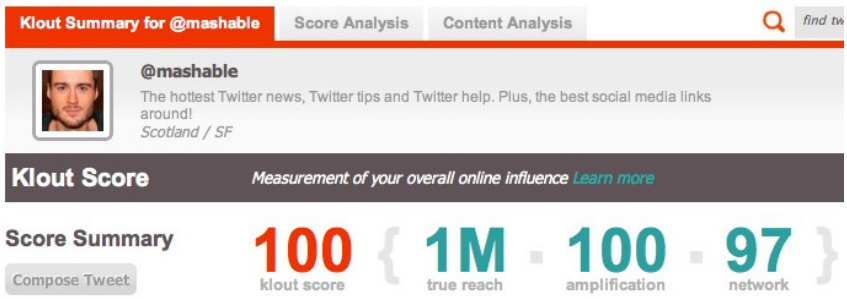


Figura 19 - Schermata di Klout

a. *true reach* – rappresenta la dimensione dell'audience raggiunta. Non coincide esattamente con il numero di *follower* in quanto, oltre a non includere nel conteggio account spam e inattivi, Klout valuta in modo differente la influence per ciascuna relazione individuale;

b. *amplification probability* – misura la probabilità con un valore da zero a cento che ci sarà un qualche tipo di azione altrui legata al contenuto creato. L'abilità di indurre gli altri utenti a rispondere e la velocità di propagazione all'interno del network sono elementi chiave;

[26] http://klout.com

c. *network influence* – indica, con un punteggio da zero a cento, il livello di influence dell'audience con cui si interagisce.

Riassumiamo con la Tabella XIV tutti gli indicatori analitici che è possibile ottenere attraverso l'utilizzo di *Klout*.:

Klout Score		
true reach	***amplification probability***	***network analysis***
total followers (n)	*total retweets (n)*	*list inclusions (n)*
mutual follows (n)	*@mention count (n)*	*follower/follow ratio (n)*
follower mention (%)	*unique msg retweeted (n)*	*follower back (%)*
follower retweet (%)	*outbound msg ratio (n)*	*unique @senders (n)*
-	-	*unique retweeters (n)*

Tabella XIV - Indicatori presi in considerazione per le metriche di Klout

TunkRank

TunkRank[27], è un servizio free e accessibile via Web finalizzato alla misurazione dell'autorità di un determinato utente. Le due idee principali su cui si basano il tool e l'algoritmo che gli sta dietro sono le seguenti:

a. il totale dell'attenzione che si può fornire è distribuito tra tutti gli utenti che vengono seguiti (*following*); più questo valore è altro, minore sarà l'attenzione dedicata a ciascuno.

b. la *influence* di un utente dipende dal totale dell'attenzione che i *follower* possono dargli.

Il modello che definisce il punteggio e la metrica pone alcune assunzioni:

- *Influence (X)* = il numero atteso di individui che leggeranno un *tweet* che *X* ha postato, includendo anche tutti i *retweet* di quel *tweet*;

[27] http://tunkrank.com

- se *X* è un membro dei *Followers(Y)*, allora c'è una probabilità di *1/||Following(X)||* che *X* leggerà un *tweet* postato da *Y*, dove *Following(X)* è un insieme di individui che *X* segue;

- se *X* legge un *tweet* postato da *Y*, c'è una probabilità costante *p* che *X* ne farà il *retweet*.

Da questo modello, la influence di un utente si misura ricorsivamente, ponendo di sapere la probabilità costante di *retweet p*:

$$Influence(X) = \sum_{Y \in Followers(X)} (1 + p * Influence(Y)) / \| Following(Y) \|$$

Il significato e gli elementi della metrica, essendo il punteggio interamente legato al grafo dei *follower*, rendono questa molto simile al funzionamento dell'algoritmo di ordinamento dei risultati di ricerca utilizzato da Google, il *PageRank* (Brin & Page, 1998).

Twitalyzer

Twitalyzer[28] è un servizio freemium Web-based creato dalla Web Analytics Demystified, che fornisce una dashboard con alcuni indicatori analitici per la valutazione di un account Twitter. Il tool ha delle offerte di abbonamento a pagamento le quali variano per le opzioni di supporto e di personalizzazione del tracking. Così come altri servizi, i dati estratti ed analizzati appartengono ad un range di tempo pari ai trenta giorni precedenti alla richiesta. La versione base e gratuita permette di ottenere i seguenti indicatori:

a. *impact score* – punteggio, su una scala da uno a cento, che rappresenta la *influence* dell'utente nel network e viene calcolato come combinazione dei seguenti fattori:

    - numero di *follower*;
    - numero di *mention* uniche ricevute;

[28] http://twytalizer.com

- la frequenza con cui l'utente riceve *retweet* da utenti unici;
- la frequenza a cui l'utente effettua *retweet* a utenti unici;
- la frequenza relativa di aggiornamento dello stato (i.e attività);

b. *influencer type* – categorizzazione dell'utente come *sun, spider* o *source,* esattamente come quella utilizzata da Forrester e descritta nel Capitolo 2 (Sezione 2.6).

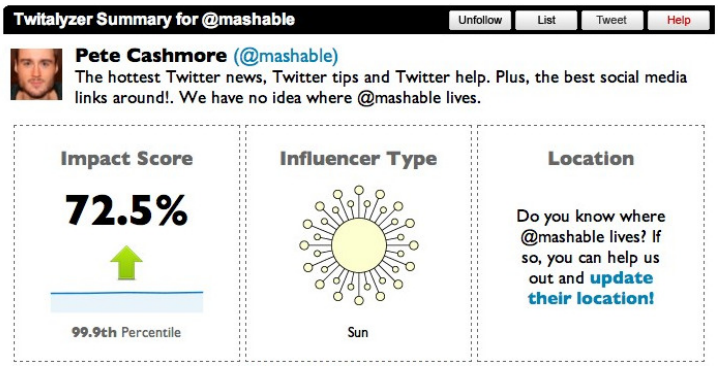


Figura 20 - Schermata di Twitalyzer

4.5 L'estrazione dei dati da Twitter

La seguente Sezione è dedicata a una spiegazione dettagliata degli elementi tecnologici necessari al processo di estrazione dei dati dalla piattaforma di microblogging, concentrandosi in particolar modo sugli aspetti riguardanti il *crawling*, i.e. API e database, con alcuni accenni sulla visualizzazione dei dati, tematica fondamentale per il *look-and-feel* della dashboard e per l'utilizzabilità dei tool.

4.5.1 Il crawling

L'estrazione di dati dal Web richiede l'impiego di tecnologie che, benché standardizzate, hanno delle evidenti problematiche in fase di implementazione, dovute in modo particolare alla diversità delle fonti che si vogliono analizzare ed ai contenuti a cui si desidera accedere. E' possibile distinguere due tipologie principali di crawler.

Parser HTML

I crawler basati su *parser HTML* (*hypertext markup language*) ottengono i dati analizzando la struttura della pagina Web, in un modo che è più o meno assimilabile al comportamento di un utente che naviga su Internet: così come accade per l'individuo, il quale visualizza una pagina alla volta e si sposta da una all'altra aprendo dei collegamenti ipertestuali, così il crawler analizza pagina per pagina e sceglie la successiva sulla base degli attuali collegamenti (i.e. link e URL) presenti. L'analisi della pagina Web è appunto effettuata attraverso dei *parser*, i quali possono essere descritti come degli algoritmi che identificano i tag HTML presenti e, in base alle scelte effettuate dall'utente, esamina il testo presente al loro interno. E' esattamente questo il contenuto che viene estratto e attraverso opportune connessioni alla base di dati memorizzato per il riutilizzo.

Questo approccio consente una ricerca molto ampia e completa a livello di sito Web, ma di contro richiede tempi di attesa piuttosto lunghi, tendenzialmente dal secondo ai venti secondi per pagina scannerizzata, paragonabili quasi alla velocità di lettura manuale ed una connessione al sito sempre attiva. Questi aspetti assumono una certa rilevanza da tenere in considerazione nel momento in cui le fonti hanno un tasso di aggiornamento elevato, esattamente come nel caso dei social network e delle fonti 2.0. A proposito della costante attività della connessione, si tratta di un elemento non gestibile esclusivamente lato tecnologia, ma è legato piuttosto alle policy del sito riguardanti il trattamento dei dati.

API

Le API (*application programming interface*) sono servizi offerti dal gestore della piattaforma o del sito con lo scopo di permettere l'interfacciamento con i dati e/o le funzionalità del sito stesso. L'utilizzo rende l'ottenimento dei dati più veloce, più facile da gestire, ma ristretto sia da limiti che possono riguardare aree del sito e/o limiti di tempo (e.g. numero di chiamate orarie fisse).

A differenza di un crawler basato su parser HTML, quello basato su API ha sì performance più elevate, ma non essendo a disposizione di default, richiede che il

sito abbia sviluppato e messo a disposizione interfacce ad hoc. Tra gli esempi più noti di utilizzo di questa tecnologia, ci sono i social network Facebook e Twitter.

Di seguito la Tabella XV riassume brevemente vantaggi e svantaggi delle due tipologie di crawler viste:

Tecnologia	Vantaggi	Svantaggi	Che cosa spetta al gestore del sito
Parser HTML	- Ampiezza dei contenuti estraibili.	- Lentezza - Rischio di essere bloccati.	- Permessi di accesso dei bot/spider.
API	- Connessione sempre attiva; - Performance elevate.	- Limiti chiamate; - Contenuti limitati.	- Implementazione API.

Tabella XV - Caratteristiche riassuntive delle metodologie di crawling

4.5.2 Le API di Twitter

Le API di Twitter sono al momento composte da due differenti tipologie, due appartenenti al tipo REST e una al tipo *streaming*. Per la maggior parte delle applicazioni gli sviluppatori le utilizzano e combinano tutte e tre.

a. metodi *REST*:

   - *REST API* – consentono l'accesso ai dati core di Twitter, quali gli aggiornamenti della timeline, i dati di aggiornamento dello status e le informazioni utente;
   - *search API* – permettono l'interazione con il search di Twitter e l'estrazione dei dati sui trend dei volumi;

b. *streaming API* – forniscono l'accesso ad un elevato volume di *tweet* in realtime in un modulo campionato e filtrato.

La presenza di due distinte REST API è dovuta esclusivamente alla storia riguardante il loro sviluppo, inizialmente affidato ad una società esterna ed

indipendente, la Summize Inc., la quale si occupava di fornire le capacità di ricerca per i dati di Twitter. La Summize fu in seguito acquisita e brandizzata come *Twitter Search*. Il rebranding fu semplice, ma l'integrazione completa del *Search* e delle sue API nel codice base di Twitter è stato più difficile. Benché sia nel dichiarato interesse del servizio unire le due API, finché le risorse lo permetteranno le REST e le *search* rimarrano entità separate.

Le *streaming* API sono a loro volta distinte e supportano le connessione *long-lived* su un'architettura differente.

REST

L'acronimo REST sta per *representational state trasfer* ed indica precisamente uno stile di implementazione di architettura del software per sistemi *hypermedia* distribuiti (e.g. il Web). Le architetture in stile REST consistono in client e server, dove il client effettua delle richieste che il server processa ritornando delle risposte appropriate. Richieste e risposte sono costruite attorno al trasferimento di *rappresentazioni di risorse*. Una risorsa può essere essenzialmente qualsiasi coerente e significativo concetto che può essere indirizzato. La rappresentazione di una risorsa è tipicamente un documento che cattura lo stato corrente o inteso di una risorsa. In qualsiasi momento, un client può essere o in transizione tra stati o "*at rest*" (i.e. a riposo). Un client in stato di riposo è in grado di interagire con l'utente senza generare carico o consumare *storage per-client* sul set di server o sulla rete. Il client inizia ad inviare richieste quando è pronto ad entrare in transizione in un nuovo stato. Mentre uno o più richieste sono in corso, il client è considerato in transizione.

Formati di Output

Le API al momento supportano e restituiscono i seguenti formati di dati, con alcuni metodi che accettano solo alcuni di questi.

- *XML* – acronimo di *extensible markup language*, indica un metalinguaggio di markup, cioè un linguaggio marcatore che definisce un meccanismo sintattico che consente di estendere o controllare il significato di altri

linguaggi marcatori. A differenza dell'HTML, il quale definisce una grammatica per la descrizione e la formattazione delle pagine Web, l'XML è un metalinguaggio utilizzato per creare nuovi linguaggi, atti a descrivere documenti strutturati e con il quale è possibile definire dei tag propri a seconda delle esigenze;

- *JSON* – acronimo di *javascript object notation*, è un altro formato adatto per lo scambio di dati in applicazioni client-server. A differenza dell'XML non è un linguaggio di marcatura, ma un formato di interscambio di dati.;

- *RSS* – acronimo di *RDF site summary* ed anche del più noto *really simple syndication*. Conosciuto come uno dei più diffusi formati di distribuzione di contenuti Web, è basato su XML, con cui condivide le caratteristiche di semplicità, estensibilità e flessibilità. La fruizione di un documento RSS è un processo molto semplice effettuabile attraverso un'applicazione che effettuando il *parsing*, converte i contenuti decodificati nel formato utile all'obiettivo;

- *ATOM* – l'*atom syndication format* è un formato di documento basato su XML per la sottoscrizione di contenuti Web. L'Atom è considerato "il fratello minore" dell'RSS e possiede caratteristiche ed utilizzi molto simili a quest'ultimo.

I limiti delle API

Le API REST di Twitter permettono ai client di effettuare solo un limitato numero di chiamate per ora, variabile a seconda del metodo di autorizzazione utilizzato.

- le chiamate anonime basate sull'indirizzo IP dell'host possono essere al massimo 150 per ora;

- le chiamate concesse attraverso l'autenticazione *OAuth* (i.e. nuovo protocollo di autenticazione utilizzato da Twitter) sono al più 350 per ora.

Esiste anche un limite riguardante l'arco temporale, cioè a quanto indietro nel tempo corrispondono i *tweet* estratti dal servizio, vincolo pari ad un massimo di 10 giorni.

4.5.3 Il database e i diagrammi E/R

La figura seguente riassume il processo di interazione client-server necessario ad effettuare le ricerche all'interno del database e all'estrazione dei dati, mettendo in evidenza tutti gli elementi descritti nei precedenti paragrafi.

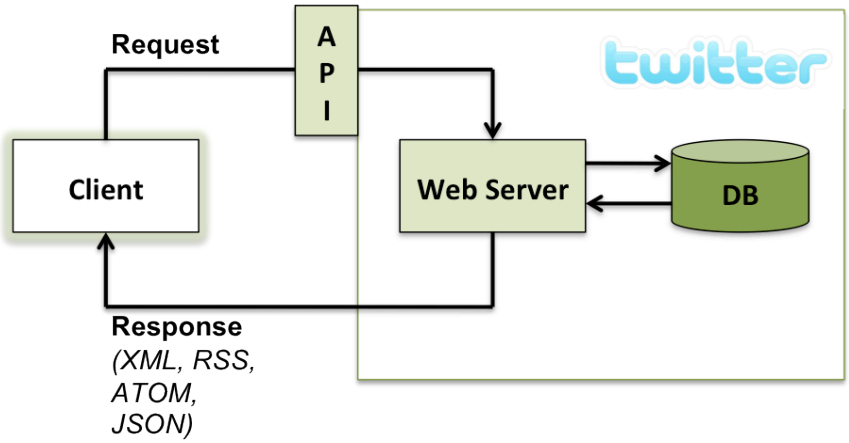


Figura 21 - Interazione client-server in Twitter

L'unico elemento su cui è possibile aggiungere qualche ulteriore dettaglio è appunto il database di Twitter, per il quale è presentato di seguito nella Figura 21 il relativo diagramma ER.

Il *modello entità relazioni* (*ERM*) è una rappresentazione concettuale ed astratta dei dati ed i diagrammi creati attraverso questo processo prendono il nome di *diagrammi ER*. Gli elementi costituenti sono le *entità,* che rappresentano classi di oggetti (i.e. fatti, cose, persone) che hanno proprietà comuni ed esistenza autonoma ai fini dell'applicazione di interesse, le *associazioni* (o anche *relazioni*), impiegate per collegare due o più *entità* ed infine gli *attributi*, finalizzati alla descrizione delle e*ntità.*

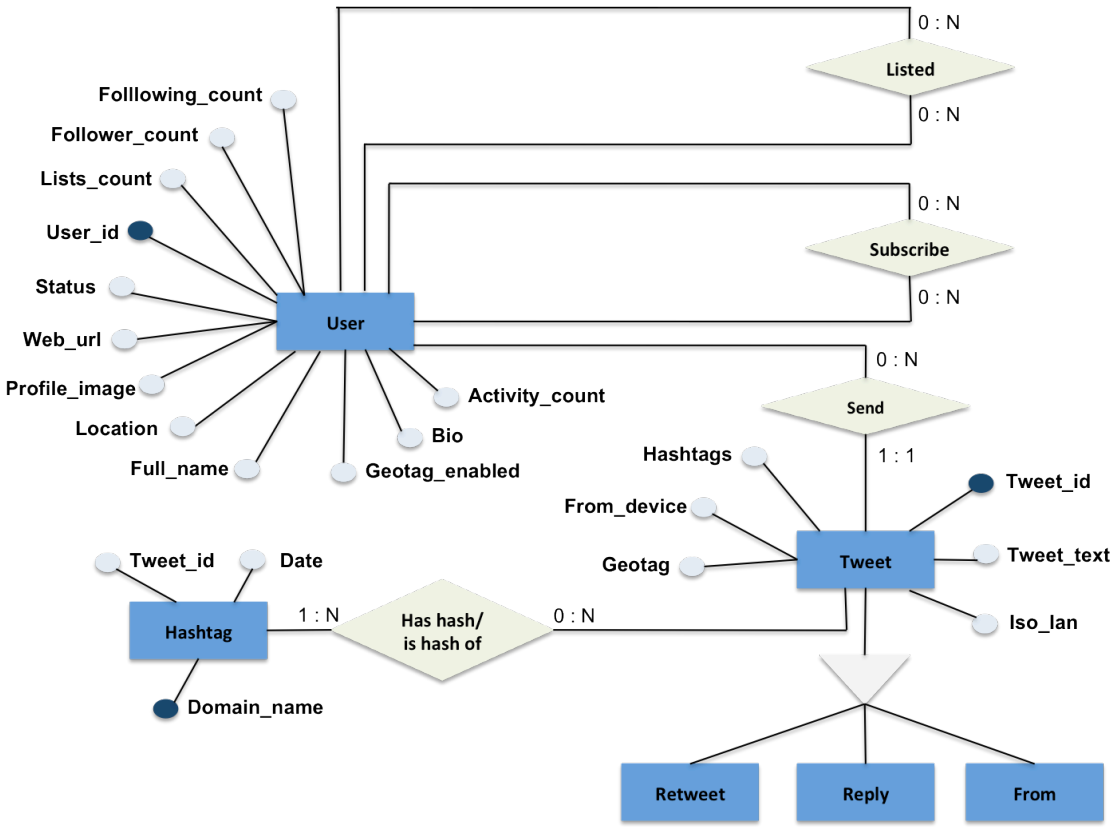


Figura 22 - Diagramma ER di Twitter

Dalla Figura 22 è possibile vedere che le *entità* fondamentali sono l'utente (i.e. *user*), il contenuto (i.e. *tweet*) con le relative *entità* figlie che definiscono le tipologie di post ed infine l'*hashtag*, che come abbiamo visto è la sintassi utilizzata specificare i topic. La relazione *subscribe* indica la sottoscrizione di utente ai contenuti altrui (i.e. *following*) e l'azione opposta (i.e. *follower*). In questo modo non c'è la necessità di specificare *user* in ulteriori sottoclassi, avendo inoltre a disposizione tra gli attributi il *count* di *following* e *follower*. La relazione *listed* ha lo stesso significato del *subscribe*, ma si riferisce all'inclusione degli utenti all'interno delle liste. L'*entità tweet* invece, è meglio specificata dalle sue *entità* figlie, che come detto defiscono le tipologie chiave di messaggio, descritte in dettaglio nella Sezione 3.3. I legami gerarchici "padre-figlio" possono essere di più tipi, dati dalla differenti combinazioni di due elementi determinati: l'appartenenza o meno degli oggetti dell'*entità* padre alle sottoclassi (i.e. figli),

che distinguono rispettivamente un legame *totale* da uno *parziale*; la condivisione o meno di oggetti tra le *entità* figlie, rispettivamente gerarchia *sovrapposta* oppure *esclusiva.* Nel caso in esame ci troviamo di fronte ad un tipo di gerarchia *totale*, in quanto ogni oggetto della superclasse appartiene ad una delle sottoclassi, ed *esclusiva*, dato che le *entità* figlie non hanno oggetti in comune.

4.5.4 La Data visualization

La *data visualization* è sostanzialmente lo studio della rappresentazione dei dati, il cui scopo principale è la comunicazione chiara ed efficiente dell'informazione attraverso mezzi grafici. Per ottenere i migliori risultati possibili, la parte estetica (i.e. *look-and-feel*) e quella relativa alle funzionalità necessitano di andare di pari passo, facilitando la comprensione di un data set potenzialmente molto complesso, riuscendo a comunicarne gli aspetti e i fattori chiave in modo intuitivo. Nel mondo Web stanno nascendo tante tipologie di rappresentazione dei dati, ma non esiste ancora un vero e proprio standard, la scelta dipende prevalentemente dalle informazioni che devono essere mostrate e dall'uso che se ne vuole fare.

Un esempio è dato dalla Figura 23[29], in cui la visualizzazione è in grado di rappresentare dinamicamente ed in modo molto efficace e comprensibile le conversazioni degli utenti Twitter su un topic predefinitio. I contenuti, indicati dalla *profile pic* di ciascun utente che ha partecipato alla discussione, sono impilati verticalmente e organizzati orizzontalmente rispetto al tempo. La centralità di un utente è in base all'attenzione, determinata dalle *mention* e *retweet* che questo riceve.

Altri esempi molto interessanti di visualizzazione sempre riguardanti Twitter sono Social Collider[30], Twitt3D[31], TweetWheel[32], Mention Map[33] (Revisit: visualizing the temporal dynamics of Twitter, 2010).

[29] http://moritz.stefaner.eu/projects/revisit/
[30] http://social collider.net
[31] http://www.twitt3d.com
[32] http://tweetwheel.com
[33] http://apps.asteriqs.com/mentionmap

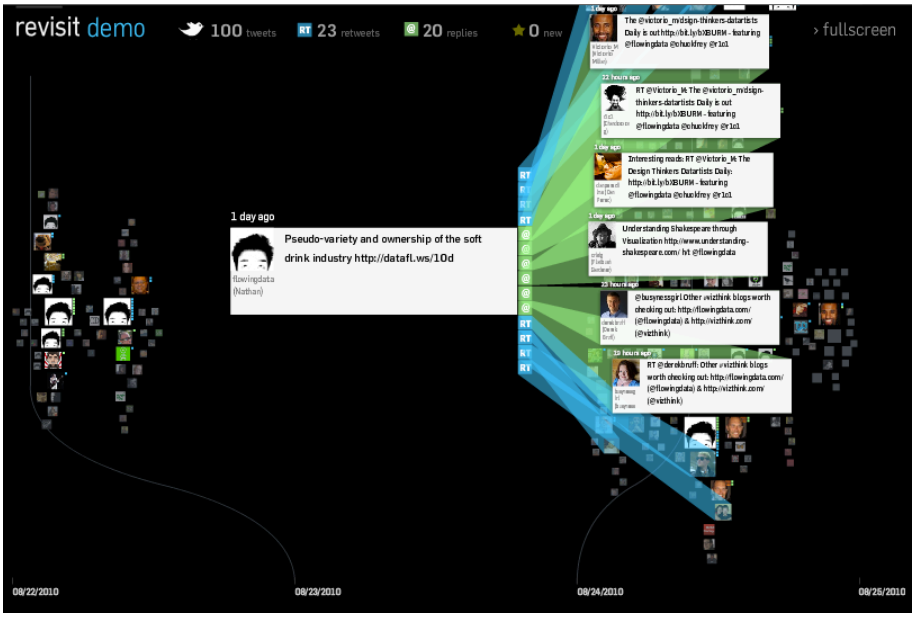


Figura 23 – Schermata del tool Revisit

Dal punto di vista tecnologico bisogna innanzitutto applicare le basi del pattern architetturale *Model View Controller* (MVC), che mira a tenere separata la parte di modellizzazione, i.e. la base di dati, dalla parte logica, i.e. il controller, quale le classi Java o C++ che richiamano il database ed eseguono algoritmi di analisi, e dalla parte di visualizzazione.

Un altro aspetto tra i più rilevanti riguarda la velocità di trasmissione dei dati alla parte grafica, tanto più importante tanto maggiore è la dimensione del data set, con l'obiettivo di non perdere il controllo delle informazioni estratte. E' utile quindi tendenzialmente usare una logica a servizi, in cui solo nel momento del bisogno una rappresentazione fa richiesta di dati a un server, che con le sue elevate capacità di elaborazione permette di alleggerire il carico del client di visualizzazione, che avrà quindi il compito esclusivo di mostrare i dati già filtrati e organizzati in base alle richieste.

Ogni architettura ha i proprio vantaggi e svantaggi. L'introduzione di diversi layer, i.e. livelli software indipendenti, permette una maggior modularità e

possibilità di personalizzazione, ma il tutto avendo un sistema con un'ottima scalabilità, capace comunque di supportare ampie quantità di dati.

Lato client ciò che è richiesto è solo una buona connessione, dato che ormai ogni applicazione viaggia attraverso un browser e di conseguenza deve necessariamente essere leggero.

Applicazioni pratiche che sfruttano l'approccio a servizi sono i *mashup* (letteralmente "poltiglia"), in grado di includere dinamicamente informazioni e contenuti provenienti da più fonti.

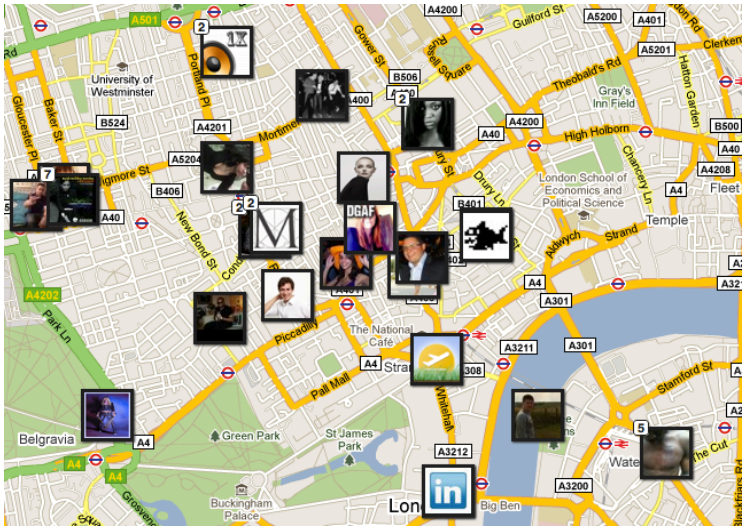


Figura 24 – Schermata del tool Streamdin

La Figura 24 mostra un esempio di *mashup*, nel quale sfruttando le API di Twitter e Google Maps è possibile rappresentare dei contenuti geolocalizzati creati sul microblogging su una mappa geografica.

4.6 Il progetto per il comune di Milano

Non sono unicamente i brand e le aziende private ad essersi accorte della rilevanza dei dati e contenuti create dagli utenti dei servizi Web 2.0, i quali rappresentano, come detto, una fonte di estremo interesse per effettuare analisi di intelligence sociale, ma una sempre più crescente attenzione viene rivolta anche da parte delle istituzioni pubbliche, in particolar modo se l'ambito di analisi è quello del turismo. Il progetto pilota realizzato dal Politecnico e CommStrategy per il comune di Milano si inserisce esattamente all'interno degli aspetti di *social media analysi*s finora trattati, con un preciso orientamento al monitoraggio e alla valutazione della reputation online del *city brand* Milano.

I servizi *social* selezionati come fonti da cui estrarre informazione sono stati molteplici, per la precisione due verticali sul settore travel, e cioè Tripadvisor, la più estesa Web community sul turismo con oltre 40 milioni di utenti registrati, e Lonely Planet, combinati con la piattaforma di microblogging Twitter e Facebook, introdotto nel secondo semestre di lavoro.

In fase di definizione del progetto è stato sviluppato un complesso modello di *city branding*, mutuato da quello realizzato da Anholt[34], e composto di fatto dai cinque fattori competitivi largamente accettati come caratterizzanti il vissuto e l'attrattività di un centro urbano: *places, pulse, people, presence* e *prerequisites.*

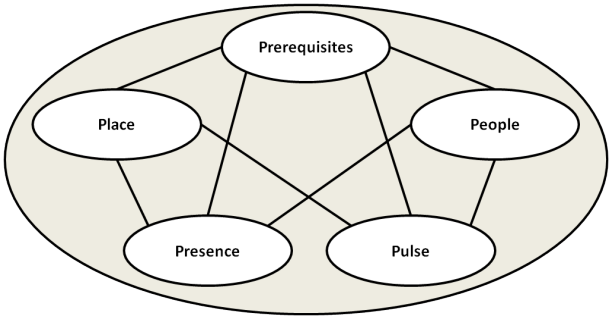


Figura 25 - I cinque fattori competitivi secondo il modello di Anholt

[34] Anholt, S; *Places: Identity, image and reputation*. Palgrave Macmillan, December 2009

Questi fattori rappresentano il punto di partenza di una struttura ad albero che si delinea attraverso nove macro categorie – *arts&culture, events&sports, fashion&shopping, weather&environmental, life&entertainment, night&music, ticket, food&drink* e *services&transports* – andando a costituire una griglia di label a tre livelli che comprende oltre cento domini. L'applicazione di una categorizzazione che aiuti a catalogare i contenuti in vari cluster conversazionali simili tra di loro è indispensabile per ottenere dai dati grezzi un significativo incremento della qualità dei dati stessi. La suddivisione per tag, nel caso specifico con un modello assolutamente verticale sul turismo, consente inoltre di navigare attraverso i messaggi senza scontrarsi ogni volta con una granularità troppo fine e di avere una classificazione ad elevata precisione per ogni post estratto dalle fonti.

La Figura 26 di seguito mostra un sezione del modello, indicando alcune sotto-categorie dei driver *food&drink* e *weather&environmental*, entrambe appartenenti al fattore *place.*

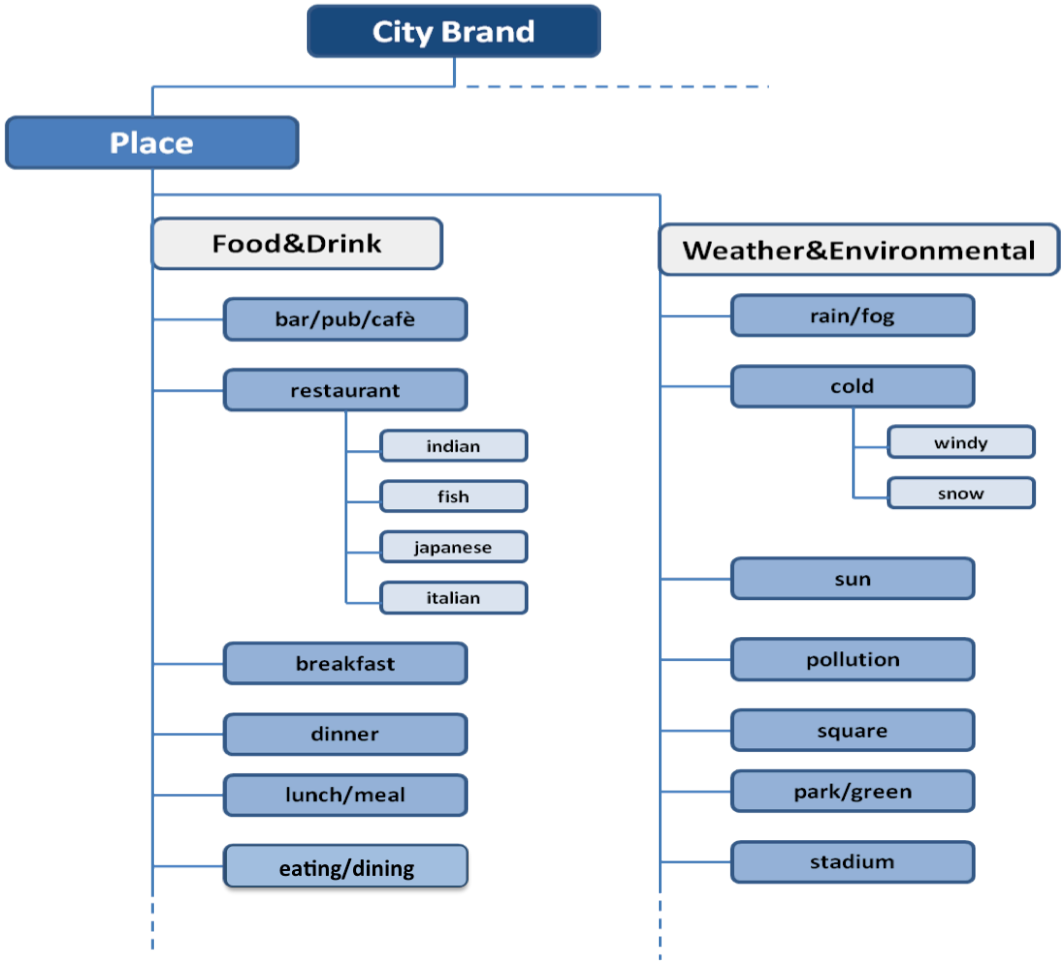


Figura 26 – Sezione dell'alberatura con alcune label

Con l'obiettivo di permettere analisi comparative, il flusso di estratti non è esclusivo alla città di Milano, ma anche a quelle di Londra, tenuta in considerazione come benchmark, Berlino, introdotta nel secondo semestre di progetto, e Madrid, quest'ultima selezionata per la sua forte somiglianza in termini di caratteristiche al capoluogo lombardo.

La gestione operativa del processo del lavoro è suddivisa in tre passaggi, esattamente come visto e descritto in precedenza nelle Sezioni 4.2 e 4.3.

Se per quanto riguarda il *crawling* non ci sono aggiunte significative rispetto a quanto già detto in precedenza, è invece necessario specificare espressamente la fase di *processing*, suddividendo e descrivendo al meglio i due aspetti metodologici e tecnologici differenti che la compongono: la categorizzazione dei

contenuti (i.e. *tagging*) svolta da un motore semantico e l'analisi del sentiment (i.e. *scoring*), anche questa automatizzata.
Una volta estratto il flusso di dati non strutturati, per ora processati esclusivamente in lingua inglese, solo una percentuale di questo verrà categorizzato all'interno del modello di reputation impiegato. Ovviamente il primo filtro applicato è quello relativo alla disambiguazione del nome delle città, per il quale la semantica, intervenendo con opportune tecniche, è in grado di risolvere le problematiche di polisemia, in modo che i post che contengono "*Milan*" si riferiscano esclusivamente alla città e non per esempio all'omonima squadra di calcio o all'attrice Alyssa Milano, tra l'altro molto popolare su Twitter.

Questa procedimento, a differenza di quella di costruzione del modello (i.e. *mapping)* che ha richiesto necessariamente un intervento *analyst-intensive,* viene svolto assolutamente in automatico da un motore semantico.

L'enorme progresso tecnologico degli ultimi anni viene incontro all'esigenza di analizzare contenuti prettamente *user-generated*, permettendo di ricavare gli elementi essenziali per la ricostruzione del brand partendo dall'analisi soggetto-azione-oggetto, soggettività-oggettività di una frase, analisi dei concetti (i.e. persone e luoghi) più importanti presenti nei testi. Tutte e tre queste funzionalità sono conseguenze di un *parsing* sintattico e semantico delle frasi. Grazie a degli alberi delle dipendenze (*dependency tree*), è possibile mostrare i concetti del discorso e le relazioni che intercorrono tra questi con una certa sicurezza, e nonostante la tecnica non sia esente da errori, essa garantisce risultati con un'accuratezza notevole ed è di aiuto a chi ha il compito di costruire una mappa di tag, in quanto grazie all'identificazione dei topic di cui si sta principalmente parlando, si ottiene un modello di brand guidato dagli utenti (i.e. i veri interessati) e non staticamente generato da istituzioni pubbliche e/o private.

Una volta categorizzati i contenuti, il passo immediatamente successivo riguarda lo *scoring* del sentiment, i.e. l'attribuzione di un valore di sentiment ai messaggi in cui viene registrata una polarizzazione.

Benchè tra i due metodi ci sia un grosso margine di dati inutilizzati, questi sono comunque ritenibili in qualche modo complementari: se con il *tagging* si riesce a

raggruppare per similarità, con lo *scoring* viene suddivisa l'informazione che non porta significato da quella che ha valore.

La *sentiment analysis* è una materia ancora in fase di perfezionamento, che garantisce risultati performanti ma non ancora il 100% in termini *precision* e *recall;* se poi si restringe il suo campo d'azione su categorie di dettaglio, la bontà del risultato tende a perdere qualche punto percentuale.

Questo step, relativamente al quale le ricerche dimostrano che la percentuale di testi che hanno una polarità si attesta su meno di un decimo del totale, include in aggiunta una serie di elementi di distorsione che è necessario tenere bene in considerazione, un esempio dei quali è la differenza tra la neutralità e la mancanza di opinione, differenziazione necessaria ma che la maggior parte dei tool così come dei consulenti generalmente tralascia, e la forte tendenza a parlare positivamente di un avvenimento, oggetto o evento, per cui è essenziale studiare la reputazione della fonte per valutare che peso dare a ogni sito preso in analisi e per riequilibrare il divario dovuto all'eccesso solitamente di commenti positivi.

Dal punto di vista tecnologico, l'analisi può essere affrontata a più livelli: documento, frase e *snippet*[35]. Il primo è la cosiddetta classificazione a livello di documento, l'assegnazione di un voto positivo o negativo ad un testo. Questo è stato il punto di partenza, da cui poi si è capita la necessità di volere estrarre informazioni a granularità più fine rispetto all'intero documento, arrivando alla classificazione a livello di frase, che ottiene valori più dettagliati, ma a discapito di maggiori costi. Più precisamente per determinare la soggettività/oggettività sono usate tecniche quali la somiglianza delle frasi e/o i classificatori bayesiani singoli e multipli. La polarità di una frase soggettiva invece è determinata basandosi su una lista di parole o sensi (in caso di analisi semantiche) aventi associati già un sentimento a priori. La classificazione a livello di *snippet* è la terza granularità, la più fine, per il calcolo del sentiment. A livello di documento si fa generalmente un'assunzione, che ci sia solo un valore di polarità per tutto il testo, quindi uno e un solo utente che descriva positivamente o negativamente un solo oggetto, mentre lo *snippet* punta a catturare tutti contenuti, distinguendo

[35] http://en.wikipedia.org/wiki/Snippet_(programming)

all'interno delle stesse frasi più sezioni, ognuna delle quali è composta da un'unica informazione specifica.

Nella Figura 27 è presentata l'architettura complessiva del tool.

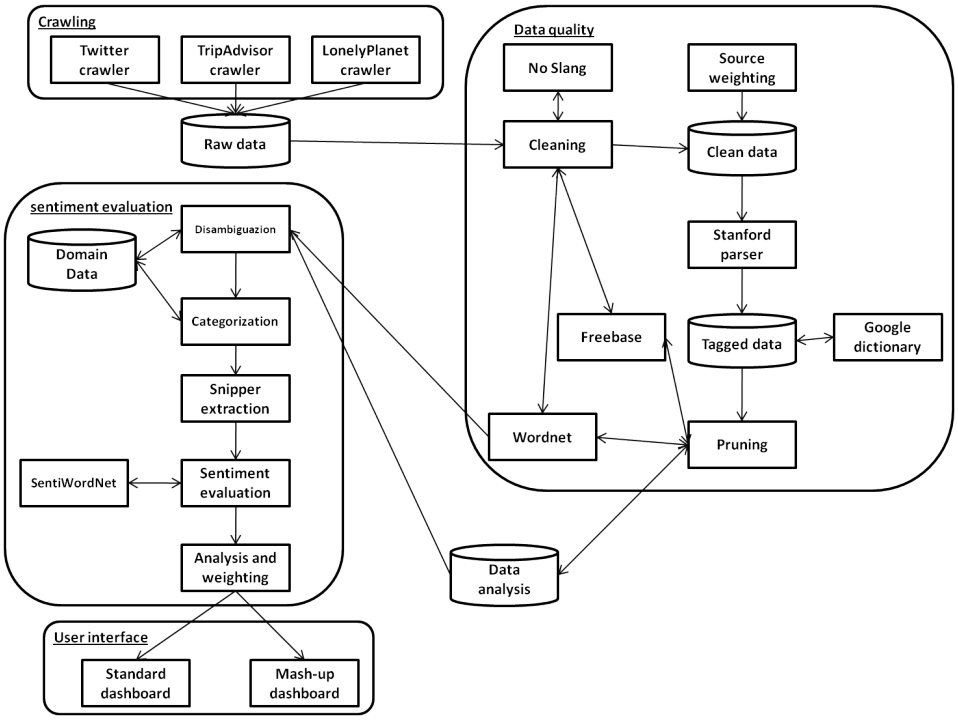


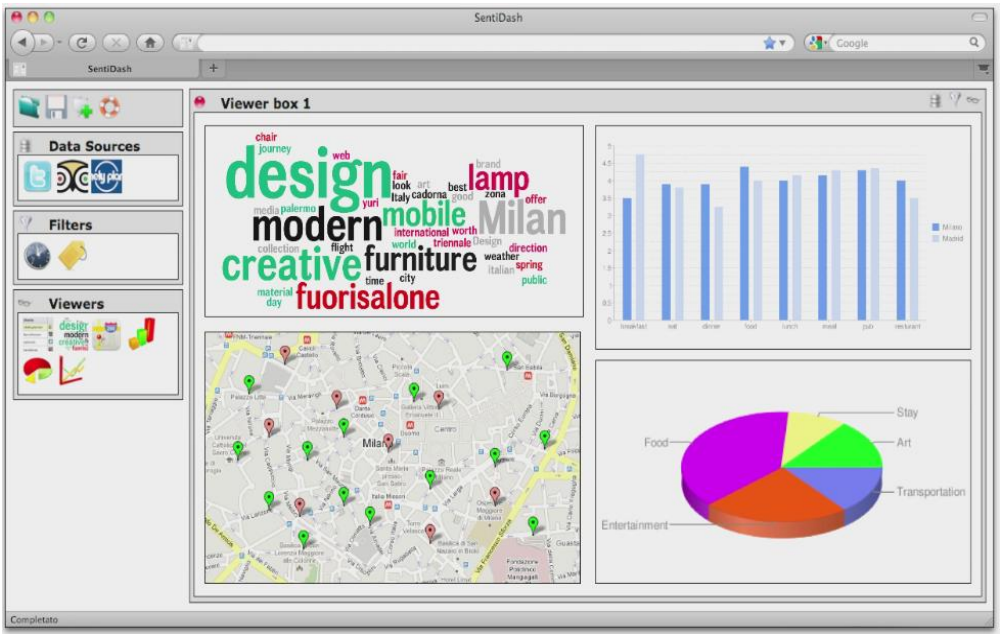
Figura 27 - Architettura del tool del progetto

Il tool sviluppato è, almeno per il momento, un'applicazione accessibile in locale il cui database contenente i dati viene progressivamente aggiornato con l'estrazione e analisi di contenuti freschi.

Le funzioni principali riguardano l'analisi dei volumi delle categorie della mappatura rispetto a range di tempo impostati dall'utente, in modo da consentire un'analisi dei trend; l'andamento del sentiment, a più livelli, rispetto al tempo e per differenti label; la distribuzione delle categorie rispetto al totale dei post con setiment positivi o negativi; la tagcloud per descrivere le associazioni dei contenuti all'interno di un subset di dati.

L'interfaccia utente, così come l'architettura, costituisce un elemento di assoluta innovatività. L'applicazione delle tecnologie *mashup,* che consentirà via via l'introduzione di funzionalità aggiuntive con relativa semplicità, e la gestione

(i.e. caricamento e impostazioni delle funzioni e dei grafici) di tutte le componenti in *drag&drop*, rendono da un lato molto piacevole il *look-and-feel*, ma soprattutto favoriscono notevolmente l'usabilità e la facilità di interazione con lo strumento e le sue funzionalità. Nella Figura 28 è proposto un screenshot tratto dal prototipo dell'interfaccia.



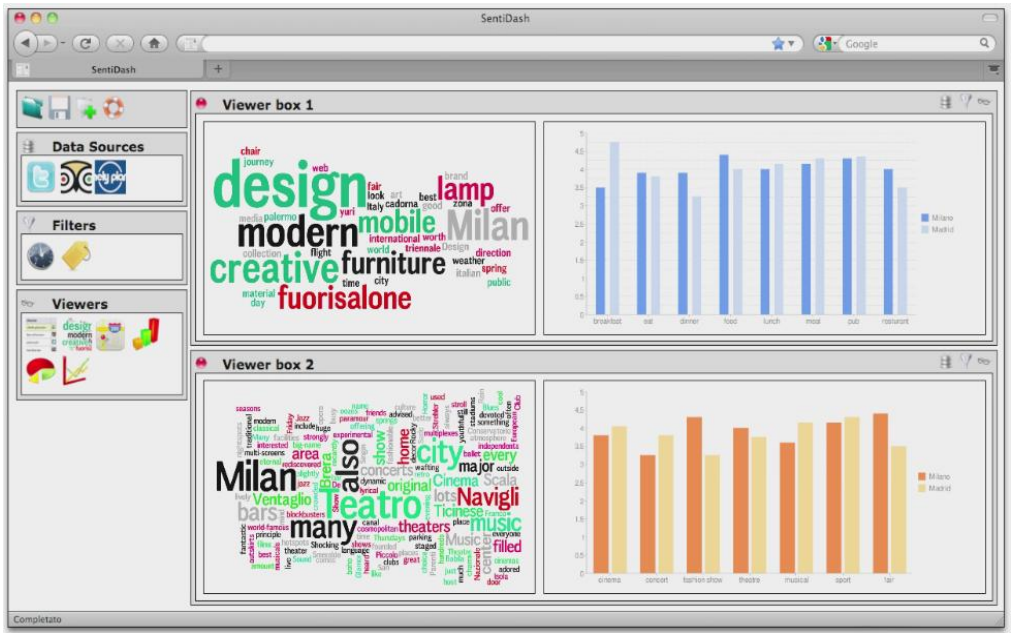


Figura 28 - Interfaccia prototipo del tool

Il primo semestre di progetto è stato impiegato, per quanto riguarda gli elementi tecnologici, ad un miglioramento delle prestazioni sia per quanto concerne il *crawling* dei dati, con lo scopo di migliorare la *recall* del tool, che, ed in particolar modo, per la fase di *processing*, al fine di aumentare la *precision* dei risultati ottenuti dalla categorizzazione e dall'analisi del sentiment.

Nel periodo Marzo-Giugno sono stati raccolti circa un milione e 100 mila post, di cui circa 500 mila inclusi attraverso il tagging nel modello ad albero. Nel caso di Milano, nonostante sia il city brand che mostra le maggior problematiche a livello di disambiguazione, è stato possibile raggiungere una copertura praticamente totale rispetto alle conversazioni che citano effettivamente la città, e cioè i contenuti significativi per l'e-traveller. Per Londra, in cui l'utilizzo del microblogging è circa 15 volte superiore in termini di volumi, la copertura si aggira al momento intorno al 5%, campione secondo la letteratura comunque sufficientemente significativo.

Il lavoro *analyst-intensive* di consulenza e gestione del cliente svolto principalmente da CommStrategy si è dedicato invece ad un monitoraggio costante dei flussi, alla distribuzione dei volumi rispetto ai driver del modello, valutando trend e cause di eventuali picchi. Lo stesso approccio è stato svolto a livello di sentiment, con un controllo costante dei volumi dei post, dell'accuratezza dei risultati automatizzati restituiti dal tool, individuando e verificando i picchi all'interno degli andamenti.

Ulteriore ambito del servizio di consulenza è stata un'analisi qualitativa e di alto livello degli influencer, individuando, in particolare su Twitter, gli utenti più popolari localizzati sulle città incluse, collocandoli all'interno delle categorie mappate.

Sono stati inoltre affrontati differenti casi di studio, riguardanti soprattutto la città di Milano e ad alcuni eventi avvenuti, monitorando ad-hoc per ciascuno di questi l'andamento del volume dei messaggi, valutando il sentiment e aggiungendo altri aspetti e metriche di analisi sia di tipo qualitativo che quantitativo. Esempi sono il Salone Internazionale del Mobile ed il Fuori Salone svoltosi ad Aprile; il blocco aereo ed i disagi trasporti causati dalle polveri del

vulcano Islandese che ha portato a più di mezzo milione di tweet in soli 9 giorni e la gestione della crisi da parte di Milano e Londra; lo show per il ventennale del noto marchio del fashion Dolce&Gabbana tenutosi a Giugno al Palazzo della Scala.

Capitolo 5

Analisi e risultati

5.1 Introduzione

L'obiettivo della tesi è lo studio e l'analisi degli indicatori che possano contribuire alla valutazione della opinion leadership e alla individuazione di utenti riconosciuti come autorevoli e influenti all'interno del network di Twitter.

Il capitolo è strutturato come segue. La Sezione 5.2 espone le caratteristiche del data set di utenti impiegato; nella Sezione 5.3 vengono descritti gli indicatori e le metriche per la valutazione della influence e i risultati di una prima serie di analisi statistiche; nella Sezione 5.4 viene presentata una possibile categorizzazione degli account del set e le ulteriori analisi svolte partendo da quella; infine nella Sezione 5.5 vengano dettagliati i risultati ottenuti e indicate alcune linee guida per la valutazione dell'autorità nel servizio di microblogging.

5.2 Il data set di utenti

Come descritto nello Stato dell'arte, la letteratura ha ampiamente trattato e studiato le proprietà topologiche delle reti sociali, sia online che offline. Uno dei risultati più comuni e condivisi è quello secondo il quale la ripartizione dell'autorità tra i nodi, i.e. gli utenti, comunemente misurata attraverso il numero di connessioni o identificandone la posizione all'interno del network, segue una distribuzione Paretiana, mostrando come alcuni utenti di fatto posseggano una maggior rilevanza all'interno della rete, tale da far sì che vengano identificati con il nome di *hubs* (Sarshar, Boykin, & Roychowdhury, 2004; Adamic, Lukose, Puniyani, & Huberman, 2001).

Recentemente alcuni lavori hanno posto dei dubbi sulla validità di queste considerazioni e risultati, specialmente nel caso ci si trovi di fronte a reti sociali geograficamente localizzate (Sala, Zheng, Zhao, Gaito, & Rossi, 2010; Naruse & Kubo, 2006; Askira Gelman & Barletta, 2008).

Gli utenti appartenenti al data set sono stati selezionati con un preciso criterio di collocazione geografica, includendo nelle analisi esclusivamente quelli che presentassero come *location* nelle informazioni di profilo la città di Londra. Utilizzando il noto servizio di ranking e analytics per Twitter, Twitaholic[36], è stata prelevata la lista dei 1000 utenti "londinesi" considerati più popolari, i.e. con maggior numero di *follower*, e quindi potenzialmente influenti. Al fine di poter ottenere i dati utili relativi agli utenti, è stato sviluppato un crawler ad hoc in grado di gestire il data set e il totale delle richieste necessarie al server di Twitter, sfruttandone le API e rispettando i limiti imposti, inseriti nella Sezione 4.5.2.

In seguito al *crawling* dei dati, eseguito in maniera continuativa per un arco di tempo pari a 30 giorni, dal 01/06 al 30/06, è stata svolta una pulitura del set, escludendo gli account inattivi nel periodo indicato, i.e. nel caso non avessero postato nessun messaggio, e eventuali account sospesi, situazione in cui solitamente ricadono i bot generatori automatici di spam. Sempre durante l'arco di riferimento è stata registra in maniera semi-automatica l'attività, i.e. numero

[36] http://twitaholic.com

effettivo di *tweet*, compiuta dagli utenti, grazie all'utilizzo del tool di analytics TweetStats[37].

L'estrazione si è focalizzata su *mention* e *retweet* ricevuti da ciascuno degli account inclusi nel sample, ottenendo attraverso le API tutti quei messaggi che presentassero al loro interno le sintassi[38]:

- *"[messaggio] @username [messaggio]" – mention*;
- *"[messaggio] RT @username [messaggio]"* oppure

  "[messaggio] via @username [messaggio]" – retweet;

modalità di comunicazione specifiche del microblogging e spiegate in dettaglio nella Sezione 3.2.

La Tabella XVI mostra un estratto del set completo con gli indicatori raccolti e a disposizione per le analisi.

Name	Nickname	Follower	Following	Activity	Retweet	Mention
Guardian News	@guardiannews	67.208	964	1.266	5.036	1.854
Fearne Cotton	@fearnecotton	845.541	121	213	538	11.247
Mtv Uk	@mtvuk	59.627	7.289	677	988	1.376
…						

Tabella XVI - Estratto di esempio del data set

[37] http://tweetstats.com

[38] come spiegato nella Sezione 3.2, la parte identificata con *[messaggio]* può essere presente sia all'inizio che alla fine del *tweet*, oppure esclusivamente in una delle due posizioni.

In seguito alla pulitura secondo i criteri descritti, il data set finale è risultato composto da un totale di 837.401 messaggi crawlati e 813 utenti , i quali, come mostra la Tabella XVII, posseggono un certo grado di eterogeneità, con il minimo valore per *mention* e *retweet* che può essere pari a 0, e con una differenza di circa quattro ordini di grandezza tra l'utente più connesso, i.e. con maggior numero di *follower*, e quello meno.

	Minimum	Maximum	Mean	Std. Deviation
follower	129	2.833.657	32.142,33	162.739,228
mention	0	83.853	830,16	3.952,191
retweet	0	11.520	119,85	736,121
activity	1	4.870	275,43	393,530

Tabella XVII - Statistiche descrittive del data set

5.3 I paramentri della influence

Così come descritto dalla letteratura precedente e come appreso dalla practice dei tool di *social media analysis* menzionati nel Capitolo 4, sono state prese in considerazione le metriche di influence diffusamente ritenute come le più valide su Twitter, i.e. il numero di *follower*, il numero di *retweet* che un utente riceve e il numero di volte che questo viene menzionato nei messaggi postati da altri, i.e. numero di *mention*.

La misurazione dell'attività, che compare dalla Tabella XVII delle statistiche descrittive, è stata effettuata con il preciso scopo di calcolare le metriche relative che misurano l'efficienza dello sforzo, i.e. il numero relativo delle *mention* e dei *retweet* rispetto al totale dei *tweet* postati da ciascun user, identificati rispettivamente con il nome di *mention_rate* e *retweet_rate*.

Di seguito è riassunto il set di metriche utilizzate:

Volume	Efficiency
follower = # of follower	N/A
mention = # of mention received	mention_rate = $\frac{mention}{activity}$
retweet = # of retweet of own tweet	RT_rate = $\frac{RT}{activity}$
activity = # of messages tweeted	N/A

Tabella XVIII - Metriche per la misura della influence in Twitter

Successivamente, attraverso un controllo manuale di tutti gli utenti, è stata effettuata una categorizzazione che rispecchiasse delle precise tipologie di utilizzo del servizio, nello specifico suddividendo gli account secondo *brand*, *news* e *people*, identificati ed intesi così come descritto in seguito:

- *brand* – Twitter viene sostanzialmente impiegato come un canale di promozione e di diffusione di aggiornamenti relativi alle attività (e.g. di business) in cui l'account è coinvolto, e.g. Mtv Uk, *@mtvuk*;

- *news* – l'attività è focalizzata alla pubblicazione e diffusione di news, si tratta quindi di fonti di informazione sia online che offline, come quotidiani, riviste, magazine e blog, e.g. The Guardian, *@guardiannews*;

- *people* – l'account è gestito da individui che comunicano principalmente fatti ed esperienze personali, e.g. Fearne Cotton, *@fearnecotton*

A differenza di altre tassonomie illustrate alla Sezione 3.6, si è deciso di adottare una categorizzazione che non fosse troppo fine e che si potesse adattare sufficientemente bene alla dimensione del sample utilizzato. E' importante menzionare il fatto che, vista la natura intrinseca del *blogging* come mezzo di auto-promozione, le cosiddette *celebrity* presenti non sono state incluse a priori nella categoria *brand*, ma nel momento in cui la loro attività comprendesse

soprattutto aspetti e fatti relativi alla vita personale e non solo professionale, si è deciso di etichettarli come *people.*

Studi recenti non considerano nel caso specifico di Twitter il numero di *follower* di per sé significativo e capace di ben rappresentare l'autorità di un utente, in quanto metrica scarsamente correlata con il numero di *retweet* e i *mention,* così come mostrato da (Cha, Haddadi, Benvenuto, & Gummadi, 2010). Quello che però manca al lavoro appena citato è una correlazione che includa anche l'efficienza della influence, intesa e misurata rispetto allo sforzo effettuato, i.e. al numero di *tweet* postati.

La ripartizione del numero di *follower* degli utenti appartenenti al data set è ben rappresentata con una distribuzione Paretiana, come mostra chiaramente la relazione lineare nel grafico log-log in Figura 29.

Per contro, dalla stessa analisi svolta rispetto alla distribuzione di *mention* e *retweet*, rispettivamente nei grafici in Figura 30 e 31, emerge una situazione differente: la relazione non è più lineare, ma è possibile individuare un punto di rottura all'incirca intorno alla posizione 300 del rank.

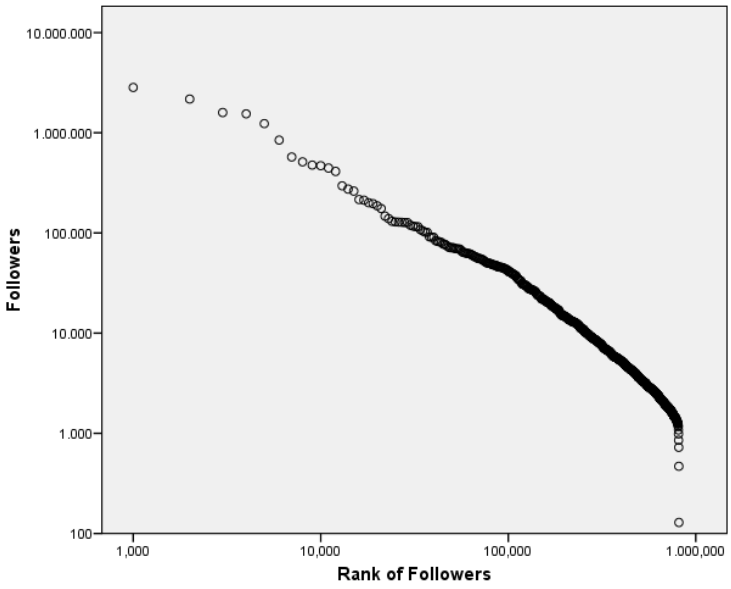


Figura 29 - Grafico log-log di distribuzione dei follower

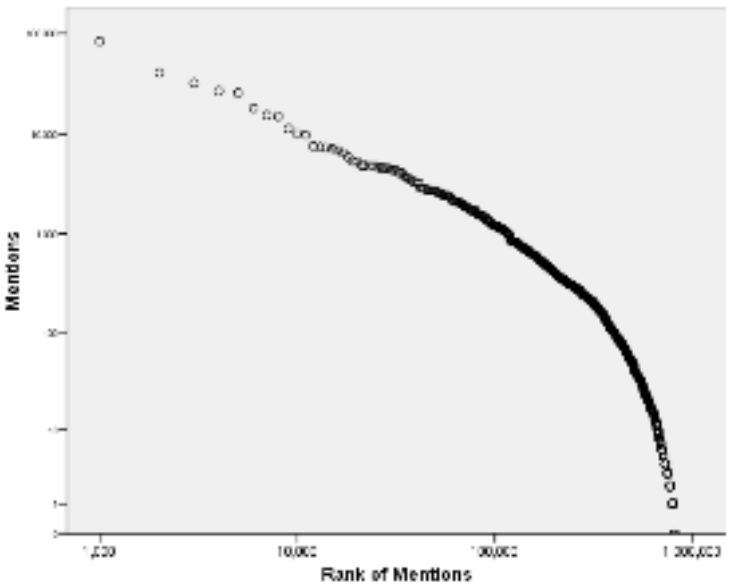


Figura 30 - Grafico log-log di distribuzione delle mention

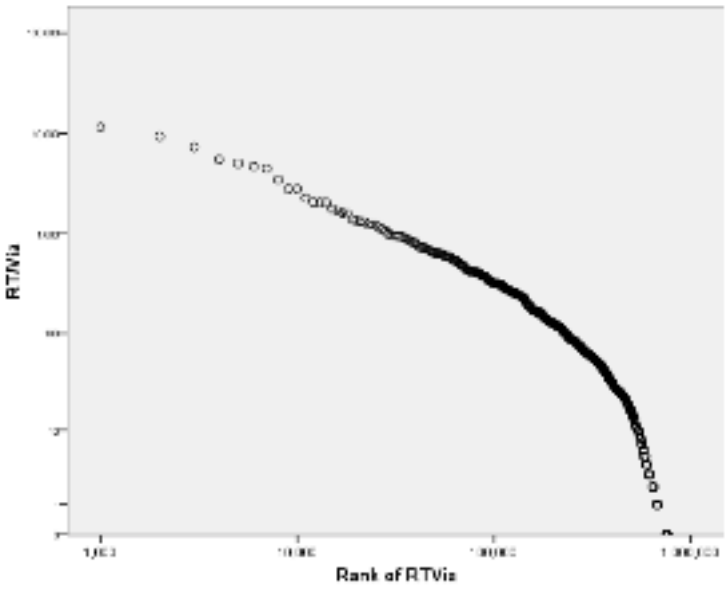


Figura 31 - Grafico log-log di distribuzione dei retweet

Un'analisi lognormale Q-Q plot effettuata sui dati invece, rappresentata in Figura 32 e 33, mostra chiaramente come la distribuzione di *retweet* e *mention* sia senza dubbio più simile ad una distribuzione lognormale.

Esattamente per questa ragione il data set è stato suddiviso in due cluster: il primo, che include gli utenti che hanno una posizione del rank più alta del trecentesimo posto basandosi su entrambe le metriche, i.e. *mention e retweet*, mentre l'altro, il cluster B, composto dal sub set rimanente.

Nel momento in cui le variabili non sono distribuite seguendo una normale, non c'è la possibilità di effettuare il test di correlazione di Pearson, rendendo quindi necessario l'utilizzo del test di Spearman su i due cluster in maniera separata, i cui risultati sono presentati nella Tabella XIX.

Le Tabelle XX e XXI riportano invece rispettivamente le statistiche descrittive per i cluster A e B e mostrano i valori ottenuti per un t-test indipendente per l'uguaglianza media tra i due cluster.

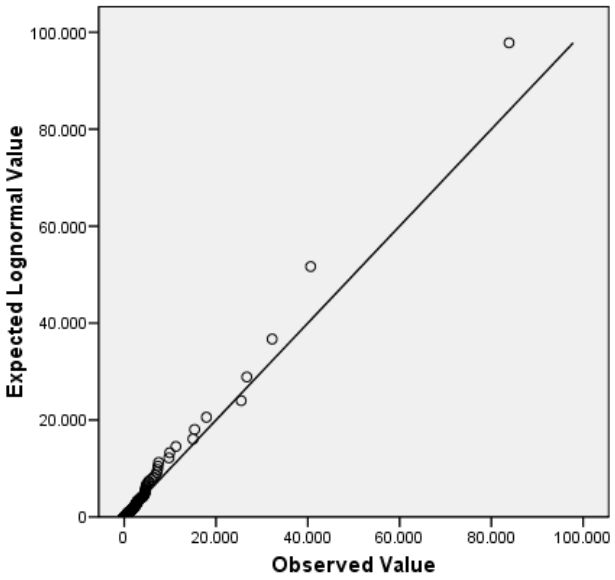


Figura 32 - Grafico lognormale Q-Q di distribuzione delle mention

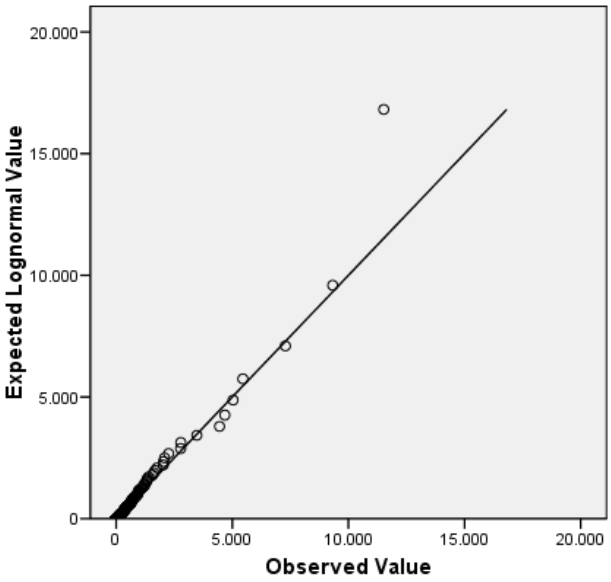


Figura 33 - Grafico lognormale Q-Q di distribuzione dei retweet

			follower	mention_rate	RT_rate
CLUSTER A	follower	Correlation coefficient	1,000	.557 (**)	.568 (**)
		Sig. (2-tailed)	.	0,000	0,000
		N	222	222	222
	mention_rate	Correlation coefficient	.577 (**)	1,000	.635 (**)
		Sig. (2-tailed)	0,000	.	0,000
		N	222	222	222
	RT_rate	Correlation coefficient	.568 (**)	635 (**)	1,000
		Sig. (2-tailed)	0,000	0,000	.
		N	222	222	222
CLUSTER B	follower	Correlation coefficient	1,000	.178 (**)	0,097 (*)
		Sig. (2-tailed)	.	0,000	0,020
		N	591	583	584
	mention_rate	Correlation coefficient	.178 (**)	1,000	.328 (**)
		Sig. (2-tailed)	0,012	.	0,000
		N	583	583	583
	RT_rate	Correlation coefficient	0,097 (*)	.328 (**)	1,000
		Sig. (2-tailed)	0,020	0,000	.
		N	584	583	584

**. correlazione è significativa allo 0,01 (2-tailed)

*. correlazione è significativa allo 0,05 (2-tailed)

Tabella XIX - Indici di correlazione di Spearman per i cluster A e B

I risultati mostrano chiaramente che il cluster A presenta valori superiori, quindi performance decisamente migliori, sia per quanto concerne il *mention_rate* che per il *reweet_rate* rispetto al cluster B, nonostante gli utenti inclusi in quest'ultimo abbiano mantenuto un'attività maggiore, potendo quindi concludere che il cluster A è dotato sicuramente di una efficienza più elevata nel processo di diffusione dell'informazione, facendo inoltre opportunamente notare che la media dei *follower* per il cluster A è significativamente superiore a quella del cluster B.

		Minimum	Maximum	Mean	Std. Deviation
A	follower	1348	2.833.657	86.667,18	291.183,740
	mention	223	83.853	2.702,93	7.225,477
	retweet	62	11.520	589,67	1.170,008
	activity	3	4.870	494,36	554,236
B	follower	129	1.234.334	11.660,91	56.132,375
	mention	0	5.862	126,68	354,973
	retweet	0	9.321	53,43	392,869
	activity	1	1.835	191,92	268,136

Tabella XX - Statistiche descrittive per i cluster A e B

	t	df	Sig. (2-tailed)	Mean difference	Std. error difference
follower	5.979	811	0.000	75,006.266	12,544.925
mention_rate	6.480	803	0.000	22.97481	3.54534
retweet_rate	9.915	804	0.000	2.51892	0.25405
activity	10.368	802	0.000	302.433	29.169

Tabella XXI - t-test per l'uguaglianza media dei due cluster

5.4 Analisi per categorie

Come detto in precedenza, è stata effettuata una categorizzazione del set introducendo tre possibili tipologie di account, i.e. *brand*, *news* and *people*, applicate in base all'utilizzo del servizio di microblogging.

La Tabella XXII riporta l'analisi ANOVA finalizzata ad una valutazione comparativa delle caratteristiche dei tre gruppi, dalla quale si può notare come la differenza in termini di *follower* non sia significativa, i.e. la distribuzione dei *follower* è abbastanza uniforme all'interno delle tre categorie. Quello che è invece importante considerare è il fatto che c'è una differenza in termini di media di *mention* e *RT*[39], calcolata attraverso le analisi post-hoc svolte con Bonferroni e riportate nella Tabella XXIII per evidenziare in dettaglio i valori.

Emerge chiaramente che le fonti di *news* hanno volumi assoluti di *retweet* molto più elevati rispetto a *people* e *brand*, mentre la differenza considerando questi ultimi non è affatto significativa. In aggiunta, i *brand* mostrano un livello i attività inferiore rispetto alle altre due categorie.

Questi risultati hanno una certa rilevanza in quanto i *retweet* sono sempre stati ritenuti come uno tra gli indicatori più importanti al fine di valutare la influence in Twitter, ma è chiaro che le *news* per loro natura abbiano un vantaggio intrinseco nella capacità di vedere i proprio contenuti riproposti da altri utenti, mettendo in luce la necessità di effettuare una distinzione tra le varie tipologie di utenti al fine di limitare i bias all'interno delle metriche.

D'altro canto, gli account *people* mostrano un valore medio in termini di *mention* superiore alle altre due categorie, dovuto senza dubbio ad una maggiore attività propriamente conversazionale rispetto a *brand* e *news*. Dal punto di vista dell'azione di marketing, come sostenuto da alcuni testi divulgativi dei principali esponenti del social media marketing (Solis, Engage: The complete guide for brands and businesses to build, cultivate and measure success in the new Web, 2010) e da qualche caso di studio (Rao, 2010; Eaton, 2009), la comunicazione *one-to-one* ed il coinvolgimento dei *follower*, effettuabile in Twitter in maniera

[39] RT = *retweet*

efficace attraverso l'uso delle *mention,* sono il miglior mezzo per stimolare l'engagement degli utenti e favorire il successo delle azioni del marketing stesso.

I valori relativi di *retweet* e *mention* invece, non mostrano differenze al variare delle categorie. Questo implica che anche le fonti di *news*, le quali mostrano volumi assoluti più elevati, non sono poi effettivamente in grado di diffondere con la stessa efficacia ogni contenuto che trasmettono, e.g. verosimilmente un numero limitato di notizie sarà in grado di ricevere valori molto elevati di *retweet*, mentre altre non riusciranno a stimolare alcun interesse nei lettori.

		Sum of square	df	Mean sqaure	F	Sig.
follower	Between groups	5.515.831.937	2	2.757.915.968	0.10390511	0,90132283
	Within groups	2.14995E+13	810	26.542.639.382		
	Total	2.15051E+13	812			
mention	Between groups	177.095.092,6	2	88.547.546,29	5.73503861	0,00336337
	Within groups	12.506.195.228	810	15.439.747,2		
	Total	12.683.290.321	812			
retweet	Between groups	11.756.367,73	2	5.882.683,863	11.1269689	1,7093E-05
	Within groups	428.236.475,1	810	528.687,0064		
	Total	440.001.842,9	812			
activity	Between groups	2,544,951,369	2	1.272.475,684	8.36742349	0,00025323
	Within groups	121.812.051.7	801	152.074,9709		
	Total	124.357.003,1	803			
mention_rate	Between groups	8.858,880596	2	4.429,440298	2.09115605	0,12421735
	Within groups	1.698.778,586	802	2.118,177788		
	Total	1.707.637,467	804			
retweet_rate	Between groups	35,62129121	2	17,8106456	1.53260778	0,2166079
	Within groups	9.331.77331	803	11,62113737		
	Total	9.367,394602	805			

Tabella XXII – Analisi ANOVA per il confronto delle medie per le diverse tipologie di account

Dependent variable	(I) Cat	(J) Cat	(I-J) Mean difference	Std. Error	Sig.
follower	brand	news	-4.099,598	17.204,722	1,000
		people	-6.267,591	13.755,944	1,000
	news	brand	4.099,598	17.204,722	1,000
		people	-2.167,994	15.000,217	1,000
	people	brand	6.267,591	13.755,944	1,000
		news	2.167,994	15.000,217	1,000
mention	brand	news	22,805	414,950	1,000
		people	-928.440(*)	331,771	0,016
	news	brand	-22,805	414,950	1,000
		people	-951.244(*)	361,781	0,026
	people	brand	928.440(*)	331,771	0,016
		news	951.244(*)	361,781	0,026
retweet	brand	news	-302.916(*)	76,785	0,000
		people	-0,498	61,393	1,000
	news	brand	302.916(*)	76,785	0,000
		people	302.418(*)	66,946	0,000
	people	brand	0,498	61,393	1,000
		news	-302.418(*)	66,946	0,000
activity	brand	news	-154.791(*)	41,417	0,001
		people	-113.940(*)	33,096	0,002
	news	brand	154.791(*)	41,417	0,001
		people	40,850	36,125	0,775
	people	brand	113.940(*)	33,096	0,002
		news	-40,850	36,125	0,775

*. La differenza media è significativa allo 0,05

Tabella XXIII - Continua alla pagina seguente

Dependent variable	(I) Cat	(J) Cat	(I-J) Mean difference	Std. Error	Sig.
mention_rate	brand	news	4,53296	4,88797	1,000
		people	-3,95711	3,90461	0,933
	news	brand	-4,53296	4,88797	1,000
		people	-8,49007	4,26218	0,140
	people	brand	3,95711	3,90461	0,933
		news	8,49007	4,26218	0,140
retweet_rate	brand	news	-0,20183	0,36166	1,000
		people	0,31237	0,28872	0,839
	news	brand	0,20183	0,36166	1,000
		people	0,51420	0,31570	0,311
	people	brand	-0,31237	0,28872	0,839
		news	-0,51420	0,31570	0,311

*. La differenza media è significativa allo 0,05

Tabella XXIII – Comparazioni multiple con il metodo della correzione di Bonferroni

5.5 Conclusioni

Le analisi effettuate e i risultati ottenuti, alcuni dei quali già esposti, hanno permesso di identificare degli interessanti spunti finalizzati all'implementazione di un modulo di *influencer analysis*.

Uno degli elementi di complessità già emersi dallo studio della letteratura, dal profilo sociale presentato nel Capitolo 3 e confermato infine dalle analisi finali, è la forte eterogeneità che contraddistingue il network di Twitter e l'assoluta difficoltà nell'individuare una metrica unica, condivisa ed effettivamente funzionante, impiegabile per misurare autorità e opinion leadership di ciascun utente. Dall'altra parte, la creazione di un numero eccessivo di complessi indicatori, come visto per esempio nella practice nel Capitolo 4, porterebbe ad un raggiungimento solo parziale dell'obiettivo di individuare univocamente gli influencer all'interno di una community, fornendo invece una serie di punti di vista analitici, utili, ma che richiederebbero comunque un'ulteriore fase di interpretazione e valutazione. La letteratura ha inoltre criticato o quantomeno sminuito la potenziale e intuitiva relazione esistente tra la popolarità, i.e. il numero di *follower*, e la influence, evidenziando invece come già detto la rilevanza dei principali strumenti di conservazione, *mention* e *rewteet*.

La categorizzazione applicata per distinguere i differenti utilizzi del servizio ha confermato altri studi (Romero, Asur, Galuba, & Huberman, 2010) secondo i quali le fonti di notizie sono quelle in grado di ricevere il maggior numero di *retweet*, possedendo quindi una discreta facilità nel raggiungimento di ampie audience e un apprezzabile potenziale di influence; mentre gli account identificati come *people* hanno una maggiore propensione e facilità nella ricezione di *mention*. Tuttavia, i *RT* di notizie non possono essere considerati esattamente come la normale attività, promozionale e conversazionale, degli utenti, in quanto l'informazione che si sta diffondendo e/o che viene inoltrata è stata effettivamente prodotta da qualcun altro e il contenuto in una grande maggioranza dei casi non si riferisce specificatamente al creatore del messaggio. Questo è confermato dal fatto che il *retweet_rate* della tipologia *news* non è risulta più elevato rispetto alle altre categorie, comportando che la loro comprovata autorità e influence non basta per

far sì che ogni notizia abbia lo stesso successo in termini di trasmissione e utenza raggiunta. A questo proposito, sembrano avere assolutamente senso da un lato gli studi che si focalizzano sui pattern di diffusione e viralità dei messaggi e conferiscono un ruolo fondamentale non solo a chi genera informazione, ma soprattutto a chi lo cerca, i.e. *opinion seeker*, recepisce e ritrasmette; dall'altro quelle categorizzazioni, e.g. Sezione 2.6, che mirano ad effettuare una sottodivisione degli influencer stessi, distinguendo il ruolo della fonte (*source*), da quello dei trasmettitore o connettore.

Sempre relativamente all'uso di *mention* e *RT*, che funzionalmente potrebbero apparire come strumenti differenti, uno focallizato alla riproposizione di contenuto, l'altro alla conversazione, i risultati, mostrando una significativa correlazione tra questi due, evidenziano come entrambi a causa della struttura non coesa di Twitter, siano assolutamente degli strumenti conversazionali equivalenti, teoria già espressa da alcuni studi (Boyd, Golder, & Lotan, 2010) e che addirittura, se da un parte il *retweet* consente il raggiungimento di un audience "lettrice" senza dubbio e per ovvi motivi superiore, la comunicazione *one-to-one* e l'engagement vengono identificate come le strategie migliori e più efficaci in ambito Web marketing. Per questo i tool analizzati che valutano esclusivamente il potenziale di un utente rispetto ai *retweet* che riceve, deficitano di un aspetto di valutazione fondamentale ed imprescindibile all'interno del servizio di microblogging.

In aggiunta, il nostro data set ha mostrato che la distribuzione dei gli utenti rispetto alla quantità di *retweet* e alle *mention* porta alla formazione di due cluster separati e ben definiti, al cui interno la correlazione tra il numero di *follower* e le metriche di efficienza è differente: mentre il meno efficiente presenta valori non correlati, comportando la necessità di considerare i *follower* come una metrica rilevante.

Rispetto alle caratteristiche specifiche del progetto e del tool prototipo, descritto nella Sezione 4.6, il modulo di *influencer analysis* non avrebbe un utilizzo prettamente legato e dedicato alla pesatura del sentiment, i.e. introdurre

una metrica per cui le opinioni positive/negative di un individuo che risultasse influencer dovrebbero avere una rilevanza maggiore di quelle espresse da un non influencer, questo essenzialmente per due ragioni:

- l'attività di un singolo individuo effettivamente rilevante in termini di *sentiment analysis*, i.e. tutti quei messaggi postati da un utente i quali, oltre a poter essere inseriti nel modello di mappatura, contengono anche un opinione, sono da considerarsi un numero davvero molto poco significativo e per cui l'introduzione di una metrica di pesatura avrebbe scarsa utilità;

- il crawler ed il *processing* svolto dal motore semantico includono anche i *reweet*, questo implica che un singolo messaggio nel caso in cui venga riproposto abbia comunque un impatto maggiore sull'overall del sentiment.

Il modulo avrebbe decisamente più senso se andasse quindi ad operare ad un livello più alto e cioè quello di categorizzazione all'interno dei driver. L'obiettivo sarebbe sostanzialmente quello di individuare dei cluster di utenti influenti da combinare con le nove macrocategorie di riferimento della mappatura, in base ai topic conversazionali più comuni di ciascuno, tenendo eventualmente come riferimento anche una categoria *general*, per i contenuti generali sulle città e non assegnabili specificatamente ad una particolare categoria, e.g. caso verosimile per delle fonti di informazione generaliste. Uno degli elementi che rafforza che questa ipotesi è il fatto che alcuni utenti molto autorevoli, grazie anche all'attenzione che generano nel network venendo menzionati o con la riposizione dei loro messaggi, contribuiscono in modo significativo e soprattutto più di altri, ad alimentare il flusso di messaggi riguardanti un particolare topic o una specifica categoria del modello, se consideriamo la mappatura del progetto trattato.

L'applicazione delle metriche di efficienza, sia di *retweet* che di *mention*, consentirebbe inoltre una valutazione più oggettiva ed in grado di limitare i bias per cui gli account di *news,* che investono la quasi totalità dei loro messaggi nella

diffusione di link a notizie e spesso con volumi elevati, risultino costantemente come i più influenti.

La capacità di individuazione degli influencer permetterebbe in questo modo, da un lato una maggiore efficacia delle azioni di marketing che impiegano il *targeting* di un definito gruppo di utenti, e per ciascuna categoria del modello un'interessante comparazione tra il sentiment espresso al cluster di influencer e quello di non influencer, i.e. denominato nel seguito come *rest of us*. L'utilizzo della pesatura e valutazione comparativa troverebbe inoltre un proficuo utilizzo, non solo nel processo di monitoraggio costante, ma anche in caso di analisi e gestione di eventi e alert particolari.

Un esempio potrebbe essere il monitoraggio dei messaggi scambiati online durante un evento in ambito fashion, valutando l'*overall* del sentiment attraverso il confronto dei messaggi degli influencer clusterizzati nel driver *fashion&shopping* con il sentiment generale espresso.

La matrice seguente, Figura 34, è esattamente finalizzata a spiegare questo concetto, descrivendo la casistica all'interno del quale è possibile ricadere e ipotizzando dei possibili provvedimenti.

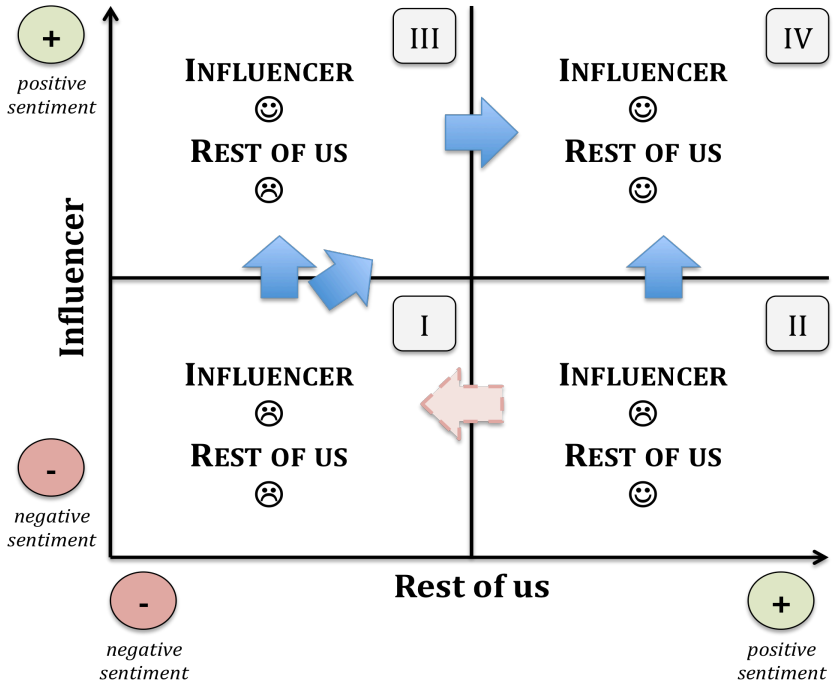


Figura 34 - Matrice di confronto del sentiment influencer/rest of us

E' opportuno precisare alcuni aspetti riguardanti la valutazione del sentiment. Quello che è definito come *overall*, indica una media pesata tra i due gruppi *influencer/rest of us,* così come il sentiment del singolo cluster è da intendersi come una media espressa dal totale delle opinioni degli utenti che fanno parte di ciascun gruppo. Le modalità in cui queste singole medie vengono calcolate o vengono inseriti dei particolari parametri e/o pesi per una valutazione del sentiment esulano dagli obiettivi e dalle tematiche principali di questo lavoro.

Vediamo ora in dettaglio ciascuno dei casi inclusi nei quadranti della matrice qualitativa:

I. Sia gli *influencer* che *i rest of us* esprimono un sentiment negativo. Questa è ovviamente la situazione peggiore, in cui l'*overall* è a sua volta necessariamente negativo. Verosimilmente l'azione più efficace comporta l'engagement degli *influencer,* coinvolgendoli al fine di un miglioramento

delle opinioni da loro espresse, con la possibilità che questo abbia poi un effetto esteso anche ai *rest of us,* con una crescita progressiva dell'*overall* dal quadrante III al IV.

II. I *rest of us* esprimono un setiment positivo, mentre l'opinione degli *influencer* è negativa. In questa situazione c'è il pericolo che l'opinione di questi ultimi incida direttamente sull'*overall*, impattando e modificando il sentiment dei *rest of us*. Il coinvolgimento degli *influencer* permetterebbe potenzialmente uno spostamento nella situazione di *overall* positivo.

III. Gli *influencer* esprimono sentiment positivo, mentre l'opinione dei *rest of us* è negativa. La situazione è quella opposta alla precedente e il rischio connesso da gestire è relativamente inferiore. Gli opinion leader, coinvolti con l'obiettivo di renderli promotori attivi, potrebbero avere un effetto diretto sulla parte restante dell'audience, facendo sì che anche l'*overall* si sposti nel IV quadrante, in cui entrambe i cluster esprimono un sentiment positivo.

IV. Sia il sentiment degli *influencer* che quello dei *rest of us* è positivo. E' la situazione migliore, in cui l'*overall* complessivo è necessariamente positivo. Questo è il caso ottimo, obiettivo da raggiungere attraverso le azioni di engagement se l'evento o topic specifico da gestire hanno mostrato che un disallineamento delle opinioni tra *influcenr* e *rest of us*, o, peggio ancora, un *overall* complessivo negativo.

Capitolo 6

Conclusioni e sviluppi futuri

"...in an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it"

Herbert Simon

L'obiettivo iniziale della tesi è stato lo studio in ambito Web reputation degli indicatori che possono aiutare a definire e analizzare il concetto di influence e autorità all'interno della piattaforma di microblogging Twitter.

Come discusso e approfondito nel corso del lavoro, la presenza di numerosi studi sottolinea la crescente importanza ed interesse dell'argomento e il notevole potenziale di sviluppo futuro, sia per quanto riguarda l'impiego di questi servizi per iniziative di Web marketing, sia dal punto di vista tecnologico per l'accuratezza della fase di *processing* dei tool, i.e. la ricerca semantica e la valutazione del sentiment delle opinioni. A livello accademico e anche commerciale, sono state effettuate differenti proposte riguardo la misurazione

della influence, nessuna delle quali però largamente condivisa e considerata come davvero valida ed efficace.

Le analisi si sono concentrate sullo studio dell'interazione e l'utilizzo dei principali strumenti conversazionali del microblogging, *retweet* e *mention*, all'interno di una rete di utenti con caratteristiche definite e coerenti con i fini del progetto in corso, i.e. la collocazione geografica degli utenti.

Il lavoro di tesi ha permesso di verificare e confermare alcune teorie sulla valenza e sulla rilevante correlazione statistica che intercorre tra queste modalità di comunicazione e di ampliare inoltre il punto di vista sul tema, introducendo delle metriche di efficienza ottenute attraverso la registrazione dell'attività effettiva degli utenti del sample nel periodo di analisi, giungendo a risultati parzialmente contrastanti con altri studi. Se da una parte infatti la correlazione tra i *retweet_rate* e i *mention_rate*, i.e. le metriche di efficienza, è dimostrabile, dall'altra, se ci si concentra su quel sottoinsieme di utenti che hanno ricevuto il maggior livello di attenzione[40] da parte del network, suddivisione ottenuta grazie ad una clusterizzazione del set effettuata in seguito all'analisi della distribuzione statistica degli indicatori, emerge un'effettiva rilevanza della popolarità, i.e. numero di *follower*, la quale presenta una significativa correlazione con le altre due metriche.

In aggiunta, applicando una categorizzazione precisa, i.e. *brand/news/people*, al set di account, mirata a distinguere diverse tipologie di utilizzo del servizio in base alle finalità, sono emersi dei bias significativi tra le categorie. Confermate le ipotesi e gli studi secondo cui le fonti di news mostrano le migliori performance in termini di *reach audience* e di riproposizione, i.e. *retweet*, dei propri contenuti da parte degli altri utenti, mentre gli individui, i.e. categoria *people*, sono invece molto più orientati alle conversazioni *one-to-one* con un elevato scambio di *mention*, non si hanno però gli stessi risultati nel momento in cui viene preso in considerazione anche il livello di attività e il confronto viene effettuato utilizzando le metriche di efficienza citate. Questo suggerisce che nel momento in cui si vogliano valutare i valori assoluti degli indicatori, bisogna tenere ben

[40] Con "attenzione" si intendono *retweet* e *mention* ricevute.

presenti delle caratteristiche intrinseche e dei risultati potenziali che differenziano pesantemente le differenti tipologie di utilizzo e quindi di account.

Nell'ottica di approfondimenti futuri sarebbe importante utilizzare queste linee guida per finalizzare lo sviluppo del modulo di *influencer analysis* all'interno del tool, con un uso fortemente orientato, come detto, all'individuazione di cluster di utenti collocabili all'interno del modello di analisi e ad una valutazione del sentiment comparata *influencers/rest of us,* illustrata nel Capitolo 5.

Di seguito alcuni spunti per l'immediato futuro:

- Sviluppo del crawler. Implementare un agente software in grado di sfruttare le API di Twitter per estrarre correttamente tutti gli indicatori utilizzati nella valutazione degli utenti;

- Individuare elementi o soglie quantitative per automatizzare la categorizzazione degli account. Come visto dai risultati la partecipazione di diverse tipologie di utenza con finalità ed utilizzo specificatamente differenti ha condotto a dei risultati che presentano dei bias nel momento in cui si utilizzano per le valutazioni gli indicatori assoluti. La possibilità di introdurre delle categorie di utilizzo ben definite da alcune soglie e parametri (e.g. attività molto elevata e efficienza nel *retweet_rate*, *information broadcaster*) arricchirebbero il modulo introducendo un punto di vista aggiuntivo sull'analisi. Allo stato dell'arte, e come effettuato nel lavoro di tesi, la categorizzazione necessita di un processo *analyst intensive;*

- Raggiungimento di un compromesso efficace e funzionale tra una metrica univoca che restituisca un valore sintetico di autorità e l'impiego di un numero elevato di indicatori analitici. Da un lato infatti, visti i bias emersi dai risultati, un unico *score* finale potrebbe risultare fortemente impreciso e fuorviante, così come un eccesso di metriche richiederebbe un ulteriori e probabilmente eccessivo sforzo interpretativo all'utente finale del tool.

Una volta implementato il modulo, non rimarrebbe che verificarne l'effettivo funzionamento con un flusso continuo di dati o testarlo con un pilota per il monitoraggio di un evento in particolare. In questo modo non solo si studierebbe un campo innovativo che sta attirando molta attenzione, ma si avrebbe la possibilità di vedere messo in pratica uno stimolante lavoro di progettazione e soprattutto di analisi.

Bibliografia

Adamic, L. A., Lukose, R., Puniyani, A. R., & Huberman, B. A. (2001). Search in power-law networks. *Physical Review* (64).

Askira Gelman, I., & Barletta, A. (2008). A "quick and dirty" website data quality indicator. *In Proceedings of the 2nd ACM Workshop on Information Credibility on the Web* (pp. 43-46). New York, NY, USA: ACM.

Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth and evolution. *In Proceedings of 12th International Conference on Knowledge Discovery and Data Mining* (pp. 44-54). Philadelphia, PA, USA: ACM.

Barres, J. (1954). Class and committees in a Norwegian island parish. *Human Relations , 7*, 39-58.

Beck, T., & Davenport, J. (2001). *The attention economy: Understanding the new currency of business.* Cambridge, MA: Harvard Business School press.

Bickart, B., & Schindler, R. M. (2001). Internet forum as influential sources of consumer information. *Journal of Interactive Marketing , 15* (3), pp. 31-40.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. *In Prooceedings of the 43rd Hawaii International Conference on System Science* (pp. 1-10). Kauai, HI: hicss.

Branckaute, F. (2010 June). *Twitter's meteoric rise [infographic]*. From Blogherald: http://bit.ly/9qTIlR

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *In Prooceedings of the 7th Internation World Wide Web Conference , 30* (1-7), pp. 107-117.

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

Carlson, N. (2010 2-June). *14 Twitter charts you must see*. From Business Insider: http://bit.ly/dCt1lb

Cartwright, D. (1959). *Studies in social power* (Vol. 6). University of Michigan, USA: Reasearch Center for Group Dynamics.

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.

Cha, M., Mislove, A., & Gummadi, P. (2009). A measurement-driven analysis of information propagation in the Flickr social network. *In Proceedings of the 18th International Conference on World Wide Web* (pp. 721-730). New York, NY, USA: ACM.

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.

Clippinger, J. H. (2007). *A crowd of one: The future of individual identity.* New York, USA: PublicAffairs.

Cova, B., Giordano, A., & Pallera, M. (2007). *Marketing non convenzionale.* Il Sole 24 Ore.

Deal, R. (2009 01-January). *The ten users you'll meet on Twitter.* From Mashable: http://bit.ly/EECr

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.

Dichter, E. (1966). How word-of-mouth advertising works. *Harvard Business Review , 44* (6), 147-160.

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.

Douglas, N. (2007 12-March). *Twitter blows up at SXSW conference.* From Gawker: http://bit.ly/dxeedb

Eaton, K. (2009 8-December). *Twitter really works: Makes $6,5 million sales for Dell.* From Fast Company blog: http://bit.ly/7qpmqM
Ellison, N. B., & Boyd, D. M. (2007). Social network sites: Definition, history and scholarship. *Journal of Computer Mediated Communication , 13* (1).

Engel, J. F., & Blackwell, R. D. (1994 December). Consumer behaviour. *The Dryden Press Series in Marketing* .

Esch, F. R., Langner, T., & Schmitt, B. H. (2006). Are brands forever? How brand knowledge and relationships affect current and future purchases. *Journal of Product and Brand Management , 15* (2), 98-105.

Flynn, L. R., Goldsmith, R. E., & Eastman, J. K. (1996). Opinion leaders and opinion seekers: Two new measuerement scales. *Journal of the Academy of Marketing Science , 24* (2), 137-147.

Frank, A. (2008). *Social media delivers marketing intelligence.* Gartner.
Freeman, L. C. (1979). Centrality in social networks: Conceptuali clarification. *Social Networks , 1* (3), 215-239.

Friedman, V. (2007 2-August). *Data visualization: Modern approaches.* From Smaghing Magazine: http://bit.ly/9U4Age

Gladwell, M. (2002). *The tipping point: How little things can make big difference.* Big Bay Books.

Golder, S. A., & Wilkinson, D. &. (2007). Rhythms of social interaction: Messaging within a massive online network. *Proceedings of 3rd International Conference on Communities and Technologies* (pp. 28-30). East Lansing, MI, USA: HP Labs.

Gotz, M., Leskovec, J., McGlohon, M., & Faloutsos, C. (2009). Modeling blog dynamics. *In Proceedings of Internation Conference of Weblogs and Social Media.*

Granovetter, M. S. (1973). The strenght of weak ties. *The American Journal of Sociology , 78* (6), 1360-1380.

Granovetter, M. S. (1983). The strenght of weak ties: a network theory revisited. *Sociological Theory , 1*, 201-233.

Guya, Y. (2006 Ottobre). *Viral Marketing - Miti, leggende e verità del marketing non convenzionale*. From Ninja Marketing: http://bit.ly/ccvKrZ

Harary, F., Norman, R., & Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs.* New York, NY, USA: John Wiley & Sons.

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.

Hogan, B. (2008). Analyzing social networks via the Internet. In *Online Research Method* (pp. 141-160). Thousands Oaks, CA, USA: N Fielding, R.M. Lee & G. Blank.

Honeycutt, C., & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. *In Proceedings of the 42th Hawaii International Conference on System Sciences* (pp. 1-10). Big Island, Hawaii, USA: HICSS.

Huffington Post. (2010 14-April). *Twitter user statistic REVEALED.* From Huffington Post: http://huff.to/aSzASZ

Infosthetics. (2010 19-April). *Revisit: Visualizing the temporal dynamics of Twitter.* From http://bit.ly/cr2VBd

Ingelbrecht, N., Patrick, C., & Foong, K. Y. (June 2010). *User survey analysis: Consumer marketing using social networks analysis worlwide.* Gartner Industry Research.

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

Java, A., Finn, T., Song, X., & Tseng, B. (2007). Why we Twitter: understanding microblogging usage and communities. *In Proceedings of the 9th WEBKKD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis* (pp. 56-65). San Jose, CA, USA: ACM.

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.

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

Kawasaky, G. (2009 December). *The six Twitter types.* From Open Forum: http://bit.ly/4zarXE

Keller, E., & Berry, J. (2003). *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy.* Free Press.

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.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of online social networks. *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 611-617). ACM.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *In Proceedings of the 19th International World Wide Web Conference.* Raleigh, North Carolina, USA: ACM.

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.

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

Lazarsfeld, P., & Katz, E. (1955). *Personal influence: The part played by people in the flow of mass communications.* New York, NY, USA: Free Press.

Leavitt, A. (2009). *The influentials: New approaches for analyzing influence on Twitter.* Boston, USA: Web Ecology Project.

Lee, C., Kwak, H., Park, H., & Moon, S. (2010). Finding influentials based on the temporal order of information adoption in Twitter. *In Proceedings of the 19th International World Wide Web Conference* (pp. 1137-1138). Raleigh, North Carolina, USA: ACM.

Makrehchi, M. &. (2006). Learning social networks from Web documents using support vector classifiers. *Proceedings of IEEE/WIC/ACM international Conference on Web Intelligence* (pp. 88-94). IEEE.

Marlow, C. (2005). *The structural determinants of media contagion.* MIT Media Lab.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology , 27* (1), pp. 415-444.

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

Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., & Magoulas, R. (2008). *Twitter and the micro-messaging revolution: Communication, connections, and immediacy-140 characters at a time.* O'Reilly Radar.

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.

Naruse, K., & Kubo, M. (2006). Lognormal distribution of bbs articles and its social and generative mechanism. *In Web Intelligence* (pp. 102-103, 18-22). IEEE/WIC/ACM.

Nielsen Company. (2009 9-March). *Global Faces and Networked Places.* From Nielsenwire: http://bit.ly/9HtYt3

Nielsen Company. (2010 2-August). *What Americans do online: social media and gaming dominate activity.* From Nielsenwire: http://bit.ly/audwel

Ohira, M., Nakakoji, K., & Matsumoto, K. (2006). D-sns: A knowledge exchange mechanism using social network density among mega-community users. *In Proceedings of Supporting the Social Side of Large Scale Software Development* (pp. 39-42). ACM - CSCW Workshop.

Ohira, M., Ohsugi, N., Ohoka, T., & Matsumoto, K. (2005). Accelerating crossproject knoweldge collaboration using collaborative filtering and social networks. *In Porceedings of the International Conference of Software Engineering* (pp. 1-5). ACM.

O'Reilly, T. (2005 September). *What is Web 2.0 : Design Patterns and Business Models for the Next Generation of Software.* From http://o'reilly.com/: http://oreil.ly/aa1sP

Pals, L. (2010 18-May). *Huge Twitter Infographics.* From TheNextWeb: http://bit.ly/d0uWgt

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.

Rao, L. (2010 йил 17-July). *How social media drives new business: Six case studies.* From Techcrunch: http://tcrn.ch/aiN7zu

Reichheld, F. (2003 йил December). The one number you need to know.

Reynolds, F. D., & Darden, W. R. (1971). Mutually adaptive effects of interpersonal communication. *Journal of Marketing Research , 8* (4), 449-454.

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.

Rogers, E. (1962). *Diffusion of innovation.* New York, NY, USA: Free Press.
Romero, D. M., Asur, S., Galuba, W., & Huberman, B. A. (2010). Influence and passivity in social media . *ACM* .

Sala, A., Zheng, H., Zhao, B., Gaito, S., & Rossi, G. (2010). Brief announcement: Revisiting the power-law degree distribution for social grpah analysis. *In Proceedings of the 29th ACM SIGACT-SIGOPS symsposium on Principles of distributed computing* (pp. 400-401). New York, NY, USA: ACM.

Sarshar, N., Boykin, P. O., & Roychowdhury, V. P. (2004). Percolation search in power law networks: Making unstructured peer-to-peer network scalable. *In Proceedings of the 4th International conference on Peer-to-Peer computing* (pp. 2-9). Washington, DC, USA: IEEE Computer Society.

Semprini, A. (2006). *Marche e mondi possibili.* Milano: Franco Angeli.

Solis, B. (2010). *Engage: The complete guide for brands and businesses to build, cultivate and measure success in the new Web.* Wiley.

Solis, B. (2008 8-August). *Introducing the conversation prism.* From Brian Solis: Defining the convergence of media and influence: http://bit.ly/18Qr6v

Stefaner, M. (n.d.). From Revisit: http://moritz.stefaner.eu/projects/revisit

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.

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.

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).

Sundaram, D. S., & Mitra, K. &. (1998). Word-of-mouth communication. A motivational analysis. *Advances in Consumer Research , 25* (1), 527-531.

Sysomos Inc. (2010 January). *Top Twitter countries and cities (part 2).* From Sysomos Blog: http://bit.ly/5pB9Ng
Techcrunch. (2010, September 2). *Twitter now over 145 million users, almost 300.000 apps.* Tratto da Techcrunch: http://tcrn.ch/cjOXRx

Teutle, A. R. (2010 22-24-February). Twitter: Network properties analysis. *In Proceedings of the 20th International Conference on Electronics, Communications and Computer* , pp. 180-186.

Trusov, M., Bodapati, A. V., & Ucklin, R. E. (209 йил 20-April). Determining influential users in Internet social networks. *Journal of Marketing Research* .

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

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Giulia, M., & Haythornthwaite, C. (1996). Computer networks as a social networks: Collaborative work, telework, ad virtual community. *Annual Review of Sociology , 22* (1), pp. 213-238.

Weng, J., Lim, E., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influentials in Twitter. *In proceedings of the 3rd ACM Internation Conference on Web Search and Data Mining* (pp. 261-270). New York, USA: ACM.