Università
Ca'Foscari
Venezia

Corso di Laurea magistrale in Marketing e Comunicazione

Tesi di Laurea

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Le potenzialità dell'analisi dell'utilizzo dei social network a fini di marketing

Caso studio sulle sigarette elettroniche

Relatore
Prof. Stefano Campostrini

Correlatore
Prof. Roberto Franzosi

Laureando
Federica Fedrigo
Matricola 820994

Anno Accademico
2013 / 2014

Ringraziamenti

Desidero innanzitutto ringraziare il Prof. Stefano Campostrini per il Suo aiuto e la Sua disponibilità durante il mio lavoro di ricerca. Vorrei inoltre esprimere la mia sincera gratitudine al Prof. Roberto Franzosi per la Sua collaborazione nella fase operativa d'analisi e per i Suoi preziosi consigli. La mia riconoscenza va anche al Prof. Andrea Marin e al Prof. Rodolfo Delmonte per il Loro supporto tecnico.

Un ringraziamento particolare va ai miei genitori senza il cui sostegno non sarei arrivata a questo importante traguardo, a mio fratello Marco e a Simone che è sempre stato al mio fianco, a loro dedico la mia tesi.

Ed infine un grazie a tutti i nonni, parenti, amici ed alla mia amica di sempre Trejsi per avermi accompagnato durante il mio percorso.

Indice

Introduzione 1

1. Big Data 4

1.1 La definizione di Big data 6

1.2 Big Data analysis 9

1.3 Utilizzo dei Big Data: dal settore privato a quello pubblico 10

1.3.1 Utilizzo dei sensori dei cellulari 11

1.3.2 Casi nel settore economico 14

1.3.3 Casi nel settore sanitario 16

2. Social Media 25

2.1 Diffusione di Internet e dei social network 28

2.2 Le motivazioni dell'interesse dei ricercatori per i social network 35

2.3 Il potere influenzante dei social network 38

2.4 Le modalità di diffusione di una notizia all'interno dei social network 42

2.5 I diversi tipi di reti che si creano nei social network 46

2.6 I social network e le previsioni 54

3. Twitter 84

3.1 Dati sull'utilizzo di Twitter 84

3.2 I vantaggi dell'analisi attraverso Twitter 87

3.3 Che cosa viene pubblicato e cercato in Twitter 89

3.4 Metodologie per estrarre i dati da Twitter 92

3.5 Metodologie per analizzare i dati di Twitter 97

3.6 I limiti dell'analisi di Twitter 101

4. Analisi di Twitter riguardo al tema della sigaretta elettronica 110

4.1 La sigaretta elettronica 110
4.1.1 La legislazione in tema di sigaretta elettronica 111
4.1.2 Dati sul fumo e sulle sigarette elettroniche in Italia 115
4.1.3 Riflessioni mediche sulle sigarette elettroniche come mezzo per smettere di fumare 127
4.2 Analisi di Twitter 130
4.2.1 Metodologia adottata per l'estrazione dei tweet 130
4.2.2 Analisi quantitativa dei tweet estratti 131
4.2.3 Analisi linguistiche del contenuto dei tweet estratti 143
4.2.3.1 Word cloud 143
4.2.3.2 Content analysis 148
4.2.3.3 Analisi semantica 153
4.2.3.4 Quantitative narrative analysis e thematic analysis 155
4.2.4 Sintesi dei risultati ottenuti dall'analisi dei tweet estratti 165
4. 3 Confronto tra dati provenienti dal mondo online e quelli raccolti nel mondo reale 167
5. Spunti d'applicabilità per il marketing - Riflessioni finali 169
Bibliografia 180

Introduzione

La quantità di dati socio-economici generati ogni giorno è cresciuta vertiginosamente negli ultimi anni grazie alla diffusione della connessione internet su vasta scala e all'aumento della disponibilità di dispositivi elettronici, il cui utilizzo lascia un enorme quantitativo di tracce digitali di vario genere: foto, email, registri di chiamate, informazioni sugli acquisti effettuati, transazioni finanziarie, interazioni nei social network.

I Big Data sono dati caratterizzati da volume, velocità e varietà: sono cioè estratti ed elaborati ad alta velocità e raccolti in dataset di grandi dimensioni, i quali sono costituiti da dati provenienti dalle fonti più varie e pertanto non solamente da dati strutturati. '

La raccolta di dati è tipicamente difficile e dispendiosa, in termini sia di tempo che di denaro; invece, l'entusiasmo che circonda i Big Data è dovuto proprio alla percezione di grande facilità e velocità d'accesso a un elevato ammontare di dati a basso costo.

In letteratura, i casi studio che riportano le modalità di sfruttamento di tali dati sono in continua crescita. Nel Capitolo 1 ne viene fornita una panoramica generale, per focalizzarsi poi su una particolare tipologia di Big Data: quelli provenienti dai social network.

Questi ultimi appaiono particolarmente interessanti poichè permettono di conoscere ciò che le persone "pensano", dato che ormai sulle piattaforme digitali si esprimono senza censure le proprie idee e i propri sentimenti riguardo a qualunque tema e si raccontano molti eventi della propria vita, dai fatti privati più importanti ai piccoli contrattempi quotidiani.

Sono stati individuati quattro principali approcci di ricerca sui social network e per ognuno dei diversi filoni sono stati riportati casi studio inerenti agli ambiti più vari: economia, psicologia, sociologia, finanza, politica, medicina e persino sismologia e linguistica, a testimonianza della validità di questo tipo di analisi e dell'elevata versatilità che la caratterizza.

I vantaggi della scelta di Twitter quale social network da analizzare sono, principalmente, la sua natura pubblica, la presenza degli hashtag, che permettono di categorizzare i tweet in base alle diverse tematiche, e la possibilità di geolocalizzare i

messaggi postati. Tali caratteristiche hann guidato la scelta di incentrare il lavoro di ricerca su questa piattaforma sociale.

Dopo una breve presentazione degli strumenti ufficiali disponibili per estrarre i tweet pubblicati e dei metodi adottabili per analizzarli, sono stati messi in luce i limiti di questa tipologia d'analisi relativi alla violazione della privacy, alla rappresentatività dei dati ottenuti, all'autocensura praticata dall'autore, alle difficoltà di accesso alla totalità dei tweet pubblicati, alla presenza di dati mancanti, al rischio di apofenia e alla presenza di possibili errori d'interpretazione generati dalla sintassi irregolare e dal linguaggio particolare adottati in Rete.

L'utilizzo della piattaforma sociale, che a inizio 2014 registrava 645.750.000 utenti, dà la possibilità non solo di ottenere informazioni, in tempo reale e continuamente aggiornate su quello che l'opinione pubblica pensa riguardo a determinate tematiche, ma anche di monitorarne l'evoluzione per individuare eventuali trend ciclici o anomalie.

Per testare questa sua potenzialità, da Gennaio ad Aprile 2014, attraverso una Twitter REST API, sono stati estratti i tweet contenenti parole chiave facenti riferimento al fenomeno della sigaretta elettronica o inerenti la discussione sul tema dello smettere di fumare, per capire se il fumo elettronico, in Italia, è utilizzato come metodo per diminuire o eliminare il consumo di tabacco. I dati estratti sono poi stati elaborati sfruttando diversi software per analizzarli sia dal punto di vista quantitativo che del contenuto. Successivamente, sono stati confrontati con quelli reali raccolti dal sistema di sorveglianza adottato dal Ministero della Salute per capire in quale misura i messaggi pubblicati nel social network riflettessero comportamenti adottati nel mondo reale.

La ricerca ha permesso di porre in evidenza le potenzialità, le difficoltà e le limitazioni dell'analisi dell'utilizzo dei social network suggerendone interessanti spunti d'applicabilità. Il presente lavoro è stato focalizzato, in particolare, sull'utilità che l'analisi del comportamento online degli utenti delle piattaforme digitali può avere nell'ambito del marketing. In questo campo, potrebbe essere adottata per capire meglio i gusti e le idee dei consumatori, per indagare le motivazioni del loro comportamento d'acquisto fino a riuscire quasi a prevederlo, per analizzare come e quanto un'impresa viene paragonata ai competitor o per comprendere l'immagine che un'azienda o un marchio si sono creati, meglio di quanto permettono di fare i soli sondaggi tradizionali. In questo modo, possono essere ricavati indicazioni e spunti creativi sfruttabili nella

progettazione di una campagna di marketing (per renderla adattiva all'ambiente in cui si inserisce), o informazioni utilizzabili per la valutazione della stessa. Questa risulta essere una fase sempre importante e complessa, soprattutto nel caso di campagne di marketing sociale, per le quali il giudizio riguardante l'efficacia non può nemmeno avvalersi del supporto di indicatori economico-finanziari o di mercato.

È stato infatti proprio su un caso sociale, in particolare di sanità pubblica, che si è andati ad applicare questa tipologia d'analisi.

1. Big Data

L'ultimo decennio è stato caratterizzato da innumerevoli innovazioni tecnologiche, dall'aumento esponenziale del collegamento alla rete Web e dalla crescente influenza dei dispositivi digitali sulla nostra vita quotidiana; tutto ciò ha comportato un enorme incremento della quantità di dati prodotti ogni giorno.

Ogni anno vengono creati più dati economico-sociali che non in tutta la storia precedente dell'umanità (Vespignani, 2013). I dati generati in Internet ogni secondo equivalgono alla produzione d'informazioni sull'intera Rete di venti anni fa; Facebook, ad esempio, ingloba circa 500 Terabyte al giorno (Dello Iacovo, 2013).

Negli ultimi due anni è stato creato circa il 90% dei dati mondiali e si prevede che nel 2020 la quantità di dati memorizzati potrebbe essere 50 volte più grande rispetto a quella del 2010 (Barton, 2013).

E' stato inoltre stimato che l'ammontare di dati digitali disponibili a livello globale aumenterà del 40% ogni anno; in questo modo tra il 2007 e il 2020 esso aumenterà di 44 volte raddoppiando ogni 20 mesi.

Sempre di più ogni nostra azione quotidiana lascia dietro di sé una traccia digitale: le nostre abitudini di consumo, le nostre ricerche sul Web, i video caricati o visualizzati online, le opinioni condivise nei social network, i registri chiamate, persino i movimenti delle auto o dei nostri stessi corpi. L'utilizzo di qualsiasi dispositivo elettronico genera, infatti, informazioni, sia dirette che indirette, sulle persone e sulle loro interazioni che rimangono registrate e possono andare a creare dataset dalla dimensioni enormi; si stimano circa due quintilioni e mezzo (2.500.000.000.000.000.000) di byte creati ogni giorno nel mondo.

Quest'enorme disponibilità di dati in tempo reale, per chi è interessato a studiare il comportamento umano, è paragonabile ad una miniera d'oro da sfruttare in qualche modo.

Sociologi, scienziati, fisici, economisti, matematici, politici, informatici e altri ricercatori vedono in essi l'opportunità di avere strumenti predittivi per il corpo sociale.

La sociologa Zeynep Tufekci paragona l'impatto dell'applicazione della Big Data analysis applicata ai social media per studiare il comportamento umano all'introduzione del microscopio o del telescopio nei campi della biologia e dell'astronomia.

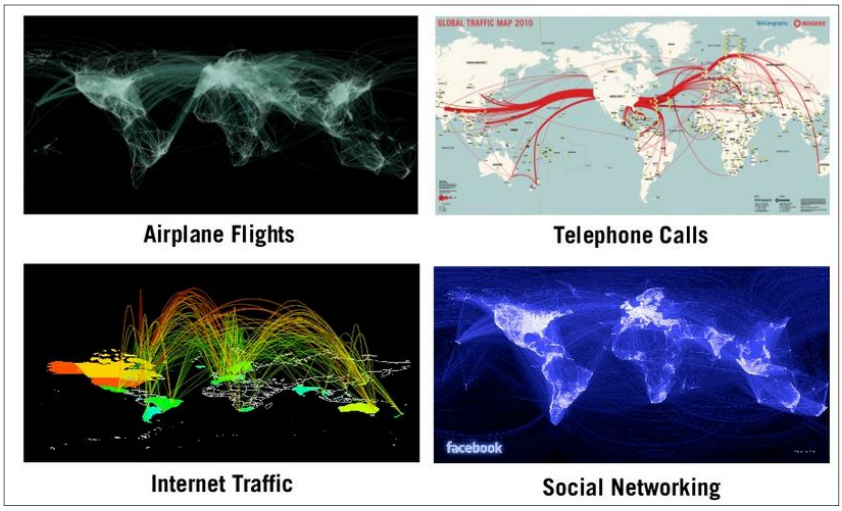


Figura 1.1 : Visualizzazioni di diverse tipologie di reti che creano flussi di dati digitali
Fonte: Kirkpatrick, 2012

E' importante sottolineare che tutti questi dati non sono prodotti solo nei Paesi industrializzati, ma è sempre più in crescita il flusso proveniente dai Paesi in via di sviluppo, basti pensare che nel 2010 i cellulari nel mondo erano oltre cinque miliardi di cui l'80% nei Paesi in via di sviluppo e che, secondo la Mc Kinsey, il traffico in Internet si prevede aumenterà del 25-30% tra il 2011 e il 2015 in Nord America, Europa e Giappone, mentre in America Latina, Medio Oriente e Africa più del 50%.

Il boom di dati è considerato il nuovo petrolio e i più fiduciosi affermano che siamo agli albori di un'importante nuova era nella storia dell'umanità e che questa rivoluzione, ormai imminente, migliorerà il modo in cui prendiamo le nostre decisioni facendo cambiare in un decennio le nostre vite; sostanzialmente non perché disporremo di un nuovo strumento tecnico, bensì perché la nostra comprensione della realtà migliorerà nettamente (Mayer-Schonberger, 2013).

Fino ad oggi si è tentato di convalidare teorie ed interpretazioni della realtà attraverso modalità che richiedessero il minore quantitativo possibile di dati poiché la loro raccolta, memorizzazione ed analisi sono sempre state molto dispendiose sia in termini di tempo, che di risorse economiche. Per questo le ricerche di marketing utilizzano piccoli campioni estratti dal proprio target per testare prototipi di un nuovo prodotto e per stimarne la domanda, nelle fabbriche si fanno test su prodotti prelevati casualmente dalle linee di produzione per garantire la qualità della produzione e per prevedere il

risultato delle elezioni si fanno sondaggi su una determinata percentuale della popolazione.
Attraverso questi procedimenti si esplorano solo interrogativi già posti, ma non se ne generano di nuovi sui quali non abbiamo ancora riflettuto. Con la disponibilità attuale di dati, invece, è possibile studiare vari fenomeni disponendo di un numero notevolmente maggiore di informazioni da analizzare e reperibili a costi e tempi inferiori. Con i Big Data possiamo non solo utilizzare grandi database per convalidare un'ipotesi già formulata, ma anche dar voce ai dati stessi, facendoci così non solo illuminare da essi, ma addirittura ispirare.

1.1 La definizione di Big data

Qual è la grandezza che può farci definire "big" un dataset? La dimensione definibile grande si è col tempo modificata. Se nel 1975, durante la prima conferenza internazionale sui VLDB (Very Large Databases), ci si preoccupava di come gestire i milioni di dati ottenuti attraverso il censimento americano perché la loro grandezza creava problemi nell'organizzarli ed analizzarli, oggi numerosi dataset hanno dimensioni di gran lunga maggiori.
Shneiderman durante una conferenza internazionale, la ACM SIGMOD International Conference on Management of Data, nel 2008, definiva invece un dataset come "big" quando è troppo grande per esser inserito in uno schermo.
Oggi la crescente disponibilità di dati rende sempre più diffuso il termine Big Data il quale necessita, pertanto, di una definizione più precisa: un grandissimo dataset che raccoglie dati relativi a uno stesso argomento che presentano contemporaneamente tre particolari caratteristiche, le cosiddette 3 V, ossia *Volume, Velocity, Variety* (Volume, Varietà e Velocità)**.** Per velocità s'intende la rapidità della loro generazione, ma soprattutto la necessità di raccoglierli in tempo reale ed elaborarli immediatamente poiché la loro analisi diventa in breve obsoleta. Quando si parla di dati in tempo reale, ci si riferisce ad informazioni su un fenomeno disponibili abbastanza velocemente da permettere sia un'accurata valutazione del suo stato attuale, che un'effettiva azione in risposta (UN Global Pulse, 2013).

Per volume s'intende la sommatoria delle dimensioni dei singoli dati che costituiscono il dataset, la loro misura va dai petabytes fino ai zettabytes; unità di misura enormi se si considera che un zettabyte corrisponde ad un miliardo di terabytes, cioè un triliardo di bytes.

Infine con la caratteristica varietà si fa riferimento alle fonti di provenienza dei dati che possono essere tra le più disparate; questo a conferma che un dataset costituito da Big data non è quindi composto esclusivamente da dati strutturati.

I dati strutturati hanno tutti lo stesso formato e lunghezza predefiniti e s'inseriscono senza problemi in campi fissi di database relazionali o in righe e colonne di fogli di calcolo; sono pertanto facili da raccogliere, comparare ed analizzare. Dati provenienti da fonti varie sono invece non strutturati, non hanno cioè formati e lunghezze uguali, non si possono confrontare e le fasi di organizzazione e analisi diventano molto più complicate. Essi possono essere foto, immagini, articoli, video, flussi di dati real-time da sensori o computer o altre macchine, messaggi di testo, tweet, registrazioni audio, dati di localizzazione geografica, mappe, registri di chiamate, email, tracce di interazioni nei social media, transazioni finanziarie, registrazioni relative al pagamento di tasse e altre tracce digitali lasciate dalle persone durante la loro vita quotidiana sia in modo consapevole che non.

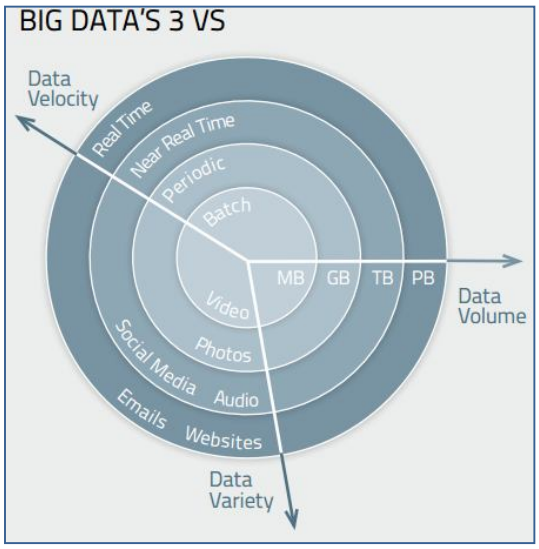


Figura 1.2: Le 3 V per definire cosa sono i Big Data

Fonte: The Center for Digital Government, 2013.

I settori da cui i dati provengono sono tra i più disparati: finanza, aereonautica, telefonia, servizi bancari e assicurativi, e-commerce, trasporti, pubblica amministrazione, salute, social network, supermercati, astronomia e meteorologia.
Alcuni esperti aggiungono una quarta "V", *Veracity* (precisione), intendendo che i dati devono esser affidabili.
Altri ancora aggiungono ulteriori "V": *Visualization* e *Value* (visualizzazione e valore), che rispettivamente suggeriscono l'importanza della presentazione dei dati e la loro rilevanza, *Vocabulary* (vocabolario), che si riferisce alla creazione di metadati, e *Venue*, che sottolinea l'importanza della geolocalizzazione.
A fornire una definizione standard rimangono le tre "V" originali (*Volume*, *Velocity* e *Variety*): i Big Data sono un enorme ammontare di dati provenienti da fonti diverse e aventi differenti formati e lunghezze che vengono estratti ed elaborati ad altissima velocità con strumentazioni differenti da quelle usate per i classici dataset costituiti da dati strutturati.

L'organizzazione internazionale UN Global Pulse elenca cinque caratteristiche che generalmente accomunano quelli che chiama i "Big Data for Development"[1], ossia i Big Data adatti a permettere in qualche modo un miglioramento nelle scelte prese dai decisori pubblici e i cui effetti ricadono sulla società:

- sono generati digitalmente e non sono digitalizzati in un secondo momento manualmente, possono quindi essere manipolati immediatamente da computer;
- sono prodotti passivamente, senza che il soggetto se ne renda conto, sono il frutto delle sue azioni quotidiane che prevedono l'uso di servizi digitali;
- sono collezionati automaticamente non appena vengono generati;
- sono tracciabili sia per quanto riguarda la loro provenienza geografica che il momento della loro generazione;
- sono analizzati continuamente ed in tempo reale.

I dati generati dalle persone in seguito a loro interazioni attraverso servizi digitali vengono anche definiti "*digital smoke signals*".

[1] UN Global Pulse, Maggio 2012. *Big Data for Development: Challenges & Opportunities*

1.2 Big Data analysis

Big Data analysis significa applicare specifici strumenti e tecniche d'analisi a dataset aventi le caratteristiche precedentemente descritte per estrarre da una moltitudine di dati di bassissimo valore e non strutturati (come per esempio una lista di tweet) un dato dall'altissimo valore informativo che può essere utilizzato per guidare un'azione concreta. Ciò richiede tecniche e strumenti, magari già sviluppati in altri campi, adatti a lavorare su questo tipo di dati e capaci di rivelare andamenti e correlazioni inter e intra dataset, che sarebbero altrimenti rimasti nascosti.
Inoltre, una caratteristica importante della Big Data analysis è il ruolo della visualizzazione che può offrire nuove prospettive sui risultati ottenuti che sarebbero difficilmente acquisibili diversamente.

Questo intenso processo di raccolta e analisi di dati socio-economici è conosciuto come "*reality mining*". Esso può esser svolto principalmente in tre modi:

- analisi ininterrotta di un flusso continuativo di dati come ad esempio la raccolta sistematica dal Web di prezzi di prodotti in tempo reale;
- assimilazione di dati semistrutturati e non strutturati online per evidenziare i bisogni, le percezioni e gli *hot topic* del momento;
- correlazione ed integrazione di un flusso di dati in tempo reale con un repertorio di dati storici accessibile in tempi più lunghi.

Si possono identificare due diversi ambiti in cui gruppi di esperti nell'analisi di grandi quantità di dati utilizzano le loro conoscenze in statistica, *data mining*, *machine learning* e visualizzazione: uno è quello aziendale in cui il fine è rispondere a domande poste dai dirigenti, l'altro è quello della ricerca accademica e scientifica.
Nel primo caso, partendo da informazioni aziendali, come ammontare delle vendite o numero di clienti, e alcune volte combinandoli con dati pubblici disponibili, come tweet o informazioni anagrafiche, si arriva a risultati il più possibile affidabili che vengono riportati ai livelli gerarchici superiori in modo che essi possano agire di conseguenza. Questo lavoro d'analisi è di tipo esplorativo e *demand-driven*, ha cioè uno specifico

obiettivo precedentemente stabilito. Limitazioni a questo modo di procedere sono la possibile non aderenza tra le fonti di dati disponibili e quelle ideali .
Nel mondo accademico, invece, gli scienziati analizzano i dati per testare ipotesi e formulare teorie, pertanto decidono autonomamente i loro temi di ricerca in base anche alle fonti a loro disposizione.

1.3 Utilizzo dei Big Data: dal settore privato a quello pubblico

I Big Data possono essere utilizzati per complesse ricerche interdisciplinari e la loro analisi sta interessando molti campi tra loro diversi. Il loro potere predittivo è stato al momento ampiamente testato in campo economico e scientifico, mentre in quello delle scienze sociali e delle politiche pubbliche sta attirando molta attenzione per la sua possibilità di offrire una migliore comprensione dei comportamenti umani.
Molti sono gli esempi di sfruttamento dei Big Data, utili a sottolineare la loro potenzialità intrinseca ed elevata versatilità: offerte di lavoro postate online sono una fonte per analizzare l'andamento del mercato del lavoro; similmente, i dati sulle transazioni finanziarie e sulle vendite online sono sempre più utilizzati per fare previsioni sull'output mondiale.
Studi hanno dimostrato che l'andamento del volume di ricerche relative al tema della compravendita di immobili in Google prevedeva in modo più accurato le vendite di case nel trimestre successivo rispetto alle previsioni fatte da economisti specializzati nel settore immobiliare (Lohr, 2012).
Ricercatori della MIT hanno stimato l'inflazione raccogliendo e analizzando giornalmente i prezzi di beni venduti o pubblicizzati nel Web, tanto che hanno lanciato una start-up in Cambridge, la PriceStats, la cui missione è quella di diventare una risorsa fondamentale per le statistiche sull'inflazione nel mondo[2].
Il fattore chiave che dà valore a questo progetto è che sul Web i prezzi sono disponibili ogni giorno mentre gli indici di prezzo al consumo sono pubblicati in molti Paesi solo su base mensile. In questo modo si possono individuare picchi d'inflazione in anticipo

[2] Cavallo, A., Maggio 2011. "BPP and PriceStats". *The Billion Prices Project @ MIT*

rispetto ai metodi tradizionali, oppure offrire nuove visioni di come le fluttuazioni di prezzo si trasmettono tra vari beni di diversi settori.
Dipartimenti di Polizia americani, primo fra tutti quello di New York, usano mappe computerizzate e analisi di variabili come i giorni di paga, eventi sportivi, acquazzoni, dati storici di arresti e ferie per predire i luoghi più probabili per il verificarsi di crimini e collocare in anticipo delle pattuglie proprio in quelle parti della città.
Ancora, riportando in una mappa il numero di donazioni ai partiti a Mahanattan, con colori differenti a seconda che siano dirette al partito democratico o a quello repubblicano, si è potuto ottenere una mappa del pensiero politico della città.
Si è persino provato a stimare il prodotto interno lordo di un Paese basandosi sulle emissioni di luce durante la notte raccolte in tempo reale da un sensore remoto (Helbing e Balietti, 2011). Disponiamo di questi dati in quantità sempre maggiore e questi riguardano non solo la nostra localizzazione e i nostri spostamenti, ma addirittura i nostri pensieri; si può fare una tomografia della società mettendo insieme a strati questi dati che dicono cosa pensiamo, cosa facciamo, dove andiamo, come interagiamo (Vespignani, 2013).

1.3.1 Utilizzo dei sensori dei cellulari

Alcune interessanti ricerche sono state condotte utilizzando i sensori collegati ai cellulari per monitorare gli spostamenti delle persone e ottenere così informazioni utilizzabili per varie finalità: dalla previsione sulla diffusione di un'epidemia all'organizzazione degli aiuti dopo un disastro naturale.
Uno studio di ricercatori della Karolinska Institute di Stoccolma e della Columbia University (Linus Bengtsson et al., 2011) ha analizzato la mobilità delle persone che risiedevano nella zona di Haiti colpita dal terremoto e dall'epidemia di colera nel 2010 utilizzando i sensori presenti all'interno di ogni cellulare.
Dopo disastri naturali la mobilità delle persone in cerca di cibo, acqua e cure mediche, può causare un aumento della diffusione di malattie come il colera, e ciò rende più difficile sia raggiungere le perone in stato di bisogno per poterle aiutare, sia arrestare il diffondersi dell'epidemia poiché se non si sa la localizzazione delle perone infette risulta impossibile curarle. Non esistendo un metodo rapido ed accurato per tracciare gli

spostamenti delle persone dopo disastri naturali e poter prevedere i potenziali focolai di epidemie, questi studiosi hanno tentato di utilizzare la posizione delle carte SIM della maggiore compagnia telefonica di Haiti, la Digicel, per stimare la grandezza ed il trend degli spostamenti della popolazione in seguito al terremoto e all'esplosione del colera. Ogni volta che una chiamata partiva da una carta SIM, veniva registrato quale trasmettitore inoltrava la chiamata, in questo modo veniva fornita la posizione geografica delle persona che effettuava la telefonata. Linus Bengtsson e gli altri studiosi che hanno collaborato al progetto hanno seguito la posizione di 1.9 milioni di carte SIM per i 42 giorni precedenti il terremoto e i 158 seguenti. Hanno stimato che 630,000 persone, collegate a 197,484 carte SIM Digicel, presenti a Port-au-Prince il giorno del terremoto nei 19 giorni successivi se ne erano andate, corrispondenti al 20% della popolazione della città prima del cataclisma.

I dati sugli spostamenti geografici ottenuti analizzando i sensori dei cellulari sono stati confermati da un sondaggio condotto dalle Nazioni Unite.

Inoltre, per dimostrare la flessibilità di tale metodologia d'analisi della mobilità, i ricercatori hanno anche registrato in tempo reale gli spostamenti di 140.000 SIM durante i primi otto giorni dello scoppio dell'epidemia di colera, iniziata vicino a Saint-Marc, e hanno potuto comunicare ai soccorsi entro 12 ore dove le persone erano andate e quindi, dove potevano nascere altri focolai dell'epidemia. I risultati di questo studio hanno confermato che i movimenti delle popolazioni durante epidemie o catastrofi naturali possono essere stimati rapidamente e in modo attendibile, utilizzando i sensori collegati ai telefoni cellulari, in aree che presentano un elevato utilizzo di questi apparecchi.

Lo studio suscitò particolare interesse e comparve non solo nel PLOS Medicine, giornale di medicina, dove venne pubblicato nella sua interezza, ma anche in testate giornaliste rivolte al grande pubblico come il New York Times e la BBC News.

Anche in Messico durante l'epidemia dell'influenza H1N1 del 2009 sono stati usati i sensori dei cellulari per misurare l'impatto di allerte governative relativamente al tema sulla mobilità della popolazione e sulla diffusione della malattia. I registri analizzati dal primo gennaio al 31 maggio 2009 hanno mostrato come le campagne informative del governo avessero ridotto la mobilità dal 10 al 30% e posticipato di 40 ore il picco dell'epidemia riducendo il numero degli infetti del 10% (vedi Figura 1.3).

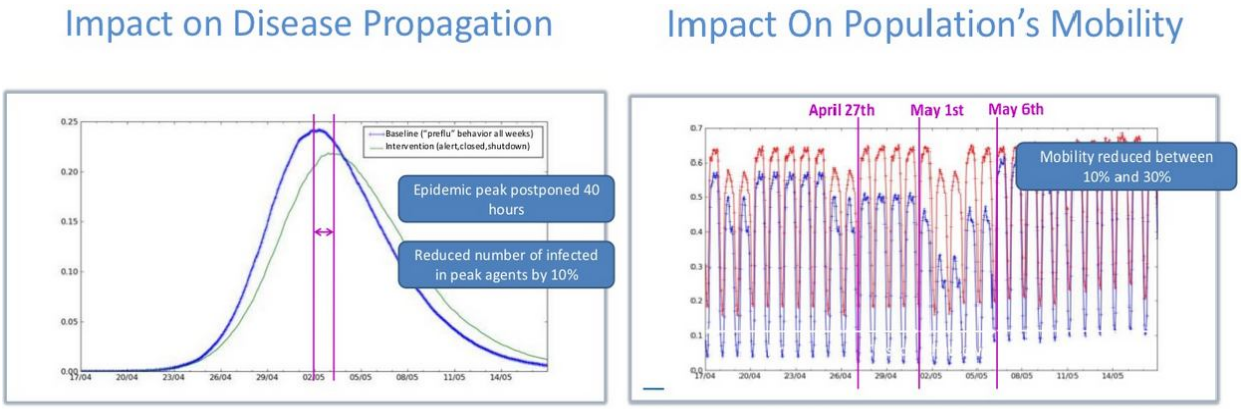


Figura 1.3: Grafici sull'impatto delle allerte governative sulla mobilità della popolazione e sulla propagazione della pandemia in Messico durante l'epidemia dell'influenza H1N1 del 2009
Fonte: UN Global Pulse, 2012

Ancora, durante la notte della musica a Parigi, sfruttando i sensori degli smartphone, si è potuto capire, associando ogni telefonino ad un soggetto, come le persone si muovevano per la città e questo è servito a monitorare il traffico e i luoghi di principale interesse durante la manifestazione. Lo si è potuto fare sfruttando solamente il telefonino che tutti ormai portano con sé, utilizzando quindi un dato che è nato per un'altra finalità ma che è in grado di rispondere a questo e ad altri scopi. (Vespignani, 2013)

Partendo dai dettagli delle chiamate (CDRs), preventivamente rese anonime, si può arrivare a dedurre variabili comportamentali ed indicatori relativamente a:

- Consumi, attraverso per esempio il numero di sms, mms o chiamate e relativi durata e costo;
- Mobilità;
- Società, correlando ad esempio la frequenza delle comunicazioni ed il peso dei contatti con il diverso uso che uomini e donne o adulti e ragazzi fanno del cellulare;
- Censo.

Si possono correlare, ad esempio, alti livelli di mobilità ad alti livelli socioeconomici (SELs). In questo modo utilizzando i registri di dati, ottenuti dai sensori collegati ai cellulari, è possibile creare anche mappe di censo risparmiando notevolmente tempo e denaro. Gli istituti statistici nazionali svolgono sondaggi per ottenere dati relativi al

livello socioeconomico di una zona solo in alcune sub-regioni e poi deducono il livello dell'area complessiva, utilizzando i registri delle chiamate e l'uso dei cellulari è invece possibile fare predizioni sui livelli socioeconomici delle sub-regioni non sottoposte ai sondaggi per avere delle stime complessive migliori, riducendo così i costi ingenti che un sondaggio di tutta l'area comporterebbe e aiutando i decisori pubblici a procedere nelle loro scelte strategiche.
Analisi di questo tipo, che qualche anno fa sarebbero state possibili solo impiegando ingenti risorse umane ed economiche, oggi sono realizzabili analizzando un enorme ammontare di dati digitali generati in seguito ad attività umane svolte per altri scopi.

1.3.2 Casi nel settore economico

In campo economico da tempo si stanno sfruttando i Big con ottimi risultati.
Andrew McAfee, scienziato ricercatore alla *MIT's Center for Digital Business*, durante il suo intervento ad una conferenza mondiale, la *SAS Global Forum Executive Conference*, nell'aprile 2013, afferma che escludendo l'avvento di Internet, i Big Data stanno avendo il più grande impatto sull'economia rispetto ad ogni altro evento.
Durante il *GigaOm Structured Data* di marzo 2013, una serie di conferenze tenute a New York a cui hanno partecipato più di 900 esperti per esaminare come i Big Data possono contribuire al successo di un business, Paul Maritz, direttore generale della EMC Corporation, ha affermato che le aziende *data-driven*, che basano cioè la loro politica strategica su una preventiva analisi di dati disponibili in tempo reale, stanno creando nuove forme di competizione. Ciò che è determinante non è tanto la tecnologia quanto il modello di business adottato.
I Big Data possono essere usati da un'azienda per predire trend economici futuri, per creare nuovi prodotti, per aumentare i ricavi o per migliorare la propria offerta adattandola alle esigenze del cliente in tempo reale.
Basta citare qualche esempio: Zynga, azienda fornitrice di giochi online, studia i dati relative a come i suoi utenti giocano al fine di aggiornare i suoi giochi in tempo reale per accontentare sempre più chi li utilizza; Ken Rudinm vicepresidente del dipartimento che si occupa di analisi dell'analisi dei dati afferma "Noi siamo un'azienda di analisi travestita da azienda di giochi online".

Amazon, invece, formula i propri consigli per gli acquisti elaborando le statistiche su quali libri vengono spesso acquistati insieme ad altri, ed in questo modo ottiene il 30% dei suoi proventi.

Allo stesso modo Netflix, utilizzando l'algoritmo CineMatch, consiglia ai suoi utenti i film da vedere online, a partire dalle preferenze di altri iscritti con interessi affini.

Distributori, come Walmart e Kohl's, analizzano dati sulle vendite, demografici, meteorologici e andamento dei prezzi per personalizzare i prodotti presenti nei diversi punti e per determinare le tempistiche dei ribassi di prezzo. Compagnie navali, come la U.P.S., analizzano dati sui tempi delle consegne via terra e sul traffico stradale e marittimo per definire al meglio i loro percorsi.

I siti d'incontri, come Match.com, costantemente setacciano le loro liste Web di profili e di caratteristiche personali per migliorare l'algoritmo che utilizzano per organizzare appuntamenti tra uomini e donne che non si conoscono.

Il progetto EconoMining dell'Università di New York ha evidenziato il valore economico degli *user-generated content*, cioè dei contenuti online creati dagli utenti stessi. Le opinioni espresse in Rete dai consumatori possono influenzare i volumi delle vendite, i giudizi su eBay o Amazon creando la reputazione online dei venditori incidono sulle loro possibilità di vendita; tutto ciò significa che gli *user-generated content* hanno un impatto economico e quindi un loro valore intrinseco. Ovviamente è il contenuto che deve interessare più che il totale delle recensioni online del prodotto, quindi è fondamentale decomporre il testo e comprenderlo per avere un input importante durante la definizione del prezzo.

Numerosi sono gli esempi di start up che nascono da idee imprenditoriali che sfruttano i Big Data, come Empatica che ha progettato e creato un braccialetto per il monitoraggio diretto di alcuni parametri fisiologici. Questo bracciale raccoglie continuamente moltissime informazioni dal nostro corpo, come il battito cardiaco, e le salva in un archivio personale online così da poter metter in luce problematiche nel loro nascere, individuare cambi di abitudini consigliabili per migliorare il proprio benessere fisico e fare confronti nel tempo del proprio stato di salute.

Dopo che ad Helsinki è stato possibile accedere ad informazioni in tempo reale sul trasporto pubblico, una delle prime città in cui questo tipo di informazioni è stato reso

disponibile, alcune imprese innovative hanno creato applicazioni e pagine web dove osservare gli spostamenti dei mezzi in diretta, come in una mappa vivente.
Molti altri progetti sono in fase di sviluppo e molte sono pure le iniziative che supportano la nascita di start up che utilizzano tecnologie innovative applicate a dati digitali, come l'Open Data Institute (Odi), un acceleratore per start up che avviano iniziative imprenditoriali sostenibili utilizzando i dati open source, con sede in Inghilterra, fondato da Tim Berners Lee, già promotore del Web Index.
Una ricerca pubblicata nel 2011 da Erik Brynjolfsson in collaborazione con altri della MIT durante la quale sono state studiate 179 grandi aziende americane, ha rivelato che le imprese *data-driven* hanno un tasso di produttività maggiore di circa il 5/6% rispetto a quelle che non lo fanno.
Uno dei temi principali del Gartner Customer 360Summit, annuale conferenza in cui ricercatori ed esperti di tecnologia e di business si confrontano sulle ultime innovazione nel campo del CRM, di maggio 2013 a San Diego, è stato lo sfruttamento da parte delle aziende dei Big Data che esse possono estrarre dalle loro attività di vendita, marketing e servizio al consumatore. Durante la conferenza è emerso che l'ammontare di dati che le aziende raccolgono, conservano ed analizzano sta aumentando del 40% ogni anno e che, entro il 2016 i ricercatori di Gartner si aspettano che le spese in IT per sfruttare i Big Data saranno di circa 232 miliardi di dollari. Michael Maoz, vicepresidente alla Gartner Research, ha fatto notare che al momento solo il 2/3% delle organizzazioni ha tra gli addetti al servizio al cliente un soggetto esperto nell'analisi di Big Data, ma che circa il 54% delle aziende identifica il *data mining* e il *data analysis* come una delle loro principali obiettivi futuri.
L'analisi dei dati, quindi, sta diventando sempre più una competenza fondamentale per il successo delle aziende.

1.3.3 Casi nel settore sanitario

Particolare attenzione merita il settore della salute poiché rispetto ad altri, e soprattutto rispetto a quello economico, presenta meno casi studio in cui l'analisi dei Big Data è stata applicata e, quindi, le sue potenzialità non sono ancora state appieno comprese ed

adeguatamente sfruttate. Le ricerche già svolte però confermano la possibile utilità del loro utilizzo da parte dei responsabili di decisioni in ambito di pubblica sanità.

La prima analisi di dati in tempo reale con la finalità di prevedere la diffusione di un virus è stata fatta nel 2009 con l'influenza H1N1.

I nove ricercatori partecipanti al progetto, applicando il GLEAM, un modello di simulazione stocastico globale di diffusione di un'epidemia, integrato a tre tipologie di dati, sono riusciti a creare nei computer un mondo sintetico e a prevedere come, dove e quando il virus dell'H1N1 si sarebbe diffuso in 220 Paesi. La prima tipologia di dati integrati riguardava la distribuzione della popolazione mondiale e si basava sul database progettato dal centro di applicazione dei dati socioeconomici (SEDAC) della NASA, il "*Gridded Population of the World*". Relativamente alla mobilità analizzando sia i viaggi aerei globali, ottenuti dai database dell'International Air Transport Association e dell'Official Airline Guide, sia i trasferimenti giornalieri tra località adiacenti, è stato simulato il numero di passeggeri che giornalmente viaggiavano. La terza tipologia di dati inseriti nel modello, infine, riguardava l'approccio comportamentale specifico per la malattia studiata.

Il modello ha fornito una stima della trasmissione potenziale stagionale durante la fase iniziale della pandemia H1N1 e ha generato una previsione dei picchi di contagio nell'emisfero nord nell'autunno-inverno 2009/2010. I risultati sono stati convalidati da dati empirici raccolti in 48 Paesi con metodi di sorveglianza tradizionali, sia per quanto riguarda i picchi della pandemia, sia per quanto concerne la localizzazione spaziale dei contagi.

I metodi di sorveglianza tradizionali solitamente si basano sul numero di individui che presentano i sintomi e ricercano cure, ciò porta ad una sottostima dei casi clinici perché alcuni non ricercano cure mediche. In ogni caso i risultati di questo tipo di sorveglianza sono sempre troppo lenti per permettere un aumento delle vaccinazioni nelle regioni in cui si verificano i picchi e dove intervenendo si potrebbe accelerare il declino della pandemia.

Il risultato dello studio ha incoraggiato l'uso di approcci computazionali di grandi database per stimare in tempo reale picchi di contagi di epidemie, sottolineando comunque che la qualità delle previsioni dipende sempre dalla qualità e di dati a disposizione.

Un altro caso che merita di esser riportato è uno studio condotto dalla Microsoft Research riguardante le ricerche effettuate con la parola chiave “stop smoking” nel Regno Unito nel 2012 nel motore di ricerca Bing.

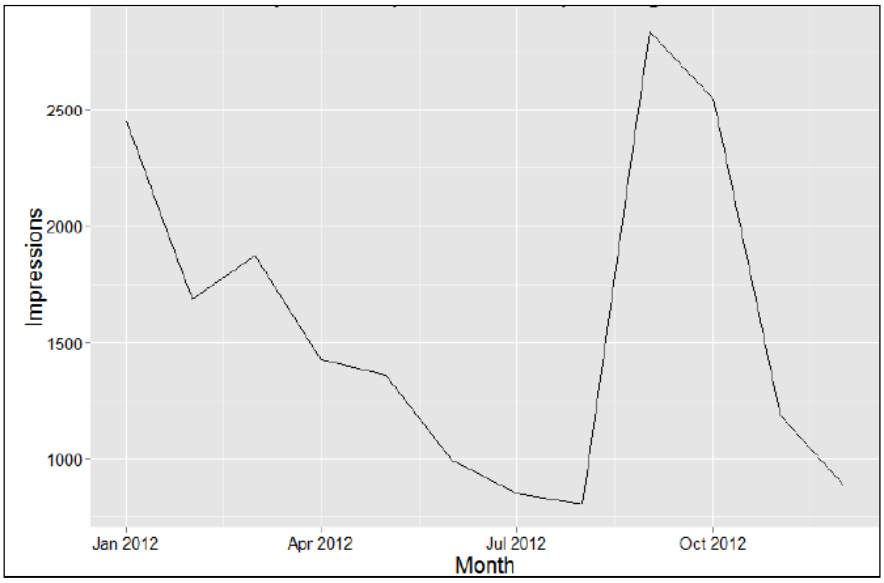


Figura 1.4: Grafico relativo al numero di ricerche fatte in Google nel 2012 nel Regno Unito per la parola chiave “stop smoking”
Fonte: Microsoft Research-WHO-UN Global Pulse, 2013

Come mostra il grafico riportato in Figura 1.4, in Ottobre è stato registrato un picco considerevole di persone che cercavano online informazioni riguardanti l’argomento smettere di fumare; la domanda che sorge spontanea è: perché proprio in Ottobre molti inglesi hanno pensato alla possibilità di smettere di fumare, o per lo meno hanno voluto informarsi su questo tema? Con estrema probabilità l’aumento in quel mese dell’anno era stato causato da una campagna di sensibilizzazione denominata Stopoctober, avviata dall’autorità nazionale inglese competente nell’ambito della salute pubblica, la *National Health Service* (NHS). I dati raccolti hanno testimoniato come quell’iniziativa di marketing sociale abbia registrato ottimi risultati, almeno di termini di “ascolto”, poiché ha indotto molte persone per lo meno a cercare informazioni su come fare per smettere di fumare.

Un altro studio, sempre relativo all’analisi delle parole chiave inserite nei motori di ricerca, svolto nel 2009, ha rilevato un picco nelle ricerche in Google con termini legati all’influenza, come “sintomi influenzali” e “trattamenti dell’influenza”, esattamente

prima di un aumento nel numero di pazienti che si rivolgevano al medico perché contagiati dal virus dell'influenza. Monitorando le query poste ai motori di ricerca online si è riusciti a calcolare il livello settimanale di contagi d'influenza in ogni stato degli USA, con uno scarto temporale pari ad una giornata. La conclusione dello studio è stata che le ricerche online possono rilevare epidemie d'influenza in aree con un'elevata percentuale di utilizzatori del Web (Ginsberg et al., 2009). Dato che con un repentino rilevamento dell'inizio di un'epidemia si possono metter in atto misure idonee a ridurne la diffusione, nuove e più veloci modalità per identificare l'inizio di un'epidemia e stimarne l'andamento possono risultare molto utili in campo medico.

Il successo di questa modalità di prevedere la diffusione di un virus è stato confermato dal progetto avviato da Google, chiamato "Google Flu Trends", ossia una pagina Web (www.google.org/flutrends/intl/en_gb/) in cui, aggregando le ricerche, rese anonime, fatte su Google dagli utenti interessati ad avere informazioni sanitarie sui sintomi influenzali, viene stimato il numero di persone realmente contagiate dal virus dell'influenza nelle varie parti del mondo.

I risultati si sono rivelati essere molto vicini agli indicatori forniti dai sistemi di sorveglianza tradizionali dell'attività influenzale forniti pubblicamente dagli U.S. Centers for Disease Control (vedi Figura 1.5). Sembra, quindi, che utilizzando la frequenza di query online relative a sintomi influenzali si possa stimare con buona approssimazione i livelli di diffusione dell'influenza in diverse parti del mondo.

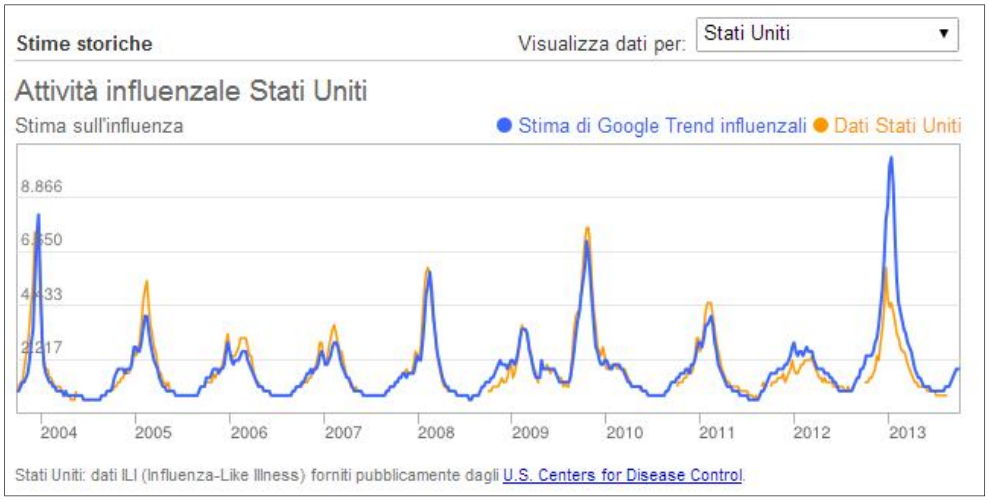


Figura 1.5: Confronto tra le stime dell'attività influenzale negli Stati Uniti ottenute aggregando le query in Google e i dati ufficiali dal 2004 al 2013

Fonte: www.google.org/flutrends

Naturalmente i risultati passati non sono una garanzia di validità per risultati futuri, ma quella appena presentata potrebbe esser considerata una metodologia complementare, considerando oltretutto che la sorveglianza tradizionale dell'attività influenzale presenta stime aggiornate settimanalmente, mentre “Google Flu Trends” viene aggiornato ogni giorno.

“Google Dengue Trends” funziona allo stesso modo, relativamente al fenomeno della febbre esotica. Ovviamente non tutti gli utenti che digitano "febbre esotica" nei motori di ricerca sono realmente malati, ma esiste una stretta correlazione tra quanti cercano informazioni relative ai sintomi della febbre esotica e il numero di persone che li manifestano effettivamente. La schermata iniziale della pagina Web di “Google Dengue Trends” riporta una cartina colorata; più la gradazione della colorazione delle diverse zone del mondo è intensa, più sono state le ricerche in Google su questo tema e quindi si presume si registreranno più casi della malattia.

In conclusione sembra che basandosi su un’analisi delle query nei motori di ricerca si riesca a delineare un modello attendibile di un fenomeno reale.

Appare corretto menzionare che, seppur pochi, ci sono stati casi in cui le stime si sono rivelate non molto accurate, come nel Gennaio 2013 quando Google Flu Trend aveva sovrastimato la diffusione del virus influenzale. In quel caso la giustificazione è stata trovata in un avvenimento esterno che aveva alterato il comportamento di ricerca degli utenti online: quando a New York è stato dichiarato lo stato d’emergenza sanitario a causa della forte epidemia influenzale, probabilmente molte ricerche online relativamente al virus H1N1 sono state effettuate anche da parte di chi non riportava alcun sintomo (Butler, 2013).

Bisognerebbe, infatti, “ricalibrare ogni anno l’algoritmo utilizzato”, come suggerisce John Brownstein, epidemologo alla Harvard Medical School di Boston.

Gli esempi precedentemente riportati dimostrano come anche nel settore della sanità sfruttare i Big Data si può rivelare estremamente utile. Ecco che organizzazioni internazionali si stanno affacciando al mondo dei Big Data, come l’OECD e UN Global Pulse, iniziativa lanciata nel 2009 dal Segretario Generale delle Nazioni Unite. Essa ha lo scopo di sfruttare le nuove fonti di dati digitali e le innovative tecnologie per un miglioramento globale aiutando i decisori pubblici a comprendere meglio ed in tempo reale cambiamenti collettivi di comportamento e crisi emergenti. In questo modo si

rende possibile sia avviare campagne per la prevenzione di danni alla salute, sia organizzare interventi mirati e progettati ad hoc per intervenire il prima possibile e nel modo migliore.

L'organizzazione spera in un futuro in cui l'accesso rapido ad informazioni migliori permetta di proteggere ed aiutare le popolazioni più vulnerabili e renda migliori gli interventi pubblici in contrasto alle crisi mondiali utilizzando i segnali digitali come primi campanelli d'allarme per organizzare in anticipo programmi di assistenza creati ad hoc.

UN Global Pulse lavora in collaborazione con la World Health Organitazion (WHO), l'autorità che coordina e dirige il sistema della salute pubblica all'interno delle Nazioni Unite, per esplorare come applicare i Big Data nello specifico al campo delle Non Communicable Diseas (NCDs). Queste sono malattie non trasmissibili generalmente di lunga durata e lenta progressione; i quattro tipi principali sono le malattie cardiovascolari, quelle respiratorie croniche, il cancro ed il diabete. Secondo i dati WHO queste patologie colpiscono soprattutto i Paesi con un reddito medio-basso, sono la principale causa di decesso in ogni continente, eccetto che in Africa, ed uccidono ogni anno circa 36 milioni di persone (WHO, 2013).

L'incidenza delle NCDs, rispetto ad altre tipologie di malattie, risente molto più profondamente degli interventi sociali, basta pensare a come nelle Mauritius facendo cambiare l'olio da cucina si è ottenuta una notevole diminuzione dell'incidenza di malattie cardiovascolari, oppure a come l'aver proibito di fumare all'interno di ambienti pubblici abbia fatto diminuire le ammissioni in ospedale per problemi respiratori. Ecco che allora, per quanto riguarda questa tipologia di malattie, l'applicazione di analisi di Big Data sembra promettere grandi possibilità per poter intervenire in modo notevolmente migliore grazie ad una comprensione molto più rapida dei fenomeni che al momento non è permessa, se non con un lasco di tempo che supera l'anno, il che comporta ritardi nell'avvio di azioni per ridurre le NCDs. Mentre i comportamenti collegati a queste patologie registrano continui e rapidi cambiamenti nel breve periodo, i dati che sono utilizzati per prendere decisioni su questo tema si riferiscono a periodi precedenti. Lo scopo è trovare fonti di dati a cui avere facile accesso a basso costo, capaci di far emergere cambiamenti di comportamento nell'arco di settimane e non anni, ed utilizzabili per evidenziare collegamenti significativi tra cause ed effetti.

Un esempio è lo sfruttamento delle tracce digitali lasciate dagli acquisti nei supermercati registrati dalle carte fedeltà: questi possono dare informazioni circa i modelli di alimentazione più diffusi e si potrebbero creare dei collegamenti significativi tra la dieta tipica della popolazione e l'insorgere del diabete. I risultati potrebbero essere migliori di quelli ottenibili attraverso questionari, molto più costosi e che richiedono molto più tempo per esser resi disponibili a chi potrebbe utilizzarli per creare campagne di marketing sociale o altre attività di prevenzione.

Alcuni dati, come quelli sui modelli di consumo raccolti nei supermercati a cui si è fatto riferimento, potenzialmente utili per la sorveglianza di NCDs sono però difficili da ottenere, sia per motivi legati alla privacy, sia perché chi li detiene, sapendone il valore, non li vuole condividere, o per lo meno non gratuitamente.

Ci sarebbero però modalità di condivisione che non intaccherebbero né la privacy né risulterebbero lesive della competitività tra aziende:

- fornire ai ricercatori dati aggregati, che non permettono di derivare quelli individuali, non rendendo così necessario un accordo di rivelazione da parte di coloro che li hanno generati;
- permettere ai ricercatori di lavorare all'interno delle aziende private ed analizzare dati, anche in questo caso senza la necessità di un *disclosure agreement*;
- fornire ai ricercatori dati provenienti da più aziende dello stesso settore per proteggerne la competitività senza intaccare la possibilità di analisi;
- fornire ai ricercatori indicatori utili, creati dalle compagnie che analizzano i dati per conto loro.

Risulta quindi necessario, innanzitutto, trovare un modo per aggregare questi dati armonizzando le definizioni ed i codici ed anonimizzarli, per proteggere sia gli individui che le organizzazioni coinvolti, e soprattutto per convincere i loro detentori a condividerli.

Il settore privato lavora ormai quotidianamente con dati real-time, perché anche il settore pubblico non può fare lo stesso?

Come il settore privato utilizza tecnologie innovative per analizzare i loro dati commerciali per capire meglio i gusti dei clienti, identificare nuovi mercati e

prendere decisioni d'investimento, così anche i decisori pubblici potrebbero utilizzare i numerosi dati digitali che sono continuamente creati ed i nuovi metodi e tecniche di analisi che si stanno sviluppando per svolgere i loro compiti e creare un miglioramento nel benessere della collettività. Quest'opportunità risulta ancora più incisiva ed allettante se si pensa che il 75% delle decisioni in campo medico sono prese senza essere in possesso dei dati rilevanti. (UN Global Pulse, World Health Organization, 2013)

Partendo dall'ipotesi che quando le persone iniziano ad affrontare una crisi cambiano il loro modo di comportarsi, e quindi anche come utilizzano i servizi digitali a loro disposizione, se questi cambiamenti vengono colti in breve tempo possono essere estremamente d'aiuto per allertare i decisori pubblici e suggerire azioni in risposta.

Una consapevolezza in tempo reale dello stato di una popolazione o dell'efficacia di un'azione politica potrebbe portare ad un approccio più adattivo e pronto da parte dei pubblici poteri, contribuendo allo sviluppo internazionale.

Compagnie assicurative già utilizzano enormi quantitativi di dati per fare profilazione dei loro clienti e modelli predittivi utili ai loro scopi commerciali, lo stesso si potrebbe fare in ambito di medico-sanitario, anzi addirittura quegli stessi dati potrebbero essere utilizzati pure per propositi legati alla salute pubblica.

Secondo Mark Headd, direttore del dipartimento dati della città di Philadelphia "coi Big Data si tratta di estrarre grandi idee e grandi visioni dai dati che possono aiutar i governi a prendere decisioni strategiche migliori".

Le 3 categorie di opportunità offerte dalla Big Data Analysis applicata alle decisioni pubbliche sono, secondo UN Global Pulse:

- *Early warning*: il rilevamento veloce di anomalie all'inizio di una crisi per permettere azioni preventive;
- Consapevolezza real-time: una maggiore granularità nella rappresentazione della realtà permette programmi politici e norme meglio targetizzati e progettati;
- Real-time feedback: comprensione dell'efficacia di un'azione in tempo reale per render possibili lievi aggiustamenti, in risposta a quanto rilevato, prima di arrivare al risultato finale, che così sarà migliore.

Il capitolo ha messo in luce come l'applicazione della Big Data Analysis al mondo della pubblica sanità, ma anche agli altri ambiti precedentemente discussi, offre interessanti modi di essere sfruttata. Il presente elaborato continuerà però focalizzandosi solamente sui dati provenienti dai social network, tralasciando lo sfruttamento di altre tipologie di fonti.

2. Social Media

Una rete sociale è una struttura formale o informale di un gruppo di individui connessi tra loro attraverso relazioni; la definizione data dall'antropologo J.A. Barnes è la seguente: "insieme di punti congiunti da linee. I punti rappresentano le persone o anche gruppi e le linee indicano quali persone stanno interagendo con ogni altra". Le relazioni possono essere implicite o esplicite, e possono avvenire non solo nel mondo reale ma anche in Rete.

I network sociali online si sviluppano attraverso i social media, applicazioni *internet-based* costruite sui presupposti ideologici e tecnologici del Web 2.0, nelle quali gli utenti possono creare e condividere contenuti di cui essi stessi sono gli autori; è proprio quest'ultima caratteristica che li differenza dai media tradizionali (Kaplan e Haenlein, 2010).

Si possono individuare sei diverse tipologie di social media (blog e microblog, siti di *social networking*, mondi virtuali di gioco e sociali, progetti collaborativi e *content community*), e non tutte permettono di creare al loro interno una rete sociale. Wikipedia, ad esempio, è un tipo di social media definito 'progetto collaborativo' perché coinvolge gli utenti a lavorare insieme per creare dei contenuti che poi saranno disponibili in Rete a chiunque sia interessato; esso non permette però di creare una relazione tra i collaboratori.

Un social media per essere anche social network deve soddisfare le seguenti condizioni:

- ci devono essere utenti specifici;
- gli utenti devono essere tra loro collegati;
- deve esser possibile una comunicazione bidirezionale tra di essi.

(Ceron, Curini, Iacus, 2014)

Le condizioni minimali appena citate fanno dedurre che i social network producono relazioni e contenuti. Le reti sociali interne al social network possono essere preesistenti o nascere proprio attraverso esso. I contenuti creati, condivi e scambiati sono testi, video, foto, applicazioni e quant'altro, lasciando molta libertà all'utente di esprimere la propria personalità ed interagire con chi condivide i suoi interessi, passioni ed attività.

Quando successivamente nell'elaborato si parlerà di social media, si farà riferimento al sottoinsieme dei social network.

Di seguito si propone ora una breve descrizione delle piattaforme sociali più diffuse.

Facebook è il più datato, creato nel 2004 ma accessibile per i primi due anni solo agli studenti universitari americani.

Gli utenti possono condividere sul loro profilo personale foto, video, testi o audio. Per accedere ai profili altrui e interagire con altri utenti attraverso messaggi privati o attraverso l'espressione di apprezzamenti, utilizzando la funzionalità 'like', si deve prima diventare 'amici', attraverso accettazione reciproca; fanno eccezione i profili resi pubblici.

Figura 2.1: Logo Facebook

Fonte: www.facebook.com

Twitter lanciato nel 2006 da Jack Dorsey permette di interagire 'twittando', ossia postando sulla propria *timeline*, un messaggio di massimo 140 caratteri, chiamato 'tweet', tradotto cinguettio, richiamando l'uccellino che compare nel logo (vedi Figura 2.2).

Figura 2.2: Logo Twitter

Fonte: Twitter, 2014

Gli aggiornamenti di un profilo compaiono nella homepage a tutti gli utenti che si sono registrati come seguaci, 'follower', dell'utente che 'twitta'. In questo caso, a differenza di Facebook, non c'è bisogno dell'autorizzazione a diventare 'follower', a meno che il profilo Twitter non sia stato reso privato, caso che avviene solo per il 10% delle

registrazioni. Gli utenti possono seguire chi vogliono senza vincolo di reciprocità, anzi si può pure avere accesso ai tweet di utenti che non si seguono poiché Twitter per impostazione è una piattaforma aperta. E' possibile inoltre menzionare un utente in un messaggio pubblico anche senza che sia un seguace, semplicemente inserendo il nome dell'utente preceduto dal simbolo @ (vedi Figura 2.3).

I contenuti testuali sono catalogati per argomento inserendo gli 'hashtag', ossia parole o frasi senza spaziatura precedute dal simbolo #; così facendo si crea un immediato collegamento a tutti gli altri tweet postati che trattano quel tema e che hanno inserito lo stesso hashtag. Un'altra azione permessa è il 'retweet', ossia inoltrare un messaggio scritto da altri ai propri seguaci, in questo modo un tweet si può diffondere attraverso il network degli utenti che seguono l'autore.

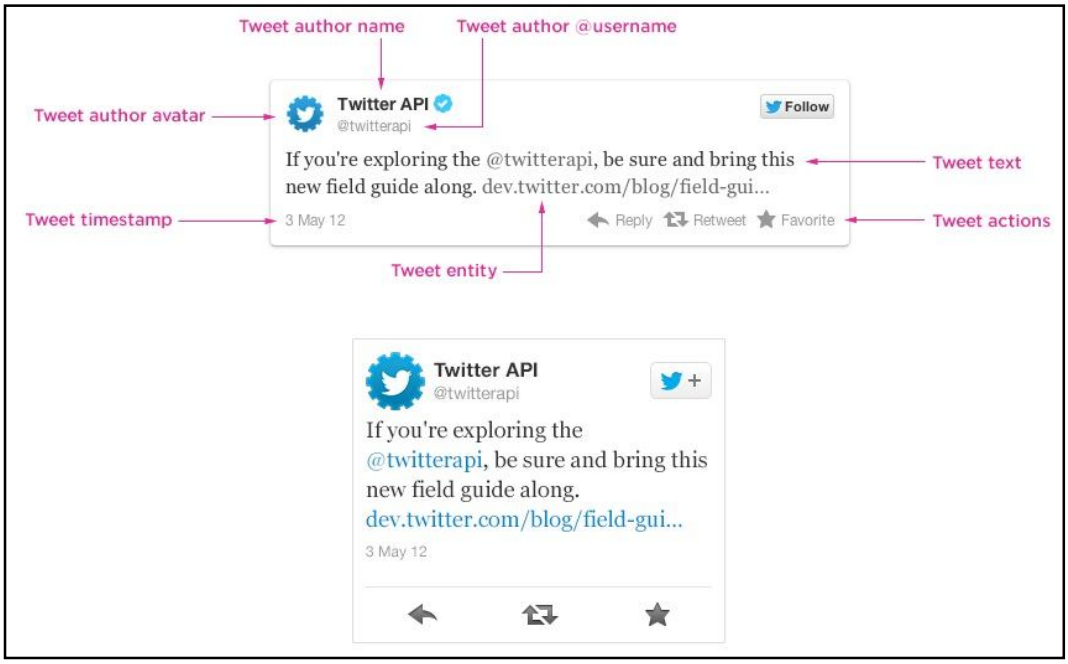


Figura 2.3: Immagine esemplificativa della terminologia relativa ad un tweet

Fonte: www.twitter.com, 2014

Twitter è considerato comunque un network che genera microblogging (Ceron, Curini, Iacus, 2014), fenomeno definito come una forma di blog che permette di pubblicare in Rete brevi aggiornamenti di testo o immagini che diventano visibili a tutti o solo a determinati contatti facenti parte della propria rete sociale.

Google+, creato da Google nel 2011, offre l'opportunità di ulteriori contenuti multimediali rispetto ai due social network precedentemente descritti, come gli 'hangouts', descrivibili come dei video ritrovi.

Google+ permette di creare collegamenti con altri utenti senza che essi diano il consenso, assegnandoli a diversi circoli, come famiglia o amici, e personalizzando quali contenuti condividere con i diversi circoli; è quindi una via di mezzo tra Twitter e Facebook per quanto riguarda l'interazione con gli altri utenti.

Linkedin ha lo scopo di sviluppare contatti professionali. Attraverso il proprio profilo personale si presentano le proprie caratteristiche rilevanti in ambito lavorativo e si creano connessioni con utenti che si reputano importanti nel proprio campo professionale per collegarsi poi ad altri soggetti interessanti per trovare od offrire un lavoro.

Pinterest è nato nel 2010, il nome deriva dalla parola inglese *pin*, che significa appendere, proprio perché in esso vengono condivise foto, immagini e video raccolte in base a tematiche precise paragonabili a post-it appesi in una bacheca reale a tema. Non si possono pubblicare testi, esclusi quelli che fungono da didascalia alle immagini, da titoli per le categorie o da commenti. Gli utenti possono anche condividere contenuti pubblicati da altri attraverso il tasto "*pin it*". Anche in questo social network si possono seguire altri utenti per poter visualizzare quello che loro pubblicano nelle loro bacheche senza doverne avere il permesso.

Flickr, di proprietà di Yahoo!, è una libreria in cui ogni minuto i suoi iscritti condividono 2.000 foto personali con tutti gli utenti internet.

Ed infine tra i più emergenti c'è **Instagram**, un'applicazione con cui si possono applicare filtri a foto appena scattate e condividerle in altri social network. Si può commentare ed esprimere il proprio apprezzamento per foto altrui condivise e seguire altri utenti, anche in questo caso senza averne il consenso.

2.1 Diffusione di Internet e dei social network

Secondo il report "Global Digital Statistics" di Gennaio 2014 gli utenti Internet nel mondo sono 2.484.915.152, pari al 35% della popolazione totale, di questi il 75%, corrispondente a 1.856.680.860 persone, ha un account su un social network.

La maggior penetrazione di Internet si registra nel Nord America (81%) seguita dall'Europa occidentale (78%), mentre la minore in India (12%) (vedi Figura 2.4), come numero assoluto di utenti primeggia invece l'Asia (oltre il miliardo).

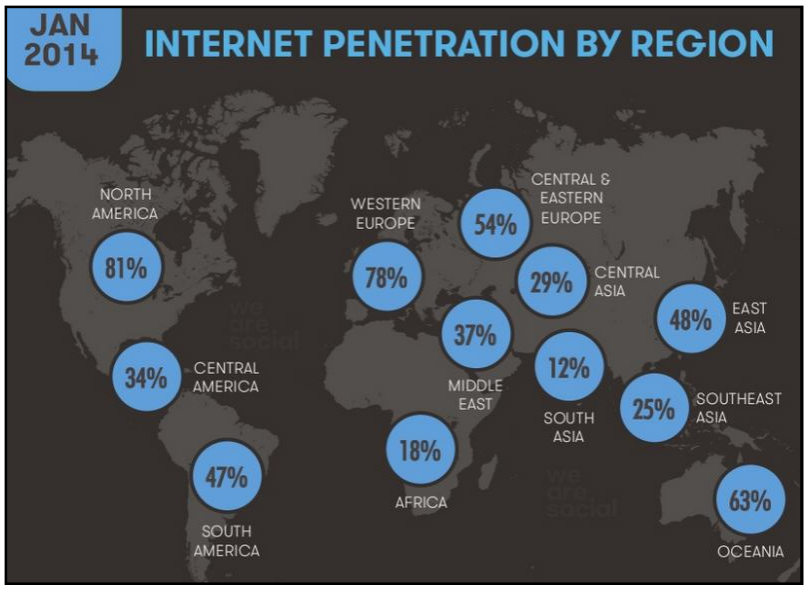


Figura 2.4: Penetrazione di Internet per Regione nel mondo a Gennaio 2014
Fonte: Global Digital Statistics, 2014

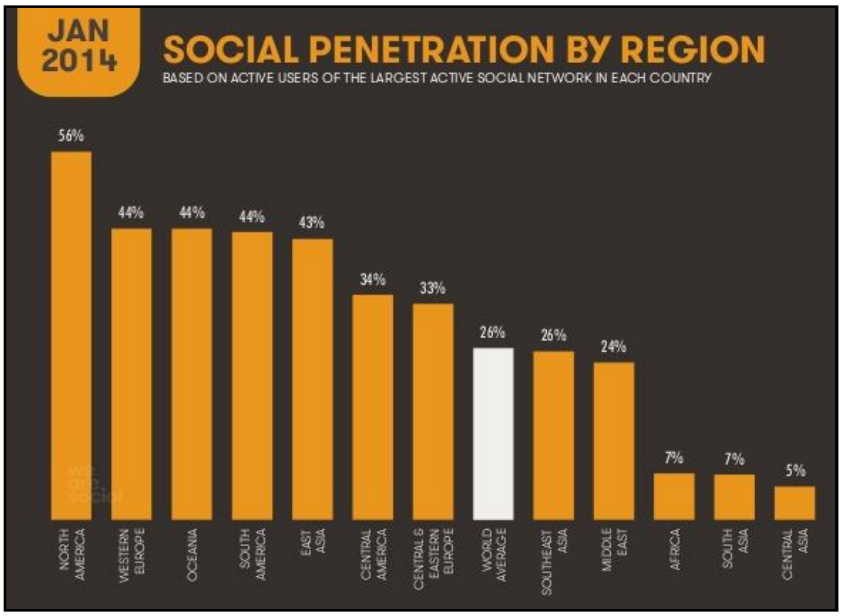


Figura 2.5: Penetrazione dei social network per Regione nel mondo a Gennaio 2014
Fonte: Global Digital Statistics, 2014

La Figura 2.5 riporta la penetrazione dei social network per Regione, anche in questo caso la maggiore si registra nel Nord America (56%), mentre la minore in Asia Centrale (5%); in Europa il 44% della popolazione ha un profilo personale in una delle principali piattaforme sociali.

La media mondiale di tempo speso online è di 7 ore di cui 4,8 tramite computer fisso o laptop e 2,1 tramite altri dispositivi mobili; del totale delle ore passate in Internet 2 sono dedicate alle attività sui social network[3].

Per quanto riguarda il mondo dei social network le statistiche del GlobalWebIndex pubblicate a Gennaio 2014 riconfermano Facebook come il social network più popolare al mondo, con il primato per utilizzo attivo (49%) e frequenza di visita (il 56% degli utenti si connette più di una volta al giorno).

Il 31 Dicembre 2013 Facebook registrava 1.189.000.000 utenti attivi mensili di cui 199 milioni erano statunitensi e canadesi, 276 milioni europei, 351 milioni di asiatici (con un aumento di 34 milioni nel secondo semestre del 2013) e i restanti 362 milioni provenienti da altre parti del mondo.

Il suo utilizzo sta recentemente registrando una lieve crescita contrapposta a quella più consistente degli altri social network, come Google+ che nel secondo semestre del 2013 è cresciuto del 6% o come Linkedin la cui utenza è aumentata del 9%[4].

Nel 2014 si prevede una crescita globale del 14% degli utenti Twitter e di solo del 3,6% di quelli di Facebook, che parte però da una base utenti maggiore.

Le crescite previste sono giustificate soprattutto dall'aumento di utenti appartenenti ad una fascia d'età oltre i 45 anni: nel primo trimestre 2013 il numero di utenti di Twitter dai 55 ai 64 anni è aumentato del 79%, mentre quelli di Google+ tra i 45 ed i 54 anni del 56%, e di Facebook del 46%.

La maggioranza dell'audience dei social network utilizza i computer per le sue attività online, ma questa percentuale sta scendendo a favore di smartphone e tablet; anzi sono proprio i visitatori più frequenti e gli utenti più attivi che utilizzano i telefonini per connettersi. Nel secondo semestre 2013 l'utilizzo dell'applicazione ufficiale per cellulari di Facebook è aumentato a livello mondiale dal 68% al 69%, mentre quella di Twitter dal 33% al 34% e quella di Google+ dal 35% al 37%.

---

[3] We are Social, 2014. *"Global Digital Statistics 2014"*

[4] VINCOSBLOG, 2013. "La mappa dei social network nel mondo-Dicembre 2013". v*incos.it* [blog] 31 Dicembre 2013

Più del 50% degli utenti attivi sulle principali piattaforme sociali nel mondo ha un'età compresa tra i 16 ed i 34 anni (vedi Figura 2.6).

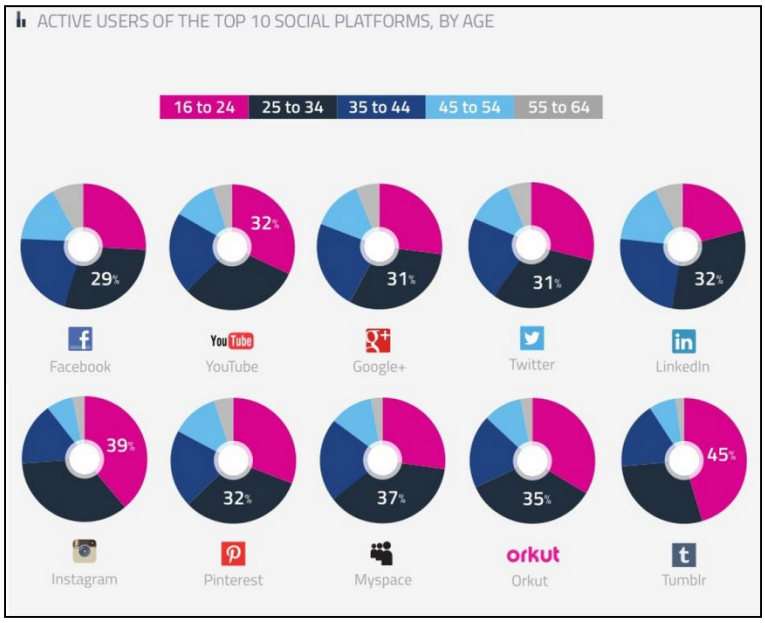


Figura 2.6 : Età degli utenti delle 10 principali piattaforme sociali nel mondo
Fonte: GlobalWebIndex Social Summary, 2014

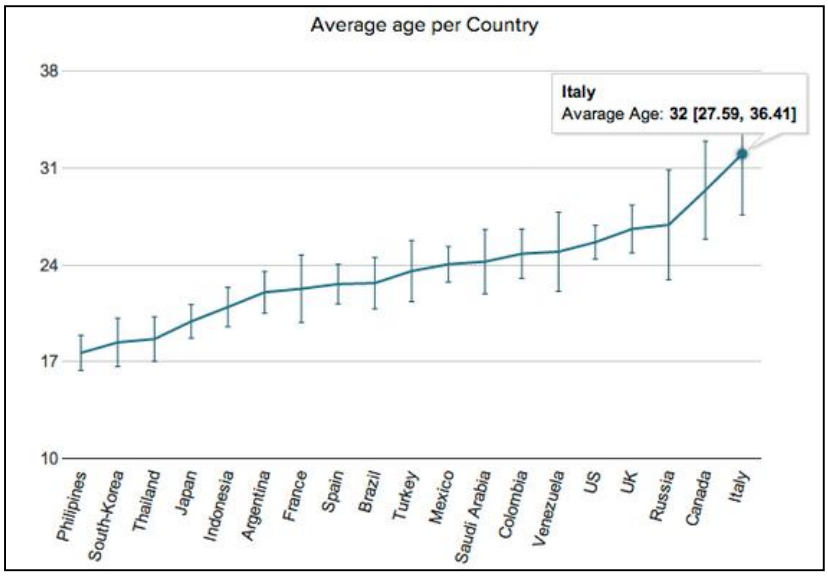


Figura 2.7: Età media degli utenti di Twitter nel mondo nel 2013
Fonte: PeerReach, 2013

L'età media degli utenti Twitter è di 24 anni e solo il 20% supera i trent'anni; l'Italia è tra i Paesi con l'età media dei 'cinguettatori' più alta (32 anni), i più giovani invece si registrano nelle Filippine (vedi Figura 2.7) con una percentuale di teenager dell'87% contro il 22% dell'Italia.

Il livello di affiliazione e di utilizzo attivo dei social sembra essere maggiore nei mercati Internet emergenti rispetto a quelli più stabili come gli Stati Uniti (vedi Figura 2.8). In linea con questa tendenza, si riscontra che in vetta alla lista mondiale si trova l'Indonesia con il 97% degli utilizzatori di Internet che ha anche un profilo Facebook e l'83% in Google+.

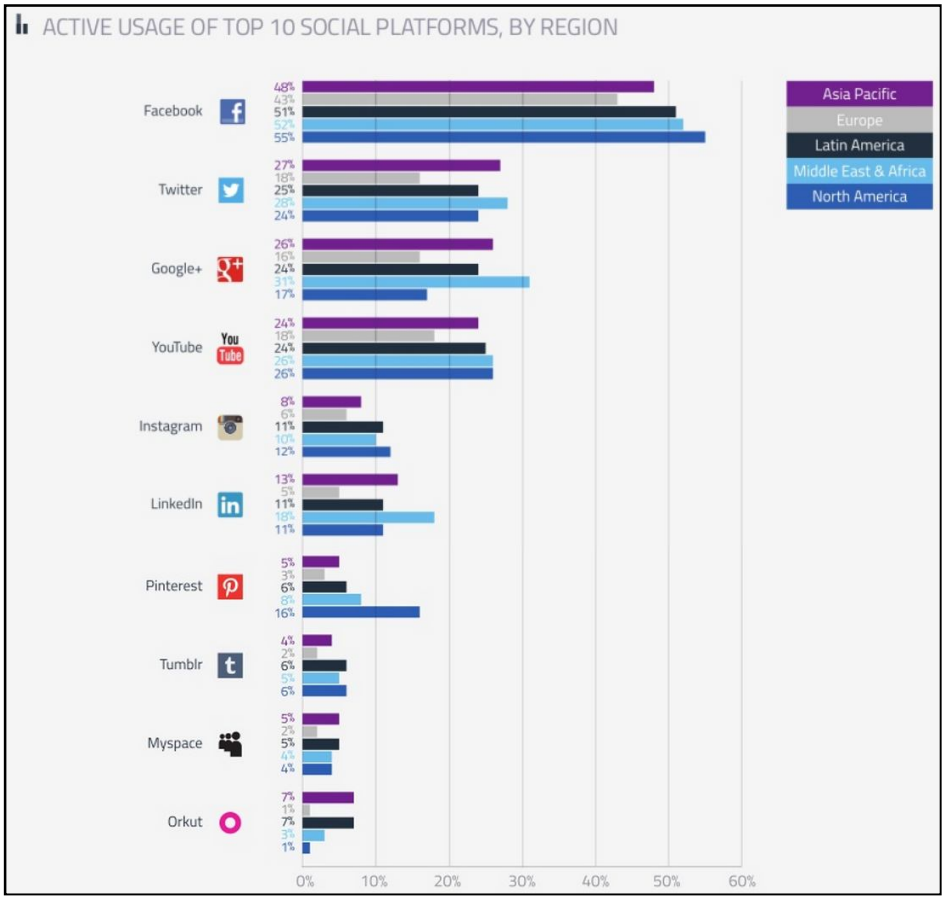


Figura 2.8: Percentuale di utenti Internet con un account nei diversi social network divisi per Regione

Fonte: GlobalWebIndex Social Summary, 2014

Concentrandosi sullo scenario italiano 35.531.527 milioni di persone, pari al 58% della popolazione, utilizzano Internet di cui il 92%, pari al 54% degli italiani, ha un proprio

account di almeno un social network, nello specifico l'83% in Facebook (23 milioni di italiani iscritti e 10 milioni di utenti giornalieri), il 53% in Google+, il 41% in Twitter ed il 24% in Linkedin[5].

Il tempo passato in media ad interagire sulle piattaforme digitali è di due ore e mezza.

Ad utilizzare la Rete è soprattutto la fascia d'età tra i 35 ed i 54 anni, seguita dagli under 25 (vedi Figura 2.9); gli uomini sono in maggioranza (7,8 milioni nel 2013 con una crescita del +3% dal 2012) rispetto alle donne (6,5 milioni nel 2013 con una crescita del +4,8% dal 2012).

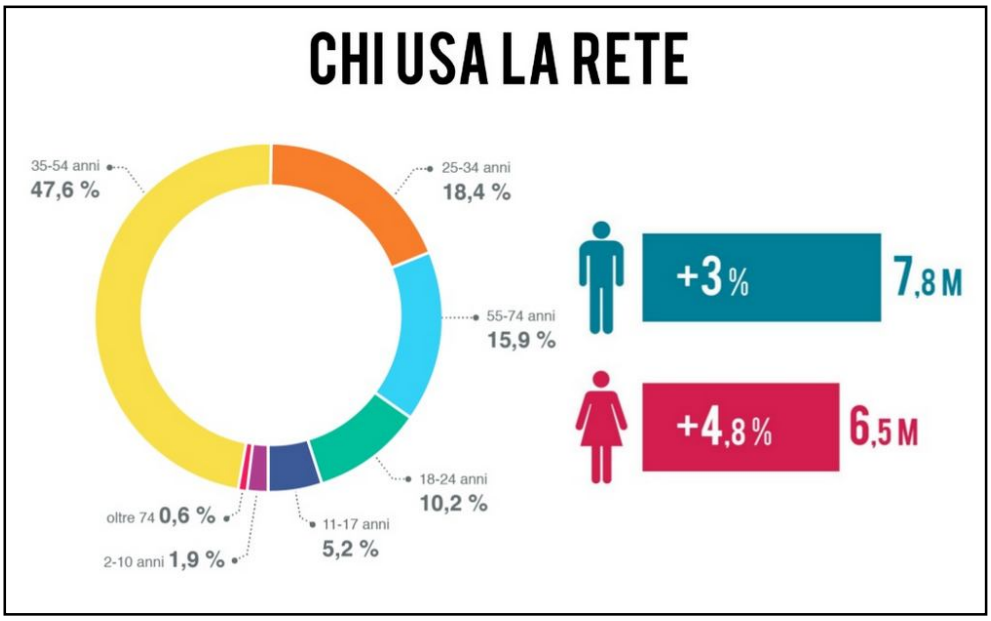


Figura 2.9: Classificazione delle tipologie di utilizzatori della Rete in Italia nel 2013

Fonte: State of the Net, 2013

Il 46% degli italiani naviga utilizzando non computer fissi o laptop ma dispositivi mobili. Il tempo speso in Internet è in continua crescita, sono state raggiunge le 4 ore e 42 minuti giornaliere attraverso computer fissi o portatili e le 2 ore attraverso gli smartphone.

Il 16% degli internauti che utilizzano il cellulare come dispositivo per connettersi accede a servizi o piattaforme che ne permettono la geolocalizzazione ed il 92% cerca informazioni geolocalizzate.

[5] VINCOSBLOG, 2013. "State of the Net 2013 Parte 1: Lo scenario italiano". v*incos.it* [blog] 1 Giugno 2013

Sul totale degli internauti italiani il 92% ha un proprio account in almeno un social network, nello specifico l'83% in Facebook, il 53% in Google+, il 41% in Twitter ed il 24% in Linkedin[6].

Nello specifico, relativamente al primo quadrimestre 2014, in Italia si farà riferimento ai dati Audiweb.

Risulta che 27.400.000 milioni di persone si siano collegate almeno una volta durante il mese di Gennaio, 27.050.000 a Febbraio, 27.800.000 a Marzo e 27.500.000 ad Aprile; con una media giornaliera rispettivamente di 13.180.000, 13.470.000, 13.318.000 e di 12.600.000. Nel quadrimestre ci sono stati in media 13.142.000 utenti attivi nel giorno medio. A gennaio la maggioranza degli utenti attivi era compresa tra i 35 e 54 anni (6,4 milioni) circa il 49% dell'utenza online in media ogni giorno; a Febbraio il range 35-54 rimane la categoria maggiormente presente online (6,6 milioni), ma sono i giovani (18-24 anni) a passare più tempo al pc con una media di 1 ora e 26 minuti al giorno online. Anche a Marzo la fascia 35-54 si riconferma come principale utilizzatrice di Internet (circa 48%), ma non anche come gruppo che passa più tempo online, che rimane quello tra i 25 e i 34 anni (1 ora e 25 minuti), lo stesso in Aprile.

	GENNAIO 2014	**FEBBRAIO 2014**	**MARZO 2014**	**APRILE 2014**
Utenti attivi al mese (000)	27,404	27,501	27,796	27,500
Utenti attivi nel giorno medio (000)	13,180	13,466	13,318	12,600
Tempo speso nel giorno medio a persona (h:m)	1:19	1:16	1:15	1:11

*Tabella 2.1:*Dati sull'utilizzo di Internet il primo quadrimestre 2014

Fonte: Audiweb, 2014

[6] My web 2.0, 2013. "Quanti utenti Twitter attivi ci sono nel mondo e in Italia?". *myweb2.0.it* [blog] 21 Novembre 2013

La distribuzione geografica relativa al mese di Gennaio 2013 è stata la seguente: la maggioranza degli utenti risiedeva al Sud e nelle isole (29,6%) mentre al Nord Est si è registrato l'utilizzo minore. Relativamente al sesso, a Febbraio il 56% degli internauti italiani erano uomini, saliti a Marzo al 57%, pari a 7,6 milioni di italiani. Questi dati, si possono ritenere simili anche negli altri mesi del quadrimestre poiché non sono stati registrati cambiamenti significativi nel totale dell'utenza.

2.2 Le motivazioni dell'interesse dei ricercatori per i social network

Sembra che ormai pochi non possiedano un account in uno dei principali social network; in essi si raccontano (quasi) tutti gli avvenimenti importanti della propria vita e i piccoli eventi della quotidianità, si esprimono liberamente le proprie idee, opinioni ed emozioni e si interagisce con gli atri utenti. Ecco che allora non c'è da stupirsi se si è tentato di scoprire le modalità migliori per sfruttare quest'oceano di informazioni allo scopo di spiegare fenomeni sociali complessi o addirittura di predirli. La letteratura su questo tema sta crescendo in modo esponenziale come pure le aziende che offrono analisi di dati estratti dai social media.

"I social network destano maggiore curiosità rispetto alle altre fonti di Big Data perché forniscono informazioni proprio su cosa la gente pensa. Analizzare i social media è come portare la voce delle persone all'interno delle organizzazioni ed il vantaggio va a quelle che sanno focalizzarsi sui giusti segnali, estrarne informazioni rilevanti, processarli rapidamente e modulare di conseguenza le proprie azioni", questo è il pensiero espresso da Stephen Rappaport, *Knowledge Solutions Director* dell'Advertising Research Foundation, in un suo rapporto presentato all'organizzazione mondiale UN Global Pulse e all'Unicef, nel luglio 2012. I social media permettono di sapere in tempo reale ciò che si dice in Rete, aggregando pezzi di informazioni separati, che nel loro complesso possono generare un mosaico coerente.

"I social media ci danno un'opportunità mai avuta prima: sapere ciò che tutti stanno dicendo riguardo a qualunque cosa", così Filippo Menczer, direttore associato del

Center for Complex Networks and Systems Research dell'Università dell'Indiana, riassume le potenzialità dell'analisi dei nuovi media[7].

L'analisi dei dati forniti dai social media può rivelarsi utile per comprendere l'opinione pubblica su determinati temi e monitorarne i cambiamenti in modo continuativo ed in tempo reale; è paragonabile a dei questionari svolti in modo passivo, intensivo e poco costoso. A differenza di questi ultimi però non si limitata alle domande predefinite, ma sono le persone stesse a decidere di cosa parlare e come parlarne, evitando così risposte mancate, indotte o strategiche, non richiedendo all'intervistato di fare appello alla memoria riportando le sue esperienze o impressioni passate e permettendo di comprendere come e quando l'opinione si è formata.

I sondaggi offline sono per definizione statici e considerando la velocità con cui opinioni e preferenze cambiano nella società attuale, affidarsi a fotografie statiche della mutevole realtà risulta sempre più limitativo. Leggendo quanto postato online, si ha invece una visione *live* di cosa sta accadendo e questo permette sia di poter prevedere risultati futuri, sia di intervenire in modo adatto per determinarne il suo evolversi.

Noah Smith, assistente Professore si Scienze informatiche alla Carnegie Mellon University, afferma che i dati di Twitter possono aiutare i ricercatori a rispondere ad una serie di domande sociologiche a cui altrimenti sarebbe difficile approcciarsi con gli altri metodi tradizionalmente adottati che risulterebbero troppo lenti e costosi per l'elevato numero di interviste necessarie[8].

I sondaggi al telefono sono il metodo tradizionale utilizzato per le ricerche nel campo delle scienze sociali per catturare l'opinione pubblica. Tale metodologia sta, però, risentendo di un calo di validità a causa della riduzione dell'uso delle linee domestiche sostituite dai cellulari, dell'aumento del tasso di non risposta e degli errori causati dalle autodichiarazioni dell'intervistato. Dal momento che le interviste telefoniche hanno dei costi organizzativi e richiedono l'impiego di risorse umane, vengono condotte con cadenza bisettimanale o mensile, pertanto le variazioni dell'opinione pubblica non sono disponibili giorno per giorno, non si riescono così a riflettere i rapidi cambiamenti di una società dinamica come quella attuale. Per superare questi ostacoli si sta cercando di capire se un'analisi semantica degli *user-generated content* può aiutare a predire i comportamenti umani dal momento che attraverso i social media un individuo esprime

[7] Savage, N., 2011 "Twitter as Medium and Message". *Communication of the ACM*
[8] Savage, N., 2011 "Twitter as Medium and Message". *Communication of the ACM*

le sue idee pubblicamente. Uno studio di Fu e Chan (2013) pone in relazione queste due metodologie confrontando i risultati da esse ottenuti ad Hong Kong nel giudicare l'operato del governo. Da inizio Aprile a fine Giugno 2011 sono stati raccolti 66.468 post contenenti termini relativi al Governo e sono stati analizzati da un classificatore (la cui accuratezza era stata stimata al 79%) che ha permesso di calcolare per ogni giorno un punteggio relativamente alla negatività dei giudizi espressi.

I risultati dei sondaggi telefonici sono invece stati forniti dai due principali siti di sondaggi del Paese che pubblicano mensilmente i risultati sulle opinioni delle persone riguardo alle principali figure politiche: il sito del 'Public Opinion Programme' dell'Università di Hong Kong (HKUPOP) e quello dell'Istituto sugli studi dell'Asia dell'Università Cinese di Hong Kong. La percentuale di risposte da cui emergeva poca soddisfazione nelle politiche del Governo, calcolata sul totale delle risposte ottenute, è stata interpolata utilizzando il metodo *autoregressive integreted moving average* (ARIMA) di Box e Jenkins, ed è stata confrontata con l'andamento giornaliero delle opinioni espresse nei social facendo emergere una correlazione significativa tra le risposte degli intervistati ed i messaggi pubblicati online.

Lo studio ha dimostrato che, nonostante il campione considerato utilizzando i social media sia elitario, se comparato con quello casuale delle interviste telefoniche, sembra comunque esser possibile adottare gli *user-generated content* per predire l'opinione pubblica, o meglio per predire i risultati mensili interpolati dei sondaggi telefonici, con uno scarto temporale di 8-15 giorni. Ovviamente non vuole sostituire il metodo tradizionale, ma potrebbe esser un approccio complementare che arricchisce i risultati ottenibili, con costi limitati ed una buona granularità temporale nei risultati.

Un'analisi la cui fonte di dati è rappresentata dai social network inizia con la definizione delle combinazioni di parole chiave più adatte per scoprire quanto più possibile sul tema della loro indagine.

Sui social media si possono monitorare:

- il volume di menzioni, visualizzazioni o post che un argomento, organizzazione o utente riceve;
- la rete sociale e canali utilizzati per scambiare contenuti;
- il livello di coinvolgimento in base a quanti utenti discutono, condividono, apprezzano o rispondono a messaggi riguardanti una precisa tematica;

- i contenuti delle discussioni categorizzando come positivo o negativo un messaggio pubblicato in rete attraverso una *sentiment analysis* o cercando di interpretare l'opinione dell'autore.

I dati così ottenuti vengono analizzati per rispondere a diversi tipi di domande.

Quattro sono le principali direzioni di ricerca sui social network indivuduate**,** le quali saranno affrontate separatamente nei paragrafi che seguono.

2.3 Il potere influenzante dei social network

Un primo approccio di ricerca sui social network indaga se e come le scelte e i comportamenti degli individui siano influenzati da quanto viene detto in Rete.

Herbst (2011) propone la suggestiva metafora che paragona la Rete ai salotti del XVIII secolo: come le conversazioni dei salotti ottocenteschi riflettevano la cultura francese ed europea del tempo, la influenzavano e la stravolgevano attraverso rivoluzioni che prendevano avvio proprio grazie a quello scambio di idee, lo stesso oggi avviene nella Rete. Anche O'Connor (2010) definisce gli utenti della Rete come degli *opinion maker*, il cui pensiero ha la capacità di influenzare quello della massa, e quindi di anticipare l'evoluzione dell'opinione pubblica, oltre che di incidere su quali argomenti l'ecosistema mediatico riverserà la sua attenzione.

I social media stanno rapidamente cambiando le modalità con cui le persone comunicano tra loro ed incoraggiano le persone a condividere informazioni personali che fino a poco tempo fa rimanevano private. Questo scambio continuo di pensieri tra gli utenti può alterare la visione del mondo che un individuo ha e di conseguenza pure il suo modo di agire; addirittura indagini confermano che sempre più i filtri professionali vengono sostituiti da richieste d'informazione alle proprie connessioni nei social. (Baresch et al., 2011).

L'impatto dei social network sulle scelte di consumo è stato ampiamente dimostrato ed oggi i social media sono diventati uno strumento fondamentale per valutare la *brand image* di un'azienda. L'opinione espressa online da parte di un consumatore soddisfatto o insoddisfatto incide sulla reputazione di un marchio, sul processo d'acquisto di potenziali consumatori e sul successo di nuovi prodotti, poiché i consumatori sempre

più si fidano di quanto leggono nelle discussioni nel Web. Nel conteggio delle raccomandazioni online, oltre a quelle esplicite, Bhatt et al. (2010) sostengono che si debba far rientrare anche quelle implicite, dove cioè ci si riferisce solo al prodotto senza giudicarlo o lo si mostra in foto o video.

Nello specifico è stato osservato che il passaparola tecnologico, il cosiddetto *electronic words-of-mouth* (eWOM), è più efficace nel primo acquisto e quando è negativo piuttosto che positivo (Park e Lee, 2009). Inoltre, più aumenta il numero delle raccomandazioni più cresce la possibilità di acquisto, ma non in modo lineare perché dopo una certa soglia, la probabilità diminuisce (Leskovec e Adamic, 2007).

Anche Bertrand (2013) ribadisce come i giudizi espressi in Rete da altri consumatori giochino un ruolo significativo nel *Zero Moment of Truth* (ZMOT), il momento zero della verità, termine coniato da Google per esprimere l'influenza che i contenuti online esercitano nelle decisioni d'acquisto: il cliente interessato ai prodotti di un'azienda interroga i motori di ricerca per scoprire se chi ha già acquistato è rimasto soddisfatto e per scegliere poi se comprare o meno.

Ciò vale ancora di più se si considera che sebbene ci siano nodi che individualmente sono più influenti, la somma di molti nodi deboli, ma che condividono lo stesso comportamento, risulta nel complesso essere molto più influenzante (Barkshy et al., 2012).

Le categorie in cui le *peer recommendation* online hanno elevata influenza sono quelle degli oggetti di alto valore (come pc o smartphone) o di alto coinvolgimento emotivo (come i prodotti fashion); se i primi alimentano conversazioni soprattutto relativamente allo loro bontà tecnica ed al rapporto prezzo-qualità, i secondi creano un ideale a cui ispirarsi.

Il ricercatore ed il suo team hanno analizzato tutte le menzioni nei social media di British Gas e di M&S, escludendo i post provenienti dai loro stessi account, durante Settembre 2012. Tra i messaggi pubblicati sui due marchi sono stati identificati quelli dai quali trapelava un sentimento di fiducia o sfiducia verso il brand. Dall'analisi dei social media emerse che M&S registrava un livello di fiducia maggiore, con un 3/5 di conversazioni positive tra tutte quelle che la riguardavano, soprattutto grazie ai suoi prodotti 'green' ed equosolidali ed alla sua politica di assunzioni. Il risultato fu confermato dal quinto posto ottenuto dall'azienda nella classifica delle marche inglesi

preferite redatta nel Dicembre 2012 in base all'indice *YouGov's Brand Index*. Analizzando cosa viene detto online spontaneamente si può, però, fornire una migliore comprensione della percezione del brand approfondendo quali siano i fattori esplicativi dei risultati quantitativi.

Joachim Mathiesen, biofisico presso il Niels Bohr Institute di Copenhagen, ha confrontato l'attività di 92 account di brand famosi (tra cui Pepsi, Apple e Toyota) con l'andamento dei rispettivi titoli sul mercato azionario ed ha notato una correlazione tra le fluttuazioni delle loro azioni ed il trend della loro attività sui social.

Sembra quindi che i social network riescano ad influenzare gli investitori finanziari; ne è un esempio il crollo improvviso del Dow Jones il 23 aprile 2013 in seguito ad un tweet falso su un attentato al presidente Obama, postato dall'account violato della Associated Press.

Qualman (2010) a sostegno della sua tesi secondo cui i social network hanno aumentato la potenzialità degli influenzatori e trend setter nell'incidere sulla formazione delle opinioni della massa, sui loro comportamenti e sui loro acquisti, riporta il dato secondo cui il 78% dei consumatori si fida delle raccomandazioni lette online postate da altri acquirenti, mentre solo il 14% dà valore a quanto affermano le pubblicità.

Per quanto concerne invece l'impatto delle opinioni espresse in rete in campo politico, sociale o medico i casi che confermano la possibilità d'incidere sulle scelte degli altri utenti diminuiscono; ciò nonostante sempre più istituzioni e leader politici stanno adottando la comunicazione attraverso i social media come strumento per influenzare l'opinione pubblica.

Se si è scoperto possibile influenzare i gusti dei consumatori sembra quindi fattibile anche incidere sui risultati di elezioni politiche. A sostegno di quest'ipotesi vi è il caso della campagna elettorale di Obama del 2008, studiato da Cogburn ed Espinoza-Vasquez (2011), i quali hanno dimostrato come il candidato attraverso i social media sia riuscito a trasformare la sua attività online in attivismo e supporto elettorale reale.

La conferma è arrivata anche dall'analisi dell'uso di Facebook da parte degli studenti di Università americane durante le presidenziali del 2008: emerse che la maggioranza dei messaggi politici erano condivisioni tratte in modo esplicito da azioni dei candidati sui

social media, dimostrando quindi come la loro attività online avesse influenzato ciò che veniva discusso in Rete (Vitak et al., 2009).

Karlsen (2009), dopo aver investigato l'effetto dell'uso dei social media da parte dei candidati alle elezioni in Norvegia del 2009, è però arrivato alla conclusione che è stato il contesto statunitense a mobilitare un tale supporto elettorale poiché negli USA la politica ha carattere fortemente personale, e risulta pertanto più adattabile all'uso dei social media, rispetto per esempio al contesto norvegese in cui i candidati non sono molto incentivati ad intraprendere campagne individuali.

Anche nell'ambito della salute sembra vi sia la possibilità che quanto letto online possa incidere sui comportamenti della gente. Studiosi, come Dearing e Kreuter (2010), sostengono che la discussione online riguardo temi relativi al campo medico, come la prevenzione al cancro, renda più probabile che nuovi metodi siano noti, accettati ed infine implementati.

A tal proposito è interessante il caso di Facebook che in America il primo maggio 2012 ha introdotto la possibilità di specificare nel proprio profilo la qualità di "donatore di organi". In seguito a questa scelta, l'utente era reindirizzato ad un registro statale nel quale poteva fare la sua nomina ufficiale; nel frattempo ai suoi amici in Facebook veniva reso noto il suo nuovo stato. Il primo giorno dell'iniziativa ci sono state 13054 nuove registrazioni, e nei 12 giorni successivi la media rimase elevata. Ciò dimostrò che far circolare informazioni sui social network può rivelarsi utile nel caso di problemi legati alla salute in cui la comunicazione e l'educazione sono essenziali per superare la riluttanza del pubblico a cooperare, come nel caso della donazione di organi.

Anche l'Unicef nel 2011 ha voluto concentrarsi sulle modalità di sfruttare i social allo scopo di divulgare informazioni scientifiche o pseudoscientifiche in grado di influenzare la percezione soggettiva sul tema delle vaccinazioni nell'Est Europa, sia per quanto riguarda la prevenzione, che gli effetti collaterali generati.

A causa di disinformazione spesso i genitori, credendo che i vaccini possano causare autismo, danni al cervello o HIV, tendono a rifiutarsi di sottoporre i figli alle vaccinazioni e come conseguenza si registrano morti di bambini a causa di malattie obsolete e prevenibili. I social media sono allora stati utilizzati sia per comprendere le cause di tale fenomeno, sia per scoprire le modalità migliori di utilizzarli per ottenere una rapida ed efficiente diffusione di informazioni mediche grazie alla loro natura

partecipativa. Dare ascolto per ottenere ascolto al fine di capire cosa causa quest'opposizione al vaccino, chi influenza l'opinione pubblica e cosa si può fare per porvi rimedio.

Sono stati raccolti messaggi pubblici contenenti parole riconducibili alla tematica postati in blog, forum, Twitter, Tumblr, Youtube e Facebook, dal 1 maggio al 30 luglio 2012, in inglese, russo, polacco e rumeno ed analizzati in termini di volume, coinvolgimento ottenuto, tipologia di utenti coinvolti, tematiche di discussione e sentimenti suscitati.

Sono stati scoperti i canali maggiormente utilizzati per veicolare messaggi contro le vaccinazioni, le caratteristiche dei principali influenzatori, gli argomenti portati a sostegno delle loro idee ed il loro evolversi nel tempo.

Basandosi su quanto scoperto, come ad esempio che gli *opinion leader* contro le vaccinazioni enfatizzano storie negative per convincere, è stato elaborato un piano di comunicazione adatto a far fronte all'atteggiamento negativo nei confronti dei vaccini che si registra nell'Est Europa.

In seguito a studi come quelli precedentemente descritti il settore della Pubblica Sanità sta valutando come i social media possono essere utilizzati per individuare comportamenti che mettono a rischio la salute dei cittadini e per comprendere la percezione di determinate tematiche riguardanti il benessere collettivo, così da poter intervenire di conseguenza diffondendo notizie e consigli medici nei modi adeguati per influenzare positivamente la società.

2.4 Le modalità di diffusione di una notizia all'interno dei social network

Twitter o Facebook diffondono non solo pensieri ed opinioni delle persone, ma anche notizie; ecco che il secondo filone di ricerca indaga come le *news* si diffondono all'interno dei social network.

"The minute news hits the wires, it becomes a conversation. It's just a question of how high the volume of that conversation will become", così Doug Frisbie mette in luce la capacità dei nuovi canali di comunicazione di fungere da megafoni amplificatori delle

notizie e da connettori tra le persone. Ecco, infatti, che molti studiosi si sono interessati allo studio degli schemi di diffusione dei contenuti pubblicati online ed alle connessioni che ciò crea tra gli utenti (vedi paragrafo seguente).

Sono due le modalità con cui una notizia si diffonde online: esogena o endogena. Nel primo caso i social media sono la cassa di risonanza di una notizia lanciata dai canali tradizionali; quando invece i social generano per primi la notizia anticipando i mass media si parla di diffusione endogena.

Seppure la prima modalità sia la più diffusa, ci sono anche esempi della seconda; uno dei principali è il caso dell'attentato alla maratona di Boston del 15 Aprile 2013. Le prime notizie sulla serie di esplosioni sono comparse sui social media prima che sui canali tradizionali, è proprio su Twitter dove hanno cominciato a circolare la foto dell'uomo sul tetto e i video dell'evento, ed è sempre su Twitter che il Boston Police Department ha lanciato un appello ai cittadini affinché postassero foto e video amatoriali in loro possesso che potevano rivelarsi utili nelle indagini, dando così il via ad indagini in crowdsourcing. YouTube addirittura creò un'apposita sezione (ancora disponibile all'indirizzo www.youtube.com/playlist?list=PLbpi6ZahtOH5T_nrvCoHfnYWbW1shGYYv) per raccoglie tutti i video sull'attentato.

Google invece mise a disposizione il 'Person Finder', già creato in occasione del terremoto di Haiti del 2010 per trovare persone disperse, che dopo poche ore registrava circa 5000 contributi caricati.

Altri esempi vengono dal calciomercato: molti calciatori 'twittano' o postano in Facebook il loro acquisto da parte di una squadra dopo la firma del contratto, prima però che la trattativa diventi ufficiale. Lo stesso fece Gad Lerner il 4 marzo 2013 quando postò in Twitter la notizia della cessione dell'emittente televisiva La7 due ore prima del comunicato stampa ufficiale della società, mentre il Consiglio di Amministrazione era ancora riunito in assemblea.

Spesso inoltre i giornalisti stessi setacciano il web in cerca degli *hot topic* del momento per capire di cosa 'parlano' le persone in Rete, per poi decidere su quali temi incentrare i loro prossimi articoli; a volte, addirittura, gli articoli pubblicati nei quotidiani stampati sono il resoconto della reazione del popolo del Web ad eventi particolari.

La Figura 2.10 mostra la correlazione positiva, con un coefficiente di correlazione pari a 0.52, tra il totale dei tweet e il numero degli articoli di giornale pubblicati sul tema della nuova legge sui rimborsi pubblici nel 2012. La notizia esogena ha forte eco nel social e avvia un dibattito che dura di più rispetto al periodo di tempo in cui i media tradizionali prestano attenzione alla notizia; anzi proprio perché in Rete si continua a parlarne alcuni articoli ritornano a trattare il tema in seguito a picchi nel quantitativo di commenti online a riguardo: più articoli pubblicati generano più tweet, ma anche tanti tweet incrementano il numero di articoli.

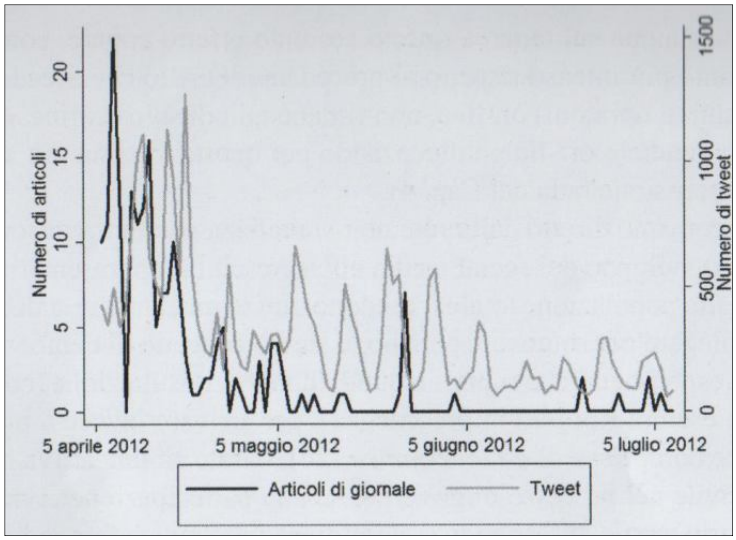


Figura 2.10: Confronto tra il numero di articoli di giornale e di tweet sul finanziamento pubblico nel 2012
Fonte: Ceron, Curini, Iacus, 2014

Un confronto tra Twitter e altri canali online che diffondono notizie è stato fatto da Leetaru et al. (2013) di cui in Figura 2.11 è riportata la mappa che compara la copertura geografica di Twitter determinata attraverso i tweet geolocalizzati e contrassegnata da puntini blu, con quella di un canale di massa come English Google News contrassegnata da puntini rossi; dove ci sono i puntini bianchi è stata rilevata la presenza di entrambi i canali ad un livello simile. I risultati suggeriscono che Twitter non è solo uno specchio dei *mainstream media*, ma ha piuttosto una sua copertura geografica distinta.

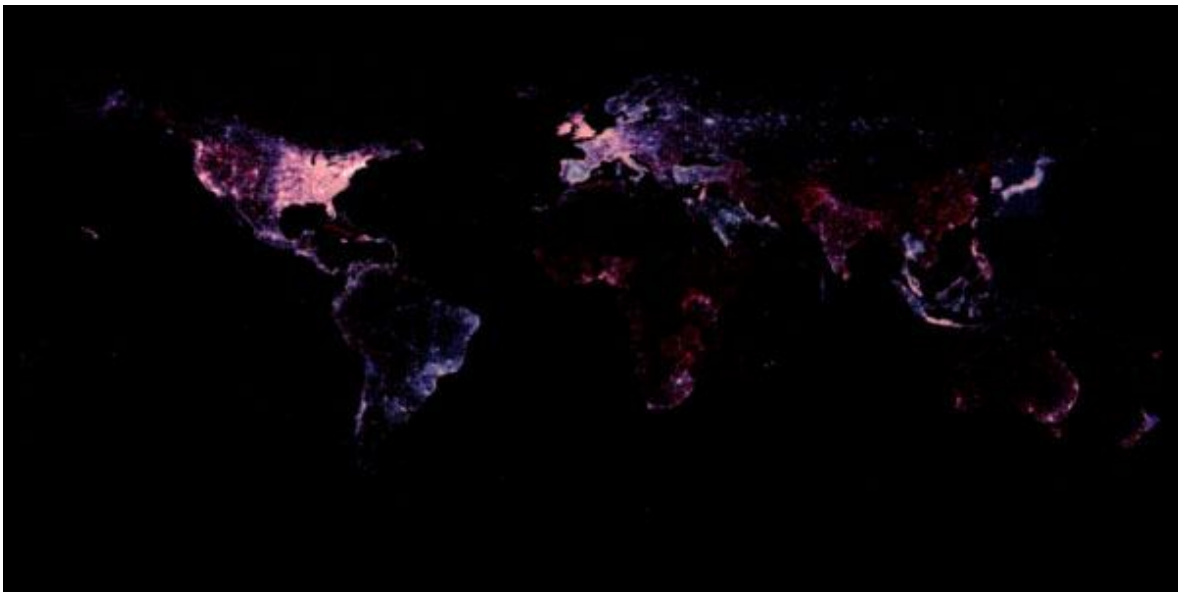


Figura 2.11: Confronto tra la copertura geografica di Twitter (dove questa è maggiore ci sono puntini blu) e quella di English Google News (dove questa è maggiore ci sono puntini rossi) tra il 28 ottobre 2012 ed il 30 novembre 2012 (dove ci sono i puntini bianchi è stata rilevata una presenza bilanciata di entrambi)
Fonte: Leetaru, 2011

Ovviamente Twitter è terreno fertile anche per notizie false o errate; è apparso, però, che esso offre anche un robusto meccanismo di autocorrezione. E' stato evidenziato come gli utenti si comportino in modo differente di fronte a *rumor* dalla dubbia attendibilità: il 50% di quelli falsi viene sfidato attraverso domande o negazioni, mentre il 90% di quelli veri viene confermato attraverso retweet (Mendoza et al., 2010).
Procter, Vis e Voss (2013) hanno analizzato i tweet inviati durante le rivolte in Inghilterra di Agosto 2011 iniziate il 6 agosto a Londra nel quartiere di Tottenham; la disponibilità dei 2,6 milioni di tweet provenienti da 700.000 utenti diversi era dovuta alla stipula di un accordo tra 'The Guardian Newspaper' e Twitter.
Lo studio ha fatto emergere un modello di diffusione di una notizia falsa:

1. Qualcuno twitta circa un presunto avvenimento;
2. Il *rumour* viene ritweetato, a volte riformulato, e vengono aggiunge alcune prove come accenni a testimoni;
3. Alcuni utenti iniziano a sfidare la credibilità del tweet fornendo nuove informazioni che mettono in dubbio l'affidabilità della notizia;

4. Un consenso inizia ad emergere relativamente alla falsità della notizia diffusa che può però ricomparire anche in seguito alla smentita perché ripresa da utenti non informati.

Il giornalismo *crowd-sourcing*, che nasce grazie ai social media, è uno sforzo collaborativo di un elevato numero di utenti che sembra essere in grado di autocorreggersi e di fornire una copertura migliore, o comunque più veloce rispetto ai canali di massa tradizionali.

I risultati di un esperimento condotto da Morris et al. (2012), durante il quale venne analizzato cosa un gruppo di persone pensava ad alta voce mentre eseguiva una ricerca interna a Twitter, ha evidenziato come le caratteristiche su cui si basano gli utenti nel determinare la credibilità di un tweet non sia tanto il contenuto quanto elementi euristici, come il numero di retweet ed il nome di chi posta il tweet, giudicando più affidabili gli autori il cui nome richiama l'argomento di cui tratta il post. Sarebbe però più utile considerare la biografia dell'autore ed il numero di menzioni ricevute, mentre basando il proprio giudizio su basi come questa, la percezione di credibilità può facilmente essere manipolata dall'autore del tweet.

2.5 I diversi tipi di reti che si creano nei social network

Le reti sociali che vengono a crearsi nei social possono essere generate dall'interazione tra gli utenti, come le amicizie in Facebook o le menzioni ed i retweet in Twitter, o dalla discussione relativamente ad una tematica. In Twitter si può ricostruire l'intera rete comunicativa soggiacente ad un argomento indagando gli scambi comunicativi avvenuti che hanno adottato determinati hashtag o parole chiave, identificando quali degli utenti coinvolti hanno avuto un ruolo centrale nel dibattito. È possibile, inoltre, estrapolare la rete semantica soggiacente la discussione, collegando con un arco le parole che compaiono in uno stesso tweet, così da fornire visivamente l'universo semantico che circonda un certo tema (Caliandro Aka Ali, 2013).

Un esempio della prima tipologia è la Figura 2.12, che riporta le connessioni (espresse sottoforma di amicizia) di un utente Facebook; i nodi rappresentano gli utenti, le linee la loro connessione; il grafico mette in evidenza attraverso la grandezza dei nodi quali

sono gli utenti che hanno il maggior impatto su gruppi multipli, fungendo da connettori tra comunità che altrimenti sarebbero disconnesse.

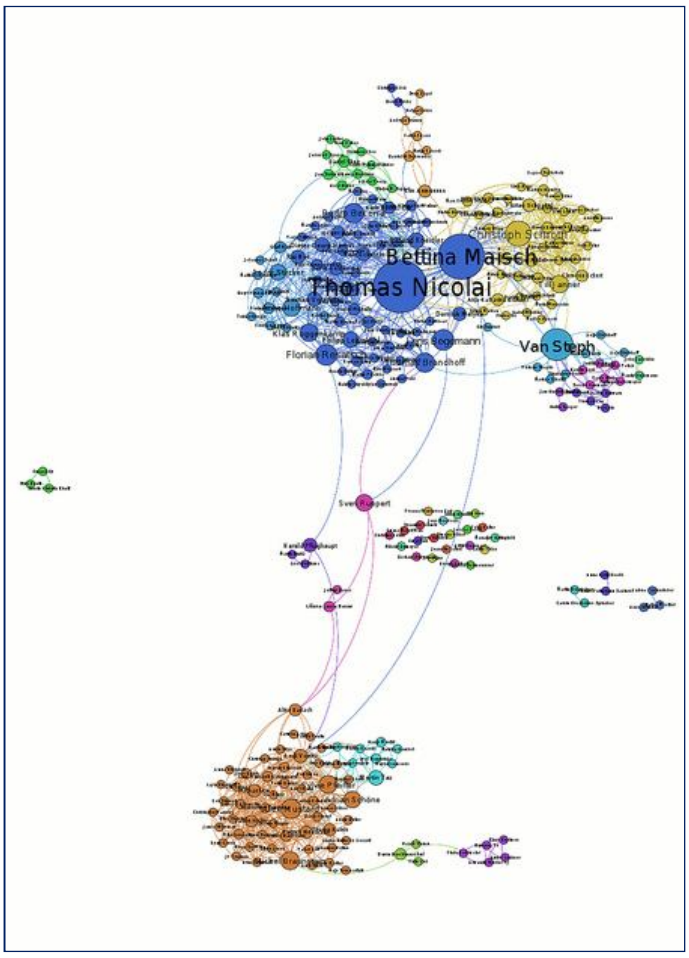


Figura 2.12: Visualizzazione delle connessioni di un utente Facebook
Fonte: persuasionradio.wordpress.com, 2010

Relativamente allo studio delle rete attraverso cui una notizia si diffonde, merita di essere citato Gilad Lotan, responsabile dell'elaborazione dei dati presso Betawork ed esperto di *data mining*, il quale ha analizzato parecchie reti sociali nate in Twitter conseguentemente alla diffusione di notizie su avvenimenti come le rivolte in Tunisia ed in Egitto e le elezioni iraniane.
Lo scopo degli studi che ha condotto era quello di individuare i principali influenzatori sul tema e le interconnessioni che lo scambio di idee ed informazioni sulla tematica generava.

Lotan (2012) si interessò di scoprire chi aveva diffuso e amplificato messaggi riguardanti la protesta in Tunisia del Gennaio 2011, che Tsotsis aveva mostrato nel suo studio (2011) aver creato un enorme traffico in Twitter pari a 28 tweet al secondo nelle ore successive alla notizia che il presidente tunisino aveva lasciato il Paese. Concentrandosi sul totale dei tweet riportanti l'hashtag #sidibouzid (la provincia in cui la protesta prese avvio), pari a 170,000, pubblicati tra il 12 ed il 19 Gennaio da 40.000 differenti utenti Twitter, individuò i principali partecipanti in termini di volume ed indagò le connessioni di questi nodi.

Il grafico in Figura 2.13 rappresenta il network sociale di uno dei nodi principali all'interno del gruppo, Sami Ben Gharbia; è evidente come il suo audience sia diviso con altri utenti chiave, come 'mfatta7' e 'jrug', rispettivamente un giornalista egiziano e un corrispondente affari esteri di Channel 4 News.

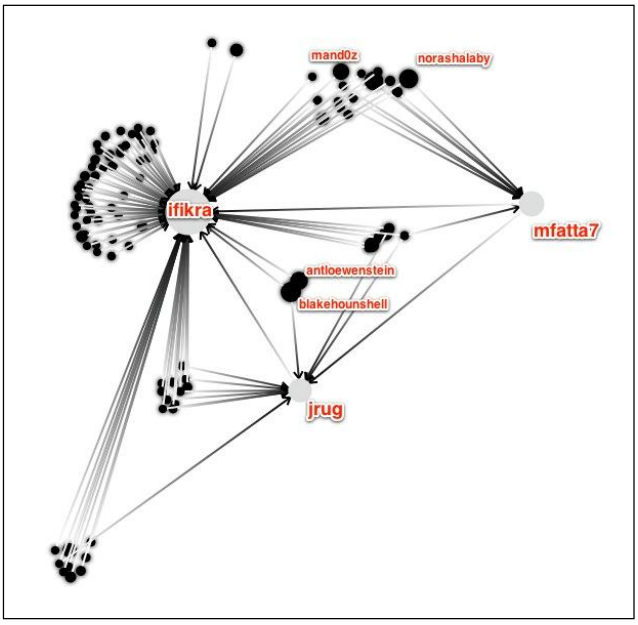


Figura 2.13: Rete sociale di Sami Ben Gharbia, uno dei principali utenti che contribuirono alla diffusione dell'hashtag #sidibouzid tra il 12 ed il 19 gennaio 2011

Fonte: Lotan, 2011

Inoltre, allo scopo di comprendere la tipologia di persone che era stata coinvolta nella diffusione dell'hashtag, creò un *word cloud* per mettere in luce le informazioni personali degli utenti Twitter che avevano partecipato alla discussione sugli

avvenimenti politici in Tunisia (vedi Figura 2.14), scoprendo che erano per la maggioranza giornalisti, studenti ed attivisti.

Figura 2.14: *Word cloud* contenente le informazioni personali dei partecipanti alla diffusione di tweet contenenti l'hashtag #sidibouzid tra il 12 ed il 19 Gennaio 2011
Fonte: Lotan. 2011

Appare chiaro da quest'esempio come la visualizzazione dei dati può rivelare qualcosa che altrimenti non sarebbe ovvio guardando semplicemente i dati grezzi, per quanto possano essere ben organizzati; questo è tanto più vero quando si studiano le reti sociali. Un'analisi simile è quella di Galea (2012) sui legami creati dai 3660 tweet e relativi retweet contenenti l'hashtag #healthinall dell'ottava Global Conference on Health Promotion avvenuta ad Helsinki tra il 10 ed il 14 giugno 2013. Nel grafico in Figura 2.15 ogni nodo rappresenta un utente nei cui messaggi è apparso l'hashtagh #healthinall, i collegamenti tra i nodi indicano l'azione di un nodo di aver retwittato un messaggio di un altro utente, senza però mostrare quale dei due fosse l'autore originale. La grandezza dei nodi e delle loro etichette dipende dalla centralità dell'utente rappresentato, mentre i colori differenziano i diversi sottogruppi. WHO appare essere il nodo maggiormente collegato, rivestendo pertanto un ruolo cruciale nell'aumentare l'esposizione dell'evento attraverso la sua attività nel social network.

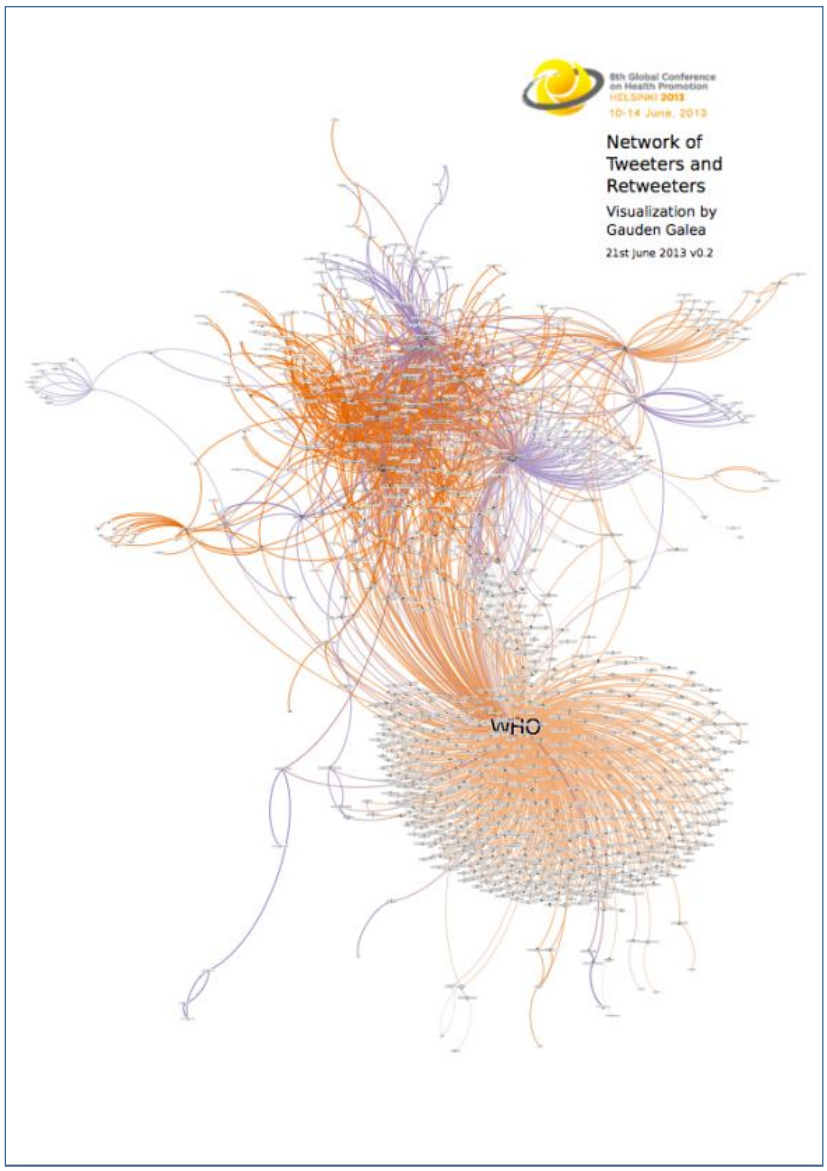


Figura 2.15: Network creato dai tweet e retweet contenenti l'hashtag #healthinall nel periodo 10-16 giugno 2013
Fonte: Galea, 2013

A sottolineare le prospettive di questo tipo di analisi c'è Vespignani, professore di Fisica, Informatica e Scienze della Salute, che durante il Wired Next Fest del 30 Maggio 2013 afferma: "Abbiamo cominciato ad utilizzare i Big Data con le epidemie per capire come si propaga l'infezione nel tempo e nello spazio. Ovviamente non riusciamo a prevedere quando ci sarà la prossima pandemia, ma a quanto appare possiamo intuire come si propagherà. Allo stesso modo si propagano le idee; il contagio del pensiero è un

fenomeno epidemico, in cui non dobbiamo più mappare la geografia, ma lo spazio sociale. Se ciò fino a pochi anni fa era impossibile, oggi si può”[9].

Infine, per quanto riguarda l’analisi delle reti semantiche sottostanti alla diffusione di una notizia online, di un hashtag o di una parola chiave Gilad Lotan afferma: “Io credo fortemente che la vera rivoluzione di Twitter sta nel come questo mezzo permette alle persone di relazionarsi alle notizie e di diffonderle in modo personalizzato”[10]. Si può così comprendere come il messaggio venga collegato ad altri argomenti perché trattato sotto prospettive differenti, e come quindi nel tempo venga anche manipolato da parte dai suoi diffusori.

Nel 2012 Lotan ha fatto emergere le diverse tipologie di rete create dalla diffusione dei tweet contenenti l’hashtag #SahelNow riferito alla campagna dell’Unicef volta a raccogliere cibo ed altri tipi di aiuti per i bambini della regione del Sahel in Africa.

I grafici seguenti sono degli screenshot delle analisi sulle reti di diffusione (Figura 2.16) e sulla rete semantica (Figura 2.17) ricavate elaborando i dati estratti da Twitter attraverso il software Ghepi.

[9] Rizzoli, D., 2013. “Vespignani: Il pensiero è contagioso. E noi possiamo prevederlo coi Big Data" Wired

[10] Lotan, G., 2012. “Analyzing UNICEF’s #SahelNow Campaign”. *giladlotan.com*, [blog] 23 Maggio 2012

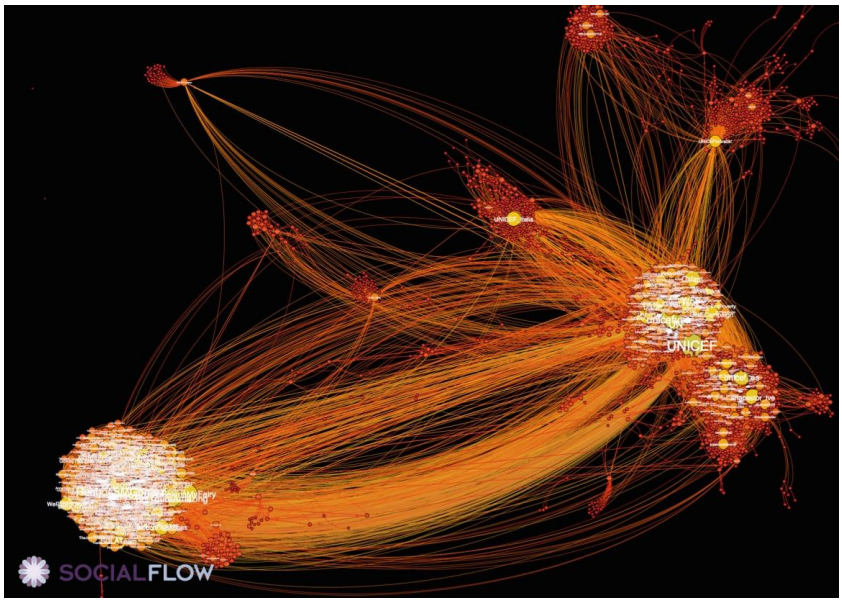


Figura 2.16: Visualizzazione attraverso Ghepi delle connessioni tra gli utenti che hanno postato l'hashtag #SahelNow durante il periodo analizzato. I cluster che emergono mettono in luce le differenti comunità a livello regionale e gli argomenti apparsi nei tweet che fanno riferimento alla campagna umanitaria Sahel Now

Fonte: Lotan, 2012

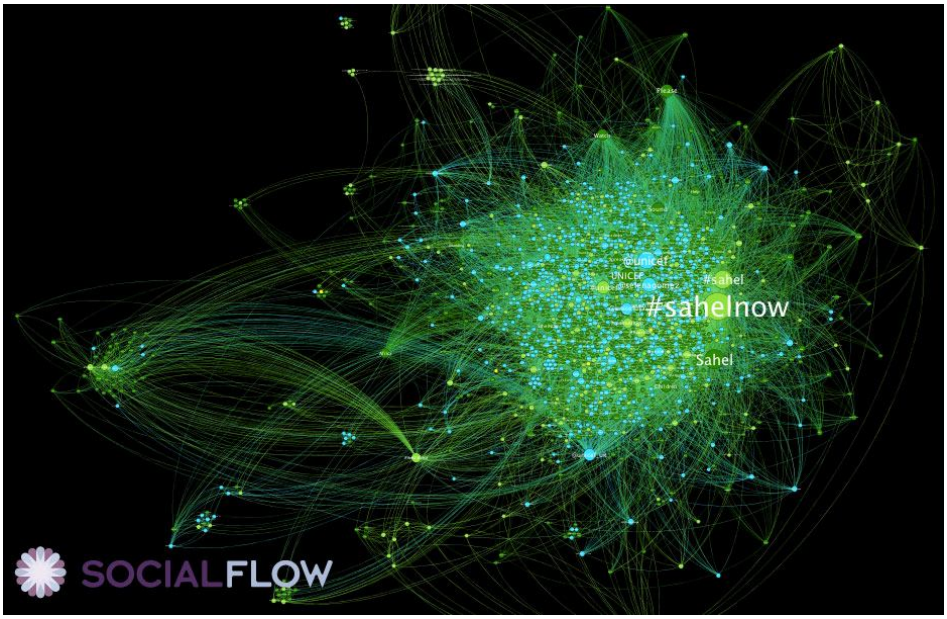


Figura 2.17: Grafico del network totale dei concetti comparsi nei tweet contenenti l'hashtag #SahelNow hashtag (blu = menzioni di altri utenti, verde chiaro = altri hashtag, verde scuro = frasi)

Fonte: Lotan, 2012

Un altro esempio di studio possibile sulle reti semantiche online è un'analisi di Lotan del 2011 sugli account Twitter dei principali canali tradizionali di notizie.

L'immagine che segue (Figura 2.18) è il grafico del network generato dagli argomenti oggetto dei tweet dei follower di Al-Jazeera English (@AJEnglish) durante il periodo di un'ora; più l'argomento è trattato e più il suo nodo è grande e gli argomenti che compaiono nello stesso messaggio sono tra loro collegati.

Si notano sottogruppi relative a notizie afferenti a diverse parti del mondo (in alto a destra ci sono temi europei, in basso a sinistra relativi al Medio Oriente, in alto a destra quelli africani), non c'è un argomento predominante ma piuttosto conversazioni multiple che avvengono nello stesso momento, i cui argomenti sono comunque eventi d'attualità a livello globale o locale. La possibilità di monitorare l'evolversi e il declino di queste discussioni in tempo reale dà ad Al-Jazeera English la capacità di interagire al meglio con il proprio audience e di concentrare i suoi post sugli argomenti che più interessano al suo pubblico in quel momento.

Figura 2.18: Grafico della rete di connessioni tra gli argomenti oggetto dei tweet dei follower di Al-Jazeera English durante il periodo di un'ora

Fonte: Lotan, 2011

2.6 I social network e le previsioni

Il terzo filone di ricerca riguarda la possibilità di fare previsioni, o più precisamente di fare *nowcasting,* termine utilizzato per indicare la comprensione di dinamiche che sono in corso di realizzazione nel momento presente per fare una previsione per l'immediato futuro, quindi a brevissima scadenza.

Sembra quindi possibile utilizzare il Web come una sfera di cristallo capace di formulare previsioni (Curini, Ceron, Iacus, 2014). Il governo americano sta, a tal proposito, finanziando il programma OSI, Open Source Indicators (www.iarpa.gov/Programs/ia/OSI/osi.html), il cui scopo è sviluppare metodi per l'analisi continua e automatizzata di dati pubblici disponibili nel Web ed in particolare sui social media per rilevare in anticipo o in tempo reale eventi come crisi politiche, crisi umanitarie, violenze di massa, instabilità economiche o scarsità di risorse monitorando come le idee dei cittadini si formano e si diffondono. Un altro progetto simile che merita di esser menzionato è l'azienda Recorded Future, fondata dalla CIA nel 2010, riceve finanziamenti da Google e dal fondo di venture capital In-Q-Tel. La start up passando al setaccio decine di migliaia di siti web, blog e profili Twitter ricerca le relazioni tra persone, organizzazioni, azioni ed incidenti, sia presenti che incombenti. Si spiega così il nome: il futuro scritto nel presente.

Nel contesto italiano relativamente allo studio dei social media, e di Twitter nello specifico, per fare previsioni a breve termine spicca Voices from the Blogs (VfB), uno spin-off dell'Università degli Studi di Milano, progetto di ricerca realizzato nel 2011 da Luigi Curini, scienziato politico all'Università degli Studi di Milano, Stefano Iacus, statistico-matematico, e da Giuseppe Porro, economista dell'Università dell'Insubria.

Vfb è un osservatorio permanente su tutto ciò che viene postato online dagli italiani e si avvale di solide basi statistiche ed informatiche dei suoi membri per monitorare costantemente la Rete al fine capire cosa pensano gli italiani riguardo ai temi più disparati: dalle elezioni al Festival di Sanremo, dal nucleare alla fine del mondo, dall'Expo2015 alla morte di Gheddafi fino agli ultimi libri pubblicati.

Adotta una tecnica innovativa per l'analisi semantica dei testi che raccoglie dal Web, l'integrated Sentiment Analysis (iSA), che si basa sulle tecnologie statistiche ed informatiche per la Blog Sentiment Analysis (BSA) sviluppate da G. King e D. Hopkins

dell'Università di Harvard. I risultati delle 220 milioni di voce che dal 2011 al 16 febbraio 2014 ha analizzato sono riportati nella pagina web www.voicesfromtheblogs.com e nel blog Sentimeter del Corriere della Sera.
Vi è una rassegna di analisi scientifiche disponibili in letteratura che hanno studiato in tempo reale fenomeni complessi o che hanno tentato di fare previsioni utilizzando i social media avvalendosi di una varietà di metodi.
I casi studio spaziano tra i campi più disparati: dal mercato finanziario, a quello medico, da casi commerciali alle elezioni politiche, per arrivare fino ai terremoti.

Le analisi previsionali si sono focalizzate principalmente su **tematiche economiche**, sia a livello macro che micro. Zhang et al. (2012) collezionando per cinque mesi, dal 15 novembre 2010 al 20 aprile 2011, i tweet contenenti le parole chiave 'dollar', '$','gold', 'oil', 'job' ed 'economy', hanno scoperto che, oltre ad essere correlate, queste variabili possono essere utilizzate per predire movimenti del mercato finanziario come il prezzo dell'oro, dell'olio crudo, tassi di cambio delle valute ed altri indicatori del mercato finanziario. Nello specifico hanno dimostrato che i cambiamenti giornalieri nel volume di argomenti economici ritwittati sembrano corrispondere al cambio di valore che avviene nel mercato corrispondente il giorno successivo.
Lo stesso team di ricercatori nel 2009, dopo aver collezionato per sei mesi, attraverso le Twitter Streaming API, un sottoinsieme pari all'1% del totale dei tweet, ha misurato la correlazione negativa tra lo stato d'animo collettivo ed indici di mercato come il Dow Jones, il NASDAQ ed il S&P 500.
Lo stato emozionale influenza le decisioni, comprese quelle relative agli investimenti, appare logico immaginare che quando le persone sono più pessimistiche ed incerte circa il futuro sono anche più caute nell'investire. Gli studiosi hanno concluso lo studio affermando che appare possibile prevedere come il mercato azionario varierà semplicemente analizzando i picchi emotivi in Twitter.
Sempre a sostegno della stessa conclusione ci sono lo studio di Bollen et al. (2011) che hanno previsto le variazioni dell'indice Dow Jones con un'accuratezza dell'87,6% basandosi sullo stato d'animo estrapolato dai dati raccolti da Twitter, e quello di Sprenger e Welpe (2010), i quali hanno associato andamenti anomali nel mercato

azionario con la polarità di 249.533 messaggi postati in Twitter relativi al mercato azionario.

Anche i gusti dei consumatori sembrano ormai essere prevedibili, come dimostrato dallo studio di Liviu e Miahela (2011) in cui attraverso 856.196 messaggi pubblicati in Twitter relativi a 30 film sottoposti ad una sentiment analysis, gli studiosi sono riusciti a prevedere con buona approssimazione la posizione dei film nell'autorevole classifica di gradimento della IMDb, nonché i loro incassi al botteghino.

Già Asur e Hiberman (2010) avevano dimostrato come le discussioni online siano correlate con gli indicatori di performance di prodotti o servizi prevedendo anch'essi i risultati ai botteghini di alcuni film nel primo weekend dopo la loro uscita.

Al fine di studiare come l'attenzione generata su Twitter per un film e la sua popolarità, nel social network creino un effetto sulla sua performance nel mondo reale gli studiosi avevano costruito un modello di regressione lineare applicato a tre milioni di tweet.

I risultati sono stati confrontati con gli incassi reali dei box-office dimostrandosi molto simili, e si sono rivelati essere buoni predittori pure dell'indice dell'Hollywood Stock Exchange (HSX index), utilizzato per aggiustare il prezzo delle azioni dei film per riflettere il reale ricavo ottenuto da questi ai botteghini. Il modello di regressione lineare che, per prevedere l'incasso dei botteghini, utilizzava come variabile predittiva la media del numero di tweet per ora che citavano il film, è risultato avere un coefficiente di correlazione di 0.90, ad indicare una fortissima relazione in termini predittivi. Successivamente si è voluto investigare l'importanza di effettuare un'analisi semantica per classificare i tweet in positivi, negativi o neutri, utilizzando LingPipe, un modello di classificazione linguistica.

Per quantificare il sentimento nei confronti di un film hanno calcolato il seguente rapporto:

PNratio =|Tweet positivi| / |Tweet negativi|.

Una nuova regressione lineare che utilizza due variabli (la media del numero di tweet per ora ed il PNratio) ha dimostrato fornire risultati migliori della precedente, con un coefficiente di correlazione pari allo 0,92.

Chaudhari (2012) propone un modello di regressione lineare che può essere adottato per prevedere il profitto di un prodotto (y) utilizzando le 'chiacchiere' online: $y = aA + pP + dD + E$, dove a,b e c sono i coefficienti, A è il parametro che cattura il volume del

vociferare relativamente al prodotto, P la polarità delle opinioni espresse online, D la disponibilità del prodotto sul mercato ed E è l'errore.

"Ascoltando i social network si può capire cosa le persone dicono, come lo dicono, che piattaforme utilizzano per comunicare e che parole usano per descrivere un prodotto"[11], queste parole di Ian Barr, general manager alla Rocket XL di Toronto, suggeriscono l'idea che i Big Data ottenibili dalle piattaforme digitali permettono alle aziende di creare non solo prodotti più vicini ai bisogni e desideri dei consumatori, ma anche creare campagne in linea con il loro immaginario e capaci di alimentarlo. Rocket XL ha indagato attraverso i social che tipo di intrattenimento cercano i teenagers, quali atleti preferiscono e cosa amano collezionare, per basare su queste informazioni l'ideazione della loro campagna pubblicitaria del 2012 che si rivelò un gran successo, con un aumento delle vendite del 24% nei sei mesi successivi e 100.000 nuovi like in Facebook da parte di soggetti appartenenti al target aziendale di riferimento.

Sempre in ambito finanziario UN Global Pulse in collaborazione con la società Crimson Hexagon ForSight ha analizzato 5317 di tweet riguardanti i prestiti in Kenya per rispondere alla domanda: "Quali difficoltà affrontano le piccole imprese in Kenya per accedere a dei prestiti?".

In Figura 2.19 viene riportato il trend dei post da Gennaio 2012 ad Agosto 2013, i picchi registrati a Febbraio-Marzo 2013 sono riconducibili al lancio del nuovo servizio di prestito e risparmio attraverso telefono cellulare, M-Shwari, disponibile per i clienti M-Pesa.; ecco che ancora una volta i dati ricavati online trovano conferma nella realtà.

Lo stesso risultato è stato ottenuto dal confronto (vedi Figura 2.20) tra il trend delle conversazioni in Twitter da Settembre 2010 a Ottobre 2011 in Indonesia relative al prezzo del riso e l'indicatore ottenuto dalle statistiche ufficiali sull'inflazione del prezzo dei beni alimentari (UN Global Pulse, 2013).

[11] Paul, J., 2012. "Big Data take center ice". *Marketing Journal*

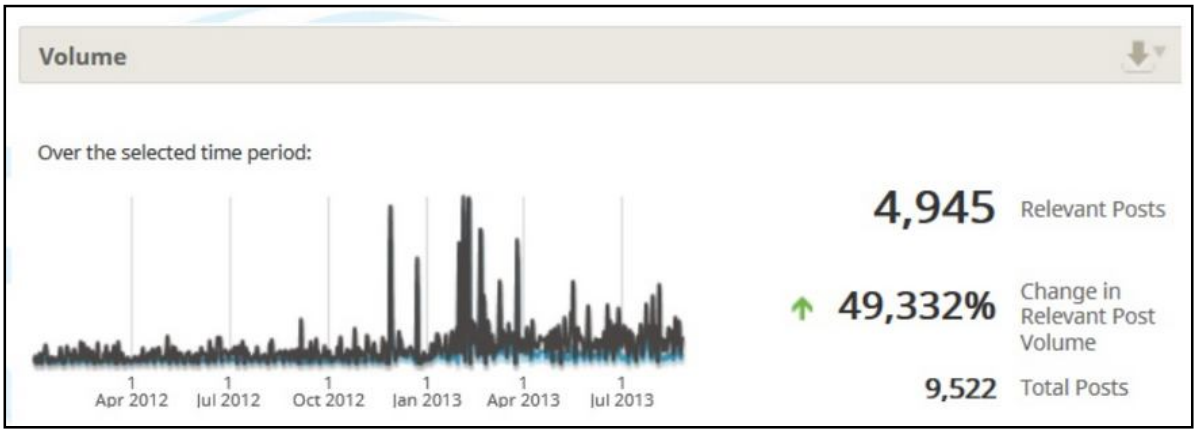


Figura 2.19: Volume di messaggi postati in Twitter relativi al tema dei finanziamenti in Kenya da Gennaio 2012 ad Agosto 2013

Fonte: UN Global Pulse, 2013

Figura 2.20: Confronto tra l'andamento dei tweet relativi al prezzo del riso da Settembre 2010 ad Ottobre 2011 ed il trend dell'inflazione sul prezzo dei beni alimentari nello stesso periodo in Indonesia

Fonte: UN Global Pulse, 2013

Se prevedere le preferenze del pubblico su film, prodotti o quant'altro non è mai stato così semplice, allora a Voices from the Blogs è apparso promettente analizzare quanto viene espresso online dai telespettatori di Sanremo, gara canora in cui è il pubblico che viene direttamente coinvolto nella votazione del vincitore, per anticiparne l'esito finale.

L’esperimento condotto sia nel 2012 che nel 2013, entrambe le volte ha sempre avuto successo.

Il risultato non dovrebbe sorprendere dato che il numero di persone che mentre guarda la televisione utilizza anche pc, smartphone o altri dispositivi collegati alla Rete è in aumento; tanto che nel Dicembre 2012 la Nielsen, società che si occupa dell’auditel negli Stati Uniti, ha stipulato con Twitter un accordo pluriennale per fornire un indicatore complementare a quelli già esistenti che correla l’audience televisiva con l’attività nei social dei telespettatori, e lo stesso verrà fatto in Italia a partire dall’autunno 2014.

Anche Ciulla et al (2012) hanno anticipato il vincitore dello show TV American Idol, competizione canora il cui vincitore era decretato dal pubblico attraverso televoto, analizzando 677.804 tweet contenenti specifici hashtag. L’esempio dello show televisivo è stato considerato dagli studiosi partecipanti al progetto come una versione semplificata di un fenomeno sociale complesso, come potrebbero esserlo anche le elezioni politiche, conducendo all’ipotesi che le informazioni presenti online possono essere utilizzate per creare indicatori in grado di comprendere ed anticipare il processo di formazione dell’opinione pubblica.

Se potrebbe risultare logico che Twitter riesca a prevedere eventi che dipendono direttamente dai gusti delle persone, poiché proprio su Twitter gli utenti esprimono le loro idee, meno immediato appare il collegamento con altri fenomeni che sono stati oggetto di vari studi. In quello di Liviu del 2011 precedentemente citato, vi era anche una parte dedicata alla previsione (rivelatasi esatta) dei vincitori dell’83° Academy Award del 27 Febbraio 2011 come miglior attore protagonista e miglior fotografia. In questo caso il vincitore degli Oscar è determinato da una giuria di esperti e non dai voti dei telespettatori ma Twitter sembra funzionare comunque, paragonabile ad un ‘cervello collettivo’ capace di aggregare con successo aspettative individuali diffuse (Ceron, Curini e Iacus, 2014). Sempre ricollegandosi a quest’immagine di Twitter come ‘cervello collettivo’ si può riportare l’esempio della previsione da parte di Voices from the Blogs delle tracce del tema di maturità del 2013 attraverso l’analisi delle chiacchiere ed opinioni degli studenti online**.**

In ambito **sportivo** lo spin off dell'Università milanese attraverso 100.000 tweet postati tra il 18 e il 25 Agosto aveva emesso il suo verdetto relativamente ai vincitori del Campionato di calcio 2012/13 indovinando tre delle prime quattro posizioni con ben 8 mesi d'anticipo, a campionato ancora da iniziare, già durante il calciomercato.

Come già precedentemente esposto per i Big Data in generale, anche i tweet si sono confermati poter essere utili nella **scienza medica** per monitorare la diffusione di epidemie ed individuare le località con la maggioranza di contagi in modo da poter intervenire in modo tempestivo per circoscriverne l'incidenza.
L'influenza è una pandemia che avviene ogni anno con cicli regolari, ma la localizzazione geografica, il momento e l'entità dei contagi variano sempre e non sono facilmente anticipabili; in letteratura ci sono vari esempi di approcci 'sintomatici' utilizzati per stimare il tasso di contagio influenzale come gli acquisti di farmaci da banco per problemi respiratori o l'assenteismo da scuola. Lo scopo di questi metodi è di anticipare i risultati delle tradizionali fonti di sorveglianza dell'influenza; l'approccio che utilizza i messaggi in Twitter vuole invece fornire una stima in real-time del numero degli ammalati e non anticipare il numero dei futuri contagi. Ciò è possibile proprio perché con questa metodologia di sorveglianza i risultati sono disponibili *live*. Sarebbe possibile avere un risultato simile anche aggregando a livello nazionale in tempo reale, utilizzando sistemi coordinati, il livello di visite ai dipartimenti d'emergenza e ai medici di base, ma la complessità e le ingenti spese che richiederebbe organizzare i dati provenienti da tutte le fonti nelle diverse aree geografiche rendono l'alternativa difficilmente praticabile.
Achrekar et al. (2011), hanno dimostrato un'elevata correlazione tra i dati disponibili online e l'indice di contagio ILI (Influenza like Illness) fornito dal Center for Disease Control and Prevention (CDC) nell'anno 2009/2010. Essi hanno concluso che con un modello di autoregressione con input esogeni (ARX), che utilizza i dati del CDC delle settimane precedenti come componente autoregressivo e i dati dai social network come input esogeno, è possibile avere una stima in tempo reale dello stato dell'epidemia che può sopperire al ritardo con cui gli indici ufficiali sono disponibili.
Volendo confermare i loro risultati nel 2010/2011 hanno collezionato 4,5 milioni di tweet e 2 milioni di post in Facebook; la correlazione tra i dati forniti dal CDC e dai

social network però era scesa da 0,98 dell'anno precedente a 0,47. La motivazione è stata rintracciata nella presenza di tweet dal contenuto fuorviante come quelli che citavano Slime Flu, il pezzo dell'artista V.A.D.O. lanciato nel 2010. Un'analisi testuale ha permesso, però, di dividere i post pertinenti da quelli non rilevanti per la ricerca e di migliorare considerevolmente le stime proposte dal modello.

Come mostrato in Figura 2.21 il confronto tra la percentuale di visite ai medici per sintomi influenzali ed il numero di utenti di Twitter che pubblicavano messaggi rilevanti ai fini della determinazione di un caso di contagio, suggerisce un buon livello di correlazione dopo che sono stati eliminati i retweet (perché non indicano un ulteriore caso di contagio) ed i successivi tweet postati da uno stesso utente (poiché ad esser stata contagiata è sempre una sola persona anche se posta più messaggi).

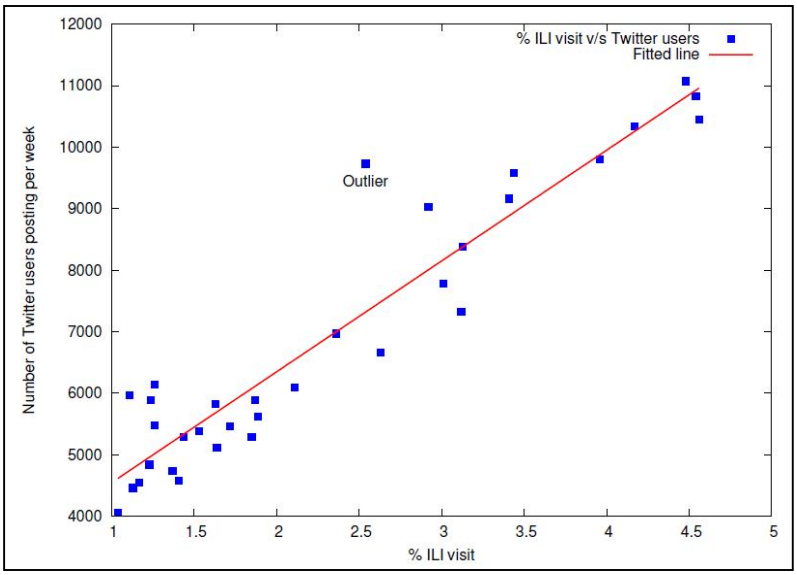


Figura 2.21: Confronto tra il numero di utenti di Twitter che cinguettavano utilizzando parole chiave relative all'influenza e la percentuale di contagi fornita dal CDC nell'anno 2010/2011
Fonte: Achrekar, 2011

Pure il confronto tra le previsioni ottenute attraverso Twitter, Facebook e la loro combinazione con il numero dei contagi reali forniti dal CDC per tre regioni del nord, centro e sud degli Stati Uniti (vedi Figura 2.22) sembra confermare che analizzare quanto viene detto nei social network può fornire le basi per una stima corretta, veloce e in real-time della diffusione della malattia.

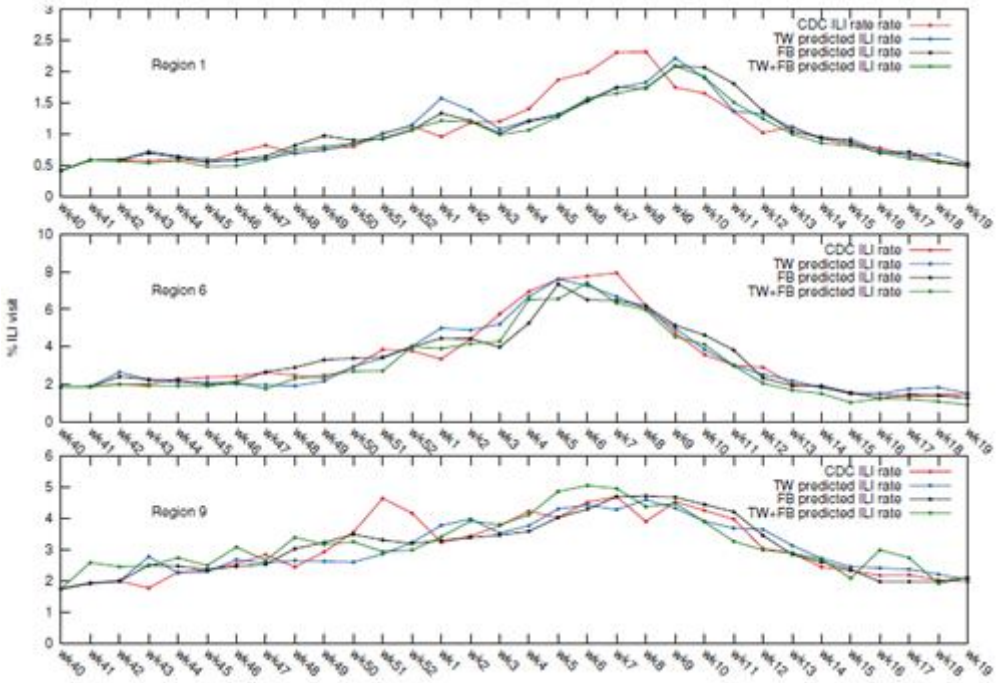


Figura 2.22: Confronto tra le previsioni ottenute attraverso Twitter, Facebook e la loro combinazione con i dati reali forniti dal CDC per tre Regioni degli Stati Uniti

Fonte: Achrekar, 2011

Anche Paul e Dredze (2011) hanno voluto cimentarsi nel tentativo di monitorare il livello di contagi influenzali negli Stati Uniti; utilizzando il modello ATAM (Ailment Topic Aspect Model) hanno analizzato settimanalmente Twitter da Agosto 2009 ad Ottobre 2010.

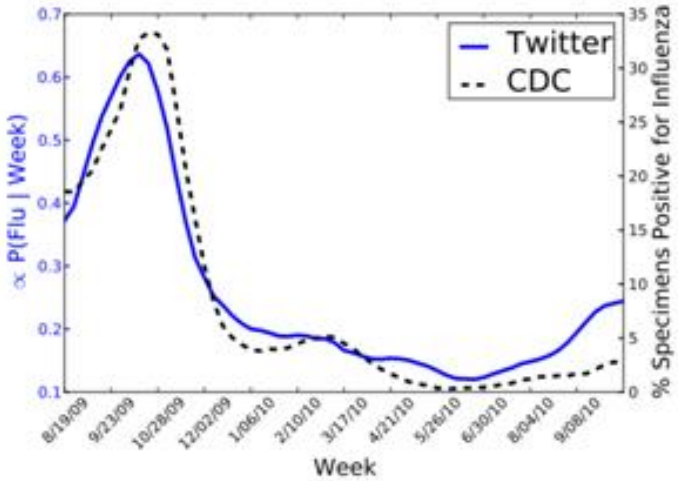


Figura 2.23: Previsioni fatte utilizzando Twitter confrontate con i dati del CDC del numero di contagi influenzali da Agosto 2009 ad Ottobre 2010

Fonte: Paul e Dredze, 2011

Come mostra la Figura 2.23, anche in questo caso le previsioni fatte utilizzando i tweet si sono rivelate essere molto vicine alle misure governative ufficiali, quelle del report settimanale FluView del CDC (Centers for Disease Control and Prevention), con un coefficiente Pearson di correlazione del 0.958. Quello adottato dai ricercatori è un modello probabilistico che associa le parole ad argomenti latenti, non è quindi focalizzato su questioni specifiche ma risponde a domande aperte e offre vari risultati: dall'osservazione geografica dell'incidenza di alcune patologie alla rilevazione di comportamenti traducibili in fattori di rischio, fino alla stima del numero di malati.

L'interessamento nell'analisi di Twitter e delle altre piattaforme sociali per la sorveglianza delle epidemie influenzali è testimoniato, ad esempio, dalla gara sponsorizzata dal dipartimento di Salute e Servizi Umani degli Stati Uniti nell'estate del 2012 in cui i ricercatori si sfidavano per progettare un'applicazione online capace di predire i maggiori picchi di alcune malattie.

Continui miglioramenti alle metodologie disponibili sono quindi in corso, come quello avvenuto grazie ad alcuni ricercatori dell'Università John Hopkins del Dipartimento di Scienze informatiche, i quali hanno creato uno strumento per estrarre non solo i tweet relativi a casi influenzali, ma persino per filtrare quali di questi siano realmente collegati a veri contagi e non a persone che parlano solo di aver paura di ammalarsi o che menzionano personaggi famosi che si sono ammalati.

I sofisticati metodi statistici utilizzati si basano su tecnologie processanti basate su studi del linguaggio umano che permettono al sistema di distinguere la differenza tra le affermazioni "ho l'influenza" e "sono preoccupato di prendermi l'influenza", limitatamente alla lingua inglese.

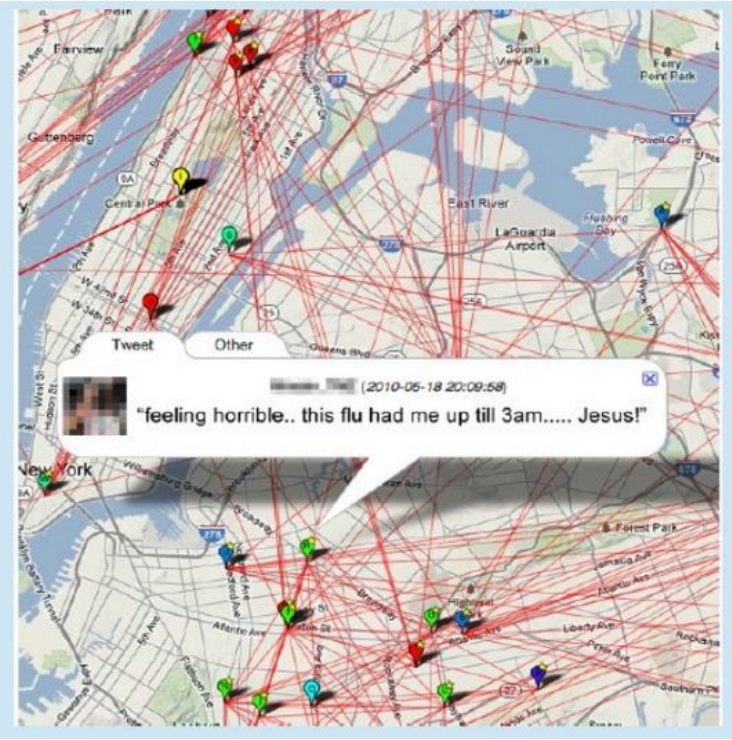


Figura 2.24: Collegamenti tra i tweet sull'influenza studiati alla Hopkins University
Fonte: Maynard, 2013

Sembra quindi che il monitoraggio e l'analisi di dati reperibili sul Web stiano diventando parte dello scenario della sorveglianza dell'influenza.
Dai tweet si possono ricavare anche altre informazioni utili in campo medico, come la variazione della percezione collettiva di temi riguardanti la salute in seguito ad una campagna sociale, le abitudini diffuse tra i cittadini relativamente alle pratiche di automedicazione o gli stili di vita che possono aiutare l'insorgere di malattie non comunicabili.
Esattamente ciò che ha fatto il Dipartimento di Scienze Informatiche della John Hopkins University nel 2010 quando ha analizzato i tweet relativi al consumo dei farmaci da banco, all'attività fisica svolta da diversi gruppi socioeconomici e all'uso-abuso di alcool tra i giovani, così da fornire alle pubbliche autorità che si occupano della salute dei cittadini previsioni in tempo reale sfruttabili nell'organizzazione delle loro azioni.
Anche UN Global Pulse, WHO e Microsoft hanno collaborato allo scopo di scoprire come sfruttare Twitter per creare degli indici utili a monitorare nel tempo l'evoluzione dei comportamenti e dei pensieri delle persone riguardo ai fattori di rischio delle

malattie non comunicabili, per poter comprendere così quali interventi potrebbero essere più efficaci per dar vita a migliori attività di prevenzione e di comunicazione.

Le parole chiave scelte per filtrare le conversazioni rilevanti su Twitter sono state categorizzate in tre gruppi principali:

- Attività fisica/obesità: *fat, work-out, lazy, yoga, gym, overweight, obese, jogging, sedentary, fast food, pizza, ice cream, chocolate, fries, fruit, vegetable, hot dog, soda, chips, fried, burgers, carbs, diet, mcdonalds, red bull, coca cola, pepsi, calories;*
- Consumo di alcool: *drunk, beer, alcohol, wine, vodka, whiskey, whisky, booze, absolut, heineken, carlsberg, Smirnoff*
- Tabagismo: *smoking, smoke, tobacco, cigarette, ecigarettes, electronic cigarette, nicotine, cigars, shisha, smokeless*

Una *sentiment analysis* è stata condotta sui tweet così estratti nei mesi di Gennaio, Aprile, Luglio e Ottobre 2011-2012-2013 per misurare il livello di positività delle emozioni associate ad ogni specifica keyword, rivelando, ad esempio, che termini generici come *fast food* sono associati a sentimenti più negativi rispetto ad altri più specifici come *chocolate*, *ice cream*, patatine o pizza.

E' stato poi calcolato l'Health Index, un indice per ognuna delle parole chiave prescelte relativo all'atteggiamento generale verso quella *keyword*, ottenuto come segue:

$$I_k = \frac{\rho P_k}{\rho P_k + \rho N_k}$$

Dove Rho è la forza del sentimento espresso, con 0<Rho<1; P ed N sono rispettivamente il numero di tweet positivi o negativi contenenti la parola chiave k di cui viene calcolato l'indice, il quale varia da zero (massima negatività) ad uno (massima positività). Di seguito si riporta l'esempio del calcolo dell'indice per la parola chiave '*beer*' effettuato dalle Organizzazioni.

Keyword	Strength Pos	Strenght Neg	Count Pos	Count Neg
Beer	0.84	0.72	11500	2768

$$I_{beer} = \frac{0.836*11500}{[(0.836*11500)+(0.716*2768)]}$$

Ottenuto l'indice è poi possibile seguirne la variazione nel tempo, come mostrato nelle Figure 2.25, 2.26 e 2.27 relativamente al tema della nutrizione/attività fisica e del fumo. Tutte le parole chiave legate a quest'ultimo argomento hanno registrato nel periodo considerato un atteggiamento sempre più negativo, solo il termine 'electronic cigarette' ha riportato un aumento del suo Health Index.

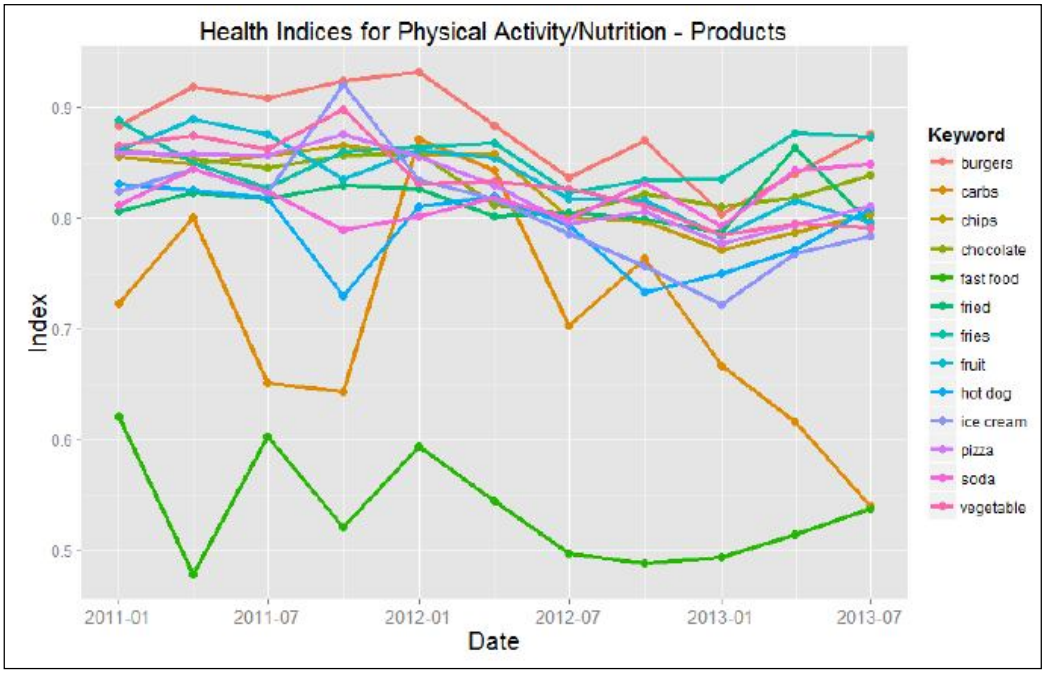


Figura 2.25: Trend dell'Health Index delle parole chiave relative ai prodotti della categoria sulla nutrizione da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

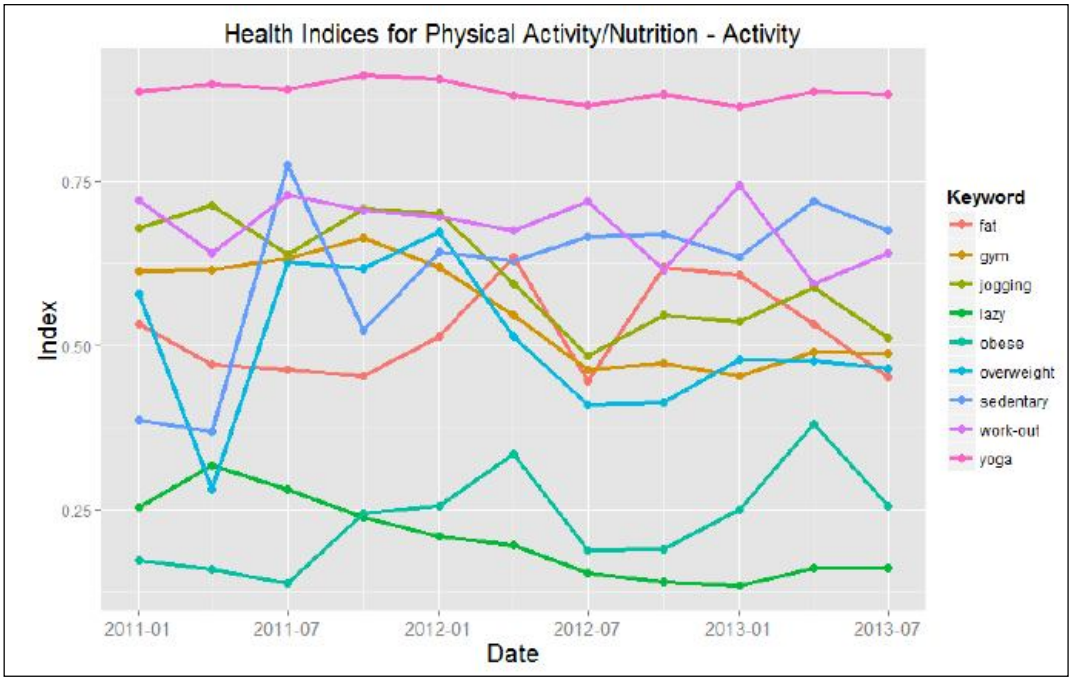


Figura 2.26: Trend dell'Health Index delle parole chiave della categoria sull'attività fisica da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

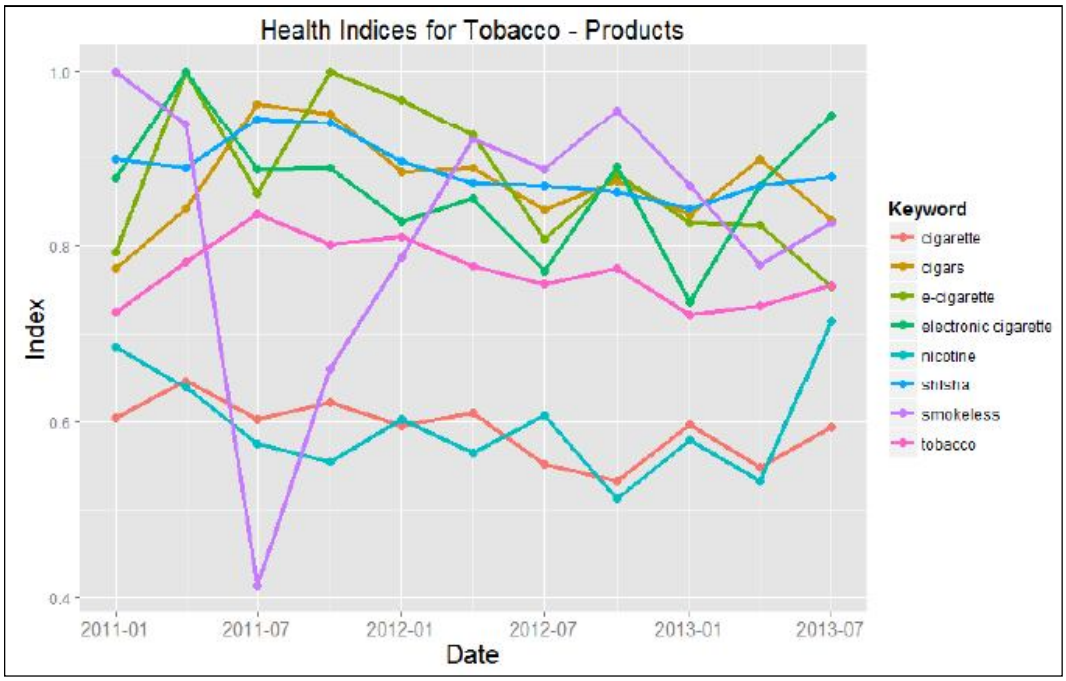


Figura 2.27: Trend dell'Health Index delle parole chiave relative ai prodotti della categoria sul tabagismo da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

Calcolando poi la differenza dell'indice ottenuto per la stessa keyword in due momenti diversi, si può ottenere una visualizzazione del suo andamento (vedi Figure 2.28-2.29.

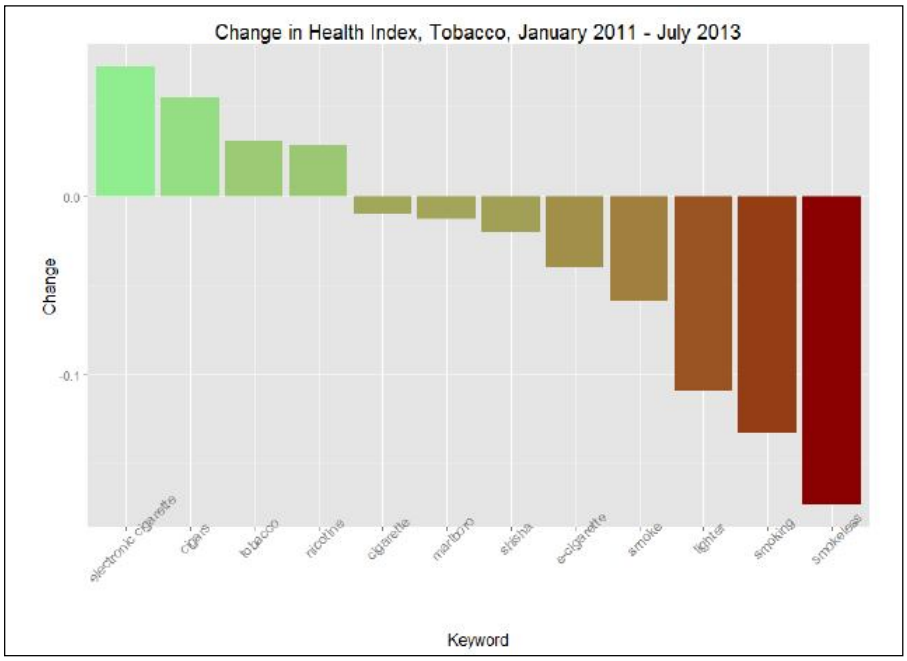


*Figura 2.28:*Variazioni Health Index delle *keyword* sul tabagismo da Gennaio a Luglio 2011
Fonte: UN Global Pulse, WHO e Microsoft, 2013

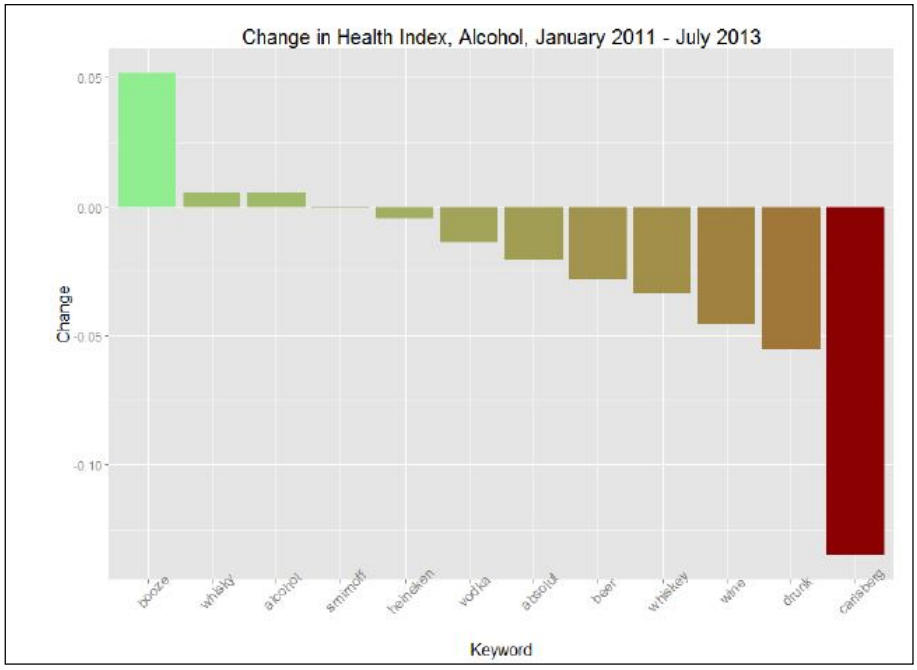


Figura 2.29: Variazioni Health Index delle *keyword* sull'alcolismo da Gennaio a Luglio 2013
Fonte: UN Global Pulse, WHO e Microsoft, 2013

Il cambiamento nella coscienza pubblica relativamente alle diverse parole chiave analizzate, ΔPC, è poi stato calcolato nel seguente modo:

$$\Delta PC_k = \Delta I_k \, (P+N)$$

Mantenendo come esempio la parola '*beer*', la modalità è mostrata di seguito:

Keyword	**Index 2011**	**Index 2013**	**Total Count 2013**
Beer	0.912	0.884	42356

$$\Delta PC_{beer} = (0.884 - 0.912) * 42356$$

La ricerca ha registrato una diminuzione significativa della coscienza pubblica da Gennaio 2011 a Luglio 2013 verso termini quali 'smoke', 'smoking' e 'drunk', come si vede in Figura 2.30 e 2.31.

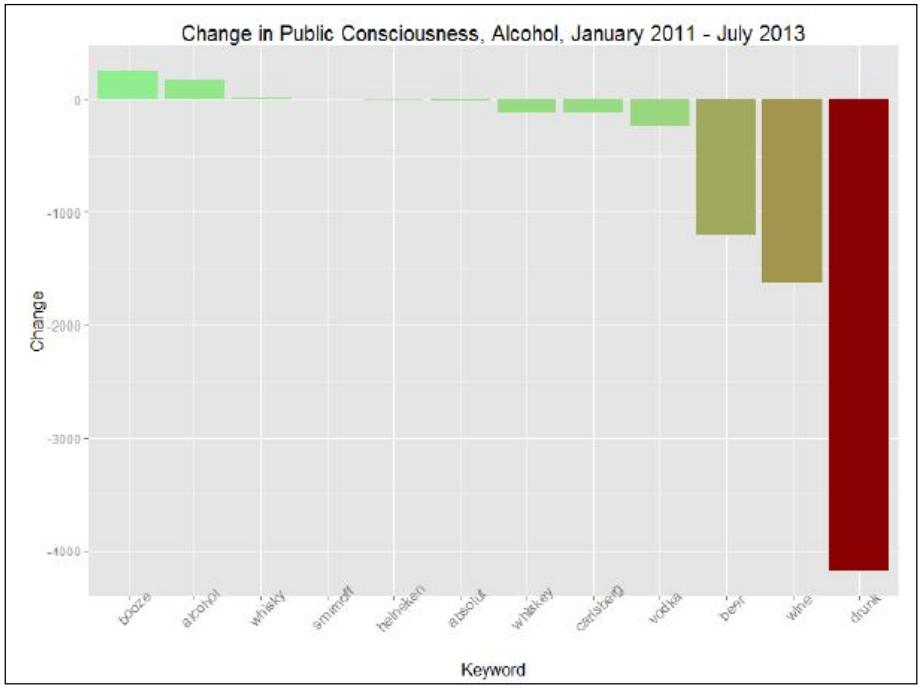


Figura 2.30: Cambiamento della coscienza pubblica relativamente alle parole chiave della categoria sull'alcolismo da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

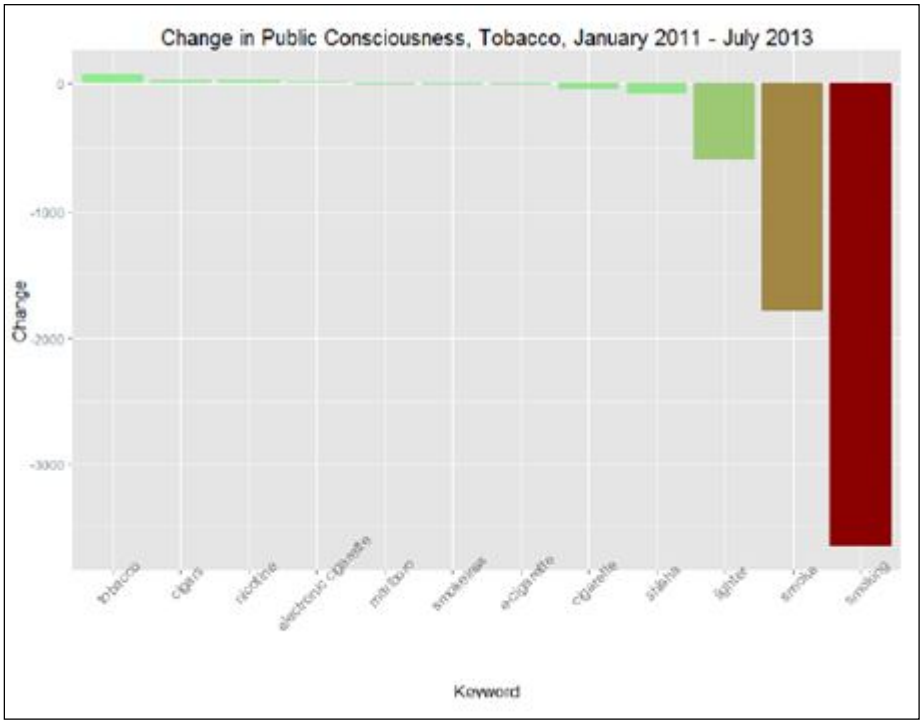


Figura 2.31: Cambiamento della coscienza pubblica relativamente alle parole chiave della categoria sul tabagismo da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

Infine il dataset è stato disaggregato per genere e localizzazione estraendo queste informazioni dalle descrizioni disponibili sugli account Twitter.

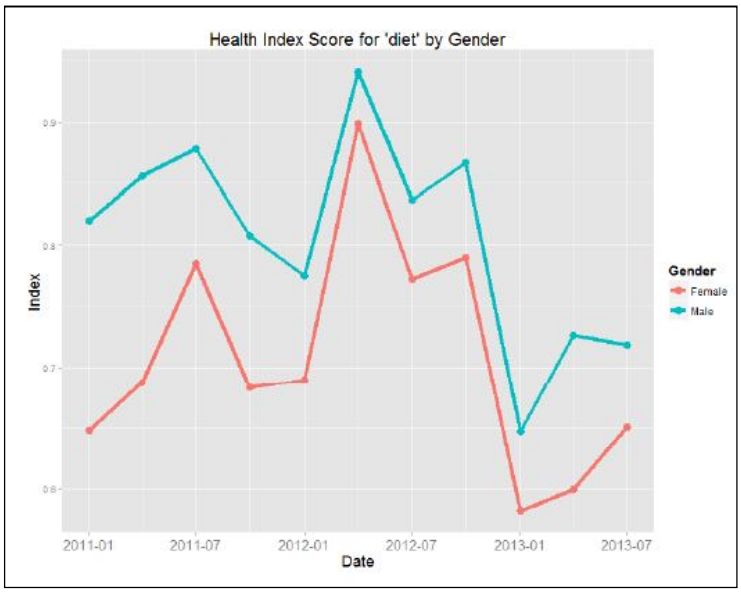


Figura 2.32: Confronto del trend maschile con quello femminile dell'Health Index della parola chiave 'diet' da Gennaio 2011 a Luglio 2013

Fonte: UN Global Pulse, WHO e Microsoft, 2013

Emerse, ad esempio, una maggiore negatività nei tweet relativi alle parole '*fat*' o '*diet*' postati da donne rispetto a quelli degli uomini sia a Gennaio 2011 che Luglio 2013, anche se la differenza nel tempo è andata diminuendo (vedi Figura 2.32).

Si è persino arrivati a **misurare l'umore di un Paese**; informazione che può essere utili sia nel campo della psicologica che per l'eventuale sviluppo di politiche pubbliche volte a migliorare il benessere dei cittadini. Guardare cosa dicono le persone nei social media sembra essere un modo per superare le difficoltà dei tradizionali questionari che indagano il livello di felicità degli intervistati attraverso una loro autodichiarazione del loro stato di benessere. Studi come quelli di Schwarz (1987) e di Akay e Martisson (2009), infatti, testimoniano la facilità con cui le persone si fanno influenzare da situazioni contingenti (come il giorno della settimana, il meteo o eventi avvenuti appena prima dell'intervista) quando rispondono a domande sul loro grado di felicità.
Nei social network ognuno condivide sia le sue piccole o grandi gioie di ogni giorno come un compleanno, la nascita di un figlio, un bel voto, una buona cena, sia i dispiaceri o i fastidi quotidiani come un furto o un treno in ritardo; questa miniera d'informazioni è sicuramente sfruttabile per capire la felicità di una comunità.
La Loughborough University ha avviato il programma Emotive (disponibile al sito emotive.lboro.ac.uk) il quale prevede l'analisi di 2.000 micropost al secondo pubblicati su Twitter per capire lo stato d'animo degli abitanti dell'Inghilterra, in modo da avere una sorta di lente d'ingrandimento sul cuore del Paese.
Voices of the Blogs dal 2012 ha voluto analizzare il fenomeno sociale complesso che è la felicità degli italiani, a livello aggregato e non per singolo individuo, come già aveva fatto Quercia et al. (2012) per l'area metropolitana di Londra. Attraverso il metodo iSA hanno condotto un'analisi semantica dei post degli italiani dal 31 gennaio 2012 al 31 dicembre 2012 per un totale di 43 milioni, classificati come post felici, infelici o 'altro'. Per ogni provincia italiana poi è stato calcolato l'indice iHappy, pari alla percentuale dei post categorizzati come felici sul totale di quelli felici ed infelici, tralasciando così quelli classificati come 'altro'. L'indice così creato varia da 0 a 100, con 100 pari alla situazione di massima felicità. Nel 2012 il valore medio dell'iHappy in Italia è risultato essere stato pari al 45,6%, con dei picchi di felicità registrati in giorni particolari segnati

da eventi improvvisi come la liberazione di Rossella Urru il 18 Luglio e il gol di Balotelli contro la Germania nella semifinale dell'Europeo di calcio il 28 Giugno.

I risultati sempre aggiornati sulla felicità italiana sono riportati in un sito appositamente dedicato (www.blogsvoices.unimi.it/) di cui in Figura 2.33 e 2.34 e 2.35 si riportano gli aggiornamenti disponibili nella giornata del 19 maggio 2014.

Figura 2.33: Mappa della Felicità in Italia il 19 maggio 2014

Fonte: Voices of the Blogs, 2014

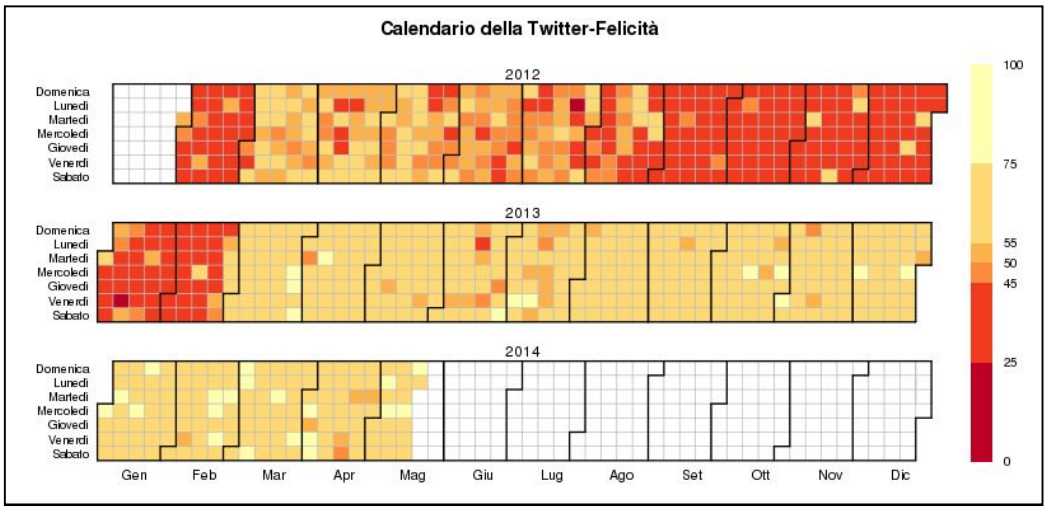


Figura 2.34: Calendario della Felicità twittata in Italia dal 1 Gennaio 2012 al 19 maggio 2014

Fonte: Voices of the Blogs, 2014

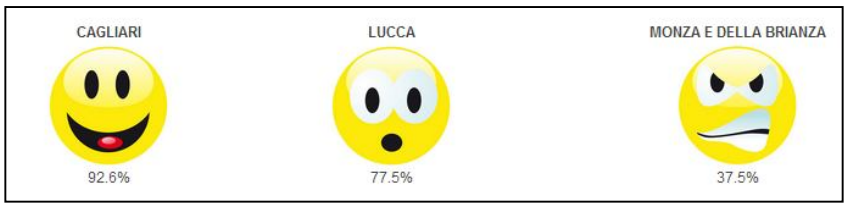


Figura 2.35: Indice iHappy distinto per Regione, risultati del 19 maggio 2014

Fonte: Voices from the Blogs

La Figura 2.33 rappresenta il calendario della felicità degli italiani: più il periodo è colorato di giallo e più significa che sono stati registrati post categorizzati come felici.
In Figura 2.35 si può vedere l'indice iHappy italiano del 19 maggio 2014 distinto per Provincia: la più felice è Cagliari con un iHappy del 55%, mentre la più triste sono Monza e Della Brianza.
Le zone più felici non corrispondono però alle parti della penisola con i migliori indici istituzionali; infatti l'indice creato da VfB riflette uno stato soggettivo e personale dei 'cinguettatori', mentre le classifiche sulla qualità della vita delle città italiane solitamente si basano su indicatori aggregati come tasso di criminalità, inquinamento, ricchezza, livello d'occupazione. Attraverso una regressione condotta con errori standard robusti i ricercatori di Vfb hanno rintracciato delle determinanti della felicità degli italiani: meteo (con 1° in più l'iHappy cresce di 0,5, ma solo sotto i 30°), giorno (il Sabato il giorno più felice, la Domenica un giorno triste, a Natale 15 punti in più) , mese, stagione (la Primavera rende tutti più felici), fatti del giorno (lo spred scende e la felicità italiana sale), variabili economiche, geografiche, sociali, demografiche (le città con una media più alta d'età hanno un indice maggiore) e istituzionali.

Sempre relativamente a studi incentrati sulla società, UN Global Pulse vuole dimostrare la potenzialità dell'analisi dei Big Data per scoprire come si sviluppa la discussione riguardo ad argomenti relativi ai *Millenium Development Goal*, anche se coloro che stanno alimentando la discussione non sono consapevoli dell'esistenza degli obiettivi ufficiali posti dall'ONU da raggiungere entro il 2015. L'interesse nasce perché UN Global Pulse sta partecipando alla *United Nation Millenium Campaign* (UNMC), una campagna avviata in seguito alla *Millenium Declaration*, risultato del summit nel Settembre 2000 dei 189 stati membri dell'ONU, la quale ha come obiettivo il raggiungimento entro il 2015 degli otto *Millenium Development Goal*, i quali vanno dal dimezzamento della povertà, al blocco della diffusione dell'HIV, alla diffusione dell'educazione primaria universale.
Al sito post2015.unglobalpulse.net è possibile visualizzare in tempo reale il totale, continuamente aggiornato, di utenti, di tweet e di retweet in inglese, francese, spagnolo e portoghese che riguardano i problemi che gli obiettivi internazionali puntano di risolvere entro il 2015, con una distinzione per genere e lingue principali (vedi Figura

2.36). Viene inoltre fornita la classifica dei 20 Paesi che da Gennaio 2013 hanno maggiormente twittato circa questi temi. Tutto ciò è possibile monitorando le conversazioni online ogni giorno utilizzando una tassonomia di 25.000 parole chiave per selezionare tutti i messaggi pubblicati in Twitter considerati pertinenti ai *Millenium Development Goal* e geolocalizzandoli.

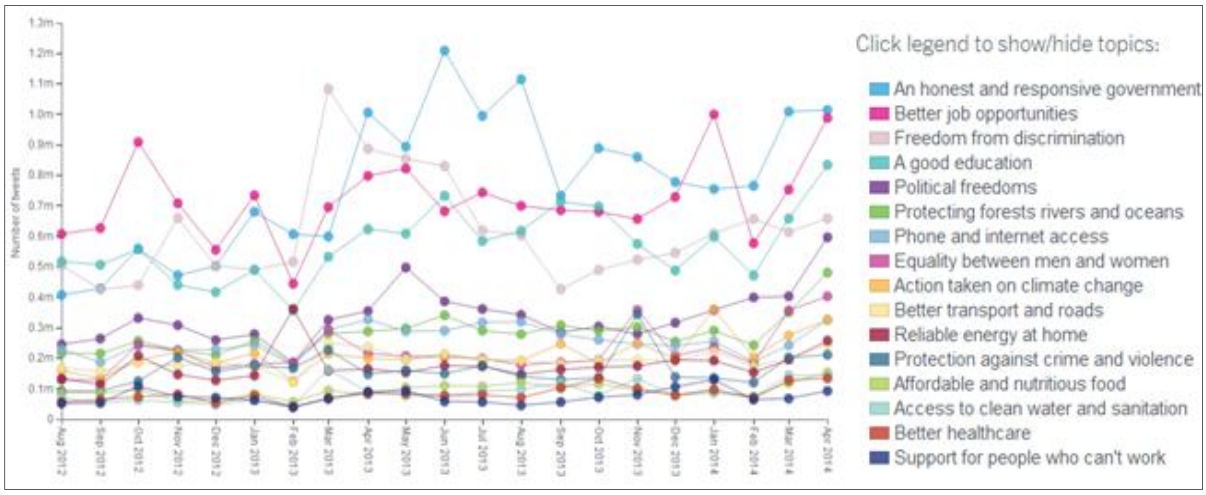


Figura 2.36: screenshot del sito post2015.unglobalpulse.net il giorno 20 maggio 2014
Fonte: post2015.unglobalpulse.net, 2014

Sfruttando le opportunità offerte dalla geolocalizzazione dei tweet, Mocanu et al. (2013) hanno investigato l'omogeneità **linguistica** e la distribuzione geografica di differenti lingue in regioni multilingue analizzando un dataset di messaggi geolocalizzati postati nel microblog.

Un altro modo di sfruttare la geolocalizzazione dei post in Twitter è quello sviluppato dalla United States Geological Survey (USGS), centro sismologico americano che monitora Twitter per rintracciare significativi picchi nel volume di messaggi su **terremoti**. Attraverso la localizzazione dei messaggi e l'estrazione di informazioni, i sismologhi della USGS verificano la reale avvenuta dei microsismi, il loro epicentro e magnitudo e valutano se innescare un'allerta per possibili terremoti imminenti. Il 90% dei casi rilevati è risultato essere convalidato da report successivi, può Twitter addirittura aiutare a salvare delle vite?

Sempre relativamente al mondo della **natura** studiato attraverso Twitter c'è 'Air Twitter', un progetto della Washington University a St. Louis che raccoglie commenti e

foto postati in Twitter riguardo ad incendi o tempeste di polvere come tecnica per ottenere indicazioni in tempo reale sulla qualità dell'aria.

Nel **settore turistico** italiano Formez, il centro nazionale di servizi, assistenza, studi e formazione per l'ammodernamento della Pubblica Amministrazione, ha collaborato con Promo Pa, fondazione che si occupa di ricerca e progetti per la Pubblica Amministrazione, per condurre un'analisi semantica, affidata al Politecnico di Milano, dei commenti dei turisti sull'Italia. I post, raccolti da dicembre 2012 a fine maggio 2012, provenivano per il 90% da Twitter, il restante 10% da Facebook, Lonely, Planet e Tripadvisor. La piattaforma di *sentiment analysis* ha analizzato tutto ciò che veniva detto riguardo a sette città, note mete turistiche. La prima lettura dei dati ovviamente ha riguardato la positività della reputazione, la quale risultata essere maggiore per la Costiera Amalfitana al 91%, per Amalfi all'88% e per Lecce e Lucca entrambe all' 82%, seguite da Roma, Napoli e Palermo al 79%, 76% e 62%.

Interessante è stato poi guardare quali erano gli argomenti dei post: la qualità del cibo era risultata essere la motivazione principale di apprezzamenti sulla penisola, mentre le critiche riguardavano spesso il traffico e i ritardi dei mezzi di trasporti pubblici, dei quali però si apprezzava la capillarità. Attraverso l'analisi dei social network è stato quindi possibile capire l'opinione dei turisti sull'Italia senza tediarli con noiosi questionari durante o dopo il loro soggiorno.

Molto interesse è stato destato dalla possibilità di avere una visione molto dettagliata dell'evoluzione dell'opinione dei cittadini rispetto alle loro **intenzioni di voto** in risposta agli avvenimenti della campagna elettorale, come scandali o dibatti, monitorando i social network.

Ascoltare quello che si dice su Twitter rende possibile rivelare le preferenze dell'elettorato in tempo reale in modo molto simile a quello reso possibile dai sondaggi, i cui tempi tecnici per fornire dei risultati sono però più lunghi. Ciò può rivelarsi uno strumento molto utile a fini di tattica elettorale: scoprire sul nascere nuovi trend nell'opinione degli elettori può aiutare a definire le prossime mosse del candidato o del partito per ottenere maggior consenso.

Tra i campi delle scienze sociali quello politico è uno dei pochi nel quale una previsione viene confrontata con un dato certo, incontrovertibile ed esogeno: il voto degli elettori, si va quindi oltre il *nowcasting* e si parla di vero e proprio *forecasting*.

Gli studi condotti per dimostrare la capacità predittiva di Twitter in ambito politico sono numerosi, come anche le perplessità suscitate dai risultati. Nel presente paragrafo s'intende presentare alcuni dei casi che hanno destato maggiore attenzione, senza tralasciare le critiche che ne sono seguite.

Tumasjan et al. (2010) hanno semplicemente relazionato il numero di tweet che citavano un partito con i voti ottenuti alle elezioni federali tedesche del 2009 per dimostrare una correlazione significativa tra di essi.

Anche Sanders e den Bosh (2013) hanno contato quanti tra i 170 mila tweet, raccolti tra il 2 settembre e il giorno delle elezioni olandesi, il 12 settembre 2012, menzionavano i partiti politici in lista per prevedere i vincitori. La Tabella 2.2 mostra per ogni partito la differenza tra i risultati delle elezioni, dei tradizionali sondaggi e la percentuale di tweet in cui il partito era stato menzionato il giorno delle elezioni (12 Settembre), il giorno precedente, i cinque e i dieci giorni antecedenti il voto. Vengono riportati nelle ultime quattro righe gli errori assoluti medi (MAE), ossia gli scostamenti medi della previsione rispetto al dato reale, ed i coefficienti di correlazione tra le colonne sovrastanti ed i risultati dei sondaggi e delle elezioni.

Partito	Election 12/09	Poll 11/09	Tweet 12/09	Tweet 11/09	Tweet 2-11/09	Tweet 7-11/09
VVD	26,6	23,7	24,6	18,9	20,7	20,6
PVDA	25,1	23,4	18,5	21,7	20,2	22,2
PVV	10,2	11,6	13,6,2	11,5	10,7	11,4
SP	9,8	13,9	8,7	9,7	12,0	10,3
CDA	8,6	8,3	6,0	7,5	8,6	8,6
D66	8,1	7,9	9,8	9,7	9,0	8,5
CU	3,2	3,7	2,6	2,9	3,0	2,7
GL	2,4	2,7	7,0	8,9	8,6	8,8
SGP	2,1	1,7	3,2	4,4	2,9	2,8
PVDD	2,0	1,8	3,6	3,5	3,2	3
50PLUS	1,9	1,7	2,4	1,3	1,1	1,1
MAE Election		1,1	2,4	2,3	2,0	1,7
Corr. Election		0.93-1,0	0,82-0,99	0,78-0,98	0,83-0,99	0,84-0,99
MAE poll	1,1		2,4	2,3	2,0	1,7
Corr. poll	0,98		0,93	0,94	0,96	0,96

Tabella 2.2: Confronto tra i risultati delle elezioni, dei sondaggi e quanto Twitter suggeriva in % in diversi archi temporali

Fonte: Sanders e van den Bosh, 2013

L'errore assoluto medio dei tradizionali sondaggi, misura della loro accuratezza, risulta essere minore in tutti e tre gli archi temporali considerati, dimostrando come questi siano dei migliori predittori rispetto al semplice conteggio delle menzioni in Twitter dei partiti senza ulteriori indagini relativamente al contenuto, nonostante la correlazione dimostrata tra risultati elettorali ed il numero di tweet sia stata comunque circa del 95%.
Jungerr et al. (2012) hanno replicato al lavoro di Tumasjan, facendo emergere che se nel conteggio del numero di menzioni dei diversi partiti fosse stato considerate anche il Partito Pirata, un partito minore, allora sarebbe risultato essere questo il vincitore.
Anche lo studio di Sang e Bos (2012) dimostra come il mero conteggio delle menzioni dei partiti candidati alle elezioni del Senato danese nel 2011 non fosse una base sufficiente per ottenere delle previsioni corrette, ma che, migliorando la qualità dei dati raccolti attraverso l'eliminazione di tweet ambigui o fuorvianti, come quelli che menzionavano più partiti o quelli postati dallo stesso utente successivi al primo, le stime ottenute fossero migliori. Sulla stessa linea sono anche le considerazioni di Skoric et al. (2012).
Andando oltre ad un'analisi solamente quantitativa altri studi hanno analizzato i tweet anche dal punto di vista semantico per poter fare una previsione più accurata e certa, come quello di O'Connor et al. del 2010.
Grazie ad una *sentiment analysis* dei tweet raccolti Tjong Kim Sang e Bos (2012) hanno previsto le elezioni del Senato danese del 2011 tanto bene quanto i sondaggi tradizionali, mentre un'analisi multi-piattaforma sul contenuto dei messaggi pubblicati su Facebook, Twitter, Google e YouTube si è addirittura dimostrata migliore nello stimare le elezioni inglesi del 2010 (Franch, 2012).
Nel tentativo di comprendere quali elementi incidano sull'errore quadratico medio (MAE) di una previsione Ceron (2012) ha analizzato differenti variabili indipendenti per arrivare alla conclusione che una *sentiment analysis* dei social fornisce previsioni più accurate quando il numero di potenziali elettori aumenta (perché una cattiva interpretazione dell'intenzione di voto di uno di essi pesa meno), ci si concentra sui partiti di massa (quelli radicali di estrema destra tendono ad essere sottorappresentati, quelli di estrema sinistra sovra rappresentati), quando vi è poco astensionismo (chi esprime la sua opinione online non per forza poi la esprime anche

ai seggi) e quando la disponibilità di tweet aumenta, anzi quest'ultimo fattore ha un effetto quadratico sulla previsione del MAE.

Gayo-Avello (2012) sconfessa, però, anche l'accuratezza dei pronostici basati su una *sentiment analysis* dei tweet raccolti e nel suo contributo critica e commenta i difetti e gli errori di studi fatti in precedenza da altri ricercatori:

- le pubblicazioni delle previsioni avvengono quando l'evento c'è già stato e non prima che si verifichi;
- vengono pubblicati solo i casi in cui le stime sono state confermate dai dati reali, mentre i tentativi falliti non sono resi noti facendo credere che l'anticipazione del risultato elettorale sia sempre corretta;
- la difficoltà di cogliere sarcasmi e humor, molto diffusi nei commenti a sfondo politico, che si somma all'ortografia e alla sintassi particolare che è adottata nei social network, rende le previsioni elettorali molto complicate. Servirebbero, quindi, degli algoritmi in grado di apprendere in continuazione per diventare sempre più capaci di cogliere le sfumature del linguaggio utilizzato in Rete, impiegando tecniche di *machine learning*;
- bisognerebbe correggere le previsioni in base agli attributi demografici degli utenti che hanno postato i tweet raccolti per superare il problema della non rappresentatività dell'intero elettorato poiché alcuni strati sono sovra-rappresentati mentre altri sottorappresentati, come già suggerito da Mislove et al. (2011).

  Ceron (2012) invece sostiene che nel caso delle previsioni in ambito politico non bisognerebbe tanto preoccuparsi di quanto la Rete rappresenti la popolazione a livello socio-demografico, ma quanto a livello di ideologie, ossia il livello di distribuzione di schieramento politico tra coloro che utilizzano i social network e sui quali esprimono le loro opinioni, utilizzate poi per fare previsioni elettorali. In questo modo sarebbe possibile pesare i risultati ottenuti per fornire delle previsioni più accurate.
- alcuni tweet sono creati in automatico da robot o da profili creati dai partiti stessi, e che quindi con la loro propaganda online sfasano i risultati facendo registrare a loro favore messaggi che non rappresentano veramente l'opinione di un elettore.

  E' però vero che i messaggi propagandistici vengono condivisi solo se c'è qualche utente che si rispecchia, ed inoltre la presenza di una buona propaganda

sicuramente si riflette in una maggiore attenzione e quindi probabilmente in un numero maggiore di voti.

A rispondere in parte a queste accuse ci sono i lavori di Voices from the Blogs, innanzitutto previsioni elettorali pubblicate prima dei risultati, e che si avvalgono del metodo di *sentiment analysis* iSA, il quale avendo uno stadio in cui è richiesta una codifica manuale riesce a superare alcuni problemi linguistici per la categorizzazione dei tweet. VfB ha applicato il metodo iSA a diverse tipologie di elezioni: le presidenziali tra Sarkozy e Hollande del 6 maggio 2012, le legislative francesi del 10 giugno 2012, le presidenziali americane vinte da Obama nel 2012, le primarie del centrosinistra italiano e le politiche italiane del 2013. Scenari diversi sotto il profilo del regime politico (presidenziale o parlamentare), per tipologia di competizione (multipartitica o con due candidati) e per legge elettorale.

I tweet selezionati contenevano parole chiave precise come il nome dei partiti o dei loro leader e gli hashtag collegati alle loro campagne elettorali. Un tweet è stato considerato come intenzione di voto se:

- dichiarava in modo esplicito l'intenzione di votare un candidato/partito:
- conteneva un commento positivo verso un partito/candidato ed anche un hashtag relativo alla campagna elettorale per rendere l'opinione più affidabile;
- conteneva un commento negativo verso un partito/candidato e un hashtag di un altro partito/candidato (a cui si presumeva sarebbe andato il voto) per diminuire il peso dei tweet che avevano solo lo scopo di insultare senza fornire alternative di voto in modo credibile;
- confrontava due partiti/candidati esprimendo un giudizio negativo per uno ed uno positivo per l'altro.

Anche i retweet delle quattro tipologie appena descritte sono stati catalogati come intenzioni di voto (Ceron, Curini, Iacus, 2014).

Nel caso delle elezioni americane del 2012, che vedevano contrapposti Romney e Obama, la vittoria schiacciante di quest'ultimo non rispecchiava un vantaggio avuto fin dall'inizio delle elezioni, ma, analizzando quanto emerge dalle conversazioni in Twitter la rielezione del Presidente americano è stata decretata da uno scatto finale poiché per tutto il periodo delle elezioni i candidati sono stati in testa a momenti alterni, in

concomitanza agli avvenimenti che hanno caratterizzato la campagna elettorale (vedi Figura 2.37).

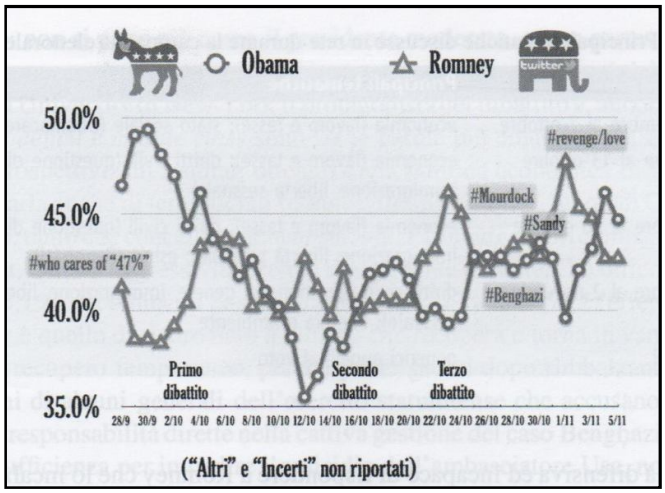


Figura 2.37: Andamento preferenze degli americani nel corso della campagna elettorale 2012
Fonte: Ceron, Curini, Iacus, 2014

La partecipazione online degli americani ha fatto registrare un numero elevatissimo di tweet, 50 milioni dal 28 settembre al 6 novembre 2012, basti pensare che nelle sole ore del dibattito televisivo del 3 ottobre ne sono stati raccolti 10 milioni.

Nel 2008 l'elezione del primo Presidente USA nero era stata determinata dall'abilità di quest'ultimo di utilizzare i nuovi canali di comunicazione per la sua propaganda elettorale. Al momento delle elezioni del 2012 il suo profilo Twitter contava 16 milioni di seguaci, mentre quello del rivale Romney non arrivava al milione. Questo divario non è stato però per nulla rilevante nella determinazione del risultato, poiché il numero di amici in Facebook o di follwer in Twitter non ha valenza di preferenza e quindi non è un indicatore di performance al seggio dato che l'atto di seguire un personaggio pubblico può essere scatenato sia da approvazione della sua ideologia, sia da contrasto con le sue idee (Crawford, 2009).

La Tabella 2.3 confronta lo scarto di voti tra i due candidati previsto attraverso Twitter (T), attraverso i sondaggi d'opinione ricavati dal sito Realclearpolitics.com (R) ed i risultati reali (V); viene riportata poi differenza tra le due tipologie di stima e i voti ottenuti veramente, ed infine l'ultima colonna dice quale dei due era il pronostico migliore. Gli 11 stati elencati sono i cosiddetti *swing state*, quelli cioè la cui preferenza era in bilico; per queste previsioni pertanto sono stati utilizzati solo i messaggi

geolocalizzati. Vince Twitter con una migliore accuratezza nei dati stimati per ben sette volte contro le 2 sole volte dei sondaggi tradizionali, e due parimeriti.

La *sentiment analysis* dei tweet ha fornito soprattutto una stima migliore per quanto riguarda il voto popolare nel suo complesso: quando il social prevedeva un gap tra Obma e Romney del 3,5% (quello reale è stato del 3,9%) le statistiche tradizionali registrando solo un lieve favoritismo per Obama dello 0,7%, tanto che Andrew Gelman, direttore dell'Applied Statistic Center dell'Università della Columbia affermava *"too close to call"*, mentre i ricercatori di VfB si pronunciavano con sicurezza.

Stato	Gap (T)	Gap (R)	Gap (V)	|T-V|	|R-V|	Stima migliore
Voto popolare	Obama +3.5	Obama +0.7	Obama +3.9	0.4	3.2	T
Florida	Obama +6.1	Romney +1.5	Obama +0.9	5.2	2.4	T
Ohio	Obama +2.9	Obama +2.9	Obama +3.0	0.1	0.1	=
Virginia	Obama +3.5	Obama +0.3	Obama +3.9	0.4	3.7	T
Colorado	Romney +1.3	Obama +1.5	Obama +5.4	6.7	3.9	R
Iowa	Obama +4.8	Obama +2.4	Obama +5.8	1.0	3.4	T
Nevada	Obama +3.3	Obama +2.8	Obama +6.7	3.4	3.9	T
New Hampshire	Obama +3.8	Obama +2.0	Obama +5.6	1.8	3.6	T
North California	Romney +3.0	Romney +3.0	Romney +2.0	1.0	1.0	=
Michigan	Obama +5.5	Obama +4.0	Obama +9.5	4.0	5.5	T
Pennsylvania	Romney +2.5	Obama +3.8	Obama +5.4	7.9	1.6	R
Wisconsin	Obama +7.4	Obama +4.2	Obama +6.9	0.5	2.7	T

Tabella 2.3: Confronto tra risultato effettivo, previsioni attraverso i sondaggi e previsioni analizzando i tweet

Fonte: Ceron, Curini, Iacus, 2014

Anche nel caso delle primarie del centrosinistra VfB, utilizzando l'analisi sentimentale iSA su 600.000 tweet raccolti tra il 6 ottobre ed il 2 dicembre 2012, ha monitorato l'andamento delle preferenze tra i candidati e ha previsto i risultati finali con buona accuratezza, sia al primo turno (vedi Figura 2.38 e Figura 2.39), che al secondo (vedi Tabella 2.4). Le primarie rappresentavano una situazione particolare in termini di elettorato poiché in questa tipologia di elezioni esso è rappresentato solo da sostenitori di una determinata area politica che decide di andare a votare, e quindi in questo caso studio il rumore di fondo, cioè le opinioni di chi poi non si recherà alle urne, è maggiore. La Figura 2.39 confronta l'errore assoluto medio delle stime effettuate

attraverso sei diversi sondaggi (Inr, Piepoli, Ipsos, CISE, SWG, Tecnè) e quelle fatte attraverso l'analisi dei tweet, dimostrando la validità di quest'ultima tecnica, le cui previsioni risultano in linea con le altre, e anzi a volte anche migliori.

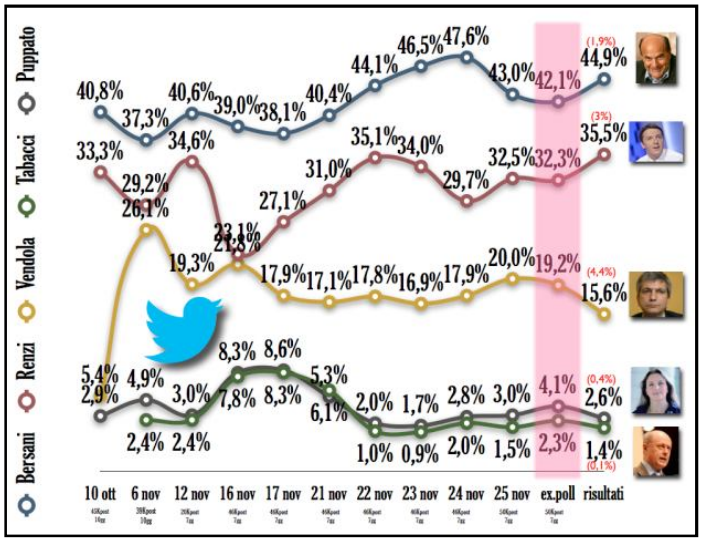


Figura 2.38: Andamento delle preferenze tra i candidati alle primarie, fino al primo turno del 25 Novembre 2012, ottenuto grazie alla *sentiment analysis* dei tweet, confrontato coi risultati reali

Fonte: Voices from the Blogs, 2012

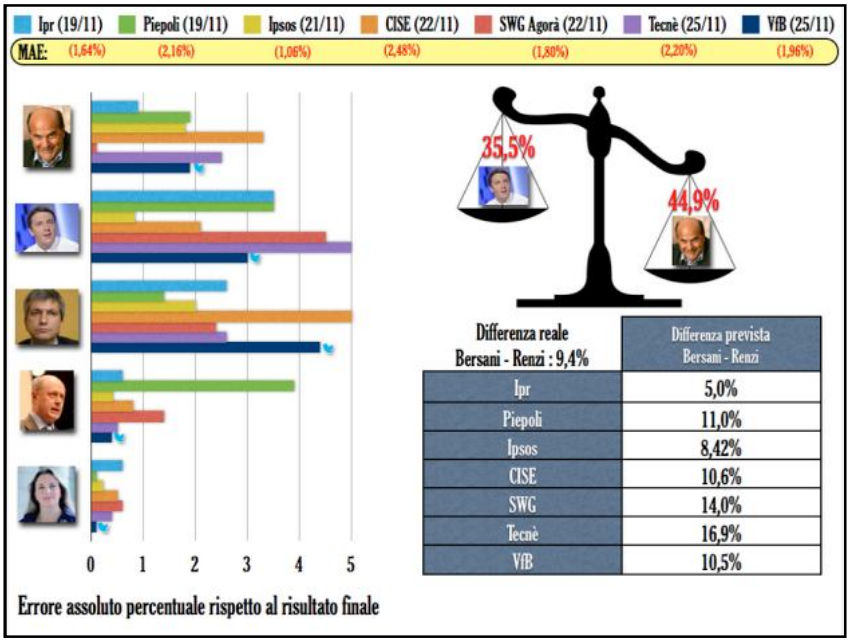


Figura 2.39: Confronto tra l'errore assoluto medio delle stime effettuate attraverso sei diversi sondaggi (Inr, Piepoli, Ipsos, CISE, SWG, Tecnè) e quelle fatte applicando il metodo iSA ai tweet al primo turno delle elezioni primarie del 2012

Fonte: Voices from the Blogs, 2012

	Giorno	Bersani	Renzi	Gap
Voti reali	-	60,9	39,1	Bersani +21,8
iSA	01/12/12	58,4	41,6	Bersani +16,8
Ipsos	29/11/12	57,5	42,5	Bersani +15
Quorum	28/11/12	56,4	43,6	Bersani +12,8
SWG	28/11/12	55	45	Bersani +10
COESIS	28/11/12	54	46	Bersani +8
ISPO	27/11/12	56,5	43,5	Bersani +13
IPR	26/11/12	56	44	Bersani +12
PIEPOLI	25/11/12	59	41	Bersani +18

Tabella 2.4: Confronto tra la previsione ottenuta attraverso i sondaggi e quella attraverso l'analisi semantica di Twitter col metodo iSA al secondo turno

Fonte: Voices from the Blogs, 2012

3. Twitter

3.1 Dati sull'utilizzo di Twitter

Twitter è uno dei 10 siti più visitati al mondo, il social network che tra il 2012 ed il 2013 ha registrato un tasso maggiore di crescita del +40%, e che a fine 2013 contava un totale di 645.750.000 utenti attivi registrati, con 135.000 nuovi iscritti al giorno. La media di tweet al secondo nel 2013 è stata di 9.100, mentre quella giornaliera di 58 milioni, con un miliardo di tweet pubblicati circa ogni 5 giorni; un numero enorme se lo si confronta con i 5.000 giornalieri registrati nel 2007, i 300.000 nel 2008, diventati nel 2009 ben 2.500.000 e raddoppiati nel 2010 (vedi Figura 3.1). Mentre le ricerche giornaliere all'interno del microblog sono state in media 2,1 miliardi nel 2013[12].

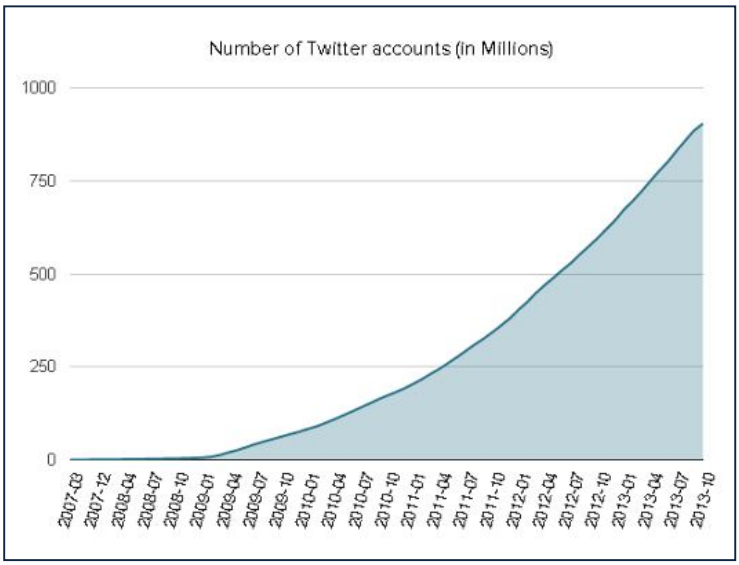


Figura 3.1: Crescita degli account Twitter nel mondo da Marzo 2007 ad Ottobre 2013
Fonte: PeerReach, 2013

Il 40% degli utenti utilizza Twitter passivamente, cioè semplicemente come mezzo per aggiornarsi sui temi di loro interesse, per curiosare tra i micro post pubblicati da altri, senza pubblicarne di personali; il dato emerso dall'analisi di PeerReach secondo cui il 56% dei nuovi iscritti a Luglio 2013 al social network non avesse twittato nemmeno una

[12] PeerReach, 2013. "4 ways how Twitter can keep growing". *Blog.peerreach.com* [blog] 7 Novembre 2013

volta fino al 25 ottobre 2013 evidenzia come molti profili registrati non rappresentino una fonte di attività (vedi Figura 3.2).

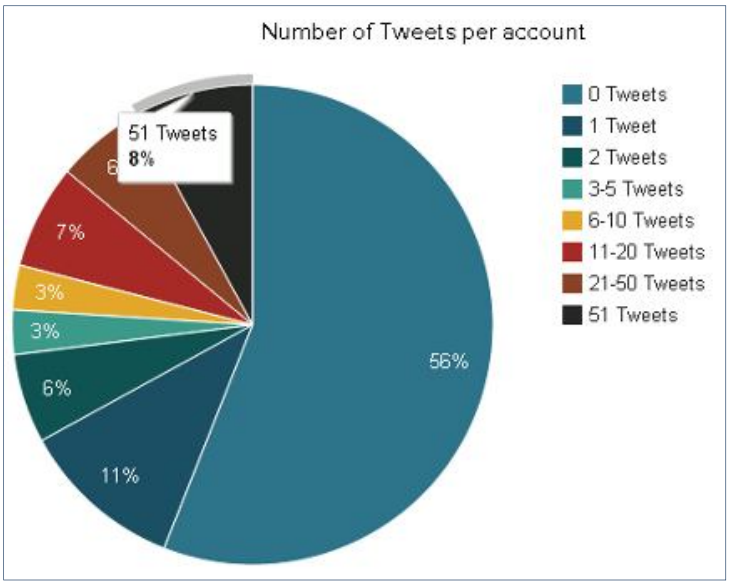


Figura 3.2: Utilizzo di Twitter da parte dei nuovi utenti registrati a Luglio 2013 dalla data di registrazione fino al 25 Ottobre 2013
Fonte: PeerReach, 2013

L'account con più seguaci è quello di Kate Perry (@kateperry) con 51.614.089 follower, seguito da Justin Bieber (50.427.066) e Barack Hobama (42.146.346).
In Cina si registra il maggior numero di utenti Twitter attivi (35,5 milioni), subito dopo c'è l'India e a seguire gli USA.
La penetrazione, calcolata come numero di utenti attivi mensilmente sul numero totale degli internauti della Nazione considerata, ad Ottobre 2013 era stata calcolata essere maggiore in Arabia Saudita (33%), con l'Italia al sedicesimo posto mondiale (5%) (vedi Figura 3.3); gli utenti più attivi si registrano però negli Stati Uniti (24,3%) (vedi Figura 3.4).

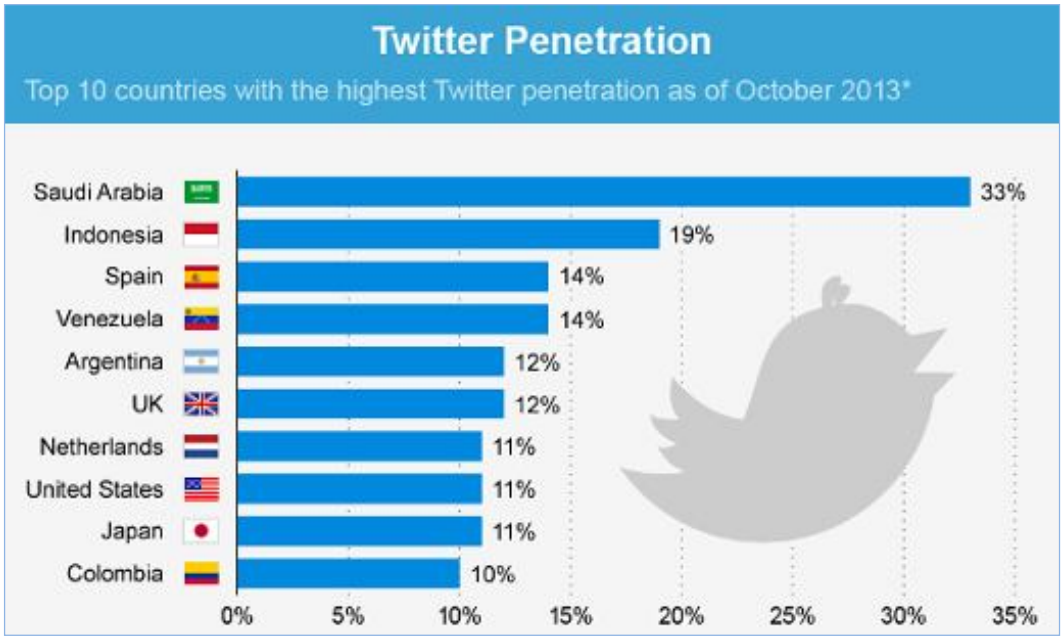


Figura 3.3: Penetrazione di Twitter nei diversi Paesi del mondo nel 2013

Fonte: PeerReach, 2013

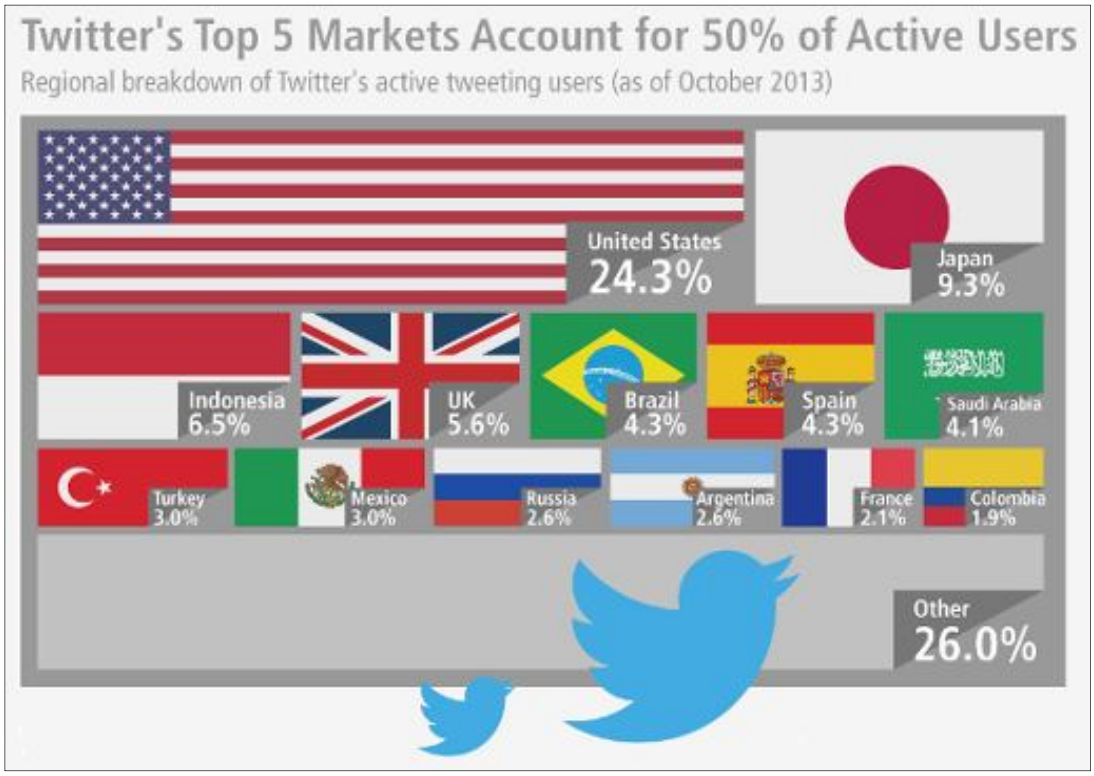


Figura 3.4: Paesi nel mondo con la maggioranza di utenti Twitter attivi nel 2013

Fonte: PeerReach, 2013

In Italia, nel primo quadrimestre 2013 sono stati pubblicati 71,6 milioni di tweet, con una media giornaliera di 653.515; di questi solo il 26% non era pubblico, il 25% è stato ritweetato, il 30% conteneva hashtag e solo l'8% era geolocalizzato[13].

3.2 I vantaggi dell'analisi attraverso Twitter

Twitter è risultato essere la fonte principale nelle analisi scientifiche precedentemente esposte nonostante le aree di studio fossero tra loro differenti, come mai?
I vantaggi dell'analisi di Twitter possono esser riassunti dai seguenti tre punti:

- i profili in Twitter sono per la maggior parte pubblici, diversamente che negli altri social network. Nel 2012, l'89% circa degli utenti non aveva profili privati, consultabili da chiunque ne fosse interessato. La stessa normativa sulla privacy del sito recita "La nostra impostazione predefinita è quasi sempre quella di rendere pubbliche le informazioni che fornisci finché non le elimini da Twitter, ma, in generale, è sempre possibile modificare le impostazioni per rendere le informazioni più private";
- i messaggi in Twitter permettono la geolocalizzazione di chi li ha postati, sempre se l'utente ha reso accessibile questo tipo di informazione, rendendo possibili interessanti osservazioni;
- attraverso gli hashtag che etichettano ogni tweet in base all'argomento a cui si riferisce, è possibile riunire discussioni relative allo stesso tema, anche se avviate da utenti che non hanno alcun legame, poiché ognuno digitando un hashtag nella barra di ricerca può vedere cosa dicono gli atri utenti rispetto a quella tematica.

Risulta però importante sottolineare che i limiti della localizzazione geografica sono notevoli poiché pochi utenti permettono l'accesso a tale informazione personale: nel primo quadrimestre 2013 i tweet che risultavano georeferenziati erano solo l'8%, rappresentando così un sottoinsieme davvero piccolo. Per sopperire a tale lacuna alcuni studiosi hanno adottato quest'alternativa: anche quando un tweet non è geolocalizzato è però possibile attribuire al suo autore una posizione geografica leggendo le informazioni

[13] Cosenza, V., 2013. "State of the Net". *Slideshare.net*

presenti nella descrizione del suo profilo, in tal modo per trasposizione si può dedurre approssimativamente da dove è stato pubblicato quel messaggio (Kumar, 2013).

Twitter può esser utilizzato per:

- mettere in luce cosa è importante per le persone in tempo reale con aggiornamenti costanti;
- scoprire anomalie sull'andamento di alcuni temi di conversazione;
- osservare trend settimanali o mensili e trovare modelli che si ripetono;
- compare sub-argomenti;
- comparare trend tra i tweet ed altri indicatori ottenuti da statistiche tradizioni ufficiali;
- registrare e analizzare comportamenti e sentimenti collettivi;
- mostrare l'impatto locale di eventi globali.

Confrontando ad esempio l'approccio di analisi basato su Twitter con quello basato sull'analisi delle query inserite nei motori di ricerca, il primo risulta migliore poiché fornisce informazioni meglio contestualizzate e permette uno studio retrospettivo ed un'analisi degli argomenti correlati. Inoltre la lunghezza dei messaggi postati nel microblog permette di distinguere meglio quelli che potrebbero essere fuorvianti da quelli pertinenti alla ricerca e che potrebbero portare a sovrastimare il fenomeno (Culotta, 2010).

De Choudhury, Ringel Morris e White (2014) invece, dopo aver confrontato i contenuti medici cercati nei motori di ricerca e quelli pubblicati online sono arrivati alla conclusione che la natura complementare dei due media (pubblico contro privato, cercare contro condividere) suggerisce una loro combinazione per una visione completa delle attività online legate alla salute.

Seppure non vi sia completo accordo tra i ricercatori, Twitter sembra essere il social network nel quale le analisi esposte nel capitolo precedente possono essere svolte nel modo più semplice e nel contempo più efficace; per questo motivo il presente lavoro di ricerca proseguirà concentrandosi d'ora in poi su questo particolare social network.

3.3 Che cosa viene pubblicato e cercato in Twitter

In Twitter chi dice cosa a chi? Prima di tutto si è tentato di capire in generale cosa viene cercato e pubblicato maggiormente in Twitter per avere un'idea di base di come la piattaforma venga utilizzata dagli utenti. Ciò che principalmente viene condiviso su Twitter sono contenuti prodotti in altre piattaforme; più della metà dei collegamenti ipertestuali contenuti nei tweet rimandano ai seguenti sei domini: Twitter, Instagram, Facebook, YouTube, ask.fm, e Tumblr. Solo il 7.8% di tutti i link condivisi rimanda a notizie provenienti dai canali mainstream inglesi, tra cui le fonti principali risultano essere la BBC, il New York Times e il The Guardian (Kumar, 2013).

Twitter, infatti, è adottato per chiacchiere quotidiane e conversazioni dirette, ma anche come mezzo per diffondere notizie e contenuti prodotti in altre piattaforme; ciò lo qualifica come vero e proprio *news medium* (Kwak et al. 2010).

La maggioranza dei contenuti presenti sul social network è generata da una piccola percentuale di utenti, in prevalenza da quelli classificati da Twitter come èlite, ossia celebrità, blogger, organizzazioni formali e media tradizionali. Diversi tipi di contenuti hanno diversa durata: quelli generati dai canali infornativi sono i principali *short-lived* URLs, mentre quelli con un ciclo di vita più lungo sono i contenuti video e musicali che sono continuamente riscoperti e condivisi, dimostrando una durata di vita indefinita (Wu, 2011).

Stieglitz e Dang-Xuan (2013) basandosi su due dataset di 165.000 tweet hanno scoperto che i messaggi dalla forte componente emotiva tendono ad essere retwettati più velocemente e più spesso rispetti a quelli neutri dal punto di vista sentimentale.

Uno studio sulla geografia della comunicazione di Twitter svolto da Leetaru et al. (2013) aveva messo in luce come le persone retwittano utenti sia vicini che lontani, senza differenza (vedi Figura 3.7), con l'Europa come luogo in cui sono creati i contenuti maggiormente retwittati nel resto del mondo. Si tende, però, a seguire, differentemente a quanto avviene per l'azione del retweet, persone che geograficamente sono più vicine o che comunque risiedono in aree facilmente raggiungibili. Quest'affermazione deriva dal fatto che la frequenza di collegamenti aerei tra due località è risultata essere un buon predittore dei legami in Twitter tra i residenti nelle due zone (Takhteyev et al., 2011).

Figura 3.7 : Mappa che rappresenta le connessioni tra utenti attraverso il retweet di messaggi geolocalizzati tra il 28 Ottobre 2012 ed il 30 Novembre 2012.
Fonte: Leetaru et al., 2013

Le persone in Twitter cercano principalmente informazioni la cui rilevanza è limitata nel tempo (come contenuti real-time o trend popolari) ed informazioni riguardanti specifiche persone, sia famose che non.
Le query in Twitter sono corte, con sintassi specifica, con riferimenti a persone e spesso sono ripetute più volte nel tempo per monitorare come le discussioni relative allo stesso argomento cambiano, mentre quelle poste ai motori di ricerca generalmente si evolvono man mano che l'utente acquisisce informazioni sull'argomento. I risultati delle ricerche in Twitter sono principalmente discussioni ed eventi sociali, mentre quelli nel Web includono contenuti più piatti ed informativi, privi di opinioni e giudizi (Teevan, Ramage, Ringel Morris, 2011).
Lo studio del 2013 di Bastos, ecologo digitale presso l'Università di São Paulo, e Zago, dottoranda presso l'Università di Rio Grande do Sul, ha fatto emergere come in base alla nazionalità tendono a cambiare gli argomenti che attirano maggior interesse nel social: in Germania l'economia, in Inghilterra le notizie globali, in Brasile lo sport, in Spagna le notizie legate alla patria.
Un'interessante analisi sui contenuti di Twitter è stata svolta da Blogmeter nel primo trimestre 2013: creando dei grafici dall'immediatezza visiva vennero messi in evidenza i contenuti più diffusi dei 71,6 milioni di tweet pubblicati in Italia in quel periodo (vedi

Figura 3.5 e 3.6); la tematica di discussione principale risultò essere la cronaca, ed in particolare le news inerenti la politica.

Figura 3.4: Word cloud delle parole più frequenti nei tweet italiani del primo quadrimestre 2013

Fonte: Blogmeter

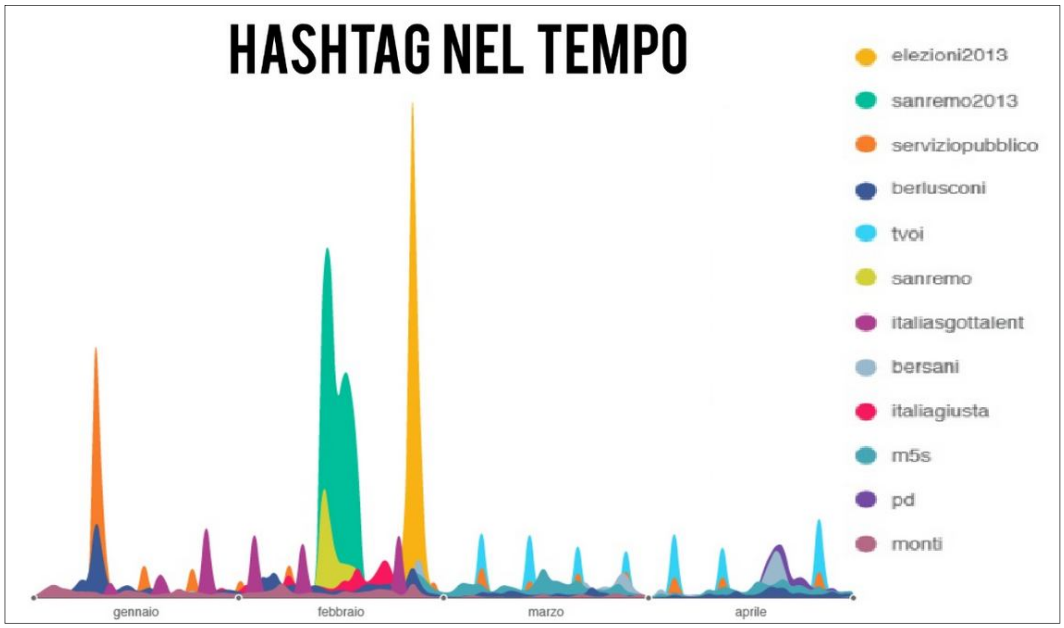


Figura 3.5: Andamento degli hashtag più frequenti nei tweet italiani del primo quadrimestre 2013

Fonte: Blogmeter

3.4 Metodologie per estrarre i dati da Twitter

Qualunque tipo di analisi di quanto viene pubblicato in Twitter richiede la disponibilità dei tweet d'interesse affinché il ricercatore possa apporvi modifiche ed studiarlo con le strumentazioni più consone al suo scopo di ricerca. Il motore interno al sito permette di leggere l'elenco dei tweet contenenti uno specifico hashtag o una determinata parola chiave, è pure possibile una ricerca avanzata filtrando i risultati per utente, luoghi e polarità dei risultati, ma non è possibile né estrarre la lista ottenuta, né avere misure aggregate, si può solo scorrere un elenco di tweet in ordine cronologico inverso. Ovviamente qualunque studio simile a quelli precedentemente esposti sarebbe impossibile senza poter estrarre i messaggi relativi al tema che si vuole studiare.

L'unico modo gratuito per ottenere un elenco di tweet in un formato modificabile e sul quale si possano svolgere determinate operazioni è utilizzare le Twitter API (Application Programming Interface), ossia procedure per interagire con i server di Twitter per la raccolta o l'invio di dati in forma automatizzata.

Ci sono due tipi di API: la REST API e la Streaming Api.

Attraverso la prima si può sia ottenere, che inviare, dati attraverso rispettivamente richieste di tipo Get e Post. Essa viene utilizzata ad esempio per acquisire informazioni relative ad un utente specifico come numero e lista dei follower, oppure per creare programmi che in automatico svolgono determinate operazioni come il retweet di post contenenti specifici hashtag.

Deriva da questa la REST Search API, la cui versione più aggiornata è la 1.1, la quale fornisce un elenco di tweet che corrispondono ai criteri specificati nella query inserita. Il filtraggio dei risultati avviene specificando i seguenti parametri, tutti opzionali escluso il primo:

- parola chiave o hashtag contenuto;
- geocode, cosicché la localizzazione dell'utente che posta il tweet risulti entro un determinato raggio dalla latitudine e longitudine specificate;
- lingua utilizzata, specificata attraverso il codice ISO 639-1;
- tipo di risultato, per default esso sarà un mix, ma può anche contenere solo i messaggi più recenti o più popolari;
- numero di tweet da estrarre, che per default è 15;

- intervallo temporale in cui i messaggi sono stati pubblicati;
- intervallo entro cui rientra l'ID identificativo dell'autore dei tweet estratti.

Invece la Streaming API non è concentrata su contenuti passati, ma estrae in tempo reale una percentuale del flusso totale dei tweet pubblicati senza specificare alcun parametro. A differenza delle precedenti, per essere utilizzata necessita di una connessione persistente, sempre aperta durante tutto il periodo di analisi.

Quando si estrae un tweet, indipendentemente dalla API utilizzata, non è possibile sapere il nome dell'autore ma vi è solo un ID identificativo per differenziare i diversi account.

Nel sito ufficiale dedicato alle Twitter API (dev.twitter.com) si legge "La Search API è focalizzata sulla rilevanza e non sula completezza. Ciò significa che alcuni tweet ed alcuni utenti potrebbero mancare tra i risultati", ed ancora "Non tutti i tweet saranno indicizzati o resi disponibili attraverso l'interfaccia di ricerca". Non sono fornite spiegazioni dettagliate sui limiti delle API. Per certo si sa che solo i tweet postati da profili pubblici sono resi disponibili da questa interfaccia ufficiale, e che i risultati delle Twitter REST API arrivano al massimo ai sette giorni precedenti. Un altro limite della REST API è che per ogni query inserita vengono restituiti al massimo 1500 risultati, pertanto se i tweet che corrispondono alla ricerca fatta sono di più non appaiono. Inoltre, non è permesso effettuare più di 180 richieste ogni finestra di 15 minuti.

Infine sembra, senza alcuna conferma ufficiale, che durante l'estrazione vengano privilegiati, tra quelli che la API riesce a raccogliere, i tweet con gli autori più popolari ed esclusi quelli dal contenuto di bassa qualità, come quelli composti da una sola parola.

Non bisogna inoltre dimenticare che i post analizzati sono comunque estratti dal social media attraverso un set di *keyword*, precedentemente stabilite dai ricercatori, e quindi il dataset ottenuto non può considerarsi completo nemmeno se Twitter permettesse di estrarre la totalità dei post corrispondenti alle caratteristiche richieste.

La Streaming API, invece, fornisce solo l'1% dell'enorme flusso totale di dati chiamato *firehose*, utile quindi per chi è interessato a un campionamento, e non necessita di un accesso integrale.

I motivi di tali limitazioni nella raccolta dei dati sono sia tecnici legati alla sostenibilità ingegneristica, sia legati alle strategie di business.

Per interfacciarsi con Twitter attraverso le API ufficiali si possono utilizzare diversi linguaggi di programmazione, come Python, per i quali sono disponibili svariate librerie dedicate, come la Python-Twitter. Per superare la difficoltà che la mancata conoscenza di un linguaggio di programmazione potrebbe creare, sono nate modalità alternative per utilizzare le API, come il pacchetto TwitteR per il software R: installando il pacchetto si possono digitare dei comandi in R affinché questo si interfacci con una Twitter API precedentemente creata dall'utilizzatore che a sua volta si interfaccia con Twitter.
Installando il pacchetto, attraverso R, si potranno ottenere i tweet contenti determinati hashtag o parole chiave, quelli provenienti da un determinato utente, quelli pubblici più recenti ed individuare gli argomenti più popolari dell'ultima settimana.
Dopo aver creato la propria Twitter API attraverso il sito ufficiale dev.twitter.com facendo riferimento ad un account, il quale diventa il proprietario dell'applicazione, si ottengono delle credenziali: *Consumer key* e *Consumer secret* necessarie per ottenere un codice d'autorizzazione da inserire nella API per iniziare ad utilizzarla. Oltre all'autentificazione appena descritta, definita OAuth signed, esiste un'altra possibilità: la *Application-only authentication*. Quest'ultimo modo di ottenere l'autorizzazione consiste nell'invio di una domanda di autentificazione per conto non di un utente, ma dell'applicazione stessa. In questo modo non è però possibile connettersi allo *stream*, fare ricerche di utenti o inserire specifiche di geolocalizzazione. La scelta tra le due alternative modifica l'esperienza e la modalità di scrittura all'interno dell'applicazione. Twitter pone particolare attenzione al passaggio di autentificazione per mantenere il totale controllo dei messaggi postati e non rischiare che qualcuno possa avervi accesso senza permesso. Il 14 gennaio 2014 la connessione alle Twitter API è stata ristretta solo a connessioni sicure TLS/SSL o HTTPS. L'utilizzo del protocollo sicuro SSL crea un canale comunicativo protetto tra i server di Twitter e le applicazioni, in questo modo nessun agente non autorizzato può aver accesso a dati sensibili nel mentre che le informazioni viaggiano attraverso Internet pubblico tra l'utente e la Twitter API. Le procedure di autentificazione OAuth non sono sufficienti poiché esse servono solo ad impedire che la password dell'utente venga estorta durante lo scambio informativo.
Queste procedure di riservatezza, in continua evoluzione, confermano non solo il tentativo di ottenere la massima privacy degli utenti, ma anche l'interesse da parte del social network di mantenere il controllo più totale sui tweet e di ridurre al minimo la

possibilità d'accesso ad essi. Ciò si spiega perché Twitter, a differentemente da altri social come Facebook, vende tutti i tweet pubblici a società esterne a prezzi elevati, ed è questa la sua unica fonte di guadagno.

Sono solo quattro i rivenditori autorizzati che hanno l'esclusivo accesso al *firehose*: GNIP, Datasift, Topsy, NTT Data. Esiste un enorme database contenente tutti i tweet generati dalla nascita di Twitter custodito dalla *Library of Congress* americana a cui solo Twitter ha accesso e ne concede la rivendita a queste quattro aziende.

Gnip è stato il primo reseller ufficiale e vanta una ricerca storica che parte dal 2006, anno di nascita di Twitter stesso, ed offre inoltre un campione pari al 10% del totale dei tweet.

Topsy ha una ricerca storica che parte solo dal 2008, ma offre servizi come analisi dei link contenuti nei tweet e *sentiment analysis* degli stessi. Acquistata nel Dicembre 2013 da Facebook, dopo questa data è stata tolta la possibilità di avere 14 giorni di prova gratuita del servizio prima di registrarsi come clienti.

Datasift registra 500 clienti, tra cui Coca-Cola, McDonald's e la Bbc, e viene così presentato da Tim Barker, Chief Product Officer di Datasift: "Invece che dar vita ad un sito che mostra gli articoli più di tendenza, ci siamo concentrati su una piattaforma a cui sviluppatori e imprese possano connettersi per estrarre i dati social di cui hanno bisogno per gestire il loro business, l'azienda di computer Dell, per esempio, usa Datasift per misurare l'opinione del pubblico sui suoi prodotti".

Datasift riceve 15.000 unità d'informazione al secondo generate dagli aggiornamenti pubblici dei social network. Entro 200 millisecondi li processa e li arricchisce attraverso algoritmi; il tutto è reso disponibile ai clienti, ma è molto costoso, anche 15.000 dollari mensili.

Infine NTT Data offre anche analisi semantiche, ma il suo accesso ai tweet è limitato a quelli in lingua giapponese.

Solo le quattro aziende appena citate possono accedere allo storico dei tweet grazie ad accordi diretti con Twitter. Altre società, e persino singoli utenti, però possono in autonomia raccogliere i cinguettii postati sul social network attraverso le API ufficiali, anche se disponibili con i limiti precedentemente esposti, e conservarli in archivi propri per poi rivenderli, come fa Hootsuite. Strumenti di questo genere sono forniti da numerosi siti come keyhole.com, hashtagtracking.com, hashtags.org, trendistic.com,

twitterfall.com, tweetarchivist.com, hashonomy.com, twubs.com e twazzup.com; i quali richiedono la creazione di un proprio account ed il pagamento di *fee* mensili per poter effettuare le ricerche d'interesse.

La presenza di una filiera della rivendita delle conversazioni passate avvenute in Twitter a prezzi elevati fa comprendere come le potenzialità dell'analisi dei tweet siano riconosciute da molti.

Un buon numero di società inoltre offre a pagamento servizi di monitoraggio dei social network e d'analisi dei dati da essi estratti, come Waggeneredstrom, Blogmeter, The Fool, Sysomos, Hashtagify.me e Socialflow.

Aggregando dati provenienti da diverse piattaforme social come Facebook, Youtube, Twitter ed altri blog e forum online riescono ad aggiungere significato ai dati grezzi estratti e questo si rivela particolarmente utile per le aziende che vogliono indagare cosa i consumatori pensano del loro brand, scoprire le tendenze ed avere una fotografia generale dell'opinione pubblica.

Anche se le analisi offerte da queste società sembrano essere principalmente sfruttate dalle aziende nella loro attività di brand management, ci sono esempi che dimostrano come anche in altri campi stia crescendo l'interesse su cosa viene 'detto' online. Ci siano realtà che attivamente stanno implementando osservatori specializzati, come BrightPlanet che filtra dal firehose informazioni rilevanti per le forze dell'ordine e per le agenzie di intelligence, o come www.symplur.com, il cui database costituito da 432.005.124 tweet storici contenenti 3.881 hashtag relativi a 9.221 temi, raccoglie informazioni interessanti relative a cure mediche, malattie, medicine, ospedali, conferenze internazionali sulla salute; utili quindi per le organizzazioni ed autorità che si occupano di pubblica sanità, per le agenzie governative, per le compagnie farmaceutiche ed ovviamente per ricercatori in campo medico.

E' notabile poi la cifra che Twitter ha incassato nel 2009 per rendere i tweet ricercabili dai motori di ricerca Google e Bing, ben 25 milioni di dollari.

Il fatto che il mercato dei dati generati dai messaggi pubblicati in Rete è sempre più fiorente e articolato sembra confermare la tesi secondo cui l'analisi dei dati provenienti dai social network sia quanto mai promettente.

3.5 Metodologie per analizzare i dati di Twitter

Dopo aver estratto i tweet riguardanti l'argomento di ricerca, attraverso le Twitter API o acquistandoli da terzi, è possibile farne un'analisi semplicemente quantitativa, oppure concentrarsi anche sul contenuto dei messaggi.
Specialmente per gli studi a fini predittivi, il mero conteggio di quanti post online contengono determinati termini non risulta essere sufficiente, come hanno dimostrato Jungerr et al. (2012).
Risulta quindi necessario scoprire non solo quante persone 'discutono' sulle piattaforme digitali di un argomento ed il loro variare nel tempo, ma anche cosa queste 'dicono'.
La sfida consiste nel comprendere il reale intento di chi ha pubblicato il messaggio in termini d'idea, polarità ed intensità, e ciò è possibile solo attraverso un'analisi linguistica del testo.

Vi sono due famiglie di tecniche di analisi testuale adottabili: quella di scoring e l'analisi testuale propriamente detta. La prima ordina un insieme di testi lungo una linea immaginaria (ad esempio ordinando un gruppo di tweet a sfondo politico su un'asse progressista-conservatrice), la seconda invece non si limita ad ordinare i testi, ma li classifica. Entrambe sono a loro volta suddivise nei due sottogruppi supervisionate e non supervisionate, mentre solo la seconda è ulteriormente differenziata in tecniche dedicate alla classificazione individuale, per attribuire il messaggio ad una categoria semantica, o aggregata, per comprendere come le categorie semantiche si distribuiscono nell'insieme dei testi analizzati.
Parlando di tecniche supervisionate di analisi testuale vera e propria si fa riferimento all'attribuzione di un testo ad una categoria semantica già definita in precedenza, azione definita 'tagging', attraverso l'intervento umano o utilizzando un dizionario ontologico, ossia una lista di parole attribuite a determinate macro aree tematiche. Il tagging automatico che si basa su dizionari ontologici richiede, per evitare classificazioni errate, la definizione di regole d'applicazione diverse per ogni lingua e anche per i vari temi di discussione.

Con il termine ***text mining*** si esprime un insieme di tecniche volte ad estrarre da un insieme di messaggi non informazioni relative ai singoli, ma all'insieme nel suo complesso, come la frequenza con cui ricorrono determinate parole chiave o il legame dei messaggi a particolari eventi; un esempio è il tentativo di scoprire gli argomenti più discussi online in un determinato momento per individuare gli *hot topic* del momento.

Un primo approccio per analizzare il contenuto di un corpo di tweet può essere il *Text summarization*, procedimento il cui scopo è appunto riassumere il contenuto dei messaggi analizzati, di cui il *word cloud* è l'esempio più comune.

Sempre basandosi sulla frequenza con cui determinate parole compaiono all'interno dei tweet oggetto dell'analisi, c'è l'approccio *Information retrieval* (IR), il quale cerca all'interno dei testi a disposizione la risposta a determinati quesiti utilizzando parole chiave, mentre quello *Information extraction* (IE), basandosi su quante volte *keyword* o frasi vengono ripetute, ha come obiettivo l'estrazione di informazioni particolari.

Fermarsi semplicemente al conteggio del numero di volte in cui determinati termini considerati rilevanti vengono utilizzati o a quali sono le parole più ripetute non permette però di comprendere il legame esistente tra esse e la tematica di cui i tweet trattano; serve quindi un passo ulteriore per comprenderne la correlazione.

L'***Opinion Mining*** è un approccio che studia come estrarre da un testo un'opinione o un sentimento utilizzando *Information Retrieval* (IR) e tecniche di NLP (*Natural Language Processing*). Termine introdotto per la prima volta nel 2009 da Dave et al. facendo riferimento ad una tecnica che indaga gli attributi associati ad una determinata parola chiave e la loro distribuzione per poter comprendere l'opinione associata a quella *keyword*.

Il *Natural Language Processing* (NLP), un campo delle scienze informatiche e linguistiche, ha come finalità scoprire la sovrastruttura linguistica ed esplorare le relazioni tra parole e contenuti utilizzando un algoritmo basato su assunzioni cognitive, per far in modo che, in modo automatico, si riesca a derivare da input costituiti da linguaggio umano naturale il significato che l'autore voleva dare al testo prodotto.

Una sottocategoria dell'*opinion mining* è la *sentiment analysis* volta a studiare la tipologia e l'intensità dei sentimenti espressi attraverso i messaggi analizzati, le cui motivazioni possono poi esser scoperte attraverso un'analisi linguistica ulteriore.

A tal scopo si fa affidamento a dei database semantico-lessicale che organizzano, definiscono e descrivono i concetti espressi dai vocaboli; *WordNet* è stato il primo elaborato per la lingua inglese dal George Armitage Miller presso l'Università di Princeton. Questo processo utilizza due risorse linguistiche:

-il *sentiment lexicon*, ossia un elenco di parole o espressioni multi parola associate al sentimento che riflettono, specificando anche il contesto linguistico in cui la loro valenza sentimentale viene annullata (come nel caso di "piacere" utilizzato nella formula "per piacere");

- regole sintattico-semantiche per trattare fenomeni come la negazione, identificare le domande o le comparazioni e per la quantificazione del sentimento espresso.

Questo procedimento dà buoni risultati nel caso di argomenti ben circoscritti, ma non raggiungerà mai l'accuratezza di un tagging manuale, anche se quest'ultimo richiede molto tempo e controlli incrociati da parte di più individui.

Vi sono *tool* disponibili su piattaforme online come www.socialmention.com o twittersentiment.appspot.com che offrono analisi sentimentali dei contenuti in un insieme di tweet. Essi funzionano abbastanza bene quando sono utilizzati per comprendere la polarità dei messaggi su un brand o un personaggio famoso, hanno però delle difficoltà a valutare post relativi a concetti astratti o su temi che non si basano su una stringa di parole chiave ben definite, ad esempio è possibile sapere il pensiero collettivo su Madonna, ma sapere quello sul vestito che ha indossato durante il suo ultimo concerto non è possibile. Questi strumenti lavorano bene ad un livello base d'analisi e principalmente utilizzano dei dizionari di parole classificate come positive o negative in generale, ma non con riferimento a quello specifico tema analizzato. Inoltre, il sentimento fa riferimento al tweet nel complesso e non specificatamente alla parola chiave utilizzata come discriminante durante l'estrazione dei messaggi dal social network. Essi possono quindi rivelarsi utili, ma anche parecchio errati, in base a quale domanda si sta cercando di rispondere. Oltre a ciò, le loro analisi non sono disponibili per periodi antecedenti il mese precedente al momento dell'analisi. Infine, un'altra loro limitazione è la scarsità di lingue con cui questi *tool* riescono a lavorare: twittersentiment.appspot.com ad esempio funziona solo per l'inglese e lo spagnolo.

I testi estratti dai social network, ed in particolare i tweet, risultano problematici perché oltre ad essere frasi brevi e spesso incomplete, creano parecchie difficoltà d'analisi per la loro sintassi particolare, la presenza di slang, dialetti locali, abbreviazioni, sarcasmo, mix di lingue differenti, iperboli, ironia, doppi sensi, giochi di parole e l'assenza di punteggiatura. L'utilizzo di strumenti tipici di pre-*processing* linguistico come POS *tagger* (la cui finalità è etichettare ogni singola parola del testo in base alla sua categoria grammaticale e sintattica, valutandone il significato, il contesto e la sua relazione con i termini seguenti o adiacenti) , *tokeniser* (che avviano un processo che suddivide un testo in parole, frasi, simboli ed altri elementi significativi chiamati *token*, i quali diventano l'input di processi seguenti) o *parser* (algoritmo di un programma applicativo che data la grammatica ed il lessico di una lingua riesce in automatico ad analizzare la struttura di un testo e se riconosce una frase in senso grammaticale produce un'analisi morfologica e sintattica delle parole) risulta ostacolato. Il loro utilizzo è cruciale per procedere con *un'opinion mining* affidabile; servono quindi dei NLP *pre-processing tool* creati ad hoc date le particolari caratteristiche dei testi pubblicati in Twitter.

Problematiche non indifferenti sono create anche dal sarcasmo che di frequente si ritrova nei post pubblicati nella piattaforma sociale. Oltre alla difficoltà della sua identificazione c'è anche il problema di come definirne la polarità poiché non sempre basta invertirne il significato.

Il metodo utilizzato per le varie analisi linguistiche, qualunque esso sia, deve essere validato. Nel caso dei metodi supervisionati, nei quali le categorie semantiche sono note a priori, un sottoinsieme dei messaggi da analizzare, il *training set*, viene codificato manualmente e poi si confrontano i risultati della codifica fatta a mano con quelli ottenuti applicando il classificatore automatizzato. Se invece il metodo non è supervisionato, e quindi le categorie semantiche emergono a posteriori, una possibile validazione è far analizzare all'algoritmo un testo di cui si conosce il contenuto e testare se esso viene classificato in modo corretto.

Gli strumenti d'analisi testuale tradizionali applicati ai dati provenienti dai social possono essere pericolosi perché sono creati per dei documenti di testo pre-classificabili. Ciò non esclude che possano dare risultati simili a quelli di sondaggi svolti per scoprire l'opinione pubblica, ma sono necessari dei miglioramenti sostanziali per avere dei risultati su cui fare pieno affidamento.

Secondo O'Connor et al. (2010) utilizzare tecniche superficiali basate sul conteggio delle parole collegabili ad un sentimento è sufficiente per trovare opinioni aggregate anche se non sono molto precisi, poiché, anche se il tasso d'errore è elevato, con un grandissimo numero di misurazioni l'importanza di questi errori può essere diminuita.
I casi studio esposti al paragrafo 2.6 hanno adottato tecniche tra loro differenti con successo, quindi, chiedendosi quale sia l'analisi semantica migliore da adottare appare difficile trovare una risposta definitiva.
In conclusione è possibile ottenere risultati confermati dai dati reali con vari approcci, anche semplicistici, ma è il ricercatore che di volta in volta deve essere consapevole della qualità del metodo e del classificatore che sta adottando per le sue analisi ed avere cautela nell'applicare metodologie che in altri scenari possono essere risultate adatte ma che non lo sono per quello in esame.
Sicuramente servono ulteriori ricerche per sviluppare degli algoritmi più sofisticati in grado di superare i principali ostacoli che i testi dei messaggi raccolti da Twitter pongono, anche se ovviamente è impossibile che un metodo automatico riesca a cogliere tutte le sfumature del linguaggio umano e non fallisca nel confrontarsi con la sua complicatezza, acuita dal limite dei 140 caratteri che la piattaforma sociale impone.

3.6 I limiti dell'analisi di Twitter

Avere a disposizione tanti dati non significa in automatico ottenere informazioni migliori; solo con la massima consapevolezza dei limiti si può riuscire a sviluppare al meglio le prospettive dell'utilizzo dei Big Data provenienti dai social network.
Questo paragrafo cercherà di mettere in luce i limiti dell'analisi dei social network in generale, e nello specifico dell'analisi dei dati provenienti da Twitter, che sono emersi nel corso della presente ricerca e durante i casi studio precedentemente presentati.

Il problema della **privacy** è uno dei principali ed il più delicato per le sue implicazioni concettuali, legali e tecnologiche. Essa è stata definita dalla *International Telecommunications Union* come "il diritto di un individuo di controllare o influenzare quali informazioni ad egli correlate possono essere svelate".

La privacy è un diritto umano, un pilastro della democrazia ed un fondamento della libertà dei cittadini, che altrimenti sarebbe a rischio, anche per quelli che "non hanno nulla da nascondere", data l'importanza di alcune informazioni sensibili per certe organizzazioni o per i Governi.

La ricchezza d'informazioni che potrebbero avere su un individuo Google, Facebook, compagnie telefoniche e gestori di carte di credito se fondessero i dati in loro possesso è già abbastanza preoccupante di per sé, se poi si aggiunge il rischio che alcuni potrebbero utilizzarle per scopi immorali diventa ancora più preoccupante.

Internet è un ambiente molto poco regolato e gli utenti al momento sono poco informati e poco tutelati poiché la protezione da usi illegali dei loro dati è molto bassa e perché spesso consentono l'utilizzo dei loro dati personali e lo studio del loro comportamento online in cambio di contenuti prestando poca attenzione a ciò che questo comporta.

Un esempio che mette in luce la problematica della privacy è il seguente: nel 2006 un gruppo di ricercatori dei Harvard studiò come gli interessi e le connessioni di 1.700 profili Facebook di studenti universitari variavano nel tempo. I risultati della loro ricerca furono pubblicati insieme ai dati che essi avevano reso anonimi, ma presto altri studiosi si resero conto che era possibile de-anonimizzare parti del dataset compromettendo fortemente la privacy degli studenti che non erano nemmeno consapevoli che quei dati su di loro erano stati raccolti (Zimmer, 2008).

La domanda che sorge è la seguente: è eticamente corretto usare certi dati, senza richiedere alcun tipo di permesso, solo perché sono accessibili?

Anche se l'autore pubblicandoli online era consapevole che sarebbero diventati di dominio pubblico c'è sempre il rischio che qualcuno li estragga dal loro contesto originale e li utilizzi per analisi che l'autore non avrebbe immaginato, o che qualcuno venga costantemente monitorato senza saperlo. Ma chi ha la responsabilità di accertarsi che nessun individuo in particolare, e nemmeno la società nel suo complesso, vengano danneggiati da processi di ricerca avviati senza che essi ne sappiano nulla?

I ricercatori non sono mai tra l'audience immaginario che si Figura l'utente medio di un social network, inconsapevole della molteplicità di agenti ed algoritmi che costantemente raccolgono e immagazzinano i suoi dati e registrano le sue azioni per eventuali usi futuri. Gli internauti devono, quindi, essere quantomeno resi consapevoli

del fatto che una volta pubblicato online un messaggio chiunque può impossessarsene con relativa facilità.

A proposito della privacy dei dati bisogna affrontare tre questioni:

- La consapevolezza da parte di chi li genera di aver consentito alla loro raccolta;
- La conoscenza di eventuali riutilizzi in altri contesti;
- La ri-identificazione dopo aver reso anonima la fonte che li ha generati.

La problematica riguarda quindi sia l'acquisizione, che l'archiviazione, l'uso e la presentazione dei dati.

Quello concernente la privacy dei Big Data è un tema che rientra tra i diritti umani. Ci sono tre regole che dovrebbero essere rispettate per rendere una loro analisi giusta dal punto di vista etico: quando si analizzano questi dataset mai utilizzare dati che sono riconducibili a qualcuno di identificabile e che siano confidenziali e mai cercare di re-identificare gli individui partendo dai dati a disposizione; nulla però garantisce che chi procede con le analisi rispetti questi paletti etici. Come per i sondaggi tradizionali sta alla moralità del ricercatore non rivelare all'esterno l'identità degli intervistati, così i ricercatori che utilizzano i Big Data devono proteggere la privacy di coloro che li hanno generati (Malhotra et al., 2007).

Interventi legislativi a tal riguardo sono già stati proposti per limitare i rischi di una violazione di un diritto così importante, come la *Do Not Track Online Act of 2011* negli Stati Uniti; con questo termine ci si riferisce ad una serie di proposte di legge che proteggono il diritto degli utenti di decidere se le loro attività all'interno di un sito possono essere o meno tracciate da terzi.

Il rischio però rimane sempre poiché Twitter, come molti altri social network o servizi online, mantiene ampio margine d'azione e si riserva il diritto di modificare a sua discrezione la sua normativa sulla privacy, cui si deve dare il consenso per crearsi un profilo. Resta all'utente il dovere di tenersi aggiornato su come e con chi il social network può condividere i dati in suo possesso. La normativa (aggiornata il 21 ottobre 2013) recita infatti: "Potremmo periodicamente rivedere questa Informativa sulla Privacy. La versione più aggiornata dell'informativa regolerà il nostro uso delle tue informazioni e sarà sempre disponibile all'indirizzo https://twitter.com/privacy. Se apporteremo una modifica a questa informativa che, a nostra esclusiva discrezione, sia sostanziale, ti informeremo con un aggiornamento di @Twitter o tramite email

all'indirizzo associato al tuo account. Se continui ad accedere o a utilizzare i Servizi dopo l'entrata in vigore di tali modifiche, accetti di essere vincolato dall'Informativa sulla Privacy modificata."

Un'altra questione che deve essere presa in considerazione quando si lavora con dataset costituiti da Big Data è la **rappresentatività dei dati**.
Nel caso specifico di Twitter è noto che i suoi utenti non rappresentino in modo omogeneo l'intera popolazione, ma appartengano soprattutto determinate fasce demografiche e socio-economiche, creando problemi nell'estendere i risultati ottenuti dall'analisi del microblog all'intera società. Le opinioni espresse online non sono rappresentative né dell'intera popolazione, perché non tutti i cittadini hanno un profilo nei social media, né dell'insieme degli utenti del social, poiché solo chi decide di esprimere la sua opinione online su quell'argomento viene rappresentato nel campione analizzato, di conseguenza la totalità della popolazione è sottostimata se si utilizzano solo gli *user-generated content*. Non tutti quelli che hanno Twitter pubblicano messaggi, ma nemmeno tutti quelli che twittano esprimono la loro opinione su ogni campo, soprattutto quello politico o medico sono i più sottoposti ad autocensura. A ciò si aggiunge che alcune volte vengono postate online cose non vere solo per far apparire agli altri qualcosa che nella realtà non c'è.
Vi sono poi casi in cui alcuni utenti hanno più account, alcuni account sono utilizzati da più utenti o in cui alcuni profili sono *bot* che producono contenuti in automatico senza coinvolgere direttamente persone reali.
Non perché i ricercatori possono disporre di una quantità enorme di dati significa che questa sia sufficiente per avere un campione rappresentativo, infatti alcuni studi hanno dimostrato che in Rete gli anziani sono sottorappresentati e vi è una prevalenza di uomini con un livello d'istruzione medio-alto (Wei and Hind-man 2011).
Mustafaraj et al. (2011) espongono, attraverso evidenze empiriche, come esistano due differenti tipi di comportamento all'interno dei social media: da un lato c'è una minoranza di utenti che produce la maggioranza dei contenuti (la minoranza parlante), e dall'altro vi è la maggioranza silenziosa che difficilmente produce contenuti ma che funge da cassa di risonanza nel diffondere opinioni che si allineano al loro pensiero. Durante le elezioni del Gennaio 2010 per il Senato americano in Massachusetts, coloro

che twittavano spesso (minoranza parlante) e coloro che avevano twittato una sola volta (maggioranza silenziosa) avevano generato contenuti differenti: i secondi inserivano più link esterni verso altri siti e retwittavano di più. Mustafaraj ed il suo team consigliano pertanto di prestare molta cautela e di separare i due gruppi quando si fanno previsioni basandosi su quanto viene postato nei social; similmente Gayo Avello (2011) suggerisce di identificare i diversi strati degli utilizzatori dei social per dare ad ognuno un peso in base alla loro porzione nella popolazione prima di generalizzare i risultati ottenuti.

Lo stesso problema emerge quando si utilizzano anche altre fonti di Big Data: ad esempio se il Governo prende decisioni basandosi su considerazioni che sono il frutto dell'utilizzo dei sensori presenti negli smartphone eliminerebbe dalla sua attenzione quella percentuale di cittadini che non possiedono un cellulare. In base alla tipologia di dati ci si può aspettare che coloro che li hanno generati possono essere i più giovani o anziani, i più ricchi o i più poveri, i più istruiti o quelli con un'istruzione bassa.

Un altro motivo per cui i dati possono non essere rappresentativi è la disponibilità offerta da chi li detiene: non tutti i tweet sono estraibili attraverso le Twitter API, non tutte le compagnie telefoniche mettono a disposizione i dati in loro possesso, ecco che in questo modo oltre a non essere gli utenti di Twitter o i detentori di uno smartphone rappresentativi dell'intera popolazione, nemmeno i dati a disposizione sono rappresentativi del sottoinsieme della società che si va a considerare (UN Global Pulse, 2012).

Avere dati da campioni non rappresentativi non significa però avere dei dati del tutto non informativi, essi lo sono sicuramente relativamente al campione, e lo sono in tempo reale e a basso costo.

Bisogna poi considerare anche che la popolazione dei social network sta via via aumentando tanto da far quasi affermare che il problema della non rappresentatività stia pian piano affievolendosi. C'è pure da tener presente che, come ha osservato O'Connor (2010), gli utenti attivi online possono agire da opinion maker capaci di influenzare le idee della massa. Sembra, infatti, che l'opinione espressa, pur essendo quella di un campione elitario, abbia la capacità di influenzare l'opinione del resto della popolazione e goda di una buona cassa di risonanza mediatica che ne permette un'ampia diffusione. In secondo luogo, partendo dall'assunto che l'opinione pubblica è influenzata dalle notizie pubblicate dai media tradizionali (Kepplinger, 2008), ed aggiungendo che online

moltissimi post fanno riferimento ad una notizia di un canale di massa tradizionale come referenza, appare chiaro come le idee dei cittadini siano influenzate da quanto si legge online.
Ecco forse perché, nonostante tutto, da alcuni studi, come quello di Fu e Chan (2013), risulta che una *sentiment analysis* degli *user-generated content* può esser in grado di predire i risultati di sondaggi telefonici sull'opinione pubblica.

L'**accessibilità** è un altro limite da tener presente. Ottenere dati estratti da Twitter senza le limitazioni imposte dalle applicazioni ufficiali, ma completi e datati richiede l'intervento di aziende specializzate che abbiano competenze ed autorizzazioni adeguate; un intervento che si ottiene a pagamento.
Come Manovich (2011) sottolinea, "solo i social network stessi hanno l'accesso diretto e completo alla totalità dei dati generati sulle loro piattaforme; un antropologo che lavora per Facebook o un sociologo che lavora per Google ha accesso a dei dati che il resto della comunità di studiosi non potrebbe mai avere".
Allargando la considerazione all'intero universo dei Big Data nonostante molti di essi siano pubblici, e quindi accessibili per chiunque ne sia interessato, ce ne sono di altri, altrettanto utili, che non sono accessibili perché detenuti da compagnie private riluttanti a condividere le informazioni sui loro clienti e sulle loro azioni per motivi legali, per proteggere la loro reputazione, per una cultura aziendale orientata alla segretezza, o più semplicemente per la mancanza d'incentivi.
UN Global Pulse sta cercando di introdurre il concetto di "*data philanthropy*", secondo cui le organizzazioni dovrebbero prendere l'iniziativa di anonimizzare i loro dataset e fornirli a degli innovatori sociali affinché questi possano analizzarli per scoprire in essi schemi ricorrenti e andamenti in tempo reale, o quasi.

Non si deve sottovalutare anche che ogni dataset, e soprattutto quelli estratti da Internet, sono inclini alla presenza di **errori** e di **dati mancati**, e quando più banche dati diverse vengono aggregati queste mancanze si amplificano.

Bisogna inoltre interrogarsi sulla **validità** dei dati a disposizione, poiché spesso questi sono poco attendibili sia perché creati appositamente falsi o perché derivano dalla

percezione soggettiva delle persone, come ad esempio le ricerche online che nascono dalla supposizione di un soggetto di avere un sintomo quando magari nella realtà non ce l'ha.
Anche qualora i dati siano considerati accurati, la loro interpretazione non è sempre così immediata poiché, come suggerisce Andreas Weigend (2013), "i dati sono il nuovo olio, e come l'olio devono essere raffinati prima di esser usati", e, come sottolinea Crawford (2009), la collezione automatica di un dato non è autoesplicativa, ma richiede selezioni ed interpretazioni accurate.

Secondo la sociologa Tufekci anche se interessante, non è sufficiente sapere il numero di 'mi piace' ad uno status in Facebook o quanti retweet ha ricevuto un messaggio; ciò che renderebbe tale informazione più completa sarebbe sapere quante persone hanno preso visione di quell'*user-generated content* e non hanno intrapreso alcuna azione, cosa al momento impossibile, per avere una visione della reazione dell'intero pubblico a quello che viene pubblicato.

Un rischio in cui i ricercatori incorrono quando lavorano su grandi dataset è l'**apofenia**, ossia la percezione di uno schema rilevante in un insieme di dati senza significato, ciò è possibile perché un'enorme quantitativo di dati può portare ad un range di possibilità di correlazione che però nella realtà non esistono. Questa problematica è possibile con ogni tipo di dati, ma in particolare con i Big Data provenienti dai social media. Un caso di questo tipo è quello messo in luce da Jim Fruchterman (2011) in cui la correlazione scoperta tra l'invio di sms e la presenza di edifici danneggiati era erroneamente emersa dai dati telefonici relativi ai giorni immediatamente successivi al terremoto di Haiti.
E' quindi sempre necessario un accertamento dell'oggettività delle interrelazioni scoperte.

Emergono ovviamente problemi **d'interpretazione** per comprendere, attraverso algoritmi adatti, cosa le persone volessero comunicare attraverso il contenuto pubblicato. Dal momento che si studia una conversazione che avviene online tra gli utenti bisogna pensare ad una situazione in cui il dialogo avviene attraverso modalità particolari che potrebbero rendere invisibile o incomprensibile agli algoritmi sviluppati

dai ricercatori il suo evolversi, o che potrebbero causarne un'errata interpretazione: un retweet ad esempio è molto ambiguo perché non esprime solo supporto, ma potenzialmente potrebbe anche essere una forma di denuncia o derisione.
La **contestualizzazione** è la chiave di lettura fondamentale per evitare errori di comprensione durante l'interpretazione dei dati a disposizione.
La contestualizzazione deve esser fatta non solo a livello di singolo dato, ma anche a livello culturale. Pratiche e norme culturali sono diverse nel mondo e queste differenze si estendono anche all'uso dei servizi digitali. La dimensione etnografica deve essere considerata dal momento che diverse popolazioni hanno norme differenti riguardo a come comunicano pubblicamente circa la loro vita privata, quindi sapere cosa viene definito 'normale' in una certa regione è un prerequisito per identificare anomalie.

Concentrandosi sui limiti relativi solo all'analisi di Twitter emergono i seguenti punti:

- selezionando preventivamente gli hashtag e le parole chiave che si vuole siano contenuti nei tweet che vengono estratti, perché si considerano rilevanti per l'analisi che si intende svolgere, può accadere che alcuni, seppur pertinenti al tema indagato, non vengono inclusi nel corpo analizzato. Al contrario tra i tweet così estratti molti di essi sono irrilevanti al fini della ricerca, secondo una stima Ceron, Curini e Iacus (2014) circa l'11%. Tutto ciò accade a causa della generalizzazione delle basi dell'identificazione.
- la sintassi irregolare e il linguaggio non standard del microblog creano maggiori difficoltà nell'analisi semantica rispetti ad altre tipologie di testi. Vi è quindi la necessità di creare specifici classificatori linguistici per condurre *sentiment analysis* o *opinion mining*, con tutte le difficoltà già esposte nel paragrafo 3.3.
- l'architettura della piattaforma ostacola il generarsi di conversazioni articolate, cosa che invece avviene ad esempio in Facebook.
- pochi tweet sono geolocalizzati, quindi anche se numerosi studi interessanti potrebbero essere svolti basandosi sulla localizzazione dell'autore dei messaggi postati in Twitter, la possibilità per l'utente di negare l'accesso a quest'informazione rende difficile un loro svolgimento. Un'alternativa escogitata dai ricercatori è estrapolare la provenienza del tweet indagando la descrizione personale dall'account che lo ha pubblicato, questo modo però può

risultare poco affidabile dato che molti profili non contengono informazioni personali aggiornate o vere.

Quando i ricercatori si approcciano ad un'analisi del social network devono essere consci non solo dei limiti del dataset a loro disposizione, ma anche delle domande a cui essi possono ragionevolmente cercare risposta attraverso i dati in loro possesso; solo con questa consapevolezza saranno in grado di dare la giusta interpretazione ai risultati ottenuti.

Anche se l'analisi dei dati ottenibili da Twitter presenta limitazioni importanti e rischia di dare risultati meno precisi di quelli provenienti dai sistemi ufficiali tradizionalmente adottati, offre interessanti possibilità, veloci e poco costose, sfruttabili in diversi ambiti.

Le problematiche elencate non sembrano quindi scoraggiare gli studiosi che ripongono nell'utilizzo dei Big Data provenienti dai social network grandi speranze di sviluppo.

4. Analisi di Twitter riguardo al tema della sigaretta elettronica

Per completare la presentazione delle potenzialità dell'analisi dell'utilizzo di Twitter si è deciso di adottare questa metodologia d'analisi per scoprire se la sigaretta elettronica è utilizzata dagli italiani come strumento per smettere di fumare.
La scelta di questa tematica è stata determinata sia dalla sua novità, caratteristica che ne aumenta la possibilità di essere discussa online, sia dalla sua rilevanza per la salute pubblica. Nel paragrafo 4.1 sono riportate informazioni generali riguardanti il fenomeno del fumo elettronico per dare al lettore un'immagine dello scenario attuale; nel paragrafo 4.2 sarà invece descritta nel dettaglio l'analisi condotta durante il primo quadrimestre 2014 sui dati estratti dal social network. Infine, nel paragrafo 4.3 sarà confrontato quanto emerso dallo studio dei tweet con dati provenienti dal mondo reale.

4.1 La sigaretta elettronica

La sigaretta elettronica, o e-cigarette, è un dispositivo elettronico che riprende le mimiche e le percezioni sensoriali delle sigarette, dei sigari o delle pipe ma non prevede la combustione del tabacco e non contiene catrame; in questo modo vengono eliminati gli effetti cancerogeni del fumo tradizionale.
Brevettata nel 1965 da Herbert A. Gilbert venne inizialmente commercializzata, dal 2003, in Cina per poi diffondersi nel resto del mondo.

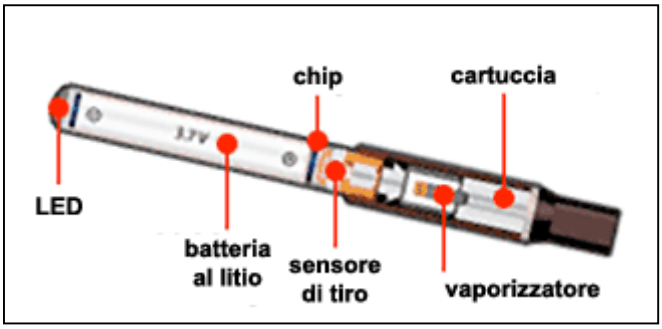


Figura 4.1: Elementi costitutivi di una sigaretta elettronica
Fonte: www.svapo.it

Il suo funzionamento è molto semplice: attivata manualmente o in automatico al momento dell'aspirazione la batteria posta all'estremità avvia l'atomizzatore che

riscaldando un liquido, costituito da sostanze innocue come glicole propilenico e glicerina vegetale aromatizzati a vari sapori, lo fa diventare vapore ed uscire dal buco della cartuccia arrivando così alla bocca del fumatore; in questo modo la sigaretta non si consuma e nulla viene bruciato. La soluzione presente nella cartuccia contenuta nel filtro può avere sapori diversi e differenti concentrazioni di nicotina o esserne anche priva. La batteria si ricarica semplicemente con un cavo USB. I modelli principali in Italia sono le 510 e le Ego.

Lo scopo della sigaretta elettronica è quello di dare a colui che la usa, chiamato svapatore, delle sensazioni simili a quelle che ha con le sigarette classiche attraverso un prodotto che le imita ma che non crea, a detta dei suoi sostenitori, danni alla salute e che è meno costoso. Sembra quindi essere una possibile soluzione per coloro che vogliono migliorare la qualità della loro vita senza però rinunciare al piacere di fumare e alla dipendenza dalla nicotina.

4.1.1 La legislazione in tema di sigaretta elettronica

Data la loro recente diffusione sul mercato, il panorama legislativo a livello europeo in tema di sigaretta elettronica è molto diversificato ed in continuo cambiamento: le leggi in vigore sono differenti in ogni Paese (vedi Figura 4.2) e sono in fase di elaborazione ed assestamento.

Se in Australia ed in Finlandia vendere sigarette elettroniche contenenti nicotina è illegale, ad Hong Kong lo è anche il possesso, mentre in Belgio, Danimarca, Estonia, Germania, Ungheria, Austria, Slovenia, Finlandia, Portogallo, Svezia e Nuova Zelanda esse sono equiparate a dispositivi medici e le cartucce di nicotina sono registrate come medicinali. In Olanda ed Inghilterra invece non vi sono restrizioni per la loro vendita e utilizzo, come anche in Italia.

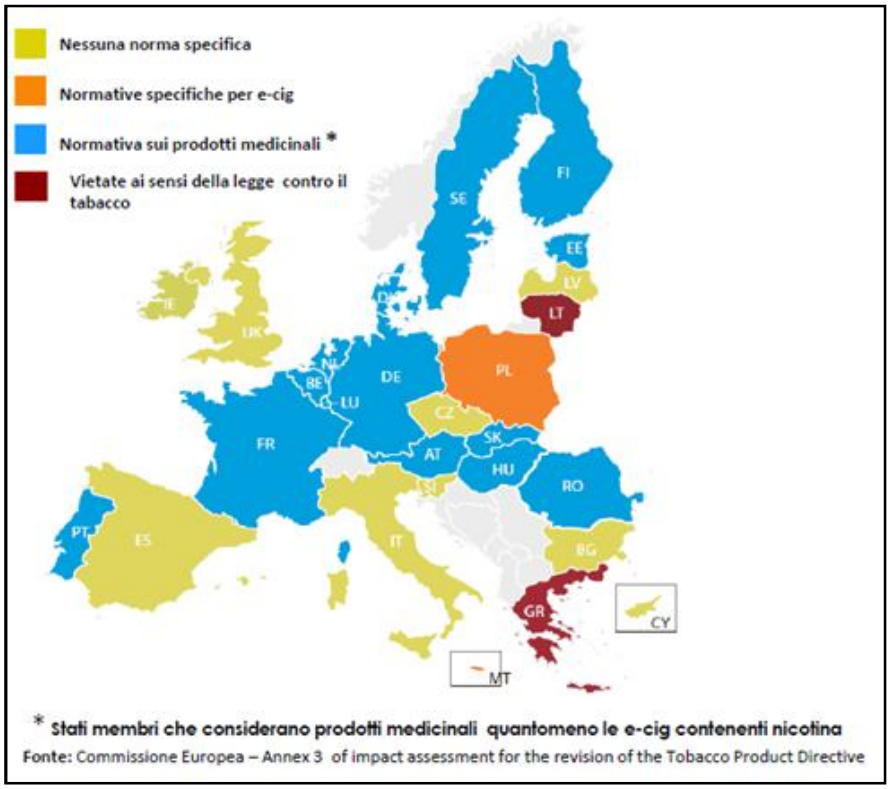


Figura 4.2: Differenze di regolamentazione della sigaretta elettronica all'interno dell'Unione Europea nel 2013

Fonte: Rapporto ISS-DOXA 2013

Con lo scopo di uniformare la legislazione vigente in Europa il 18 dicembre 2013 gli Stati membri UE ed il Parlamento europeo hanno raggiunto il compromesso finale su tutti i punti di una direttiva anti fumo, che dovrà essere trasferita in legislazione nazionale entro due anni dall'adozione, avvenuta il 26 febbraio 2014.

L'emendamento oltre ad imporre regole riguardanti il fumo tradizionale allo scopo di diminuirlo, ha introdotto regolamentazioni anche in materia di fumo elettronico: alla sigaretta elettronica non sarà dato il ruolo di farmaco, ma gli Stati che già la considerano un medicinale, come Francia e Germania, potranno continuare a farlo. Inoltre le aziende produttrici di e-cigarette dovranno seguire una procedura autorizzativa per poterle vendere.

Altri limiti enunciati dalla direttiva sono la massima concentrazione di nicotina nei liquidi concessa pari a 20 mg/ml, regole sulla pubblicità e l'imposizione di un packaging sicuro per i bambini. Infine la Commissione europea s'impegna a presentare

entro un biennio dall'adozione del testo un report sulla dannosità delle sigarette elettroniche.

Concentrandosi sull'Italia, in attesa di informazioni certe sulla tossicità e pericolosità, la legislazione italiana ne consente la vendita con alcune prescrizioni ispirate al principio di cautela: l'Ordinanza del Ministro della Salute del 4 Agosto 2011 (pubblicata sulla Gazzetta Ufficiale n. 232 del 5 ottobre 2011) ha posto il divieto della vendita ai minori di anni 16 di sigarette elettroniche contenenti nicotina, limite innalzato il 28 Settembre 2012 ai 18 anni (con l'ordinanza pubblicata sulla Gazzetta Ufficiale n. 248 del 23 ottobre 2012); vige inoltre il divieto di svapare nelle scuole e vi sono regole per le etichette e l'informativa ai consumatori (Ordinanza del Ministero della Salute 26 giugno 2013 pubblicata sulla Gazzetta Ufficiale n. 176 del 29 luglio 2013).

Ruolo centrale riveste il Decreto 76/2013 su Lavoro e Iva convertito nella Legge 99/2013 il 9 Agosto 2013, che aveva portato ad un innalzamento della tassa sulle sigarette elettroniche poiché le equiparava al tabacco tradizionale, e lo stesso per tutti gli accessori necessari, come batterie e caricatori. Con le parole del Decreto l'imposta al consumo era da applicarsi ai "prodotti contenenti nicotina o altre sostanze idonei a sostituire il consumo dei tabacchi lavorati nonché i dispositivi meccanici ed elettronici, comprese le parti di ricambio, che ne consentono il consumo". Il decreto del Ministero adottato nel Dicembre 2013 faceva entrare in vigore la normativa solo con l'anno nuovo; in questo modo era stato calcolato che nel 2014 l'erario avrebbe guadagnato 117 milioni di euro, ma, con i prezzi costanti i produttori, non avendo aggiornato i listini con preavviso, avrebbero avuto un carico fiscale pari al 146% del loro fatturato nel primo mese dell'anno.

Il Decreto inoltre imponeva che per "La commercializzazione di tali prodotti sarà assoggettata alla preventiva autorizzazione da parte dell'Agenzia delle Dogane e dei Monopoli nei confronti di soggetti che siano in possesso dei medesimi requisiti stabiliti, per la gestione dei depositi fiscali di tabacchi lavorati.

Tali soggetti sono inoltre tenuti alla preventiva prestazione di cauzione, in uno dei modi stabiliti dalla legge 10 giugno 1982, n. 348, a garanzia dell'imposta dovuta per ciascun periodo di imposta".

Emersero subito polemiche e problemi che si riferiscono alla sua applicazione dal momento che, oltre ai problemi di redditività che la tassa avrebbe causato ai produttori,

vi era un problema tecnico poiché il decreto obbligava le aziende produttrici a svolgere le pratiche di registrazione e certificazione presso i Monopoli (Aams), le cui tempistiche richieste non coincidevano con il lasso temporale a disposizione prima dell'inizio del 2014. Le aziende produttrici fecero quindi ricorso al Tar, il quale il 21 gennaio 2014 sospese solo le procedure autorizzative previste per i depositi mantenendo però la tassa prevista del 58,5%, per sospenderla poi a fini cautelari il 2 aprile 2014, avendo riscontrato «profili d'irragionevolezza», in attesa della decisione finale da parte della Corte Costituzionale.

Nonostante ora la maxi-tassa del 58,5% ed il sistema di autorizzazioni siano state sospese, essendo stato applicato quel regime fiscale nel primo trimestre dell'anno il settore ha subito gravi perdite e ha registrato un blocco; le novità imposte dal decreto facevano talmente paura ai produttori che trenta dei maggiori avevano sospeso la loro attività, e anche dopo la sospensiva era rimasto il timore che un decreto con valore retroattivo portasse loro a dover pagare pesanti sanzioni. Per queste problematiche ben mille punti vendita sui 3,5 mila hanno chiuso ad inizio 2014 ed il personale delle aziende produttrici è stato diminuito del 10%.

Le proteste provenienti dal settore delle sigarette elettroniche alla direttiva che dovrà essere implementata dagli Stati membri dell'UE sono molte. Tra esse spicca l'EFVI (Europian Free Vaping Iniziatiative) che ha l'obiettivo di raccogliere un milione di firme entro il 25 novembre 2014 per avvalersi del Diritto d'Iniziativa dei Cittadini Europei, un meccanismo attraverso cui i cittadini europei possono esprimere la loro opinione condivisa sulle modalità con cui un argomento dovrebbe essere gestito dal Parlamento, nel caso specifico il tema delle sigarette elettroniche. Il manifesto recita così: "Gli svapatori hanno preso una decisione matura, responsabile e personale, quando hanno scelto questa alternativa recente ma efficace che ha permesso di abbandonare o ridurre drasticamente un'abitudine mortale che causa centinaia di migliaia di vittime all'anno nella Comunità Europea. Il fatto che lo svapo non provochi alcun danno alle persone vicine, rende questo hobby una questione privata e non pubblica, quindi ogni tentativo di regolamentare le sigarette elettroniche è una seria violazione dei diritti privati e della sfera personale". L'iniziativa sottolinea quindi l'utilità delle e-cigarette per smettere di fumare e del diritto e chiede al Parlamento quanto segue: "Noi, gli svapatori Europei ed i sostenitori dello svapo, chiediamo che le Sigarette Elettroniche e

tutti i prodotti correlati, indipendentemente dal loro contenuto di nicotina, siano classificate a livello legislativo come prodotti di consumo, una volta per tutte, e non come medicinali, tabacco o succedanei del tabacco, od ogni altra definizione che limiti o pregiudichi l'accesso da parte degli svapatori alle sigarette elettroniche ed ai prodotti correlati."

Un'iniziativa importante per lo scenario italiana è #Svapevatelo, una campagna promossa dall'Associazione Nazionale Fumo Elettronico (Anafe), dalla Federazione Italiana Esercenti Svapo Elettronico (Fiesel) e della Lega Italiana Fumo Elettronico (Life) contro il regime fiscale che prevede la maxitassa del 58,5 % sui prodotti succedanei del tabacco tra cui rientravano anche le sigarette elettroniche dopo il decreto di Dicembre 2013. La campagna iniziata il 28 marzo 2014 si è conclusa con la sospensione cautelare della tassa in questione da parte del Tar il 2 Aprile 2014. Le associazioni di categoria promuoventi coinvolte si sono impegnate a diffondere l'hashtag #Svapevatelo nei social network per attirare l'attenzione pubblica sulla tassazione del 58,5% che è stata poi sospesa dal Tar del Lazio il 2 Aprile 2014; la campagna è stata diffusa anche nelle principali testate giornalistiche italiane a pagamento (il 28 marzo su 'La Notizia Giornale', 'Libero' e 'L'Unità', il 30 marzo su 'La Nazione Firenze', il 31 marzo su 'Il Tempo' ed il 1° aprile su 'Il Sole 24 Ore'). Il manifesto sponsorizzante l'iniziativa faceva leva sulla crisi del settore che l'introduzione del nuovo regime fiscale aveva causato, e sottolineva anche la possibilità che le sigarette elettroniche aiutino a ridurre i danni causati dal fumo.

4.1.2 Dati sul fumo e sulle sigarette elettroniche in Italia

Le fonti principali di dati sul fenomeno del fumo in Italia sono tre: il sistema di sorveglianza Passi (Progressi delle Aziende Sanitarie per la Salute in Italia), i rapporti annuali ISS-DOXA e le indagini ISTAT. In questa sede saranno prese in considerazione solo le prime due poiché la terza non si occupa, per ora, del fumo elettronico.

Il primo è adottato dal Ministero della Salute italiano per monitorare l'evoluzione del comportamento dei cittadini rispetto a tematiche di rilevante importanza per la salute pubblica, tra le quali rientra anche il fumo.

Raccoglie continuamente, attraverso indagini campionarie, informazioni su stili di vita, su fattori di rischio comportamentali connessi all'insorgenza delle malattie croniche non trasmissibili e sul grado di conoscenza e adesione da parte dei cittadini ai programmi di prevenzione o cura in atto in Italia.

Gestito dalle Asl e coordinato sia a livello nazionale dall'ISS che a livello regionale dai Dipartimenti di Prevenzione. Questo coordinamento centralizzato ne assicura procedure standardizzate di rilevazione, analisi e comunicazione dei risultati.

La popolazione di riferimento sono gli iscritti all'anagrafe assistiti dell'Asl tra i 18 e i 69 anni, facendo riferimento alle liste aggiornate al 1 gennaio dell'anno di rilevazione. Per esser inclusi nella sorveglianza Passi bastano la residenza nel territorio dell'Asl ed il possesso di un recapito telefonico; sono esclusi coloro che non conoscono la lingua italiana, non possono sostenere un'intervista o ricoprono cariche istituzionali durante il periodo dell'indagine.

Il campione è stratificato proporzionale, per sesso e classi di età (18-34, 35-49, 50-69 anni), della dimensione minima per ciascuna Asl di venticinque unità. Sono previsti 11 campioni aziendali mensili, dato che il bimestre luglio-agosto è conteggiato come singolo, per un totale annuo di almeno 275 unità campionarie per Asl e 35.000 unità per il livello nazionale. Al fine di avere un'idonea rappresentatività dei risultati a livello regionale i dati delle singole Asl vengono aggregati per Regione e opportunamente pesati.

Il sistema di sorveglianza Passi per le sue misurazioni adotta le definizioni operative dell'Organizzazione mondiale della sanità, considera:

- non fumatore colui che non fuma e che non ha fumato nell'arco della sua vita più di centro sigarette;
- fumatore chi ne ha fumate più di cento e lo sta continuando a fare o non lo fa da meno di un semestre;
- fumatore occasionale chi non fuma ogni giorno; come fumatore in astensione chi dichiara di non fumare da meno di sei mesi;
- ex fumatore chi nella sua vita ha fumato più di un centinaio di sigarette ma al momento non ne fuma nessuna e non lo fa da più di sei mesi;
- tentativo di smettere l'intenzione di cessare che ha portato ad almeno 24 ore consecutive di astinenza.

La raccolta delle informazioni avviene tramite interviste telefoniche di circa 20 minuti, effettuate nel corso di tutto l'anno dal personale dei Dipartimenti di prevenzione delle Asl coinvolte nella sorveglianza, a cittadini selezionati come unità campionarie, preventivamente avvisati tramite una lettera personale informativa.
Per l'intervista ci si avvale di un questionario standardizzato e validato a livello nazionale ed internazionale.
Le indagini ISS-DOXA, invece, sono dei rapporti annuali redatti per incarico dell'ISS (Istituto Superiore di Sanità) e OSSFAD (Osservatorio Fumo, Alcol e Droga) da Doxa, una delle più importanti aziende in Italia di ricerche e analisi di mercato, in collaborazione con l'Istituto di Ricerche Farmacologiche Mario Negri. Il campione in questo caso è costituito da circa 3000 (esattamente 3052 nel 2014) soggetti rappresentativi della popolazione italiana dai 15 anni in su, residenti in 116 diversi Comuni italiani.
Per le interviste, svolte a domicilio da intervistatori opportunamente istruiti e controllati, si adotta il sistema CAPI (Computer Assisted Personal Interviewing).
Il rapporto ISS-DOXA 2014 espone i risultati delle interviste raccolte tra il 20 Febbraio e il 18 aprile 2014, facenti quindi riferimento allo stesso periodo durante il quale è stata condotta la raccolta di dati da Twitter. La base campionaria minore porta però a considerare Passi come fonte più attendibile poiché meno soggetta a fluttuazioni legate solo al variare dei soggetti intervistati e non indicative di un reale cambiamento nell'andamento.
Ne sono stati comunque riportati i risultati per completezza e perché questi sono molto dettagliati e resi disponibili con cadenza regolare ogni 31 Maggio.
Le differenze riscontrabili negli output delle due fonti sono causate, non solo dalla diversa numerosità campionaria, ma anche dalla differente età dei soggetti che costituiscono il campione.

Anche se i danni del fumo sono a tutti noti, basta citare la dichiarazione del *National Center for Chronic Disease Prevention and Health Promotion* del 2004 che contava 69 sostanze chimiche che causano il cancro contenute nelle sigarette tradizionali, i fumatori sono ancora molti a causa della dipendenza che il loro consumo sviluppa. Bisogna tuttavia sottolineare che il consumo di sigarette in Italia è in lenta ma progressiva

diminuzione, secondo entrambe le fonti in precedenza citate (vedi Figura 4.3 e 4.4). Le percentuali riportate però differiscono, ed il trend appare essere molto più costante nei dati Passi rispetto a quelli ISS-DOXA, le cui fluttuazioni sono spiegabili dalla limitata numerosità del campione, come già esposto in precedenza.

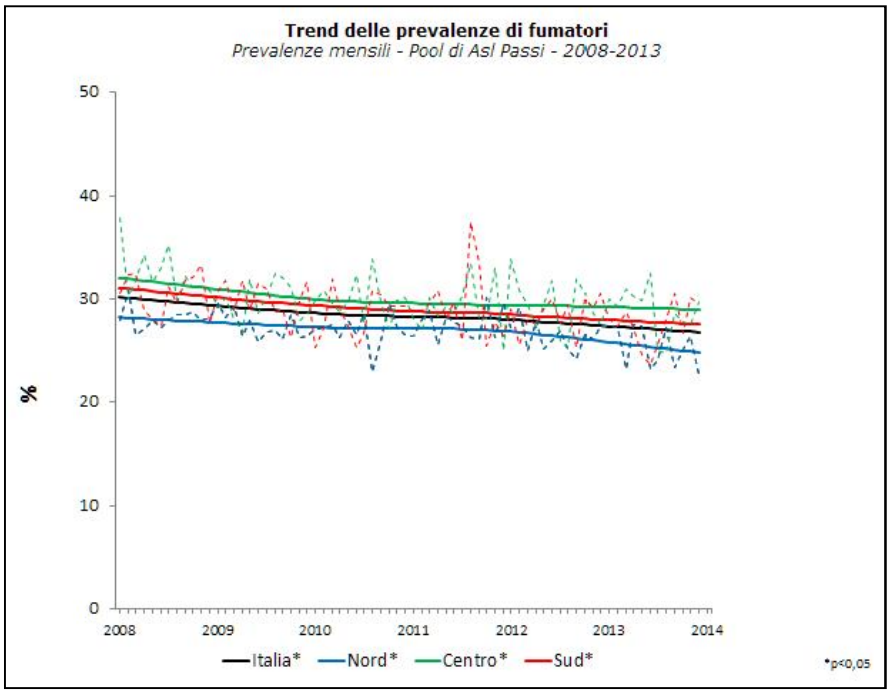


Figura 4.3: Trend del totale dei fumatori dal 2008 al 2014 in Italia

Fonte: sorveglianza Passi

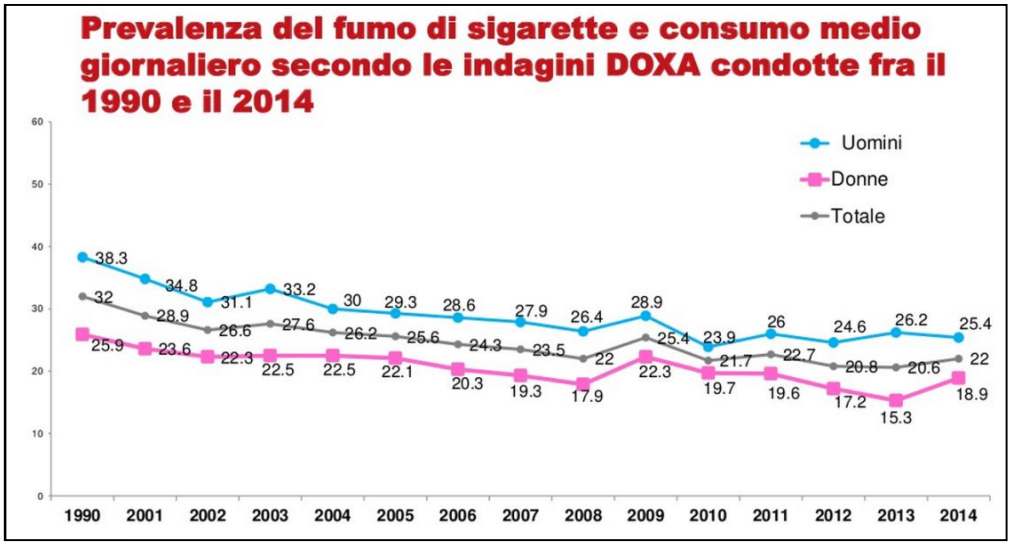


Figura 4.4: Trend del totale dei fumatori dal 1990 al 2014 in Italia

Fonte: Rapporto ISS-DOXA 2014

Secondo l'indagine ISS-DOXA nel 1990 si fumavano in media 16,3 sigarette, nel 2014 solo 12,9 (vedi Figura 4.5).

Prevalenza del fumo di sigarette e consumo medio giornaliero secondo le indagini DOXA condotte fra il 1990 e il 2014

Dom. 1 – 4b

Base: Tutti gli adulti (3000 casi)

	TOTALE		MASCHI		FEMMINE	
	Prevalenza fumo sigarette	*Consumo medio giornaliero*	*Prevalenza fumo sigarette*	*Consumo medio giornaliero*	*Prevalenza fumo sigarette*	*Consumo medio giornaliero*
1990	32.0%	16.3	38.3%	18.6	25.9%	12.1
2001	28.9%	16.4	34.8%	18.8	23.6%	12.2
2002	26.6%	16.8	31.1%	19.2	22.3%	13.1
2003	27.6%	16.1	33.2%	18.6	22.5%	12.2
2004	26.2%	14.8	30.0%	16.1	22.5%	13.1
2005	25.6%	14.0	29.3%	15.8	22.1%	11.9
2006	24.3%	13.6	28.6%	14.8	20.3%	12.0
2007	23.5%	14.1	27.9%	14.9	19.3%	13.0
2008	22.0%	14.4	26.4%	15.3	17.9%	13.0
2009	25.4%	14.1	28.9%	16.0	22.3%	11.7
2010	21.7%	13.0	23.9%	14.5	19.7%	11.5
2011	22.7%	13.6	26.0%	15.6	19.6%	11.1
2012	20.8%	13,0	24.6%	14,3	17.2%	11,3
2013	20.6%	12,7	26.2%	13,5	15.3%	11,5
2014	22.0%	12,9	25.4%	14,3	18.9%	11,2

Figura 4.5: Variazione del numero di sigarette fumate ogni giorno in Italia dal 1990 al 2014 Rapporto ISS-DOXA 2014

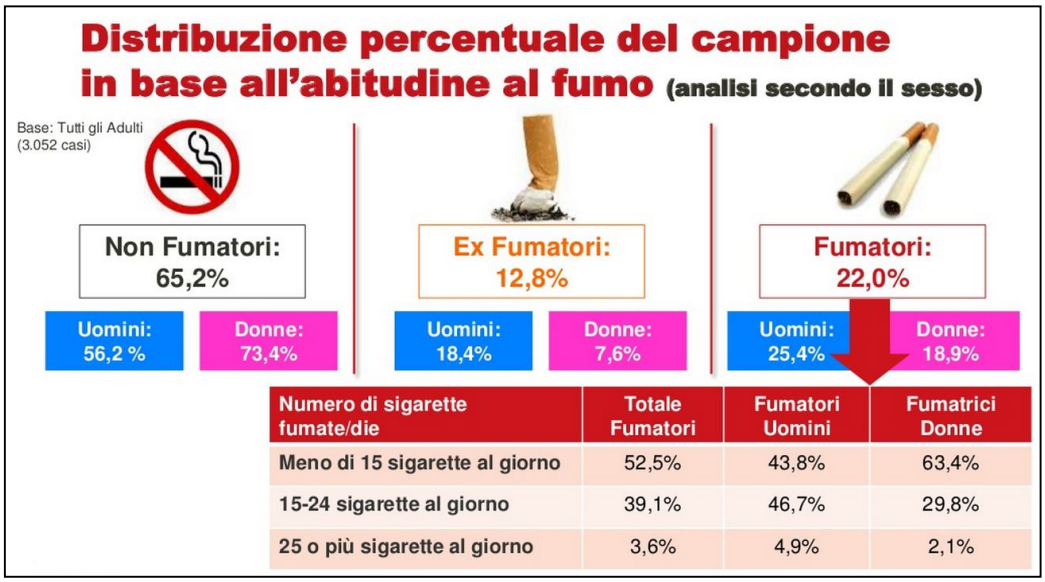
Il numero dei fumatori che diminuisce nell'ultimo ventennio ed il calo di sigarette consumate al giorno da chi fuma hanno portato ad una significativa diminuzione delle vendite dei prodotti legati al tabacco (-8% nel 2013), con un trend che negli ultimi due anni è in notevole peggioramento. Nell'ultimo decennio, infatti, si è registrata una contrazione delle vendite di circa 21 milioni di kg, dei quali oltre 10 milioni solo tra il 2011 ed il 2013.

Nel XVI Rapporto realizzato dall'Istituto Nomisma nel 2012 le superfici coltivate a tabacco in Italia sono diventate 15.106 ettari, un terzo in meno rispetto a quelle del 2011, con un calo nella produzione del 27%. La quantità di sigarette consumate è diminuita, arrivando a 78,8 milioni di kg, un livello che non si toccava da più di 40 anni. Le cause alla base sono molteplici: aumento dell'IVA, sostituzione con prodotti più convenienti, legislazione europea restrittiva, campagne antifumo.

I dati del periodo 2010-2013 raccolta da **Passi** dicono che in Italia fuma il 28% della popolazione, mentre il 18% sono ex fumatori. La prevalenza di **fumatori** è maggiore tra

coloro che hanno difficoltà economiche (37%), tra gli uomini (33%) rispetto alle donne (23%) ed è minore fra i laureati (21%).
Umbria, Lazio e Campania sono le Regioni con la maggior presenza di fumatori (31%), mentre il Veneto con la minore (24%). In generale nel Nord Italia ci sono le prevalenze più basse, ma la variabilità geografica è molto contenuta.
Dal 2008 al 2012 i fumatori sono diminuiti dal 30% al 27% (-3,4% al Nord, -3,1% al Centro e -3,5 % al Sud).
Se si analizza il trend, stratificando per difficoltà economiche, si rileva che il calo dei fumatori è significativo solo per i più abbienti, mentre rimane stazionario per quelli con molte difficoltà economiche, che presentano oltretutto la quota maggiore di fumatori: le differenze a livello economico sembrano quindi ripercuotersi in disuguaglianze nell'ambito della salute.

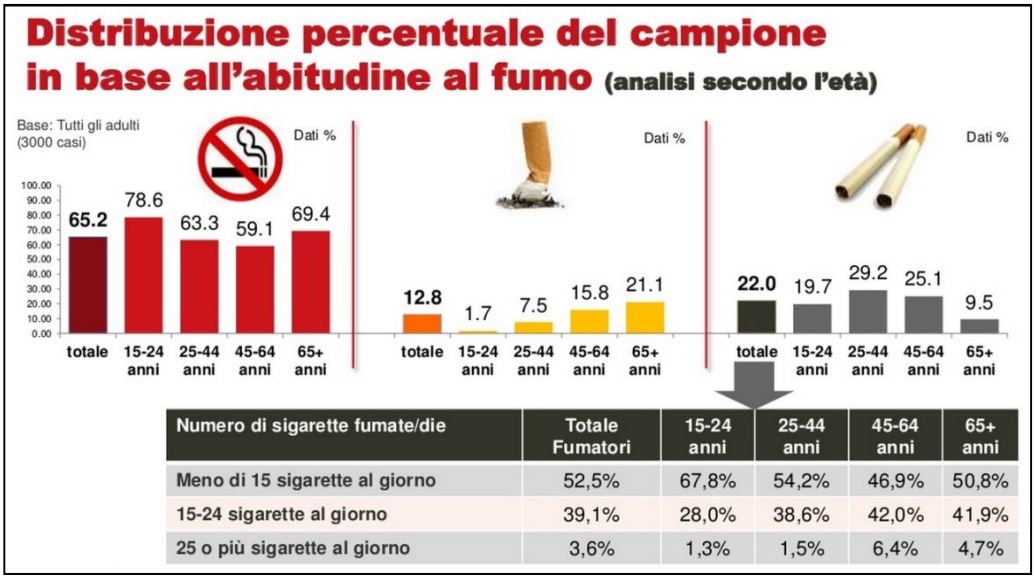
Secondo invece il **report annuale ISS-DOXA** i **fumatori** sono il 22% della popolazione, mentre il 12,8% sono ex fumatori (vedi Figura 4.6), anche in questo caso rispetto ai dati Passi il fenomeno risulta sottostimato, ma non in maniera rilevante. La fascia d'età 25-46 registra la maggioranza di consumatori di tabacco (29,2%), mentre quella over 65 la maggioranza di ex fumatori (21,1%) (vedi Figura 4.7).
Si inizia a fumare tra i 15-20 anni (72,4% del totale fumatori), soprattutto perché influenzati dagli amici (59%).
Il 94,8% di chi si dichiara fumatore consuma tabacco ogni giorno. Il 95,6% fuma sigarette tradizionali, il 18% sigarette fatte a mano (l'anno precedente la percentuale era del 9,6%; significativamente in aumento soprattutto tra i giovani, probabilmente perché costano meno di quelle confezionate), il 2,1% e-cig contenenti nicotina.



Numero di sigarette fumate/die	Totale Fumatori	Fumatori Uomini	Fumatrici Donne
Meno di 15 sigarette al giorno	52,5%	43,8%	63,4%
15-24 sigarette al giorno	39,1%	46,7%	29,8%
25 o più sigarette al giorno	3,6%	4,9%	2,1%

*Figura 4.6:*Abitudini al fumo degli italiani secondo il genere nel 2014

*Fonte:*Indagine ISS-DOXA 2014



Numero di sigarette fumate/die	Totale Fumatori	15-24 anni	25-44 anni	45-64 anni	65+ anni
Meno di 15 sigarette al giorno	52,5%	67,8%	54,2%	46,9%	50,8%
15-24 sigarette al giorno	39,1%	28,0%	38,6%	42,0%	41,9%
25 o più sigarette al giorno	3,6%	1,3%	1,5%	6,4%	4,7%

Figura 4.7: Abitudini al fumo degli italiani secondo l'età nel 2014

*Fonte:*Indagine ISS-DOXA 2014

Secondo i dati **Passi** tra il 2010 ed il 2013, il 39,1% dei fumatori ha tentato di **smettere di fumare**, riuscendo a rimanere almeno un giorno in astinenza. A tentare più spesso sono i giovani tra i 18 e 24 anni (46%), mentre non si osservano differenze significative per genere e istruzione.

È stato però evidenziato, attraverso un'analisi multivariata, come nelle donne ad un maggiore livello d'istruzione corrispondano più prove di smettere; i loro tentativi sono però inversamente associati alla disponibilità economica.

Dei fumatori che provano a smettere di fumare più dell'80% ha fallito, il 10% sta ancora tentando ma non ha raggiunto i sei mesi d'astinenza necessari per esser definito ex fumatore e solo l'8% ha raggiunto il suo scopo.

Quasi tutti i tentativi (93%) sono stati fatti senza utilizzo di farmaci o supporto medico.

Dal 2008 al 2013 si è registrato un trend costante in calo dei fumatori che provano a smettere di fumare (dal 42% al 39%.). A causare la diminuzione è stato soprattutto il Centro Italia ed il Sud, mentre al Nord non è stata raggiunta la significatività statistica. Un'inversione di tendenza è stata rilevata nel 2013, ma essa deve essere verificata con i dati dei prossimi anni per poter affermare con sicurezza che sta avvenendo un cambiamento del trend (vedi Figura 4.8).

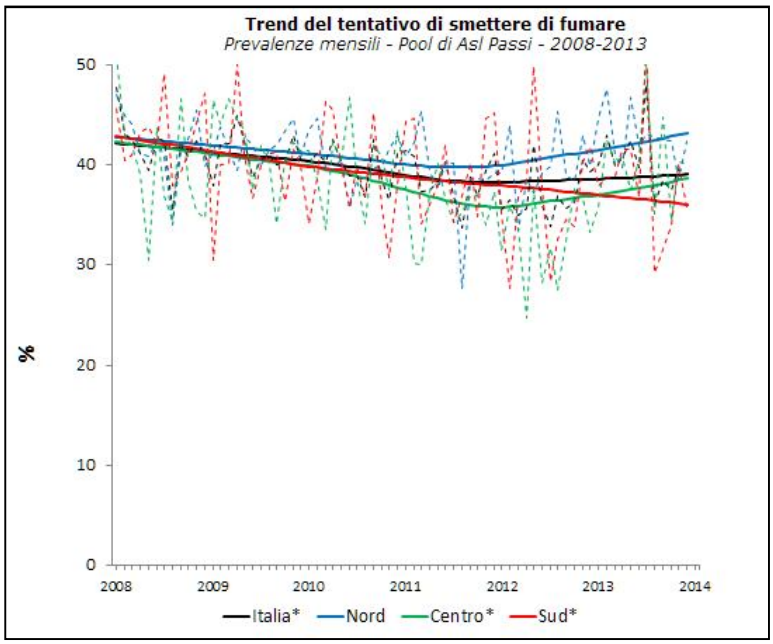


Figura 4.8: Trend dei tentativi di smettere di fumare dal 2008 al 2014 in Italia

Fonte: sorveglianza Passi

La stessa inversione nei **tentativi di smettere di fumare** si legge nel **Rapporto ISS-DOXA** 2013, seppure le percentuali differiscano (vedi Figura 4.9).

Nel 2013 sono aumentati i tentativi di smettere

	2008	2009	2010	2011	2012	2013
Fumatori che hanno tentato di smettere	29,8%	27,5%	27,1%	26,7%	23,0%	30,2%
Fumatori che hanno smesso per gg/sett/mesi	57,3%	66,7%	69,8%	72,1%	65,0%	69,2%
Fumatori che hanno smesso per anni	23,1%	18,6%	20,1%	13,4%	18,4%	19,2%
Fumatori che hanno ridotto il consumo	8,3%	5,7%	2,9%	6,7%	9.2%	9,2%
Nessun risultato	-	9,0%	7,1%	7,7%	7,4%	2,4%

Figura 4.9: Percentuale sui fumatori dei tentativi di smettere di fumare dal 2008 al 2013 in Italia
Fonte: Rapporto ISS-DOXA 2013

Nella versione pubblicata nel 2014 viene riportato il dato secondo cui una piccola percentuale (9,4%) degli attuali fumatori sta pensando seriamente di smettere nei prossimi sei mesi, di cui solo il 22,3% entro 30 giorni dall'intervista.
Quelli che invece hanno fatto almeno un tentativo di smettere sono stati solo il 28,4%, tra questi solo il 15,2% ha provato con la sigaretta elettronica, mentre tra quelli che sono ora ex fumatori nessuno riporta il dispositivo elettronico come strumento utilizzato; la maggioranza (86,9%) è riuscita senza alcun supporto (vedi Figura 4.10).

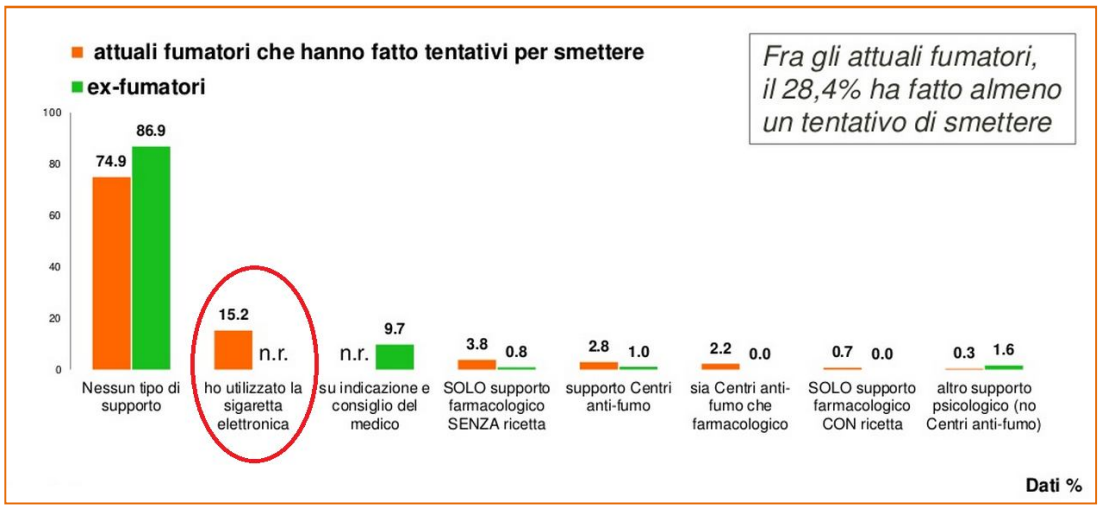


Figura 4.10: Motivi per non provare a smettere di fumare nel 2012 in Italia
Fonte: Rapporto ISS-DOXA 2012

Invece, secondo i dati Passi preliminari forniti esclusivamente a fini di questo lavoro, il 10% degli intervistati che nel primo quadrimestre 2014 ha cercato di smettere di fumare l'ha fatto utilizzando le sigarette elettroniche come aiuto, mentre l'80% ha provato da solo senza aiuto medico.
Tra gli ex fumatori solo l'1% però dichiara di esserci riuscito grazie alle e-cig, percentuale decisamente bassa.

Per quanto riguarda il fenomeno del **fumo elettronico** secondo i dati **Passi** provvisori, relativi al primo quadrimestre del 2014, il 2,2% degli intervistati (133/6030) ha dichiarato di far uso di sigaretta elettronica. Di questi il 62% consumava la variante di e-cig contenente nicotina, il 14% senza ed il 23% alternativamente una e l'altra.
Sono in maggioranza uomini gli utilizzatori di sigarette elettroniche (62%), e soggetti appartenenti alla fascia d'età 35-69 anni (76%). Si svapa di più al nord (51%) ed al centro Italia (32%).
L'80% degli svapatori è anche un fumatore.

I risultati delle **indagini ISS-DOXA** differiscono leggermente: ad utilizzare la **sigaretta elettronica** abitualmente è lo 0,5% ed occasionalmente l'1,1%. Rispetto al 2013 c'è stato un calo degli utilizzatori di 2,6 punti in percentuale (vedi Figura 4.11).

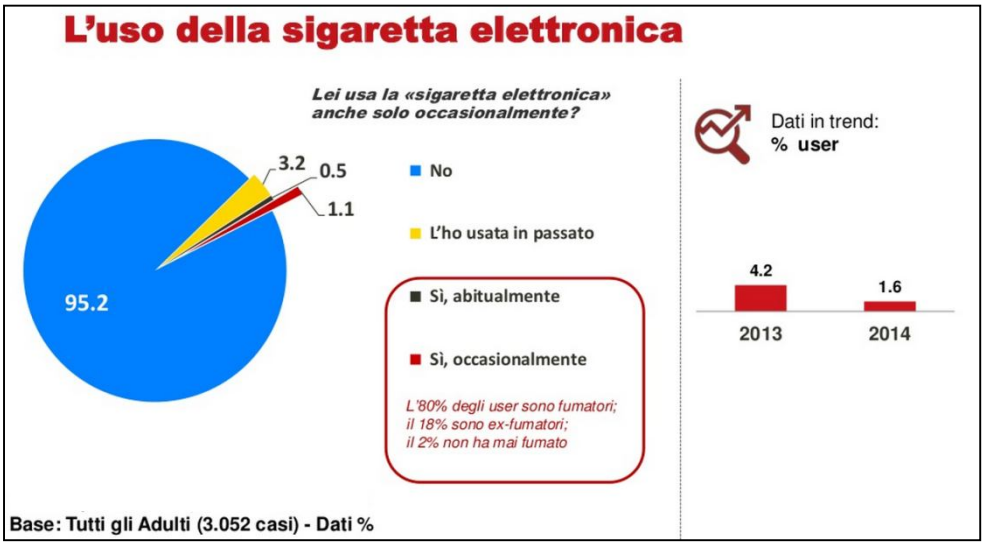


*Figura 4.11:*L'uso della sigaretta elettronica in Italia nel 2014
Fonte: Rapporto ISS-DOXA 2014

I dati che seguono sono il frutto delle risposte dei 50 intervistati che si sono dichiarati svapatori, quindi la base per le seguenti riflessioni è un campione particolarmente ridotto.
Tra gli svapatori l'80% fuma anche prodotti tradizionali contenenti tabacco (esattamente come emerge dalla sorveglianza di Passi), il 18% sono ex fumatori e il 2% sono soggetti che prima non avevano mai fumato.
Il 66% degli svapatori sono uomini, il 40% appartiene alla fascia d'età 25-44 anni, mentre il 44% a quella tra i 45 ed i 64 anni, con una media degli utilizzatori di 42 anni.
La tipologia maggiormente acquistata è quella contenente nicotina (66,2%).
Il 97,9% di coloro che la usavano ed hanno smesso (3,2% degli intervistati) sono stati svapatori per meno di sei mesi (il 44,7% addirittura per meno di un mese), a testimonianza che il loro utilizzo è stato dettato solo dalla curiosità generata dal boom delle sigarette elettroniche che però in breve tempo hanno disatteso le aspettative create.
Coloro che invece al momento dell'intervista si dichiaravano svapatori, utilizzavano una sigaretta elettronica da più di un anno per il 13,7%, mentre la stavano praticamente ancora testando perché comprata da sei mesi o meno per ben il 56,9% (di cui il 14,3% la possedeva da meno di un mese).
La sigaretta elettronica non ha cambiato le abitudini al fumo del 25% degli svapatori intervistati Il 12% degli utilizzatori di e-cigarette dichiara che prima di provarle addirittura non fumava; il 18,8% ha invece smesso di fumare grazie ad esse. Solo l'1,7% fuma un numero maggiore di quelle tradizionali, mentre il 41,8% ne ha diminuito il consumo. Si ricorda di tener presente che per queste affermazioni è stato adottato un campione caratterizzato da una bassa numerosità.
Tra gli intervistati, compresi i non fumatori e non svapatori, il 35,2% considera la sigaretta elettronica un buon metodo per smettere di fumare, percentuale che dal 2013 è diminuita di ben 9,8 punti in percentuale.

Per quanto riguarda le **vendite di e-cig**, secondo la ricerca "*Up in smoke: changes sweep the tabacco industry*" di Bloomerang Industries del 2014, le sigarette elettroniche generano un fatturato mondiale di più di tre miliardi di dollari e la crescita è stata

stimata fino a 225 miliardi di dollari nei prossimi anni, tanto da far affermare che in dieci anni il fumo elettronico supererà quello tradizionale.

Riguardo invece al mercato delle e-cigarette in Italia si riportano i dati proposti dall'Associazione Nazionale Fumo Elettronico (ANAFE): nel 2010 si contavano 750.000 e-cigarette vendute; nel 2011 il numero è più che raddoppiato per arrivando fino a 2,5 milioni, per raggiungere nel 2012 la quota di 3,5 milioni. Il fatturato del settore nel 2012 è stato di ben 350.000.000 euro, con 1500 punti vendita nel 2012, che nel primo semestre 2013 hanno raggiunto la quota dei 4.500, impegnando circa 6.800 persone. Il mercato era quindi molto promettente ed in forte espansione (nel 2012 solo il 72% della popolazione conosceva le e-cig, mentre nel 2013, dopo l'esplosione del fenomeno più del 90% degli italiani aveva avuto informazioni a riguardo) tanto che il rapporto annuale del 2013 dell'Osservatorio Nazionale su Fumo Alcol e Droga pronosticava che in 10 anni il consumo di e-cig avrebbe superato quello delle sigarette tradizionali. Questa prospettiva probabilmente sarà disattesa a causa delle normative introdotte che hanno limitato la produzione facendo registrare al settore un blocco nella crescita che aveva caratterizzato egli ultimi anni (vedi Tabella 4.1 e 4.2).

2012			**2013**		
III trimestre	**IV trimestre**	**I trimestre**	**II trimestre**	**Luglio**	**Agosto**
104	504	1198	1822	1806	1807

Tabella 4.1: Numero imprese nelle distribuzione di sigarette elettroniche dal terzo trimestre 2012 ad Agosto 2013 in Italia

Fonte: Stima Osservatorio Confesercenti

2012			**2013**		
III trimestre	**IV trimestre**	**I trimestre**	**II trimestre**	**Luglio**	**Agosto**
260	1260	2295	4555	4515	4518

Tabella 4.2: Numero punti vendita di sigarette elettroniche dal terzo trimestre 2012 ad Agosto 2013 in Italia

Fonte: Stima Osservatorio Confesercenti

Così commentava i dati dell'Osservatorio Confesercenti sul Commercio il presidente di Fiesel: "La frenata è in parte dovuta certamente a fattori fisiologici di mercato; ma dall'altra ha inciso la campagna di disinformazione condotta da alcuni media, che adduceva alle svapo presunti 'effetti dannosi' non solo mai dimostrati. Ma il vero problema è che si respira un'aria di attesa e timore per il futuro".
A Giugno 2013 le vendite sono calate del 2,4%, e di conseguenza vi è stata anche una diminuzione delle richieste di apertura pari al 99% tra Gennaio 2014 e Giugno 2014: da 200 richieste e 370 nuove aperture in Aprile, a 2 richieste e 123 chiusure a Giugno.
Mancini, il direttore dell'associazione di settore Anafe, indica come motivazioni del calo le polemiche riguardanti la nuova legge, l'insicurezza dei produttori sul regime a cui sarebbero stati sottoposti, la fine del boom iniziale che prima o poi doveva frenare, e anche la pubblicità negativa da parte delle lobby del tabacco e farmaceutiche che vedono nel fumo elettronico un rivale per i prodotti proposti come strumenti per smettere di fumare come cerotti e inalatori.

4.1.3 Riflessioni mediche sulle sigarette elettroniche come mezzo per smettere di fumare

Grazie alle sigarette elettroniche teoricamente sembrerebbe possibile diminuire la quantità di tabacco consumata e limitare il vizio alla nicotina senza rinunciare al vizio del fumo. Inoltre, dato che lo svapatore può gestire in completa autonomia quanti tiri fare ogni volta che utilizza la sua e-cig, può controllare più facilmente anche quanta nicotina consuma perché quando vuole togliersi la voglia di fumare poi non si trova costretto a finire la sigaretta che si è acceso per non gettarla via senza averla consumata del tutto.
Si potrebbe perciò essere indotti ad affermare che un aumento dell'uso delle e-cigarette possa portare una riduzione nel numero dei fumatori. Al momento però non vi sono prove mediche sufficienti e sono in corso studi medici a sostegno e contro questa tesi; anche l'Organizzazione mondiale della Sanità nel novembre 2010 ha dichiarato che per ora non ci sono prove sulla loro efficacia in questo senso, mentre ce ne sono abbastanza che rassicurano sulla loro non tossicità e sulla non pericolosità del vapore passivo

emesso che non incide sulla qualità dell'aria di un ambiente chiuso (vedi ad esempio Schober et al., 2014).

La sigaretta elettronica viene presentata dai suoi sostenitori sia come prodotto più salutare per i fumatori che come valido strumento per aiutare a smettere di fumare che potrebbe portare nelle migliori delle ipotesi alla scomparsa dei prodotti che prevedono la combustione di tabacco e che creano gravi patologie e numerosi decessi.

A contrapporsi a quest'idilliaca visione del fumo elettronico ci sono coloro i quali intravedono invece la minaccia di un aumento di dipendenti dalla nicotina tra coloro che senza le e-cigarette non avrebbero mai fumato, e la possibilità di disincentivare la totale astinenza dalla nicotina tra coloro che vogliono smettere di fumare succedanei del tabacco e che potrebbero ottenere se non continuassero poi a consumare sigarette elettroniche (Popova e Ling, 2013; Doyle, Patterson e Scott, 2014; Heavner et al., 2009).

Sembra, dunque, che le opinioni sul tema siano molte e le certezze poche.

Gli studi a sostegno degli effetti benefici e dell'aiuto che le sigarette elettroniche possono dare a chi intende provare a smettere di fumare sono numerosi (vedi Siegel, Tanwar e Wood, 2011; Polosa et al., 2011; Barbeau, Burda e Siegel, 2013; Caponnetto et al., 2013).

Uno di quelli che ha avuto un forte impatto nella discussione riguardante il ruolo delle sigarette elettroniche nel controllo dell'utilizzo tabacco è quello di Bullen e et al. (2013), uno studio pionieristico durante il quale è stato analizzato un campione di 657 fumatori adulti residenti in Nuova Zelanda che volevano smettere di fumare. Ad una parte del campione sono state fornite, come mezzo per raggiungere il loro obiettivo, delle sigarette elettroniche contenenti nicotina, ad un'altra parte e-cig senza nicotina, ed infine ai restanti del gruppo dei i cerotti di nicotina. Risultò che il 57% dei partecipanti al primo gruppo dimezzò il consumo di sigarette con tabacco per sei mesi, confrontati con solo il 41% di quelli del gruppo che utilizzava i cerotti. Lo studio presentava notevoli limitazioni, come ad esempio la mancanza di una supervisione e di un monitoraggio continuo per assicurarsi che il trattamento fosse utilizzato come previsto, fondamentale quando si testa un nuovo trattamento. Il messaggio di fondo era però che le sigarette elettroniche hanno perlomeno la stessa efficacia dei cerotti alla nicotina, e

che anzi risultano essere più attraenti per i fumatori perché ripetono la gestualità del fumare vero tabacco.

Anche le conclusioni dello studio pilota condotto da Wagener et al. (2013) riportavano come in un campione di fumatori che non erano intenzionati a smettere e che non avevano mai provato una sigaretta elettronica, dopo una settimana in cui veniva fornita loro una e-cig da usare come preferivano, la loro volontà di provare a smettere e la loro fiducia nel potercela fare fossero aumentate e come il loro uso di sigarette tradizionali fosse diminuito di circa il 44% rispetto all'inizio dell'esperimento.

In Italia la Lega Italiana Antifumo, in collaborazione con il Centro Prevenzione e Cura del Tabagismo dell'Università di Catania, ha registrato la variazione delle abitudini di 40 fumatori che non volevano smettere dopo aver provato le sigarette elettroniche, constatando che il 32,5% dei 40 partecipanti aveva ridotto di almeno la metà la quantità di sigarette fumate al giorno, dei quali il 12,5% le aveva diminuite di ben l'80%, passando da trenta a tre sigarette. Il 22,5% dei partecipanti invece aveva completamente smesso di consumare sigarette tradizionali.

Popova e Ling (2013) espongono invece la loro scetticità nel presentare le sigarette elettroniche come un valido aiuto all'astinenza dai prodotti derivati dal tabacco; mentre Sutfin et al. (2013), conseguentemente ai risultati di un esperimento condotto su un gruppo di studenti universitari americani, si pronunciano decisamente contrari.

Ecco che allora Doyle, Patterson e Scott (2014) suggeriscono la modalità per sapere la risposta: guardare l'andamento delle vendite delle sigarette elettroniche e compararlo con quelle delle sigarette normali.

Ci sono addirittura studi come quello di Gennimata, S. A. et al. (2013) che concludono affermando "Non sappiamo ancora se i prodotti per la somministrazione di nicotina, come le e-cig, sono più sicuri rispetto alle sigarette normali, nonostante il marketing punti sulla loro minore nocività".

Le opinioni sull'argomento sono quindi varie e discordanti, e non ci sono prove incontrovertibili certe da portare a sostegno di una o dell'altra tesi.

4.2 Analisi di Twitter

Da quanto precedentemente esposto emerge come il dibattito sul ruolo della sigaretta elettronica sia acceso e di elevato interesse non solo per i fumatori, ma anche per il Governo e la Pubblica Sanità.
Si è deciso quindi di mettere alla prova l'analisi dei social network per vedere cosa le persone 'dicono' online relativamente alle e-cig, ed in particolare se le ritengono un mezzo adatto per aiutare a smettere di fumare.

4.2.1 Metodologia adottata per l'estrazione dei tweet

L'analisi delle opinioni espresse online sulle sigarette elettroniche è stata condotta durante il primo quadrimestre 2014. Twitter è stato il social network scelto per il lavoro di ricerca per le motivazioni esposte al paragrafo 3.2.
Dalla piattaforma sociale sono stati estratti i tweet contenenti i seguenti termini: '#sigarettaelettronica', 'sigaretta elettronica', 'sigarette elettroniche', 'svapare', 'svapando', 'svapatore', 'svapatori', 'smettere di fumare'[14], '#smetteredifumare'[15]. Non sono state considerate come *keyword* solo hashtag poiché le statistiche riportano che molti tweet non ne contengono. A sostegno di questa scelta c'è il dato secondo cui nello stesso periodo nel 2013 solo nel 30% dei 'cinguettii' era presente il simbolo # seguito da termini identificativi dell'argomento di cui il messaggio trattava (State of the Net, 2013).
Per ottenere i dati relativi al periodo selezionato è stata creata una Twitter API ufficiale utilizzata con cadenza settimanale, per superare il limite temporale per cui i tweet estraibili non possono essere datati più di 7 giorni (vedi Paragrafo 3.4).
I termini selezionati sono stati considerati idonei sia per avere una panoramica quasi totale dei messaggi pubblicati che facessero riferimento al fenomeno del fumo elettronico, indipendentemente dal modo in cui trattassero il tema, sia per poter cercare di capire quanti di coloro che esponevano pubblicamente la loro intenzione di smettere di fumare facessero in qualche modo riferimento alla sigaretta elettronica.

[14] La raccolta per questa parola chiave è iniziata il 15 gennaio 2014
[15] La raccolta per questa parola chiave è iniziata il 15 gennaio 2014

Al fine di circoscrivere la ricerca alla situazione in Italia sono state prese in considerazione solo *keyword* in italiano. Nonostante le Twitter REST Search API permettano di filtrare i risultati impostando precise latitudini e longitudini dell'autore del tweet, dal momento che è emerso come pochi siano i profili che permettono la geolocalizzazione, si è preferito optare per un'altra modalità di filtraggio. Inserire parole chiave solo in italiano ha il risultato di far comparire nell'output solo messaggi postati da italiani in Italia o da italiani all'estero, buona parte dei quali si presume soggiornerà fuori dalla propria Nazione per breve periodo dato che continuano ad utilizzare la lingua del proprio Paese che permette loro di comunicare solo con italiani; questi quindi rientreranno comunque nelle statistiche ufficiali del proprio Paese redatte dal Ministero della Sanità italiano.

E' stato utilizzato il software R per interfacciarsi con la Twitter REST Search API creata.

L'output dell'operazione di estrazione è un file in formato csv nel quale è possibile leggere il contenuto dei tweet pubblicati resi anonimi ed altre informazioni come l'ora e la data di pubblicazione esatte, il codice ID che identifica l'autore, il numero dei commenti, se è stato inserito tra i favoriti o se è un retweet.

4.2.2 Analisi quantitativa dei tweet estratti

Sono stati estratti un totale di circa 17.000 tweet: 5049 contenevano la *keyword* 'sigaretta elettronica', 6377 'sigarette elettroniche', 795 '#sigarettaelettronica', 624 'svapare', 33 'svapando', 100 'svapatore', 647 'svapatori', 3297 'smettere di fumare'[16], 99 '#smetteredifumare'[17]. Come previsto, dal momento che non tutti i tweet contengono un hashtag che faccia riferimento alla tematica di cui tratta il messaggio, il quantitativo maggiore di post è stato raccolto per le parole chiave senza #.

I totali estratti non sono considerabili esaustivi della totalità dei tweet pubblicati poiché l'interfaccia di ricerca ufficiale, la Twitter API, si focalizza sulla rilevanza ma non sulla completezza (vedi Paragrafo 3.4), e quindi alcuni tweet potrebbero non comparire tra i risultati perché pubblicati eliminando la loro caratteristica di essere pubblici o perché

[16] Dato relativo al periodo 15 gennaio – 30 aprile 2014
[17] Dato relativo al periodo 15 gennaio – 30 aprile 2014

non resi disponibili dallo strumento ufficiale. Nonostante questa limitazione importante il dataset ottenuto può comunque ritenersi rappresentativo dell'utenza di Twitter, poiché il confronto con il numero totale di tutti i tweet pubblicati contenenti quelle parole chiave fornito da Topsy, unico dato che è disponibile gratuitamente nel sito www.topsy.com, conferma che i messaggi mancanti sono molto pochi. Inoltre, dato che le API tendono a eliminare dai risultati quei 'cinguettii' considerati meno rilevanti, la perdita d'informazione risulta ancora più ridotta.

Il limite del massimo di 1500 messaggi estraibili (vedi paragrafo 3.4) non ha causato alcun ostacolo dato che il totale dei tweet estratti per volta per ogni parola chiave era minore di 1500; ciò non ha quindi causato l'ulteriore perdita di alcun contenuto pubblicato nella piattaforma.

La media giornalieria di messaggi contenenti le *keyword* 'sigaretta elettronica' e 'sigarette elettroniche' è stata rispettivamente di 42 e 53, mentre per l'hashtag #sigarettaelettronica meno di 7 (vedi Tabella 4.3).

Parola chiave	**#sigarettaelettronica**	**Sigaretta elettronica**	**Sigarette elettroniche**	**Svapare**	**Svapando**
Totale tweet	**795**	**5049**	**6377**	**624**	**33**
Media giornaliera	**6,6**	**42,1**	**53,1**	**5,2**	**0.3**

Parola Chiave	**Svapatore**	**Svapatori**	**Smettere di fumare**	**#smetteredifumare**
Totale tweet	**100**	**47**	**3297** [18]	**99** [19]
Media giornaliera	**0.8**	**5,4**	**31,4** [20,21]	**0,9** [22]

Tabella 4.3: Totale e media giornaliera dei tweet estratti per le *keyword* selezionate nel primo quadrimestre 2014

Fonte: elaborazione personale

[18] Dato relativo al periodo 15 gennaio – 30 aprile 2014
[19] Dato relativo al periodo 15 gennaio – 30 aprile 2014
[20] Dato che maggiormente si differenzia da quello fornito da www.topsy.com secondo cui è 44,1
[21] Dato relativo al periodo 15 gennaio – 30 aprile 2014
[22] Dato relativo al periodo 15 gennaio – 30 aprile 2014

Basandosi sullo studio di Blogmeter secondo cui il primo quadrimestre 2013 i tweet in Italia al giorno era stati in media 653.515, considerando che nel 2014 era stata prevista una crescita globale del 14% del numero degli utenti, si può stimare che nel primo quadrimestre 2014 la media giornaliera sarà stata approssimativamente di 745.000.
Partendo da questa stima approssimativa si può affermare che ad aver trattato il tema del fumo elettronico utilizzando le parole chiave selezionate è stato lo 0,015% dei tweet giornalieri.
L'andamento del numero di tweet contenenti i termini selezionati è riportato nei grafici che seguono (vedi Figura 4.12 – 4.19); la fonte dei dati in essi riportati è Topsy che, essendo una delle quattro aziende a cui Twitter rivende i dati, ha a disposizione il *firehose* completo, non soggetto quindi alle limitazioni che invece colpiscono le Twitter API disponibili. Solo l'andamento durante l'ultimo mese dei tweet con determinate *keyword* è disponibile gratuitamente nel sito www.topsy.com, e ciò è stato sfruttato per riportare dei grafici che mostrassero il trend completo senza alcun dato mancante, caso che utillizando i dati estratti attraverso la API ufficiale era possibile.
L'andamento delle dei tweet contenenti le parole chiave selezionate come inerenti al tema del fumo elettronico e l'hashtag #smetteredifumare registra un picco proprio nel periodo 28 marzo – 4 Aprile. La motivazione è la campagna #sVapevatelo, promossa da Anafe, Fiesel e Life contro il regime fiscale sospeso dal Tar del Lazio il 2 Aprile 2014, descritta al paragrafo 4.1.1. Nei messaggi riportanti l'hashtag #Svapevatelo (il cui andamento è riportato in Figura 4.20) sia il tema della tassa del 58,5% considerata una minaccia per la sopravvivenza delle imprese del settore, sia quello del potenziale aiuto che le sigarette elettroniche possono fornire a chi vuole smettere di fumare sono presi in considerazione.

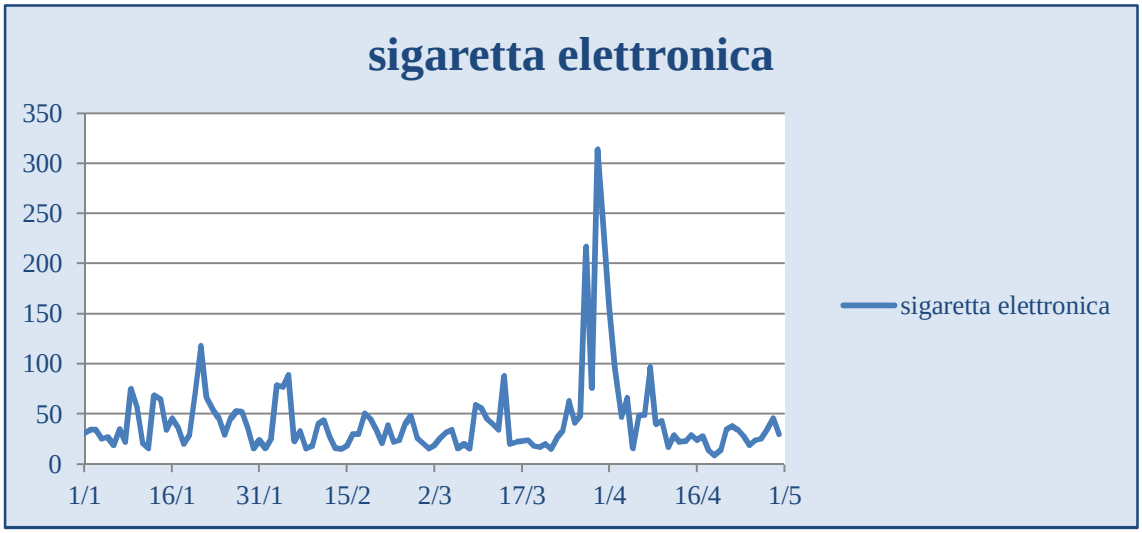


Figura 4.12: Andamento del numero dei tweet contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014

Fonte: Topsy, 2014

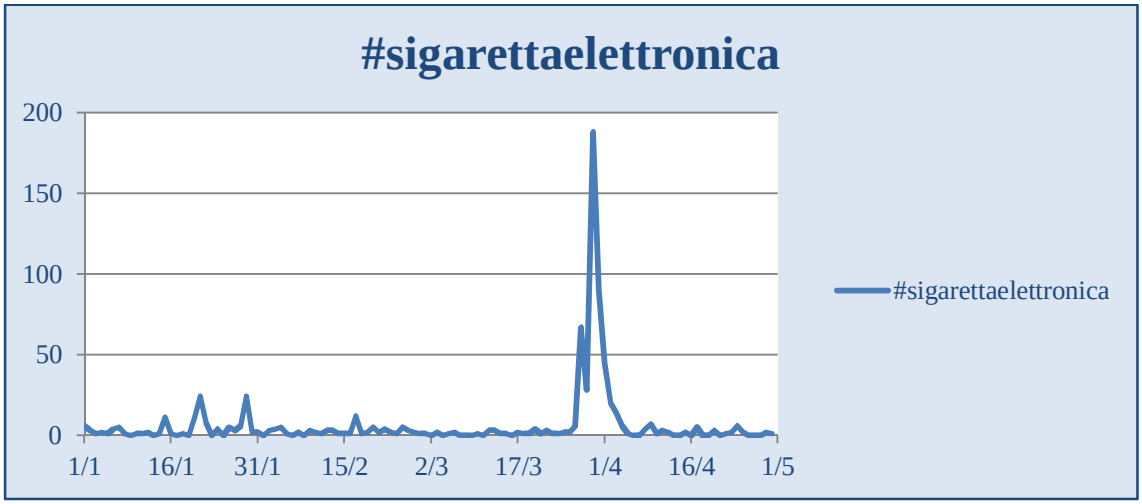


Figura 4.13: Andamento del numero dei tweet contenenti l'hashtag #sigarettaelettronica nel primo quadrimestre 2014

Fonte: Topsy, 2014

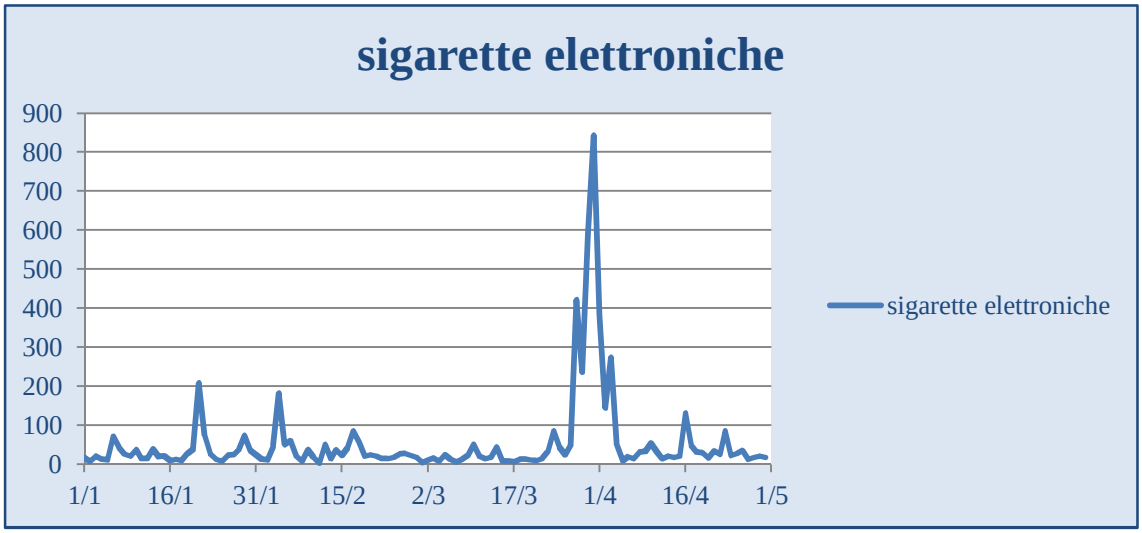


Figura 4.14: Andamento del numero dei tweet contenenti la parola chiave 'sigarette elettroniche' nel primo quadrimestre 2014

Fonte: Topsy, 2014

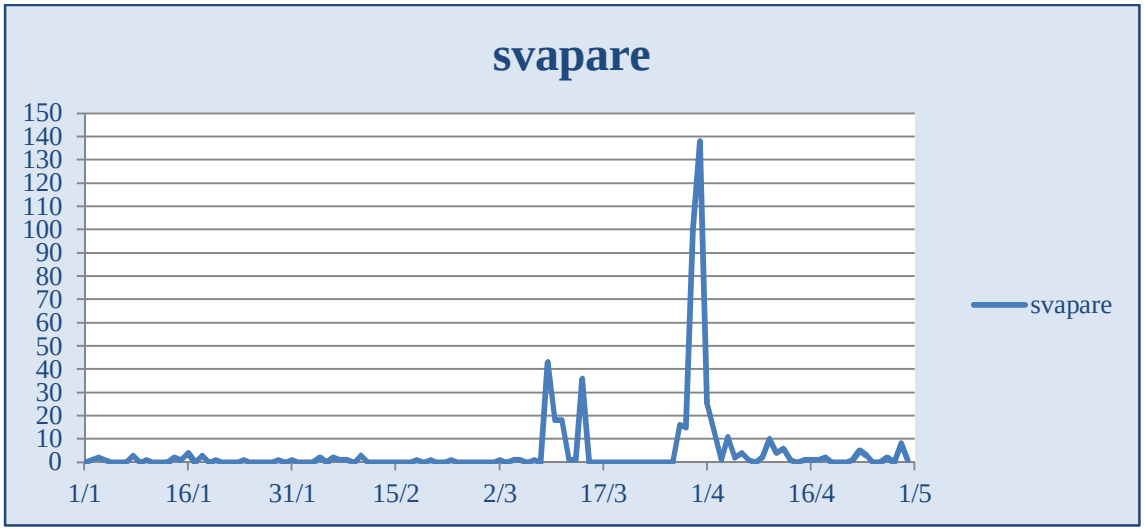


Figura 4.15: Andamento del numero dei tweet contenenti la parola chiave 'svapare' nel primo quadrimestre 2014

Fonte: Topsy, 2014

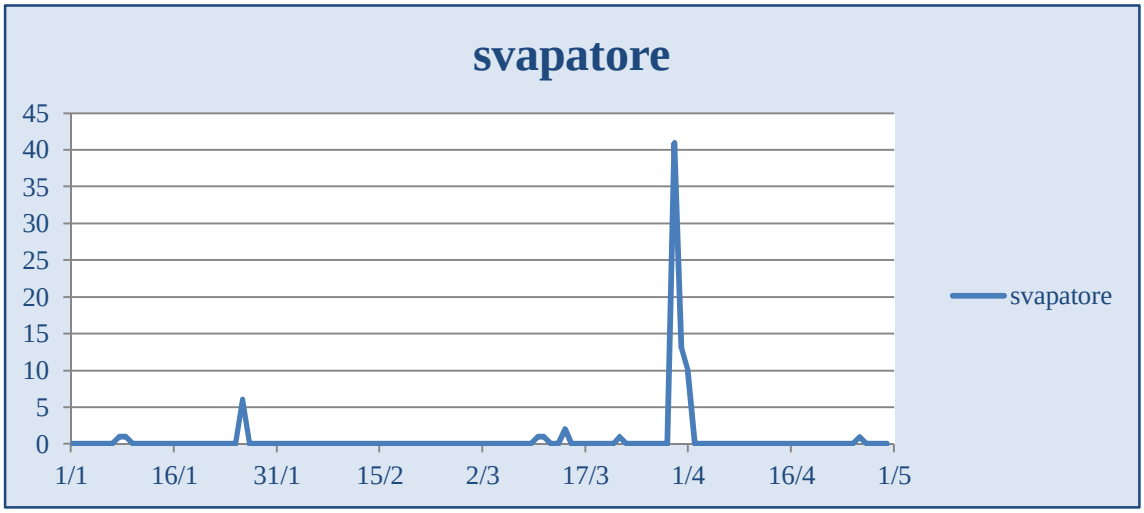


Figura 4.16: Andamento del numero dei tweet contenenti la parola chiave 'svapatore' nel primo quadrimestre 2014

Fonte: Topsy, 2014

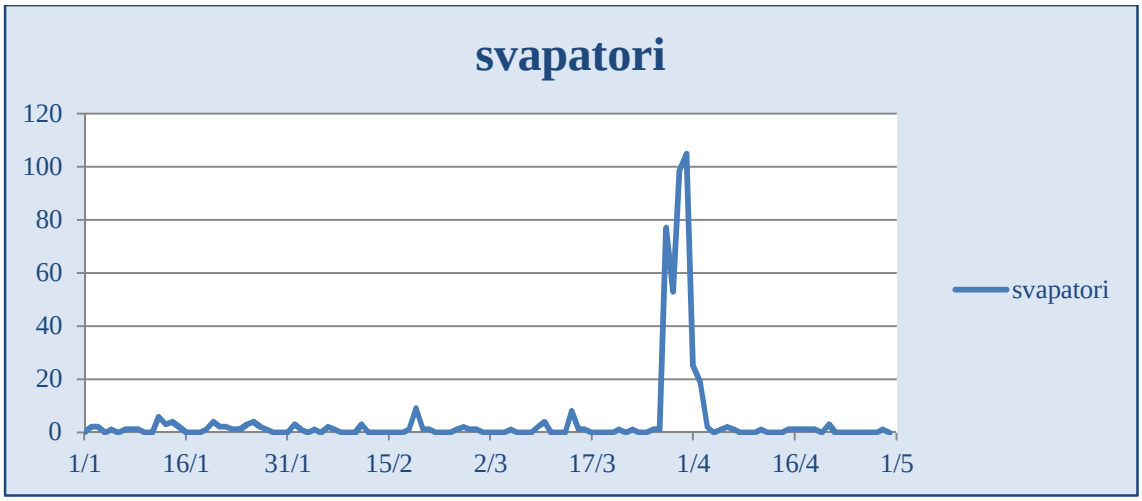


Figura 4.17: Andamento del numero dei tweet contenenti la parola chiave 'svapatori' nel primo quadrimestre 2014

Fonte: Topsy, 2014

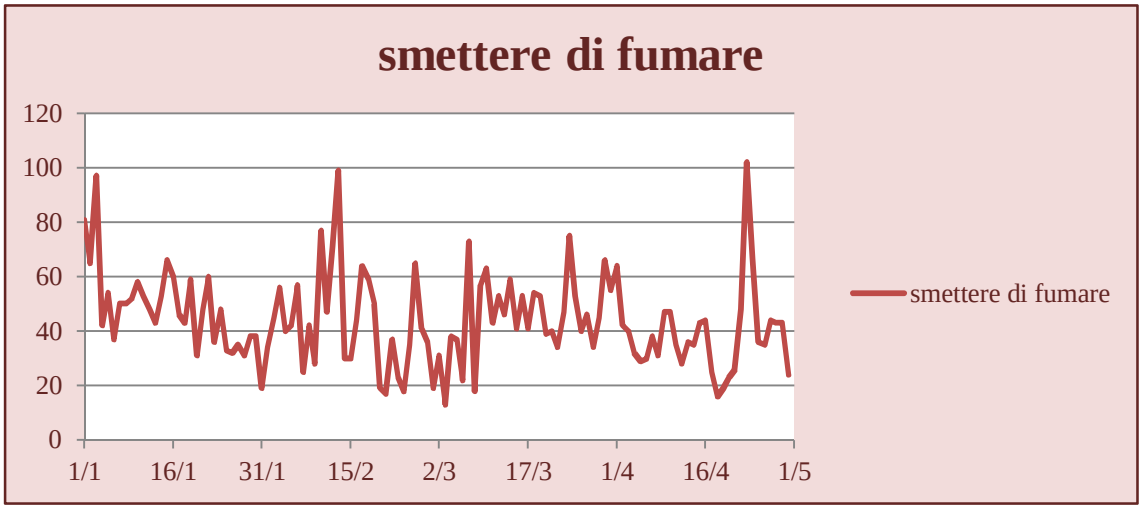


Figura 4.18: Andamento del numero dei tweet contenenti la parola chiave 'smettere di fumare' nel primo quadrimestre 2014

Fonte: Topsy, 2014

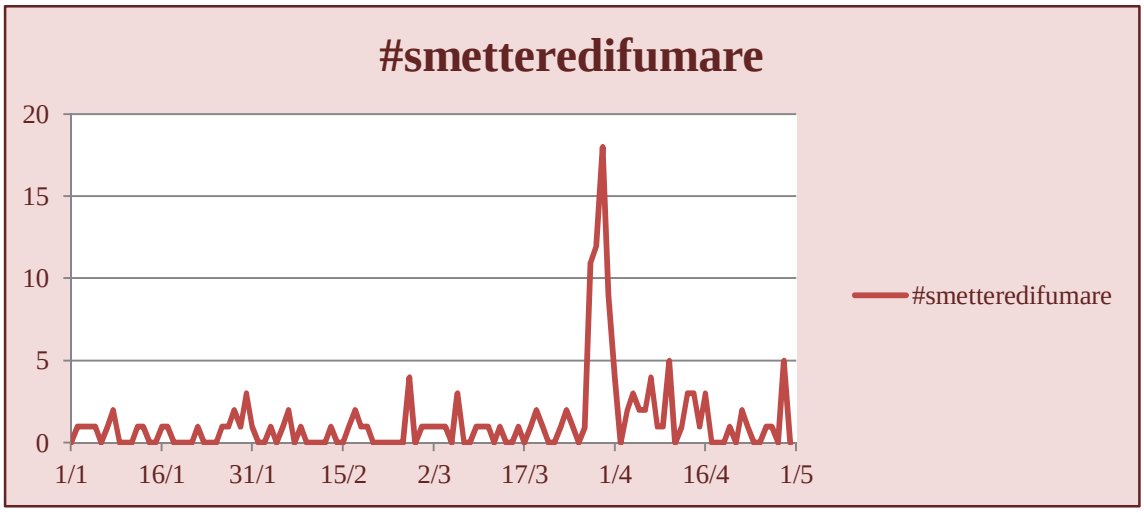


Figura 4.19: Andamento del numero dei tweet contenenti l'hashtag #smetteredifumare nel primo quadrimestre 2014

Fonte: Topsy, 2014

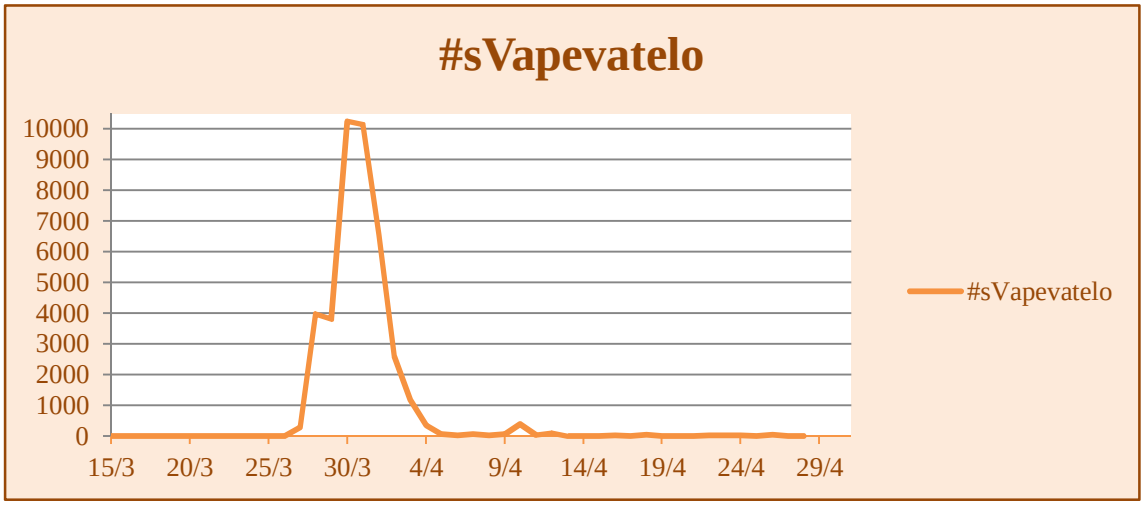


Figura 4.20: Andamento del numero dei tweet contenenti l'hashtag #sVapevatelo nel primo quadrimestre 2014

Fonte: Topsy, 2014

Oltre al picco principale appena esposto, ve ne sono altri, d'intensità minore, anch'essi collegati ad avvenimenti del mondo reale:

1. il 21 Gennaio è il giorno in cui il giudice amministrativo del Tar ha sospeso il regime autorizzativo previsto per la registrazione e la certificazione presso i Monopoli; in quella data sono stati registrati 118 tweet per 'sigaretta elettronica', 208 per 'sigarette elettroniche' e 24 per l'hashtag #sigarettaelettronica contro la media giornaliera rispettivamente di 42, 53 e 6,6;
2. l'aumento dei tweet tra il 3 ed il 5 Febbraio per le *keyword* 'sigaretta elettronica'(fino ad 89 quando la media giornaliera è di 42) e 'sigarette elettroniche'(con un massimo di 182 quando la media giornaliera 53) è spiegabile con l'introduzione nel paniere ISTAT della sigaretta elettronica il 4 Febbraio;
3. il 3 aprile il Tar decide di bloccare la tassa del 58,8% a cui era soggetta la vendita di e-cig, ecco che i tweet contenenti l'hashtag #sigarettaelettronica e che nominano 'sigaretta elettronica', 'sigarette elettroniche e 'svapare' mantengono un livello decisamente superiore alla media in quella giornata ed anche il giorno seguente;

4. l'8 aprile la notizia dell'esplode di una sigaretta elettronica di una barista in Inghilterra gira per il web e in Twitter i messaggi che contengono l'hashtag #sigarettaelettronica e lo stesso termine senza # crescono;
5. il 10 aprile un inaspettato aumento dei post contenenti il verbo 'svapare' si potrebbe spiegare con la diffusione dello studio di Park et al. (2014) presentato all'incontro annuale dell'*American Association for Cancer Research* e pubblicato proprio il 10 aprile sul settimanale scientifico Nature, secondo cui svapare nicotina ha lo stesso effetto cancerogeno del fumo tradizionale;
6. il 16 aprile un articolo viene pubblicato sul Los Angeles Times[23] in cui alcuni politici statunitensi accusano le aziende produttrici di e-cig di marketing aggressivo nei confronti degli adolescenti, il cui consumo è in netto aumento. La notizia, riportata anche in Italia, probabilmente spiega i 130 tweet contenenti la *keyword* 'sigarette elettroniche', contro una media giornaliera di 53. Lo stesso avviene il 24 marzo in cui in seguito ad un articolo del New York Times[24] che trattava del crescente numero di casi di avvelenamento causati da e-cigarette negli Usa, salgono ad 84;
7. il 14 marzo il primo caso di polmonite causata da sigaretta elettronica avvenuto in Spagna probabilmente è la motivazione del picco per i corpi di tweet estratti contenenti le parole chiave 'sigaretta elettronica', 'svpare' e 'svapatori'.

Per meglio identificare visivamente questi picchi d'intensità notevolmente minore a quello causato dalla campagna #sVapevatelo di cui si è in precedenza parlato, sono stati creati dei grafici nei quali il numero dei tweet dal 28 marzo al 2 aprile è stato ridotto a zero (vedi Figura 4.21 – 4.27).

[23] Clozel, L., 2014. "E-cigarette firms targeting young people, lawmakers say". *Los Angeles Times*

[24] Richtel, M., 2014. "Selling a Poison by the Barrel: Liquid Nicotine for E-Cigarettes". *New York Times* [online]

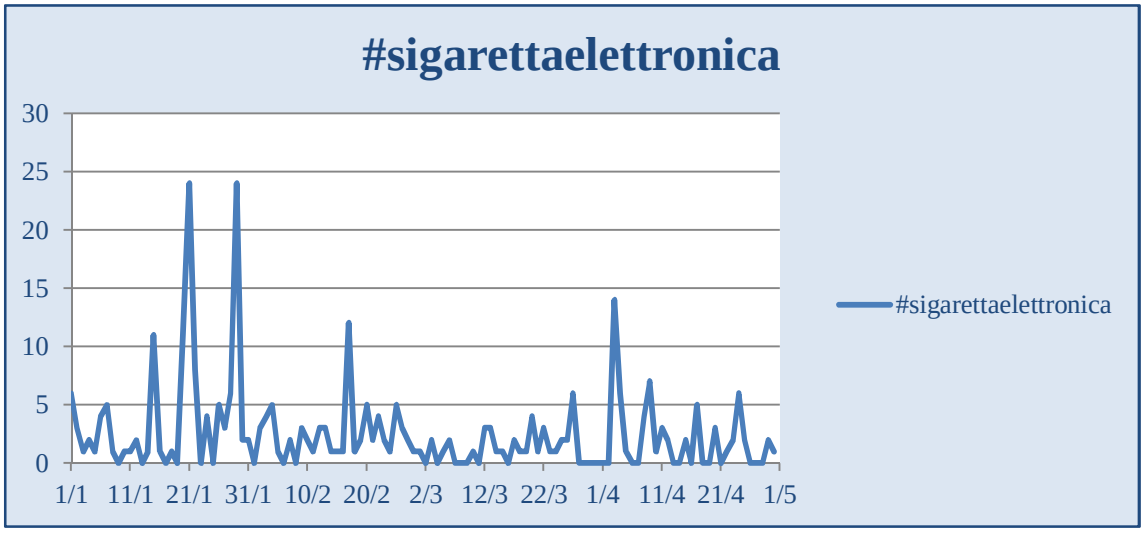


Figura 4.21: Andamento del numero dei tweet contenenti l'hashtag #sigarettaelettronica nel primo quadrimestre 2014 in cui quelli dal 28 marzo al 2 aprile sono stati eliminati per annullare il picco generato dalla campagna #sVapevatelo

Fonte: Topsy, 2014

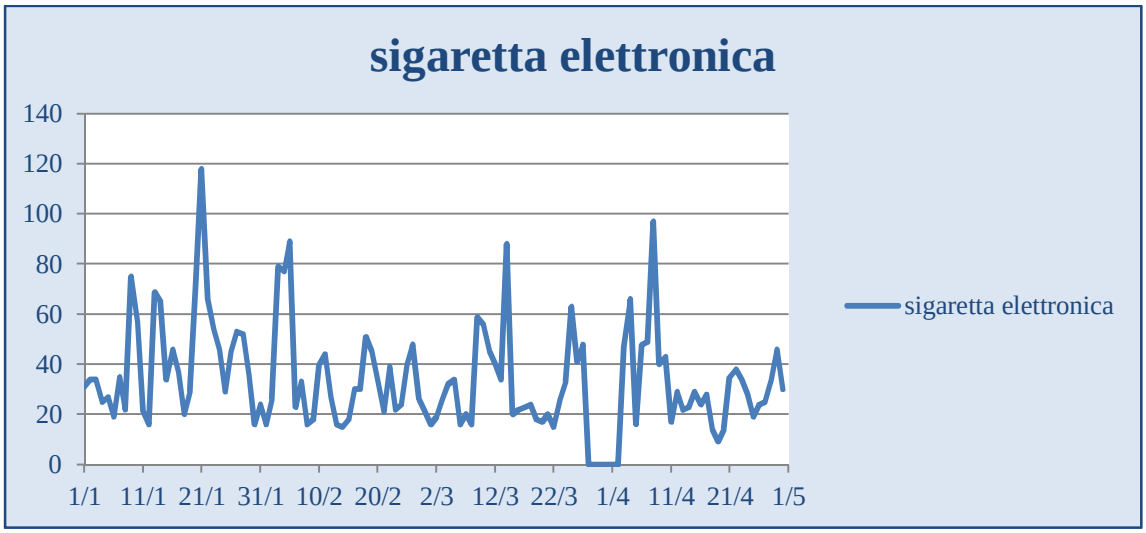


Figura 4.22: Andamento del numero dei tweet contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 in cui quelli dal 28 marzo al 2 aprile sono stati eliminati per annullare il picco generato dalla campagna #sVapevatelo

Fonte: Topsy, 2014

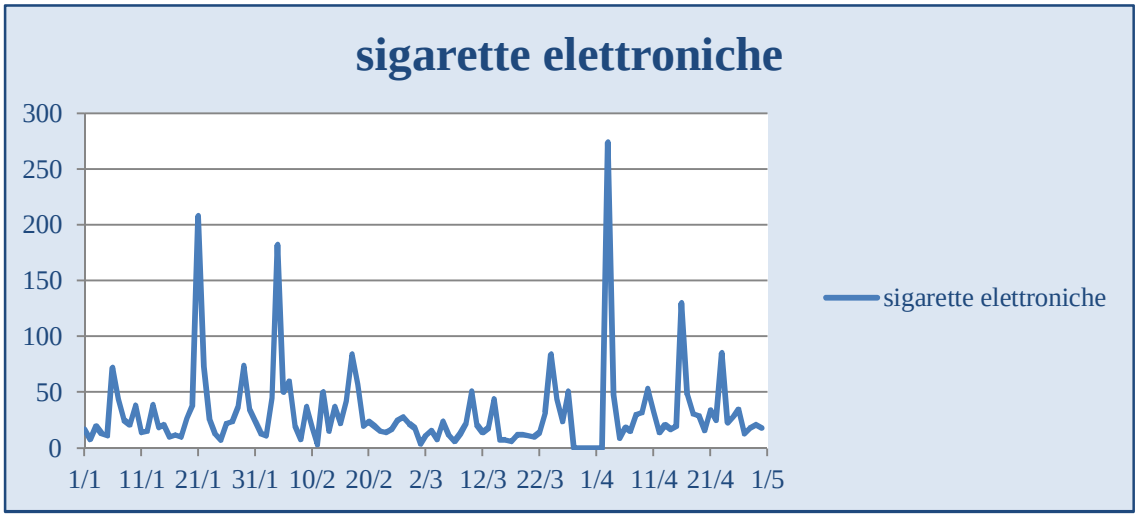


Figura 4.23: Andamento del numero dei tweet contenenti la parola chiave 'sigarette elettroniche' nel primo quadrimestre 2014 in cui quelli dal 28 marzo al 2 aprile sono stati eliminati per annullare il picco generato dalla campagna #sVapevatelo

Fonte: Topsy, 2014

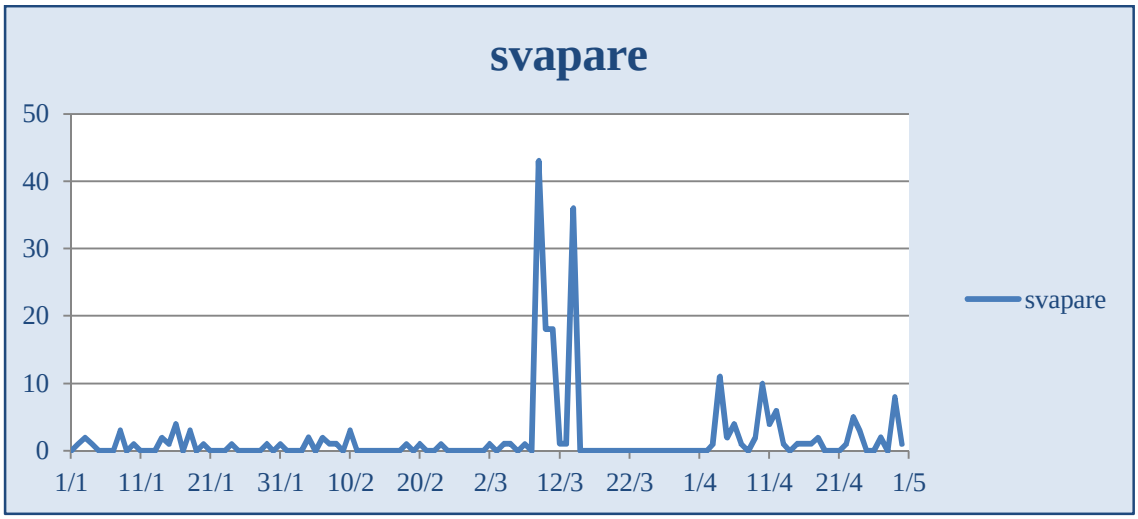


Figura 4.24: Andamento del numero dei tweet contenenti la parola chiave 'svapare' nel primo quadrimestre 2014 in cui quelli dal 28 marzo al 2 aprile sono stati eliminati per annullare il picco generato dalla campagna #sVapevatelo

Fonte: Topsy, 2014

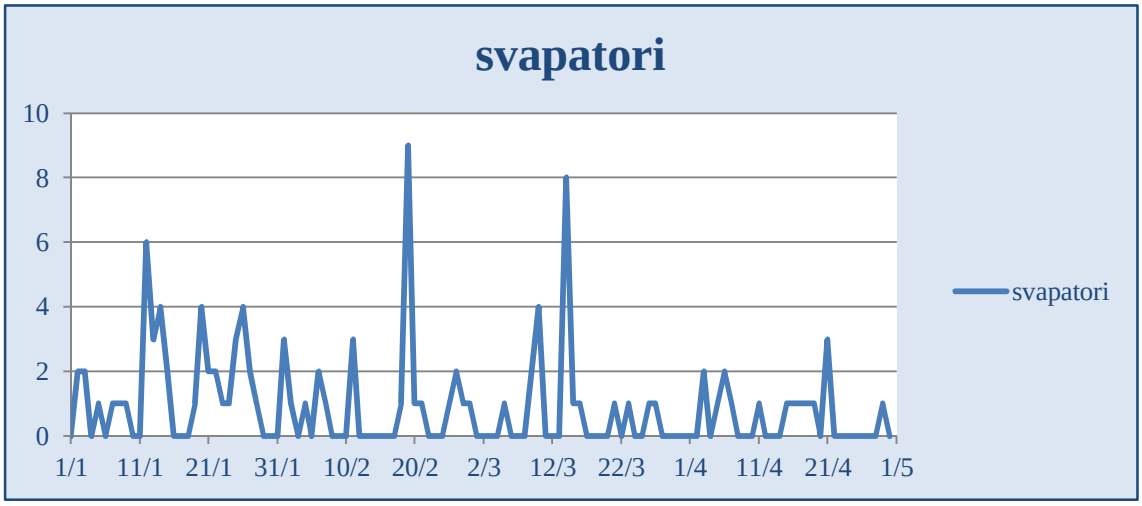


Figura 4.25: Andamento del numero dei tweet contenenti la parola chiave 'svapatori' nel primo quadrimestre 2014 in cui quelli dal 28 marzo al 2 aprile sono stati eliminati per annullare il picco generato dalla campagna #sVapevatelo
Fonte: Topsy, 2014

Gli andamenti dei messaggi pubblicati in Twitter sul fumo elettronico risentono quindi degli eventi del mondo reale, ma non sono per nulla simili a quelli relativi ai tentativi di smettere di fumare, come dimostra anche il confronto offerto da Topsy per il mese di Aprile (vedi Figura 4.26).

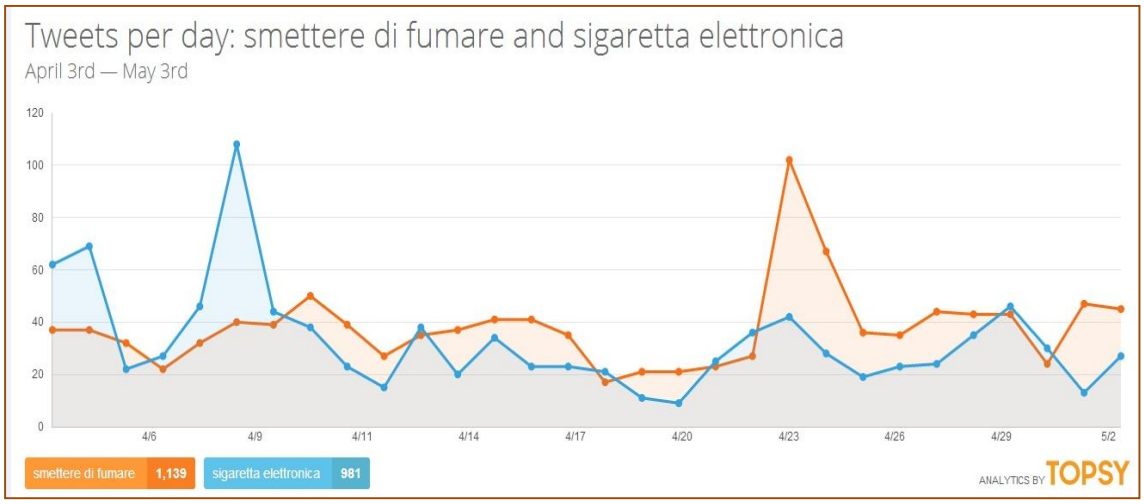


Figura 4.26: Confronto andamento dei tweet contenenti le parole chiave 'sigaretta elettronica' e 'smettere di fumare' dal 3 Aprile al 3 Maggio 2014
Fonte: Topsy

4.2.3 Analisi linguistiche del contenuto dei tweet estratti

Già dalla semplice analisi quantitativa sembra che i due temi oggetto dell'indagine non vengano collegati nelle discussioni in Twitter, si è però voluto approfondire la ricerca procedendo con un'analisi del contenuto dei testi.

4.2.3.1 Word cloud

Si è iniziato con la tecnica di *text summarization* più comune: il *word cloud*, il quale mette in evidenza le parole che con la maggior frequenza compaiono nel testo analizzato; più un termine appare grande e maggiore è la sua frequenza. Di seguito sono riportati i *word cloud* ottenuti dai tweet estratti; i termini legati al campo salute e ai tentativi di smettere di fumare sono stati cerchiati.
In Figura 4.27 si vede come in quello ottenuto dal corpo di tweet contenenti l'hashtag #sigarettaelettronica il tema maggiormente trattato sia stata la normativa sulle e-cigarette: governorenzi, bealorenzin, tassa, bocciata, news e lobby sono vocaboli che rimandano al mondo della politica. Il secondo tema che emerge è quello della salute attraverso le parole salute, scienza, tabagismo. Com'era prevedibile in base a quanto precedentemente esposto, l'hashtag #sVapevatelo appare in tutti i corpi di tweet collegati al fumo elettronico; lo stesso avviene per anafecig, che essendo l'account dell'Associazione Nazionale Fumo Elettronico è l'autore di numerosi tweet inerenti la tematica.

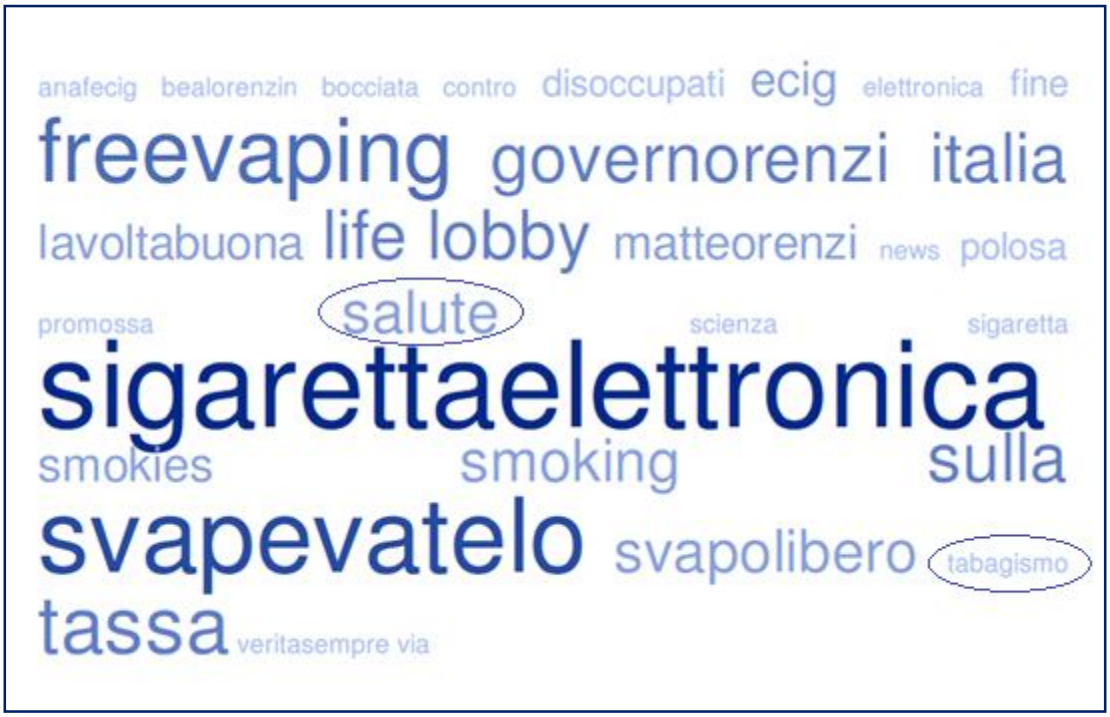


Figura 4.27: Word cloud ottenuto dal corpo di tweet contenenti l'hashtag #sigarettaelettronica estratti durante il primo quadrimestre 2014 con evidenziati termini collegabili al tema della salute

Fonte: elaborazione personale

Relativamente ai tweet estratti per la *keyword* 'sigaretta elettronica' vi sono termini di maggior frequenza collegabili ai cambi legislativi (lorenzin, governo, tax, governorenzi, tassazione, tasse, Sposetti, Lazio (in riferimento al Tar del Lazio), ricari, lobby, stato), in misura minore alla salute (tabacco, Veronesi, smettere, salute, dannosa), ed infine agli eventi di cronaca del periodo (barista, esplode, Inghilterra, paniere). Emergono anche vocaboli come 'ragazzini' e 'nuova', quest'ultimo forse inserito in messaggi pubblicitari pubblicati da negozi di sigarette elettroniche. Il termine 'ragazzini' invece si è riscontrato apparire in molti messaggi ironici facenti riferimento all'uso da parte dei più giovani della e-cig (vedi analisi successive).

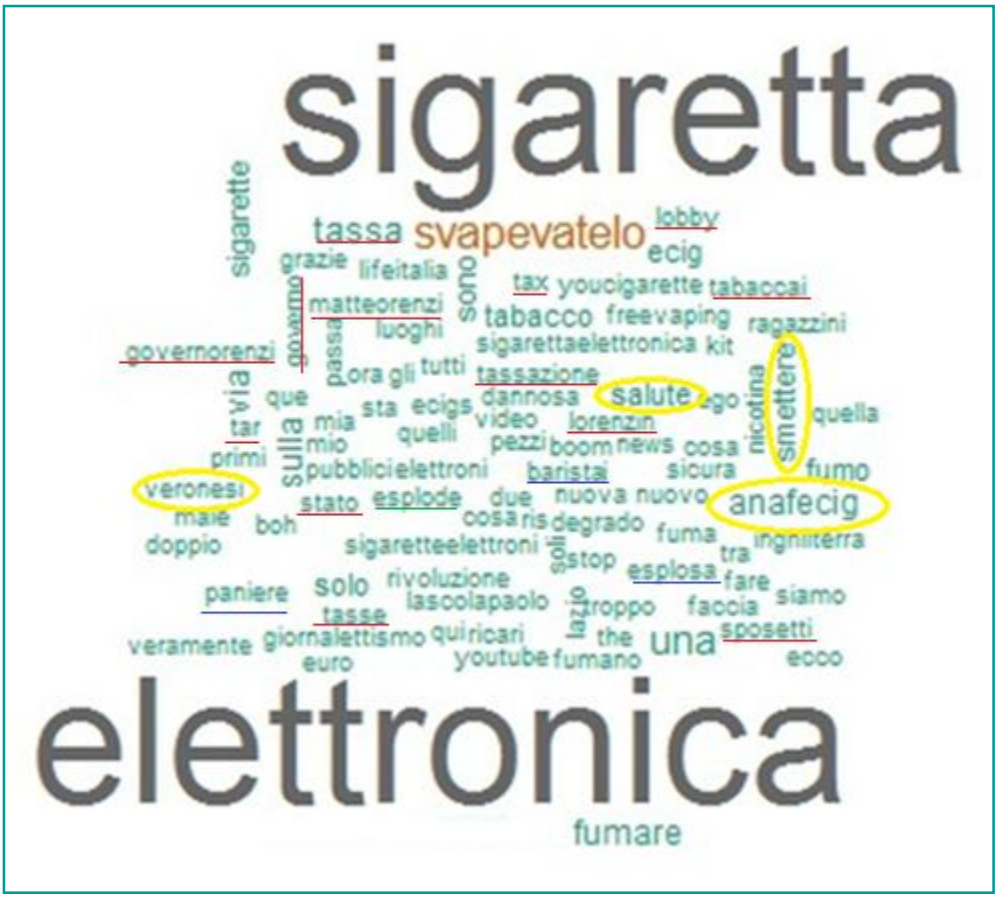


Figura 4.28: Word cloud ottenuto dal corpo di tweet contenenti la parola chiave 'sigaretta elettronica' estratti durante il primo quadrimestre 2014 con evidenziati in giallo i termini relativi alle sigarette elettroniche, in blu all'attualità ed in rosso alla politica

Fonte: elaborazione personale

Anche al verbo 'svapare' i tweet collegano principalmente vocaboli a sfondo politico: corrotte, governorenzi, Marco Pannella, matteorenzi, noecigtax, tassa, tassare, tassano, libertà, libero. Ci sono poi l'aggettivo 'bello' che potrebbe rimandare al piacere dell'atto di svapare, e 'smettere' che potrebbe esser collegato ad un tentativo di smettere di fumare, ma su questo punto non è possibile pronunciarsi con sicurezza data l'impossibilità di disambiguare, attraverso questo tipo d'analisi, i diversi utilizzi delle parole fatti nel testo, data che l'output informa solo sulla loro frequenza.

Figura 4.29: Word cloud ottenuto dal corpo di tweet contenenti la parola chiave 'svapare' estratti durante il primo quadrimestre 2014 con cerchiato il termine 'smettere' dall'ambigua interpretazione

Fonte: elaborazione personale

I tweet che parlano di 'smettere di fumare' sono collegati soprattutto all'espressione di un buon proposito per l'anno appena iniziato (newyear, l'inizio, newyearresolution, deciso, dovrei, riesco, smesso, devi). Compare anche il termine elettronica, ma con una misura, e quindi una frequenza, decisamente minore.

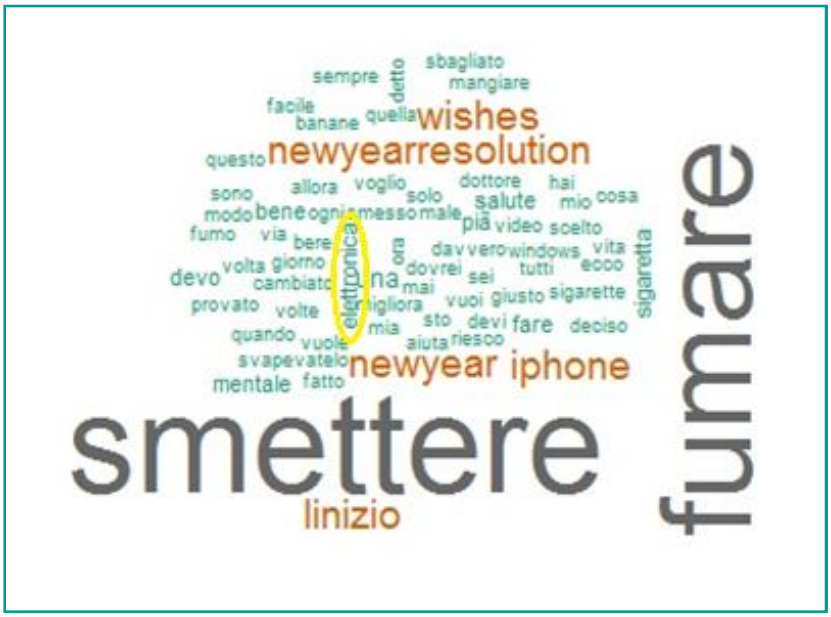


Figura 4.30: Word cloud ottenuto dal corpo di tweet contenenti la parola chiave 'smettere di fumare' estratti durante il primo quadrimestre 2014 con evidenziato l'unico riferimento alle e-cig *Fonte:* elaborazione personale

Infine nel *word cloud* dell'hashtag #smetteredifumare compare solo 'noecigtax' come termine che collega al fumo elettronico. L'elevata frequenza di 'usl' e 'modena' è spiegabile dalla campagna avviata appunto dall'USL modenese per spronare i cittadini ad abbandonare il vizio del fumo, la quale è stata attivamente pubblicizzata nel social network da parte dell'account dell'ente pubblico.

Figura 4.31: Word cloud ottenuto dal corpo di tweet contenenti l'hashtag #smetteredifumare estratti durante il primo quadrimestre 2014 con evidenziato l'unico riferimento alle e-cig *Fonte:* elaborazione personale

I *word cloud* permettono visivamente di vedere come i tweet relativi al fenomeno sigaretta elettronica del primo quadrimestre 2014 riguardavano principalmente il regime legislativo in materia e le diverse modifiche che si sono succedute ad inizio 2014 durante la fase di assestamento delll'imposizione fiscale a cui esse sono sottoposte in Italia. Dei messaggi pubblicati in cui l'autore si esprimeva sul tema smettere di fumare pochi parlavano anche di e-cig, la maggioranza invece sottolineava la volontà di prendere tale decisione, come questa scelta porti dei benefici fisici e mentali, chiedeva o consigliava agli altri utenti i metodi per raggiungere l'obiettivo desiderato o raccontava dei traguardi provvisori di un tentativo in corso.

4.2.3.2 Content analysis

Si è poi voluto approfondire l'analisi del contenuto dei tweet attraverso una *content analisys,* termine comparso per la prima volta in un documento di Lasswell (1941 in Franzosi, 2004). Essa è stata definita da Kaplan (1943 in Franzosi, 2004) come una tecnica che tenta di caratterizzare il significato di un testo in modo sistematico e quantitativo. Per Holsti (1969 : 14 in Franzosi, 2004) ha lo scopo di fare inferenze identificando oggettivamente e sistematicamente specifiche caratteristiche dei testi. In Krippendorf (2004 : 21 in Franzosi, 2004) si legge che la content analysis è una ricerca delle inferenze valide e replicabili sui contesti in cui i testi analizzati sono stati creati.
Essa è quindi definibile come una tecnica di ricerca quantitativa che utilizza strumenti e approcci diversi per ottenere il risultato di quantificare, standardizzare e sistematizzare, in altre parole fare scienza (Franzosi, 2004 : 549).
I numeri sono il risultato di conteggi di parole, temi, referenze, attori, azioni; la scelta dipende da cosa si sta analizzando e cosa si vuole scoprire. Questi poi per diventare significativi devono ovviamente essere analizzati. Paradossalmente la ricchezza e la limitazione della tecnica sono generate proprio dalla vasta libertà di scelta dello schema di codifica poiché non esiste un'unica via percorribile predefinita per cogliere il significato di un testo (Franzosi, 2004 : 550).
Dato che anche l'elementare conteggio di parole può offrire informazioni di valore riguardo ad un testo (Weber, 1990 in Franzosi, 2004), si è deciso di contare quanti dei tweet raccolti per le diverse *keyword* relative al fumo elettronico contenessero termini

facenti riferimento ad un tentativo di smettere di consumare tabacco. Viceversa sono state contate le volte in cui comparivano dei sinonimi di sigaretta elettronica e dell'atto di svapare tra i messaggi riportanti #smetteredifumare e 'smettere di fumare'.

Lo scopo era quello di individuare quanti tweet collegassero l'uso delle sigarette elettroniche ed un tentativo di smettere di fumare.

E' stato possibile, digitando dei comandi in R, ottenere il risultato dei conteggi di quanti dei tweet, contenuti nel file csv precedentemente ottenuto, contenessero i vocaboli prescelti.

Nello specifico (vedi Tabella 4.4) tra i 795 tweet contenenti l'hashtag #sigarettaelettronica, solo in 5 di essi si legge 'smettere di fumare' (0,6%), mentre in 74 il termine 'salute' (9,3%) ed in ben 200 'tassa' (25,2%).

Tra i 5049 raccolti per 'sigaretta elettronica', il 2,6% (129) conteneva 'smettere di fumare', il 3,4% (173) 'salute' ed il 10,9% (552) 'tassa'.

Nei tweet contenenti la parola chiave 'sigarette elettroniche' (6377) invece veniva digitata la *keyword* 'salute' 145 volte (2,3%), 'smetter di fumare' 65 volte (1%) e 'tassa' addirittura 1830 (28,7%).

Nei tre gruppi di tweet sopra citati l'hashtag #smetteredifumare non compare nemmeno nell'1% del totale.

Infine tra i 3297 messaggi estratti perché riportanti 'smettere di fumare', in 127 di essi compare 'sigaretta elettronica' (3,9%), in 7 #sigarettaelettronica (0,2%), in 27 'sigarette elettroniche' (0,8%), in 46 e-cig (1,4%), in 4 ecigarette (0,12%) e in 7 svapare (0,2%). Mentre in quelli estratti per l'hashtag #smetteredifumare solo uno faceva riferimento alla sigaretta elettronica, un altro usava il termine 'sigarette elettroniche' e tre riportavano l'hashtag #sigarettaelettronica (vedi Tabella 4.5).

	Gruppo di tweet estratti per l'hashtag #sigarettaelettronica	**Gruppo di tweet estratti per la *keyword* 'sigaretta elettronica'**	**Gruppo di tweet estratti per la *keyword* 'sigarette elettroniche'**
Totale	795	5049	6377
Tweet contenenti l'hashtag #smetteredifumare	6 = 0,8%	1 = 0,02%	1 = 0,02%
Tweet contenenti il termine 'smettere di fumare'	5 = 0,6%	129 = 2,6%	65 = 1%
Tweet contenenti il termine 'salute'	74 = 9,3%	173 = 3,4%	145 = 2,3%
Tweet contenenti il termine 'tassa'	200 = 25,2%	552 = 10,9%	1830 = 28,7%

Tabella 4.4: Numero di tweet contenenti determinate parole considerate rilevanti per la ricerca svolta tra il totale estratto per le *keyword* relative al fenomeno della sigaretta elettronica nel primo quadrimestre 2014

Fonte: Elaborazione personale

	Gruppo di tweet estratti per la *keyword* 'smettere di fumare'	**Gruppo di tweet estratti per l'hashtag #smetteredifumare**
Totale	3297	99
Tweet contenenti l'hashtag #sigarettaelettronica	7 = 0,2%	3 = 3%
Tweet contenenti il termine 'sigaretta elettronica'	127 = 3,9%	1 = 1%
Tweet contenenti il termine 'sigarette elettroniche'	27 = 0,8%	1 = 1%
Tweet contenenti il termine 'ecig'	46 = 1,4%	0 = 0%
Tweet contenenti il termine 'ecigarette'	4 = 0,12%	0 = 0%
Tweet contenenti il termine 'svapare'	7 = 0,2%	0 = 0%

Tabella 4.5: Numero di tweet contenenti determinate parole considerate rilevanti tra il totale estratto estratto nel primo quadrimestre 2014 per le *keyword* relative al fenomeno della sigaretta elettronica

Fonte: Elaborazione personale

E' stato poi ripetuto il conteggio eliminando i tweet contenenti l'hashtag #sVapevatelo (vedi Tabelle 4.6 e 4.7). Ciò perché, nonostante questi messaggi evidentemente

colleghino i due temi su cui è incentrata la ricerca, è presente una forzatura: i promotori dell'iniziativa vogliono difendere il fumo elettronico e calcano su questo suo possibile utilizzo positivo. Non rappresenta però una prova che sia veramente utilizzato a questo scopo. Bisogna però tener presente che, com'è emerso da alcuni studi esposti al capitolo 1, sembra che quello che si legge nei social network influenzi poi ciò che nella realtà le persone fanno. Per questo motivo il conteggio è stato condotto considerando sia il corpo totale dei tweet estratti, sia all'alternativa in cui i messaggi contenenti l'hashtag #sVapevatelo sono stati rimossi.

L'hashtag #sigarettaelettronica è stato ampiamente utilizzato in moltissimi dei tweet creati per la campagna; lo conferma il dato che il 65,3% conteneva pure l'hashtag #sVapevatelo. Dei restanti, dopo l'eliminazione di quelli con l'hashtag della campagna, in 5 si legge 'smettere di fumare', in 20 'salute' ed in 42 'tassa'.

Queste tre parole chiave appaiono rispettivamente 105, 163 e 356 volte tra i messaggi raccolti per la *keyword* 'sigaretta elettronica' dai quali sono stati rimossi quelli contenenti #sVapevatelo risultati essere il 14,7% del totale degli estratti.

	Gruppo di tweet estratti per l'hashtag #sigarettaelettronica eliminati quelli con l'hashtag #sVapevatelo	**Gruppo di tweet estratti per la *keyword* 'sigaretta elettronica' eliminati quelli con l'hashtag #sVapevatelo**	**Gruppo di tweet estratti per la *keyword* 'sigarette elettroniche' eliminati quelli con l'hashtag #sVapevatelo**
Totale tweet senza l'hashtag #sVapevatelo (% sul totale degli estratti)	276 (34,7%)	4308 (85,3%)	4502 (70,6%)
Tweet contenenti l'hashtag #smetteredifumare	6 = 2,2%	1 = 0,02%	1 = 0,02%
Tweet contenenti il termine 'smettere di fumare'	5 = 1,8%	105 = 2,4%	61 = 1,4%
Tweet contenenti il termine 'salute'	20 = 7,2%	163 = 3,8%	106 = 2,4%
Tweet contenenti il termine 'tassa'	42 = 15,2%	356 = 8,3%	964 = 21,4%

Tabella 4.6: numero di tweet contenenti determinate parole considerate rilevanti tra il totale estratto estratto nel primo quadrimestre 2014 per le *keyword* relative al fenomeno della sigaretta elettronica eliminati quelli contenenti l'hashtag #sVapevatelo

Fonte: Elaborazione personale

Mentre, degli *user-generated content* in cui era stato digitato ‘smettere di fumare’ solo il 2,5 % era stato creato per l’iniziativa patrocinata da Anafe, Fiesel e Life. Senza questo 2,5% ce n’erano 109, 27, 34, 4,e 7 contenenti rispettivamente ‘sigaretta elettronica’, ‘sigarette elettroniche’, ‘ecig’, ‘ecigarette’ e ‘svapare’.

	Gruppo di tweet estratti per la *keyword* ‘smettere di fumare’ eliminati quelli con l’hashtag #sVapevatelo	**Gruppo di tweet estratti per l’hashtag #smetteredifumare eliminati quelli con l’hashtag #sVapevatelo**
Totale tweet senza l’hashtag #sVapevatelo (% sul totale degli estratti)	3217 (97,6%)	83 (83,8%)
Tweet contenenti l’hashtag #sigarettaelettronica	7 = 0,2%	3 = 3,6%
Tweet contenenti il termine ‘sigaretta elettronica’	109 = 3,4%	1 = 1,2%
Tweet contenenti il termine ‘sigarette elettroniche’	27 = 0,8%	1 = 1,2%
Tweet contenenti il termine ‘ecig’	34 = 1,1%	0 = 0%
Tweet contenenti il termine ‘ecigarette’	4 = 0,1%	0 = 0%
Tweet contenenti il termine ‘svapare’	7 = 0,2%	0 = 0%

Tabella 4.7: numero di tweet contenenti determinate parole considerate rilevanti tra il totale estratto estratto nel primo quadrimestre 2014 per le *keyword* relative al fenomeno della sigaretta elettronica eliminati quelli contenenti l’hashtag #sVapevatelo
Fonte: Elaborazione personale

Le Tabelle 4.4 – 4.7 mostrano come ben poco gli utenti di Twitter nel primo quadrimestre 2014 abbiano avviato discussioni in cui venisse affrontata la possibilità di smettere di fumare con le sigarette elettroniche.

4.2.3.3 Analisi semantica

Le precedenti analisi sulle distribuzioni di frequenza delle diverse parole all'interno dei testi analizzati possono essere utili a collegare alcuni temi; non spiegano però la relazione esistente tra di essi.

Le analisi linguistiche automatiche dei testi richiedono preventive analisi semantiche attraverso dei *pre-processing tools,* come *parser* o *tokenizer,* in grado di catalogare ogni termine in base alla sua categoria grammaticale (ad esempio sostantivo, verbo o aggettivo). Purtroppo non è stato possibile seguire questo approccio per varie ragioni:

1. molti di questi *pre-processing tool* disponibili open-source riescono a lavorare solo su testi in inglese;
2. l'output di R che fornisce l'elenco dei tweet estratti secondo le query impostate nella API offre anche altri informazioni (come retweet, data ed ora, se è stato contrassegnato come *favourite*, ..) che rendono il testo del tweet troppo "sporco" per analisi automatiche;
3. la brevità della stringa di testo e la mancanza di una corretta struttura sintattica mettono alla prova la capacità di tali programmi di analizzare il testo correttamente.

Per questo tipo d'operazione sarebbero necessari dei software creati ad hoc, che al momento non sono disponibili *open-source*.

Relativamente alla pulizia dei messaggi, nel presente lavoro di ricerca è stato utilizzato PC-ACE per lavorare sulle stringhe in SQL del file contenente i tweet estratti per la parola chiave 'sigaretta elettronica', la cui pulizia è però stata perfezionata manualmente. Invece, i tweet contenenti l'hashtag #sigarettaelettronica sono stati puliti con l'*editor* di testo Notepad ++ ed intervento manuale.

Su questi ultimi è stato poi testato GETARUNS, un *parser* per l'analisi semantica delle parole contenute in un testo. Esso presenta un buon livello di robustezza rispetto ad altri disponibili per la lingua italiana poiché tiene in forte considerazione il contesto generale del testo quando procede con la catalogazione delle parole. È, infatti, capace di effettuare una buona disambiguazione della categoria grammaticale dei termini in base al testo in cui essi sono inseriti. Inoltre, se incontra tra le parole contenute qualche

elemento che non riesce a riconoscere passa oltre e non si blocca, a differenza di altri la cui analisi è più superficiale. Quest'ultima caratteristica si è rivelata molto importante perché, come già detto, i tweet anche se sono costituiti da frasi molto brevi hanno una struttura sintattica e una terminologia molto particolare che crea molte più difficoltà a un semplice testo.
La parte d'interesse per la presente ricerca dell'output di GETARUNS applicato ai testi dei tweet contenenti l'hashtag #sigarettaelettronica, è stato il seguente:

```
counts([factivity(433/1125),negation(54),subjectivity(87),diathe
sis(47),point_of_view(0.9454314720812182)])
```

Quanto riportato significa che della totalità delle 1125 frasi semplici che il software ha rilevato all'interno del corpo di tweet analizzati 433 di queste (38,5%) erano non fattive, facevano cioè riferimento ad un fatto non realmente accaduto ma potenziale (utilizzando ad esempio un tempo verbale condizionale) o futuro. 54 sono risultate contenere una negazione all'interno della struttura verbale, mentre 87 riportavano un'opinione personale. Le frasi passive rilevate sono state 47 (4,2%).
508 frasi semplici sono risultate essere senza soggetto o con soggetto nullo; il 45% è una percentuale molto elevata rispetto allo standard dell'italiano scritto, ma anche di quello parlato (Delmonte, 2014), probabilmente la spiegazione è il vincolo dei 140 caratteri di un tweet che spinge l'autore ad eliminare articoli, congiunzioni e a volte persino i soggetti. Altra motivazione possibile è l'intenzione da parte dell'autore di rendere il testo più impersonale possibile per dare un'idea di diffondere un messaggio dalla maggiore autorevolezza.
L'output ottenuto ha permesso le precedenti considerazioni a livello grammaticale, ma costituisce soprattutto un primo passo per ulteriori analisi linguistiche automatiche i cui risultati potrebbero rispondere alle domande più varie e da cui potrebbero poi essere estratte una moltitudine d'informazioni utili per diversi scopi. Ciò perché, oltre che misure aggregate, ad ogni singola parola delle frasi rilevate all'interno dell'input inserito nel software viene attribuita una categoria grammaticale (con la precisazione del genere e della persona) e viene fornito il numero di quante volte essa compare nel testo. Di seguito ne viene riportato un esempio, ricavato dall'output ottenuto:

```
i(1,il,art,il,sn,[sems=def,mfeats=fs],2,-)
i(2,nh,Tar-
Tar,[sems=def,mfeats=fs,[3,any,sing]],[inst,polsem=neut],5,subj/agent)
i(3,del,part,di,sp,[sems=det,mfeats=ms],2,-)
i(4,Lazio,nh,Lazio,[3,any,sing],[geo,polsem=neut],3, (mod)/specif)
i(5,sospende,v,sospendere,ibar,[cat=verb,pred=sospend+ere,scat=tr,mood
=indic,tense=pres,pers=3,num=s],0,[tr,activ,exten])
i(6,la,art,la,la,sn,[],7,-)
i(7,n,tassazione-
tassazione,[[3,fem,sing]],[[cse,dyn],polsem=neg],5,obj/theme_unaff)
i(8,al,part,al,a-[sems=det,mfeats=ms],sp,a-[sems=det,mfeats=ms],9,-)
i(9,nw,58/5%-
58/5%,[def=def,3,neu,s],[umano,istituzione],5,obl/_G599219)
```

Tra i risultati forniti vi è inoltre la classifica delle parole più ripetute, di seguito riportata, e che combacia con quanto già emerso attraverso il *word cloud* (vedi Figura 4.27), anche se manca dell'immediatezza visiva di quest'ultimo: sigaretta, tassa, governo, lobby, Matteo Renzi, fine, disoccupato, 58,5%, salute, tabagismo, scienza.
Si può quindi concludere che anche quest'analisi mette in evidenza come la tematica principale sia quella politica.

4.2.3.4 Quantitative narrative analysis e thematic analysis

Una semplice lettura dei tweet che contengono la parola 'sigaretta elettronica', non preceduta da hashtag, ha fatto notare come in questo caso vi sia una minor focalizzazione sul tema politico, e come invece compaiano parecchi messaggi ironici, racconti personali, opinioni sugli svapatori ed allusioni sessuali.
Questo probabilmente perché quando gli utenti pubblicano un messaggio realmente riguardante il fumo elettronico e contenente notizie serie ad esso riferite tendono ad inserire l'hashtag che rimanda all'argomento preciso. Quando, invece, il dispositivo elettronico compare come termine di paragone in una metafora, in racconti personali o in allusioni a sfondo sessuale, l'etichetta non viene inserita perché non si ritiene il tweet adatto a tale categorizzazione.
Per meglio comprendere questa dinamica si è deciso di sfruttare le potenzialità di **PC-ACE** (Program for Computer-Assisted Coding of Events), software per *Content Analysis* e *Quantitative Narrative Analysis*, allo scopo di indagare più in profondità il contenuto del corpo di tweet contenenti la *keyword* 'sigaretta elettronica'.

La ***quantitative narrative analysis*** è una tecnica di analisi linguistica *bottom-up* che vuole trasformare le parole in numeri, i quali però continuano a mantenere un forte legame con le parole da cui sono stati estratti. Essi sono basati, non su schemi di codifica derivanti dagli interessi teorici dell'investigatore, come avviene nella *content analysis*, bensì su proprietà linguistiche invarianti possedute dal testo originale. *"In keeping the roots of the technique in linguistics and literary theory, it does not turn against the words once the words have yielded the numbers"* (Franzosi, 2009).

La QNA con l'aiuto di tre discipline (criticismo letterario/linguistico, scienze informatiche e statistica) rende possibile andare *"from word to numbers"* (Franzosi, 2009) preservando la ricchezza informativa presente nel testo narrativo originale, ma nel contempo permettendo un'analisi statistica dello stesso. Si presenta pertanto come un modo per superare la dicotomia qualità-quantità collegando parole a numeri.

Linguisticamente un evento narrativo è tipicamente espresso attraverso la struttura linguistica di base costituita dalla tripletta Soggetto-Verbo-Oggetto, dove in narrativa il soggetto e' tipicamente un attore sociale e il verbo un'azione sociale. Ognuno dei tre elementi poi può avere dei qualificatori (definiti *modifier)* propri: per esempio caratteristiche relative al soggetto o all'oggetto possono essere il genere, la razza, l'ideologia politica o la professione (Franzosi, 2012), mentre *modifier* del verbo possono essere il luogo, il momento, lo strumento, la motivazione o il risultato (Franzosi, 2014).

In narrativa questa struttura SVO+ *modifiers* è anche definita *story grammar*, ossia grammatica del racconto. La *quantitative narrative analysis* lavora su questa tipologia di dati: attori relazionati ad altri attori attraverso una serie di azioni avvenute in un preciso luogo e momento. La natura relazionale delle categorie della *story grammar* rendono possibile il loro inserimento in un *Relational Database Management System* (RDBMS) ed una loro elaborazione che permette di trasformare le parole presenti nel testo di narrativa in numeri sui quali è possibile applicare tecniche statistiche d'analisi focalizzate sugli attori e sulle loro relazioni. I dati relazionali resi disponibili attraverso la QNA sono inoltre ideali per implementare analisi svolte con strumenti GIS (*Geographic Information Systems*) o *Network Analysis*; anche in questo caso la centralità degli attori e delle loro azioni rimane preservata. Si può così ottenere una spiegazione socio-scientifica dei fenomeni sociali che si vogliono indagare partendo dai

testi a disposizione e si riesce ad identificare chi fa cosa, a chi, quando, dove e perché (le 5 W + H).

Franzosi, negli anni ’80 ha iniziato a sviluppare un software specializzato disponibile open-source, PC-ACE (Program for Computer-Assisted Coding of Events), realizzato per organizzare, conservare e recuperare informazioni caratterizzate da strutture complesse che comprendono gerarchie e relazioni tra gli elementi costituenti le *user-table*. Si presenta come strumento adatto, quindi, anche per svolgere una QNA date le sue proprietà relazionali che permettono un’archiviazione delle componenti principali delle storie (attori, azioni, spazio e tempo) secondo strutture rigorose ed un veloce recupero delle stesse in molteplici e complesse modalità. Senza un programma di questo tipo un’implementazione della QNA a progetti consistenti nell’analisi di un elevato numero di testi sarebbe impossibile, data la complessità dello schema di codifica (la *story grammar*).

Attraverso le SQL query vengono messe in luce le proprietà relazionali dei dati inseriti, si possono ottenere frequenze di distribuzioni e si può procedere con altre tecniche statistiche multivariate standard.

Inoltre, a partire dagli attori coinvolti e dalle loro reciproche relazioni all’interno della storia narrata, è possibile ottenere una rappresentazione grafica dei network esistenti tra i diversi soggetti coinvolti e la loro variazione nel tempo, grazie al supporto del software Gephi. Infine, la disponibilità di dati spazio-temporali attribuibili a ciascun evento permette di mapparne la diffusione e distribuzione nello spazio e nel tempo adottando Google Earth.

Allo scopo di utilizzare QNA sui tweets, sono stati codificati manualmente tre tweet per ogni giorno del quadrimestre, scelti casualmente tra tutti quelli pubblicati nella stessa giornata, utilizzati poi come input per PC-ACE.

La codifica ha avuto lo scopo di individuare le triplette SVO presenti nei tweet e di assegnare ad ognuno di essi una tematica, resa poi più specifica dalla voce “contenuto del tema” che precisa la declinazione assunta all’interno della macro area indicata.

Procedendo con la catalogazione è emerso come non tutti i tweet presentassero delle triplette semantiche. Quelle individuate, inoltre, sono risultate essere tra loro completamente diverse. Purtroppo, QNA è applicabile a testi che presentano una struttura linguistica narrativa basilare.

Alla fine, si e' presa la decisione di non condurre una quantitative narrative analysis, ma di limitarsi solamente ad una **thematic analysis**, uno degli approcci più comuni della content analysis, in cui lo schema di codifica vuole catturare i temi dominanti presenti nel testo (Franzosi, 2004 : 550). Per ottenere una lista di temi dominanti si e' proceduto in maniera induttiva. Dopo una prima lettura di un sottoinsieme dei tweet che si voleva catalogare è stata creata una lista provvisoria, la quale è stata poi testata adottandola per altri messaggi. Ciò ha fatto emergere accorgimenti e modifiche necessarie che hanno portato ad una riformulazione della lista. Il procedimento è stato più volte ripetuto fino ad ottenerne la versione definitiva costituita da otto temi:

1. politica;
2. opinione personale;
3. esperienza diretta;
4. esperienza indiretta;
5. notizia d'attualità;
6. salute;
7. ironia;
8. sessualità.

Di seguito ne viene fornita una definizione:

- *politica*, attribuito ai tweet che commentano la legislazione vigente in Italia o il Governo le sue scelte. Il contenuto del tema precisa a quale ambito della politica fa riferimento il tweet: governo, tassa, emendamento o a quale personaggio politico in particolare. La disambiguazione dei seguenti casi esemplificativi come "odio la sigaretta elettronica, dovrebbero tassarla di più", "l'imposta sulla sigaretta elettronica è una vergogna" e "il Tar sospende l'imposta del 58,5%" avviene attraverso la precisazione del contenuto del tema. Nel primo esso sarà "a favore della tassa", nel secondo "contro la tassa" e nel terzo "tassa", etichetta che indica un riferimento nel messaggio alla tassazione italiana, senza dimostrare alcun disappunto o sostegno da parte dell'autore. È stato invece utilizzato "legislazione estera" quando il tweet informa sui recenti cambiamenti in ambito legislativo avvenuti in altri Paesi;

- *opinione personale*, riporta un pensiero dell'autore sul tema del fumo elettronico. Il contenuto precisa poi se l'opinione è neutrale, a favore o contraria alle e-cigarette;
- *esperienza diretta*, utilizzato quando viene raccontato qualcosa vissuto in prima persona da chi twitta. Alcuni dei messaggi con questo tema riportano come contenuto "tentativo di smettere di fumare" o "fallimento tentativo di smettere di fumare" quando chi scrive racconta la sua prova, in corso o finita male, di diminuire o eliminare il consumo di tabacco;
- *esperienza indiretta*, è stato utilizzato per catalogare tweet facenti riferimento ad episodi che implicano l'utilizzo della sigaretta elettronica non da parte dell'autore, ma di persone a lui vicine;
- *notizia d'attualità*, fa riferimento ad una *news* il cui contenuto specifico può essere ad esempio "incidente" quando la notizia riguarda eventi spiacevoli recentemente avvenuti, oppure "crisi del settore" quando sono messi in luce i trend negativi registrati dal settore delle sigarette elettroniche;
- *salute*, è il tema indicato quando il tweet discute delle conseguenze sulla salute che ha la sigaretta elettronica: "effetti positivi", "effetti dannosi", "fumo passivo", "effetti" sono le precisazioni che si leggono nella colonna contenuto del tema. Il primo caso viene attribuito ad un messaggio che sostiene le e-cig per i loro effetti benefici, il secondo al contrario ad uno che ne espone possibili danni, il terzo riguarda la salute delle persone che stanno vicine allo svapatore ed infine il quarto espone senza particolari schieramenti ideologici le conseguenze dello svapare. Quando l'opinione è espressa da un soggetto competente in materia o si riportano i risultati di uno studio sono state utilizzate le etichette "parere dell'esperto", "esperto a favore" o "esperto contrario";
- *ironia*, tema attribuito a quei messaggi che avevano lo scopo di deridere un determinato target specificato nella categoria contenuto. Esso è "sigaretta elettronica" quando la presa in giro è riferita non a persone ma al dispositivo elettronico in sé. Spesso l'ironia adotta metafore a sfondo sessuale, ecco che allora nel contenuto si legge, oltre che al target, anche il mezzo adottato per ottenere l'effetto ironico;

- *sessualità*, contraddistingue messaggi a sfondo sessuale nei quali compare la sigaretta elettronica, senza però che vi siano ironia o metafore. Il contenuto precisa anche in questo caso il target del riferimento sessuale;

In Tabella 4.8 viene riportato l'esempio della catalogazione di sei tweet.

Data	Testo del tweet	Tema	Contenuto del tema
18/04/14	Sigaretta elettronica: due società di pneumologi contrarie	Salute	Esperto contrario
08/04/14	Inghilterra, sigaretta elettronica esplode in faccia alla barista	Notizia d'attualità	Incidente
20/03/14	Bambini di undici anni che fanno i fighi con la sigaretta elettronica quando non sanno nemmeno cos'è una sigaretta. Ma!	Ironia	Ragazzini
06/03/14	Fumare la sigaretta elettronica e come mangiare al McDonald e bere una CocaCola light	Ironia	Sigaretta elettronica
14/02/14	Sigaretta elettronica, Lorenzin risponde a Veronesi: Le tasse le decide Saccomanni	Politica	Tassa
13/02/14	24 ore senza fumare: e se la sigaretta elettronica funzionasse? #pagoisallin #stopsmoke	Esperienza diretta	Tentativo di smettere di fumare
06/01/14	Fumate le sigarette elettroniche, scopate con le bambole gonfiabili, bevete la birra analcolica, il caffè decaffeinato	Ironia	Sigaretta elettronica con metafora a sfondo sessuale

Tabella 4.8: Esempio della catalogazione di sei tweet utilizzati come input per PC/ACE
Fonte: elaborazione personale

Una volta utilizzati i tweet così catalogati come input per PC-ACE, sono state analizzate le frequenze di distribuzione dei diversi temi. È emerso come i temi principali siano stati ironia, politica ed esperienza diretta (vedi Tabella 4.9 e Figura 4.32).

Tema	Tot. tweet	Percentuale
Ironia	103	28,6%
Politica	80	22,2%
Esperienza diretta	46	12,8%
Notizia d'attualità	35	9,7%
Salute	34	9,4%
Pubblicità	27	7,5%
Opinione personale	20	5,6%
Esperienza indiretta	13	3,6%
Sessualità	2	0,6%
Totale	360	100,0%

Tabella 4.9: numero di tweet contenenti i diversi temi tra quelli contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 selezionati come input per PC-ACE

Fonte: Elaborazione personale

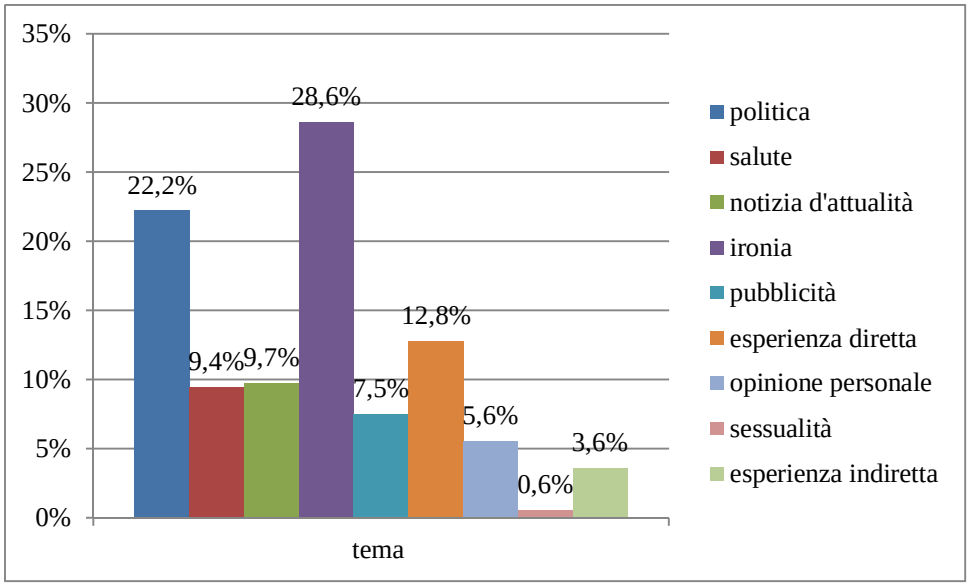


Figura 4.32: Percentuale dei diversi temi dei tweet utilizzati come input per PC-ACE

Fonte: Elaborazione personale

La politica era un risultato abbastanza prevedibile dati i cambiamenti normativi che si sono succeduti nel periodo analizzato. La presenza di numerosi racconti di esperienze direttamente vissute dall'autore del tweet si spiega con il carattere molto personalistico dei messaggi pubblicati su Twitter e con la tendenza generale a rendere pubblici avvenimenti privati sui social network.

A sorprendere è stato invece il primato ottenuto dal tema dell'ironia.

In Figura 4.33 si può vedere un'analisi più specifica della declinazione del tema politico nei tweet esaminati: è la tassa introdotta ad inizio anno l'argomento principale.
Il 35% dei tweet politici informa sugli sviluppi relativi ai cambiamenti fiscali in corso, il 24% li giudica negativamente, il 6% fa riferimento all'emendamento nel suo complesso ed l'1,3% si schiera contro; l'8% infine accusa le lobby del tabacco di favorire la tassa del 58,5% per proteggere la loro posizione.

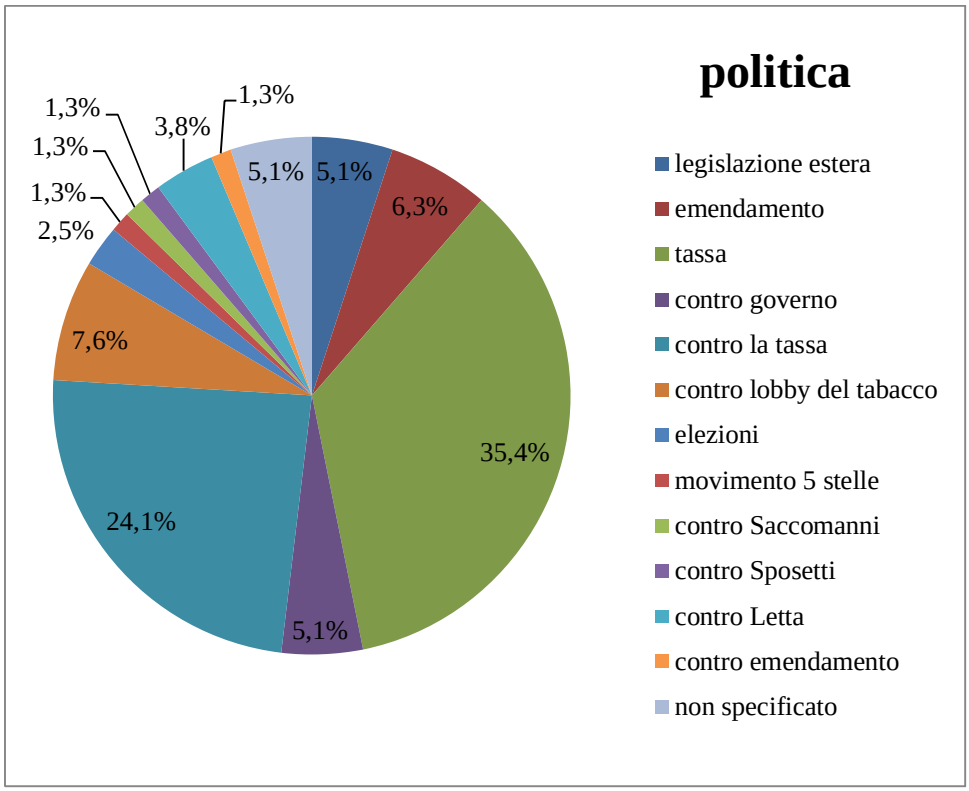


Figura 4.33: Argomenti dei messaggi riguardanti la politica contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 selezionati come input per PC-ACE
Fonte: Elaborazione personale

Tra le esperienze dirette descritte dai tweet selezionati il 20% raccontava di un tentativo di smettere di fumare in corso, ma solo il 2% di un tentativo riuscito (vedi Figura 4.34).

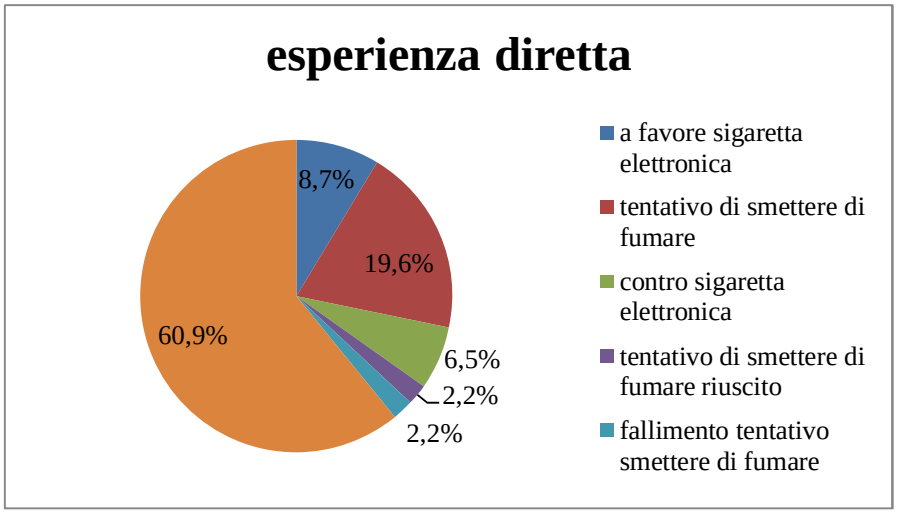


Figura 4.34: Argomenti dei messaggi riportanti esperienze dirette dell'utente che twitta contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 selezionati come input per PC-ACE

Fonte: Elaborazione personale

Focalizzandosi sui messaggi ironici si è potuto notare come molti di questi abbiano lo scopo di deridere dispositivo elettronico in sé (43,6%), i ragazzini che svapano per atteggiarsi da adulti (12,6%) o gli svapatori in generale (26,2%) (vedi Figura 4.35).

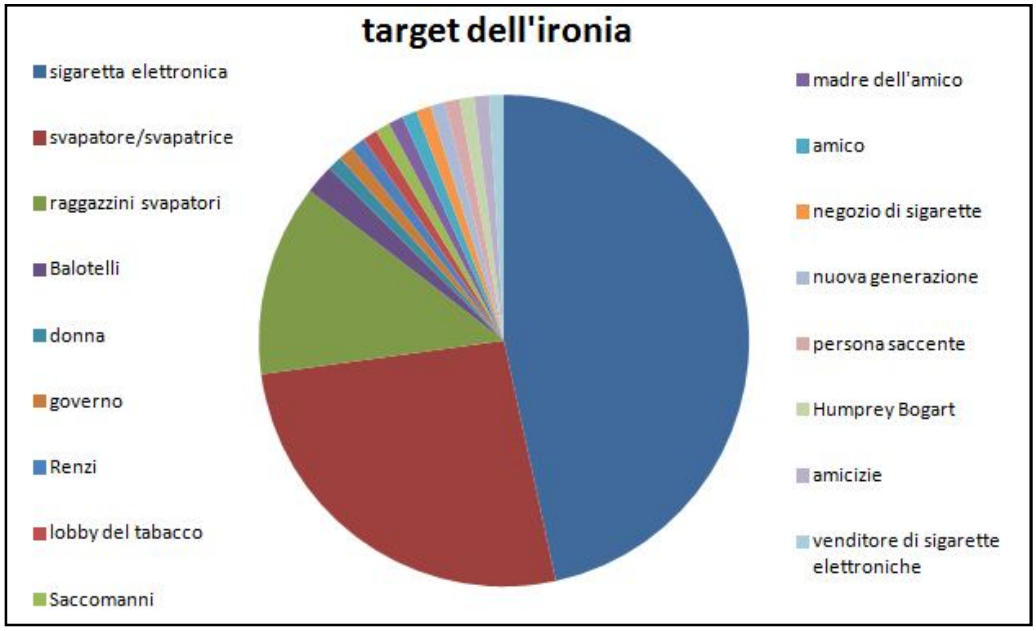


Figura 4.35: Target dei messaggi ironici contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 selezionati come input per PC-ACE

Fonte: Elaborazione personale

In numerosi casi l'effetto ironico è stato ottenuto attraverso metafore a sfondo sessuale (13 tweet pari al 12,6% dei messaggi catalogati come ironici) che, aggiunti ai tweet che sono stati fatti rientrare nel tema sessualità, hanno fatto emergere come spesso la sigaretta elettronica sia collegata e paragonata al piacere sessuale. Questo però è sempre presentato come un piacere virtuale, non dato da qualcosa di reale, ma da un surrogato di quello che realmente si vorrebbe avere. Ecco che allora nei tweet compaiono la bambola gonfiabile che sostituisce la donna per poter ottenere il piacere sessuale, la nutella senza calorie che non fa ingrassare ma nemmeno rinunciare al piacere del cibo, proprio come la sigaretta elettronica che non toglie il vizio del fumo ma permette di evitare i danni causati dal consumo tabacco. La lista continua con il fuoco virtuale, l'ebook, le amicizie elettroniche, la birra analcolica e il caffè decaffeinato. In tutti questi casi è la scienza che crea delle fonti alternative di piacere per eliminare le eventuali conseguenze negative che quelle reali potrebbero creare o per sopperire ad una loro mancanza.

È stato poi possibile ottenere, col supporto del software Gephi, la rappresentazione grafica del network tematico sottostante il corpo di tweet scelto come input. Riportato in Figura 4.36, il grafico mette in evidenza quali siano stati i temi dei messaggi pubblicati nel social network, e quali di questi sono stati i principali (lo spessore delle frecce è indicativo della frequenza con cui il tema è stato rilevato).
Avendo a disposizione anche la data relativa ad ogni tweet si è potuto creare un network dinamico il quale, con forte impatto visivo, permette di capire come i temi si sono susseguiti nel tempo.
Nel caso specifico i temi individuati si sono alternati senza evidenziare particolari trend.

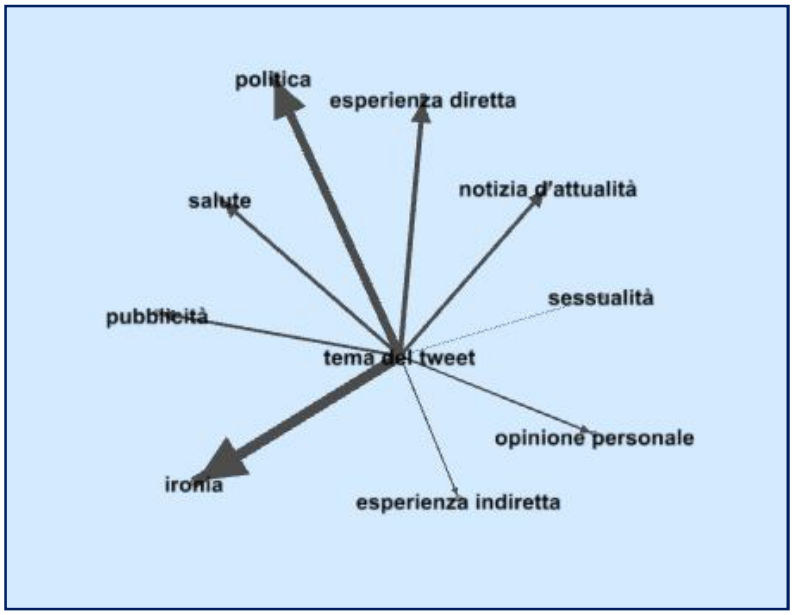


Figura 4.36: Grafico relativo ai temi dei tweet contenenti la parola chiave 'sigaretta elettronica' nel primo quadrimestre 2014 selezionati come input per PC-ACE
Fonte: Elaborazione personale

Relativamente a quanto indagato, comunque, si ha un'ulteriore conferma di la salute non sia tra i temi principali trattati da coloro che twittano, e come lo smettere di fumare non si sia dimostrato essere un elemento rilevante nemmeno nell'output di quest'analisi. Anche se fondamentalmente alla domanda di ricerca era già stata data risposta attraverso le analisi precedenti, quest'ultimo approfondimento sul contenuto linguistico è stato condotto per dimostrare gli svariati sviluppi che può avere l'analisi di un corpo di tweet estratti e le svariate possibilità che mette a disposizione.

4.2.4 Sintesi dei risultati ottenuti dall'analisi dei tweet estratti

Sui tweet estratti è stata effettuata inizialmente un'analisi puramente quantitativa, che ne ha evidenziato l'andamento durante i primi quattro mesi dell'anno. Sono stati registrati dei picchi nel numero di messaggi pubblicati sulla piattaforma digitale in concomitanza con eventi del mondo reale attinenti al fumo elettronico.
Il passaggio successivo è stata un'analisi volta ad approfondire la comprensione del contenuto dei tweet.

Sono stati, innanzitutto, creati dei *word cloud* per mettere in evidenza i termini che con la frequenza maggiore comparivano all'interno del corpo di tweet oggetto dell'analisi. Attraverso la visualizzazione dei dati si è potuto scoprire che i vocaboli più ricorrenti facevano riferimento ai cambiamenti susseguitisi in ambito legislativo.
Successivamente, attraverso R è stata contata la presenza di determinati vocaboli all'interno dei tweet estratti al fine di capire quanti di questi associassero l'utilizzo delle sigarette elettroniche ad un tentativo di smettere di fumare.
Nonostante già a questo punto dell'analisi fosse emerso come i due temi indagati non fossero collegati negli *user-generated content* pubblicati in Twitter nel corso del quadrimestre, si è voluto procedere con ulteriori analisi linguistiche dei testi dei messaggi a disposizione al fine di evidenziarne ulteriori possibili modalità di sfruttamento.
L'analisi semantica condotta con Getaruns e la *thematic analysis* hanno ulteriormente confermato come la salute non sia stato uno dei temi principali delle discussioni riguardanti la sigaretta elettronica avviate dagli utenti di Twitter, e tantomeno la possibilità dell'utilizzo delle e-cig come metodo per smettere di fumare.
Molti dei messaggi pubblicati nel social network contenenti il termine 'sigaretta elettronica' riguardavano la tassa del 58,5% introdotta nel sistema legislativo italiano, o avevano lo scopo di ironizzare sul dispositivo elettronico e i suoi utilizzatori.
Questi ulteriori approfondimenti condotti, oltre ad aver convalidato i risultati delle precedenti analisi, hanno permesso di sottolineare la varietà di quesiti a cui l'analisi dei dati provenienti dai social network permette di rispondere, ampliando ulteriormente la visione delle sue potenzialità.

Si può concludere che per quanto riguarda la domanda di ricerca "Le sigarette elettroniche sono usate per smettere di fumare?" l'elaborazione dei dati raccolti da Twitter nel primo quadrimestre 2014 ha fatto emergere come gli italiani non considerino le e-cigarette uno strumento utilizzabile per smettere di fumare, quanto piuttosto un dispositivo da usare come qualcosa di aggiuntivo o diverso dalla sigaretta tradizionale, confermando la visione proposta da Popova e Ling (2013) e da Doyle, Patterson e Scott (2014).

E’ importante sottolineare che tutte le analisi precedentemente esposte sono state condotte adottando strumenti gratuiti e open- source, e quindi sono state soggette a limitazioni ulteriori rispetto ad indagini condotte sfruttando le possibilità offerte a pagamento.
Strumentazioni diverse e maggiori competenze nel campo informatico e linguistico potrebbero portare ad output migliori sia in termini di precisione, che in termini di varietà. Nonostante ciò, i risultati ottenuti sono stati comunque utili per confermare le numerose potenzialità di questo tipo d’analisi.

4. 3 Confronto tra dati provenienti dal mondo online e quelli raccolti nel mondo reale

Per testare se quanto ricavato dall’analisi dei tweet rispecchiava ciò che poi nella realtà avveniva si è voluto fare un confronto tra i dati raccolti da Twitter e quelli reali riportati nei report pubblicati dalle Organizzazioni che si occupano della Pubblica Sanità in Italia, raccolti attraverso interviste a campioni rappresentativi della popolazione.

Riassumendo quanto esposto nel Paragrafo 4.1.2 il mercato delle sigarette elettroniche si è presentato in forte crescita dal 2010 fino al Giugno 2013, quando ha subito una brusca frenata.
Nel 2012, anche se i tentativi di smettere di fumare erano in calo, come pure la vendita di sigarette tradizionali, il consumo di e-cigarette registrava una crescita non indifferente, fornendo come possibile lettura della situazione una conferma dell’affermazione che le sigarette elettroniche non sono utilizzate da chi tenta di smettere di fumare.
Anche i dati relativi al primo quadrimestre 2014 raccolti da Passi confermano come solo il 10% dei fumatori che vogliono smettere pensa alla sigaretta elettronica come strumento da adottare. Questa poi si rivela esser stata utile solo per l’1% degli ex fumatori. Percentuali molto basse per poter affermare che le e-cigarette sono utilizzate come metodo per eliminare il consumo di tabacco e prodotti derivati.

Analogamente, dal rapporto ISS-DOXA 2014 emerge come tra i fumatori intervistati che in passato avevano tentato di smettere di fumare solo il 15,2% di questi ci aveva provato utilizzando le sigarette elettroniche, mentre nessuno di quelli che ci è riuscito le nomina.

Inoltre, sempre secondo il report annuale commissionato dall'Istituto Superiore di Sanità, il numero dei fumatori che provano a smettere di fumare era stato in declino dal 2008 al 2012, mentre nel 2013 ha registrato un aumento del 7%, cambio d'inversione che peraltro emerge, seppur con differenze nelle percentuali giustificate dalla diversità dei campioni utilizzati, anche dai dati raccolti dal sistema di sorveglianza Passi.

Anche il numero degli svapatori è cresciuto dal 2012 al 2013, ma il confronto dei due andamenti dimostra come non vi sia correlazione tra le due crescite: il boom del fenomeno delle e-cigarette registrato ad inizio 2013 aveva fatto aumentare il numero dei suoi utilizzatori ben più del 7%.

Inoltre, se si paragona il lieve, e da verificare, aumento dei tentativi di smettere di fumare ad inizio 2014 riportato da Passi con il calo del numero degli svapatori registrato dal report ISS-DOXA, emerge nuovamente una discordanza tra i due trend, che risultano quindi non collegabili.

I dati presentati nel Rapporto ISS-DOXA 2014 sugli svapatori sono stati presentati al paragrafo 4.1.2 per completezza d'informazioni, ma si possono considerare poco rappresentativi dello scenario italiano dato che, come lo stesso report sottolinea, il campione adottato era composto da solo 50 individui. Essi non saranno quindi presi in esame in questa sezione.

Si può concludere che, sebbene le sigarette elettroniche possano aiutare chi vuole smettere di fumare (vedi studi di Bullen et al., 2013), i dati raccolti online, ed anche i dati provenienti dal mondo reale, sembrano però portare alla conclusione che esse al momento in Italia non sono un metodo utilizzato da coloro che tentano di ridurre od eliminare il consumo delle sigarette tradizionali o di altri prodotti affini contenenti tabacco.

Appare però importante sottolineare che, poiché i dati collezionati sono relativi solo ad un quadrimestre, peraltro interessato da particolari eventi in ambito legislativo, ci vuole cautela nel generalizzare i risultati ottenuti.

5. Spunti d'applicabilità per il marketing - Riflessioni finali

L'analisi condotta sul tema delle sigarette elettroniche e l'indagine dei casi studio disponibili in letteratura (vedi Capitoli 1 e 2) hanno fatto emergere come i Big Data, ed in particolare l'enorme quantitativo di dati generati quotidianamente sui social network, possano essere utilizzati in vari ambiti e con molteplici finalità conoscitive.

Il presente lavoro si è focalizzato sull'uso di una piattaforma sociale in particolare: Twitter. La ragione risiede nei vantaggi che esso offre e che sono riportati al paragrafo 3.2. L'applicazione dell'analisi dell'utilizzo dei social network ad un caso concreto ha permesso di individuarne meglio le possibilità di sfruttamento e di ricavarne interessanti riflessioni.

Per quanto riguarda, in particolare, la sua utilità a fini di marketing, sono emersi alcuni spunti d'applicabilità che verranno presentati di seguito. Particolare attenzione sarà, inoltre, prestata alle opportunità che tale analisi offre nell'ambito del marketing sociale, in quanto il caso studio condotto riguarda una tematica di sanità pubblica.

- <u>Progettazione di nuovi prodotti o di una campagna di marketing</u>

Considerando quanto sia fondamentale un'analisi dell'ambiente nel momento in cui si progetta una campagna di marketing o un nuovo prodotto, la possibilità di ottenere, attraverso i social network, una fotografia molto dettagliata e disponibile in tempo reale di quello che la società pensa riguardo ad uno specifico argomento può sicuramente essere di grande aiuto. Queste informazioni, se estratte in modo adeguato e combinate con dati quantitativi e tecniche tradizionali, possono essere utilizzate per ottimizzare le proprie decisioni, in termini creativi ed operativi, e per ottenere nel modo più efficace ed efficiente l'obiettivo prefissato.

Leggere e analizzare i messaggi pubblicati nei social network può rivelarsi utile alle imprese per comprendere cosa i loro consumatori cercano, per studiarne gusti e preferenze, così da creare un'offerta *customer-tailored* che meglio risponde alle esigenze e ai desideri della clientela; il tutto senza tediare il consumatore chiedendogli direttamente una descrizione di ciò che vorrebbe, ma deducendolo semplicemente da ciò che "dice" in Rete ai suoi contatti virtuali. Un suggerimento non sollecitato ma ascoltato è molto più di valore; come recita il Cluetrain manifesto, i mercati sono conversazioni,

ed è quindi fondamentale saper ascoltare il consumatore per poterlo comprendere, soddisfare e farsi a sua volta ascoltare.

Se si esaminano i commenti e le opinioni disponibili online per ricavarne idee per i prodotti futuri, allora in un certo senso i post online creano il futuro e, quindi, pure lo prevedono, come una profezia che si auto adempie (Ceron, Curini, Iacus, 2014).

Già molte aziende sono state capaci di sfruttare la Rete per facilitare il contatto col consumatore e creare un dialogo bidirezionale, elevando il consumatore da attore passivo a *prosumer*. Attraverso l'analisi degli *user-generated content* pubblicati nei social network ogni utente che esprime una sua opinione sul Web ai suoi amici in Facebook o ai *follower* in Twitter, viene considerato come un soggetto capace di apportare idee creative che l'azienda può sviluppare, anche se il suo contributo alla realtà aziendale rimane involontario ed inconsapevole.

Le informazioni che un'azienda può ottenere da uno studio dei messaggi che appaiono sulle piattaforme digitali possono essere utilizzate sia come indizi nel corso della progettazione dei prodotti, sia come spunti creativi nella fase di sviluppo di una campagna di marketing, allo scopo di rendere le azioni dell'impresa il più possibile coerenti con l'ambiente in cui vengono implementate.

Emblematico, in questo senso, è il caso dell'azienda canadese Rocket XL che per progettare la sua campagna di marketing ha cercato di scoprire attraverso i social network quali atleti fossero maggiormente apprezzati dai teenager in quel momento, le cose che essi amavano collezionare ed il tipo di intrattenimento che cercavano. Sulla base di queste informazioni è stata creata una campagna pubblicitaria il cui successo fu' tale da far aumentare le vendite di ben il 24% in un semestre.

Una buona comprensione dello scenario di riferimento permette di avere un'offerta migliore, una comunicazione più adatta e di creare dei significati culturali in cui il consumatore possa riconoscersi.

Un'azienda che sa comprendere e catturare i messaggi latenti e decodificarli, dando risalto a qualcosa che già esiste, ma che ancora non è un bisogno o desiderio determinato, sarà in grado di parlare lo stesso linguaggio del tempo e quindi di farsi ascoltare e capire dai consumatori.

Soprattutto, per progettare un'azione di marketing non convenzionale, tale comprensione approfondita della realtà diventa fondamentale.

Inoltre, l'analisi dei social network, grazie alla capacità di previsione che sembra avere, potrebbe facilitare le aziende nel sintonizzarsi sugli scenari futuri ed individuare i nuovi soggetti, eventi e luoghi che saranno di tendenza e che potranno essere adottati per delle azioni di marketing.
Ancora, durante la fase di progettazione di una campagna può essere utile una network analysis che metta in evidenza quali sono gli influenzatori e i diffusori sociali riguardo determinati temi, per poter pianificare degli interventi specifici diretti a questa categoria che riesce a formare l'opinione della massa, come ad esempio il cosiddetto *seeding localizzato* delle campagne di comunicazione virali (Cova, Giordano, Pallera, 2008).

- Analisi dell'immagine di marca e della sua evoluzione nel tempo

Ascoltando cosa il consumatore stesso "dice" spontaneamente nelle piattaforme online, è possibile capire l'immagine di marca come è effettivamente percepita. Ad esempio, un *word cloud* che metta in evidenza le parole utilizzate più frequentemente nei tweet in cui compare il nome del brand, potrebbe aiutare a comprendere cosa le persone dicono quando parlano del marchio aziendale. Questo può essere un ottimo modo per conoscere quali sono i punti di forza da evidenziare nella futura campagna di comunicazione, quali sono i lati negativi da migliorare, cosa maggiormente colpisce l'immaginario pubblico e su cui, dunque, bisogna puntare di più durante le azioni di comunicazione e di marketing.
Monitorare la variazione nel tempo di come le persone "parlano" in Rete dell'azienda, di un suo marchio o di un suo prodotto permetterebbe di capire cosa e come influenza l'opinione dei consumatori, o di comprendere come un evento specifico, sotto il controllo aziendale o meno, abbia influenzato quello che le persone pensano.

Tenendo presente, poi, che qualunque attività aziendale, non solo di comunicazione, incide sull'immagine che il consumatore si crea, appare fondamentale riuscire a seguirne l'andamento per individuare cosa la influenza e sapere come agire per migliorarla; è proprio questa una delle principali potenzialità dell'analisi dell'utilizzo dei social network sfruttabile dalle aziende.

- Valutazione di una campagna di marketing

Confrontare i messaggi pubblicati nei social network prima e dopo una campagna di marketing è un buon metodo per valutarne l'efficacia e comprendere come è stata percepita e ha influenzato i consumatori, attività sempre complessa e problematica.
Sebbene le aziende possano utilizzare indicatori quantitativi oggettivi, come quelli economico-finanziari e di mercato, rimane sempre l'incognita su quale sia stato l'effetto, soprattutto sull'immagine aziendale percepita, generato direttamente dalla campagna e non da altri fattori contingenti. Con un'analisi qualitativa degli *user-generated content* che commentano proprio quella determinata campagna, potrebbe risultare più semplice scoprire la relazione causa-effetto e comprendere come determinate scelte comunicative abbiano influenzato l'immagine di marca che i consumatori si sono creati e le loro scelte d'acquisto. Se poi la campagna comunicativa stessa proponesse un preciso hashtag con cui etichettare le conversazioni che la riguardano, la reazione del pubblico sarebbe ancora più semplice da indagare.
La copertura netta e lorda di un annuncio pubblicitario, la penetrazione del media adottato, il *click through rate* o il numero di visualizzazioni di un banner sono parametri quantitativi, ma non sono sufficienti per valutare come una campagna di comunicazione ha modificato l'immagine di marca percepita. Ciò risulta fondamentale soprattutto quando lo scopo della campagna è proprio quello di associare il marchio a determinati valori ed emozioni, rafforzare la personalità del brand, modificare determinate credenze circa le qualità aziendali, dar vita ad un immaginario collettivo o creare dei significati da immettere nel sistema culturale e simbolico e a cui poi il consumatore potrà decidere di aderire attraverso l'acquisto.
A questo fine si rendono necessarie analisi qualitative oltre che quantitative, poiché un miglioramento nella percezione del brand, dell'azienda o del prodotto da parte del consumatore non sempre si riflette in un aumento di vendite, o non almeno nel breve periodo.
Fondere ciò che emerge dai social network con altri dati aziendali a disposizione permetterebbe di ottenere una visione molto più completa per valutare un'azione di marketing.

Inoltre, potendo analizzare i commenti e le opinione espresse online non solo al termine, ma anche durante lo svolgimento della campagna di marketing, è possibile comprendere nel mentre alcuni fattori chiave che potrebbero essere migliorati per incrementarne l'efficacia. Proprio per questo, la disponibilità in tempo reale è una delle caratteristiche che ne determina maggiormente l'utilità.

- Confronti con i competitor

Ancora, dai social network è possibile estrarre dati relativi a quanto e come un marchio viene paragonato ad un altro: ad esempio, il conteggio di quante volte il nome del principale competitor compare nei tweet che contengono il nome del proprio marchio potrebbe aiutare l'azienda a capire quanto è elevata la competizione. Con un'analisi dei testi si potrebbero poi scoprire anche quali sono gli elementi su cui si viene principalmente confrontati.

- Analisi comportamento d'acquisto

L'analisi dei social network può, inoltre, essere una modalità per capire le motivazioni del comportamento d'acquisto dei consumatori fino quasi a prevederlo; la possibiltà di anticipare fenomeni sociali attraverso i tweet o i contenuti online è stata, infatti, dimostrata in precedenza (Liviu e Miahela, 2011; Asur e Hiberman, 2010; Chaudhari, 2012; Ciulla et al., 2012; Achrekar et al., 2011; Sanders e den Bosh, 2013; Tjong Kim Sang e Bos, 2012, UN Global Pulse, 2013, Voices from the Blogs, 2013).
Si potrebbe, ad esempio, spiegare un calo improvviso delle vendite analizzando gli *user-genereted content* relativi all'impresa, al marchio o al prodotto pubblicati nel periodo nel quale è stata registrata la diminuzione.
Se poi un preciso arco temporale fosse di particolare interesse, ad esempio perché sono state riscontrate nella realtà delle anomalie rispetto al consueto comportamento di consumo della clientela, è possibile condurre analisi linguistiche del contenuto dei messaggi pubblicati nei social network per indagare più approfonditamente quale sia stata la causa di quanto rilevato. Quella condotta col supporto di PC-ACE, e presentata al paragrafo 4.2.3.4, ne costituisce un efficace esempio.

Mettere in luce i network semantici sottostanti al corpo di commenti analizzati potrebbe anche far emergere aspetti interessanti che altrimenti rimarrebbero nascosti. Ad esempio, evidenziare le tematiche maggiormente affrontate quando si digita il nome del marchio e come queste evolvono nel tempo potrebbe portare all'individuazione di un preciso momento in cui l'attenzione dell'opinione pubblica è stata rivolta verso un determinato tema, e rintracciarne nella realtà la motivazione. Quest'operazione potrebbe rivelarsi particolarmente interessante soprattutto quando dall'analisi dei temi trattati nei tweet ne emergono di negativi o non inerenti all'azienda e ai suoi prodotti.
In generale indagini di base possono essere sufficienti nel momento in cui s'intende monitorare la situazione, quando invece emergono andamenti anomali o ci sono motivazioni particolari che fanno emergere delle domande, si possono condurre delle analisi linguistiche più specifiche per trovare le risposte ai quesiti emersi.

- <u>Metodo complementare ai tradizionali sondaggi</u>

Tradizionalmente, per comprendere l'immagine aziendale, la soddisfazione della clientela e la sua fedeltà, o per valutare come è stata percepita una campagna di marketing, si utilizzano sondaggi, interviste o *focus group*. Leggere direttamente i pensieri che le persone pubblicano online può rivelarsi un enorme risparmio, sia in termini di denaro che di tempo, e permetterebbe di avere un campione forse meno rappresentativo del target d'interesse, ma sicuramente di gran lunga più numeroso.
L'analisi dell'utilizzo dei social network è paragonabile ad un sondaggio svolto in modo passivo, poco costoso, che non si limita a domande predefinite, ma registra ciò di cui le persone stesse decidono di parlare e il modo che usano per farlo. Non essendoci domande precise a cui l'intervistato deve rispondere si evita il pericolo di avere risposte mancanti, indotte o imprecise perché basate su esperienze o impressioni passate che l'intervistato non ricorda bene; sono le persone stesse che parlando di ciò che vogliono e come vogliono rispondono passivamente ai quesiti del ricercatore.
Analogamente a quanto avviene nei *focus group* si lascia parlare l'individuo che, esprimendo il suo pensiero liberamente, fa emergere la sua opinione senza obbligarlo, attraverso interviste, a rientrare in categorie predefinite.

Se poi si considera che i *focus group*, generalmente condotti nella fase di valutazione di una campagna di marketing, prevedono interviste non strutturate e non individuali, ma rivolte ad un gruppo di persone, per far in modo che la discussione tra esse faccia emergere e modificare in itinere i loro pensieri ed opinioni (Bryman e Bell, 2011), allora appare chiaro come anche attraverso l'analisi dei social network si potrebbero ottenere risultati simili. Pubblicare online un proprio pensiero genera, infatti, una discussione che permette di creare, col contributo di molti, un pensiero collettivo sull'argomento (vedi concetto di "cervello collettivo" proposto da Ceron, Curini e Iacus esposto al paragrafo 2.6, pagina 59).

Se però i sondaggi offline forniscono un'immagine statica, e quindi in un certo senso limitata, della realtà, attraverso i social network è invece possibile cogliere in tempo reale l'evoluzione delle opinioni e delle preferenze, sempre più variabili e rapide nella società attuale.

Il consiglio non è quello di sostituire un metodo con l'altro, ma di affiancarli per ottenere risultati in breve tempo e a basso costo attraverso l'analisi degli *user-generated content*, che poi i tradizionali sondaggi o *focus group* possono confermare e rendere più precisi.

Spesso quando i decisori pubblici o i manager aziendali devono compiere le loro scelte, necessitano di informazioni il più possibile aggiornate, e grazie ai social network è possibile avere una visione generale della situazione in tempo reale e delle basi per prevederne l'immediata evolversi.

Ovviamente, però, è difficile avere le stesse certezze ottenibili con i metodi tradizionali, soprattutto perché non se ne sono ancora scoperte tutte le potenzialità e tutti i limiti.

Relativamente all'utilizzo dell'analisi dei messaggi disponibili nelle piattaforme sociali per indagare la percezione collettiva e le abitudini dei cittadini riguardo a temi legati alla salute, è importante precisare che, soprattutto in questo caso, le informazioni in *real-time* non devono avere l'obiettivo di rimpiazzare le metodologie d'analisi tradizionalmente utilizzate dai decisori pubblici, ma di integrarle e di suggerire eventuali aeree che necessitano di maggiori indagini, così da rendere gli output delle tradizionali ricerche più precisi ed esaustivi. Le informazioni ottenibili grazie a questa sinergia potrebbero rivelarsi risorse estremamente utili da tenere in considerazione nel

momento in cui i decisori pubblici devono fare delle scelte che incidono o che vogliono incidere sulla salute dei cittadini.

- Progettazione e valutazione di una campagna di marketing sociale

Particolarmente interessante appare l'analisi dei social network nel campo del marketing sociale, definito da Kotler, Roberto e Lee (2002) come "l'utilizzo dei principi e delle tecniche del marketing per influenzare un gruppo target ad accettare, rifiutare, modificare o abbandonare un comportamento in modo volontario, allo scopo di ottenere un vantaggio per i singoli, i gruppi o la società nel suo complesso". L'obiettivo non è quindi portare un consumatore all'acquisto di un prodotto, ma convincere il cittadino di un'idea che lo porti ad un cambiamento cognitivo (per diffondere consapevolezza e conoscenza riguardo a un problema ed alle sue possibili soluzioni), d'azione (incentivandolo a prendere determinate decisioni piuttosto che altre), di comportamento (portandolo ad abbandonare certe abitudini dannose per assumerne di più sane) e di valori (per modificare opinioni radicate, passaggio difficile ma necessario per orientare le scelte di comportamento individuale) al fine di ottenere un vantaggio in termini di benessere e salute per l'individuo o per la collettività.

Fase di fondamentale importanza prima dello sviluppo del piano è l'analisi dell'ambiente, ossia delle forze esterne alla campagna di cambiamento sociale che ostacolano o favoriscono la capacità di sviluppare e mantenere un'influenza adeguata sugli utenti designati. Grazie ai Big Data provenienti dai social network è possibile rilevare in tempo reale, o addirittura anticipare, i cambiamenti dell'ambiente in cui la campagna si inserisce. Ciò permette di creare un programma perfettamente rispondente alle esigenze del target e di adattarlo durante il suo svolgimento in base ai feedback ricevuti.

La valutazione di una campagna sociale è la raccolta sistematica e l'analisi di informazioni sulle attività, i risultati ed il loro impatto, condotte per aumentarne l'efficacia e per essere in grado di prendere decisioni migliori riguardo al fenomeno che questi programmi mirano ad influenzare (Patton, 1982 in Bruce e Tiger, 2013).

Può essere effettuata su tre livelli: processo (se il target è stato raggiunto in modo desiderato attraverso le attività della campagna), risultato (se ha indotto il target ad un

cambiamento cognitivo o comportamentale) ed impatto (se il nuovo comportamento adottato ha portato all'obiettivo desiderato in termini di miglioramento della salute e del benessere della popolazione, come ad esempio una riduzione della mortalità infantile).

La valutazione del marketing applicato a problemi sociali risulta ben più complessa di quella del marketing di un'impresa commerciale perché non può basarsi su alcun tipo di indicatore economico-finanziario.

Giudicare l'efficacia di tali attività non è semplice, poiché gli effetti sulla salute si manifestano spesso nel lungo periodo e possono essere generati da molte concause, anche estranee al progetto. L'impatto della campagna, poi, potrebbe essere piccolo, anche se comunque importante. Per esempio, si pensi ad una campagna di comunicazione ideata per convincere le persone a lavare le auto in modo da consumare meno acqua che abbia ottenuto ottimi risultati, in quanto è riuscita a convincere quasi tutti coloro che sono stati raggiunti dalla campagna. Se fosse giudicata sulla base della diminuzione di quantità di acqua consumata, questo tipo di valutazione non sarebbe adatta, poiché servirebbero metodi di misurazione capaci di registrare effetti di minore rilevanza (Hornik, Jacobs e Coffman, 2007 in Bruce e Tiger, 2013). L'analisi di quello che le persone dicono relativamente al *car-washing* sui social network nell'area in cui è stata diffusa la campagna potrebbe invece essere un metodo valido per stabilirne l'efficacia.

Tradizionalmente, al fine di valutare una campagna di marketing sociale, un insieme di individui rappresentante il target viene sottoposto a sondaggi ed interviste prima, dopo e durante la campagna e i risultati vengono successivamente tra loro confrontati, oppure si effettuano *focus group, rolling sample surveys, cohort studies* od osservazioni dirette sul campo. Con un'analisi degli *user-generated content* sulla tematica potrebbe invece essere molto più economico e veloce avere un'idea di come una campagna di comunicazione sociale abbia condotto ad un cambio di valori o di comportamento negli individui al fine di ottenere l'impatto prefissato.

Pare però giusto sottolineare come, anche se nel marketing sociale le potenzialità dell'analisi dei dati raccolti nei social network sembra essere quanto mai utile, in quest'ambito si acuiscono alcune delle limitazioni esposte al paragrafo 3.6: l'autocensura che gli utenti applicano a contenuti collegabili al campo sociale è sicuramente più elevata di quella riscontrabile nelle opinioni su marchi o prodotti;

inoltre la non rappresentatività dell'intera popolazione sulle piattaforme sociali analizzate potrebbe creare seri problemi qualora sui risultati dell'analisi dei social network si basassero poi decisioni d'interesse pubblico influenzanti la vita dell'intera comunità, e quindi anche soggetti che in Rete e nei social network sono sottorappresentati.

Dall'altro lato, programmi pubblici finanziati necessitano di dimostrare la loro efficacia, e trarrebbero enormi vantaggi dall'avere accesso a dati in tempo reale, o quasi, utilizzabili per programmare miglioramenti e permettere alle autorità di prendere decisioni migliori ed più efficienti nel più breve tempo possibile.

Quindi, l'affiancamento di metodologie tradizionali, come il sistema Passi, all'analisi delle modalità con cui la popolazione utilizza i social network, metodologia più veloce ed immediata, ma con un livello di precisione da migliorare, sembra essere una buona soluzione per ottenere una comprensione il più possibile approfondita del comportamento della collettività relativamente alla problematica analizzata, della percezione dei cittadini di un tema legato alla salute e dell'impatto che su di essi ha avuto una campagna di marketing sociale.

Sembra, in conclusione, che utilizzando i *digital smoke signals* che rivelano cambiamenti collettivi di comportamento, si possano ottenere informazioni utili da sfruttare sia per avviare campagne di comunicazione per la prevenzione di danni alla salute, sia per organizzare interventi mirati e progettati ad hoc per intervenire il prima possibile e nel modo migliore per risolvere danni causati da determinati comportamenti abituali assunti dalla popolazione. Il loro utilizzo da parte del settore pubblico potrebbe pertanto rivelarsi estremamente utile.

A sostegno di quanto affermato, c'è l'analisi condotta sui tweet riguardanti il tema delle sigarette elettroniche pubblicati nel primo quadrimestre 2014 ed estratti attraverso una Twitter API. Lo studio presentato, oltre ad aver dimostrato come gli avvenimenti reali si riflettano in determinati comportamenti online, ha permesso di arrivare alla conclusione che il fumo elettronico non è adottato in Italia come strumento per smettere di fumare (vedi nel dettaglio paragrafo 4.2.3). I dati reali, raccolti dai sistemi di sorveglianza nazionale, hanno confermato quanto emerso dallo studio dei messaggi pubblicati in Twitter, sottolineando come l'analisi dei tweet possa fornire un'immagine della società abbastanza realistica sulla quale far affidamento per prendere decisioni in ambito

pubblico. Le differenti tipologie d'analisi condotte sui testi estratti ne hanno evidenziato la capacità di rispondere a diverse tipologie di domande, a dimostrazione di come, per il settore della pubblica sanità, dati di questo tipo possano rivelarsi alquanto utili per progettare campagne di comunicazione conformi all'ambiente in cui vengono implementate e pertanto maggiormente efficaci.

Quanto precedentemente esposto dimostra come l'analisi dei dati provenienti dai social network sia promettente, anche per quanto attiene, nello specifico, al marketing.
Ulteriori ricerche sono necessarie per affinare le tecniche d'analisi dei dati disponibili sulle piattaforme digitali, al fine di renderne gli output sempre più completi ed affidabili Tuttavia le potenzialità emerse incoraggiano ad implementare questa tecnica d'analisi sia a livello aziendale, che a livello pubblico.

Bibliografia

Achrekar, H. et al., 2011. Predicting Flu Trends using Twitter data". *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, [pdf] Disponibile all'indirizzo <http://www.cs.uml.edu/~bliu/pub/CPNS_2011.pdf> [ultimo accesso 28 marzo 2014]

Achrekar, H. et al., 2012. "Twitter improves seasonal influenza prediction". *Proceedings of the Fifth Annual International Conference on Health Informatics*, [pdf] Disponibile all'indirizzo <http://www.cs.uml.edu/~bliu/pub/healthinf_2012.pdf> [ultimo accesso 28 marzo 2014]

Achrekar, H., 2013. "Online Social Networks Flu Trend Tracker - A Novel Sensory Approach to Predict Flu Trends". *Biomedical Engineering Systems and Technologies*, [pdf] Disponibile all'indirizzo <http://www.cs.uml.edu/~hachreka/SNEFT/images/SNEFT_Harsh.pdf> [ultimo accesso 28 marzo 2014]

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B, 2011. "Predicting flu trends using twitter data" *Proceedings of the International Workshop on Cyber-Physical Networking Systems* [pdf] Disponibile all'indirizzo <http://www.cs.uml.edu/~hachreka/SNEFT/images/SNEFT_Harsh.pdf> [ultimo accesso 28 marzo 2014]

Adamic, L., Buyukkokten, O., Ada, E., 2012. "A social network caught in the Web". *First Monday*, [online] Disponibile all'indirizzo <http://firstmonday.org/ojs/index.php/fm/article/view/1057/977> [ultimo accesso 28 marzo 2014]

Akay, A., Martinsson, P., 2009. "Sundays Are Blue: Aren't They? The Day-of-the-Week Effect on Subjective Well-Being and Socio-Economic Status". Working Papers in Economics 397, University of Gothenburg, Department of Economics [online] Disponibile all'indirizzo <http://ftp.iza.org/dp4563.pdf> [ultimo accesso 28 marzo 2014]

Ante, S., 2009. "Content-Search Deals Make Twitter Profitable". *Business Week*, [online] Disponibile all'indirizzo < http://www.businessweek.com/stories/2009-12-21/content-search-deals-make-twitter-profitablebusinessweek-business-news-stock-market-and-financial-advice> [ultimo accesso 22 febbraio 2014]

Asur, S., Huberman, B., 2010. "*Predicting the future with social media*". Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, Pagine 492-499. [pdf] Disponibile all'indirizzo <http://arxiv.org/pdf/1003.5699.pdf> [ultimo accesso 28 marzo 2014]

Asur, S., Huberman, B., 2010. "Predicting the future with Social Media". *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* Volume 1, Pagine 492-499, [pdf[Disponibile all'indirizzo < http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf> [ultimo accesso 28 marzo 2014]

Avellina, G., 2013. "COMUNICARE O NO LA CRISI SU TWITTER DURANTE PRESUNTI ATTENTATI COME A BOSTON?". *GQ.com*, [online] Disponibile all'indirizzo <http://www.gqitalia.it/hi-tech/articles/2013/4/comunicare-o-no-la-crisi-su-twitter-durante-presunti-attentati-come-a-boston> [ultimo accesso 20 marzo 2014]

Bakshy, E. et al., 2012. "The Role of Social Networks in Information Diffusion". Proceedings of ACM WWW 2012, Lyon, France, [pdf] Disponibile all'indirizzo <http://arxiv.org/pdf/1201.4145v2.pdf> [ultimo accesso 18 marzo 2014]

Bancone, C., 2013. "L'attentato di Boston. La rete racconta". *cinziabancone.it* [blog] 18 Aprile 2013. Disponibile all'indirizzo <http://www.cinziabancone.it/2013/04/18/lattentato-di-boston-la-rete-racconta.html> [ultimo accesso 22 marzo 2014]

Barbeau, A. M., Burda, J. e Siegel, M., 2013. "Perceived efficacy of e-cigarettes versus nicotine replacement therapy among successful e-cigarette users: a qualitative approach". *Addiction Science & Clinical Practice,* [pdf] Disponibile all'indirizzo <http://www.ascpjournal.org/content/8/1/5> [ultimo accesso 17 aprile 2014]

Baresch, B. et al., 2011. *"Friends Who Choose Your News: An analysis of content links on Facebook"*. Austin, Texas, Aprile 2011. International Symposium on Online Journalism. [pdf] Disponibile all'indirizzo <https://online.journalism.utexas.edu/2011/papers/Baresch2011.pdf> [ultimo accesso 28 marzo 2014]

Barton, D., 2013. "Chi trova i Big Data trova un tesoro". *Il sole 24 ore*, [online] Disponibile all'indirizzo <http://www.mckinsey.it/idee/rassegna_stampa/chi-trova-i-big-data-trova-un-tesoro.view> [ultimo accesso 20 febbraio 2014]

Batty, M., 2012. "Smart cities, big data". *Environment and Planning B: Planning and Design*, [pdf] Disponibile all'indirizzo

<http://www.envplan.com/epb/editorials/b3902ed.pdf> [ultimo accesso 20 febbraio 2014]

BBC, 2011. "Mobile phones help to target disaster aid, says study". *BBC News Technology*, [online] Disponibile all'indirizzo <http://www.bbc.co.uk/news/technology-14761144> [ultimo accesso 20 febbraio 2014]

Beevolve, 2012. "An Exhaustive Study of Twitter Users Across the World". *beevolve.com*, [online] Disponibile all'indirizzo <http://www.beevolve.com/twitter-statistics/> [ultimo accesso 12 marzo 2014]

Beevolve, 2012. "An Exhaustive Study of Twitter Users Across the World" *beevolve.com* [online] Disponibile all'indirizzo <http://www.beevolve.com/twitter-statistics/>

Bengtsson, L. et al., 2011. "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti". *PLOS Medicine*, [online] Disponibile all'indirizzo <http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1001083> [ultimo accesso 20 febbraio 2014]

Beraldo, D., 2012. "A lezione di Twitter. Reti dalla Rete: brevi spunti sociologici su cosa si può fare e perché farlo". *etnografiadigitale.it*, [online] Disponibile all'indirizzo < http://www.etnografiadigitale.it/2012/05/a-lezione-di-twitter-reti-dalla-rete-brevi-spunti-sociologici-su-cosa-si-puo-fare-e-perche-farlo/> [ultimo accesso 18 marzo 2014]

Bertrand, G., 2013. "*Social media research: developing a trust metric in the social age*". International Journal of Market Research, vol. 55 Issue 3, pag. 333-335

Bhatt, R., Chaoji, V., Parekh, R., 2010. "Predicting Product Adoption in Large-Scale Social Networks". *Proceedings of the 19th ACM international conference on Information and knowledge management*, pagine 1039-1048, [pdf] Disponibile all'indirizzo <http://www.cs.iastate.edu/~parekh/papers/CIKM2010.pdf> [ultimo accesso 28 marzo 2014]

Bialous, S. A, Sarma, L., 2014."Electronic cigarettes and smoking cessation: a quandary?". *The Lancet*, Volume 383, Issue 9915, Pagine 407-408 [pdf]

Bollen, J., Mao, H., Zeng, X., 2010. "Twitter mood predicts the stock market". *Journal of Computational Science*, [pdf] Disponibile all'indirizzo <http://arxiv.org/pdf/1010.3003&> [ultimo accesso 13 marzo 2014]

Borello, G., 2014. “Usa, entro dieci anni il sorpasso delle sigarette elettroniche sul tabacco”. *Il sole 24 ore* [online] Disponibile all’indirizzo <http://mobile.ilsole24ore.com/solemobile/main?fn=srchFd&id=SEARCH%2FNEWS24%2FAB55xQGB> [ultimo accesso 14 maggio 2014]

Bouryal, K., Horwitz, R., 2012. “Exclusive Agreement Creates “Nielsen Twitter TV Rating,” the Definitive Reach Metric for Social TV Audience Measurement and Analytics”. *Nielsen.com,* [online] Disponibile all’indirizzo <http://www.nielsen.com/content/corporate/us/en/press-room/2012/nielsen-and-twitter-establish-social-tv-rating.html> [ultimo accesso 9 marzo 2014]

Bruce, S., Tiger, M., 2013. “A Review of Research Relevant to Evaluating Social Marketing Mass Media Campaigns”. *NC Clean Water Education Partnership Publication,* [pdf] Disponibile all’indirizzo <http://www.nccwep.org/pdf/evaluating-social-marketing-mass-media-campaigns.pdf> [ultimo accesso 16 aprile 2014]

Bryman, A, Bell, E., 2011. *Business Research Methods.* 3° Edizione. Oxford: Oxford University Press. [online] Disponibile all’indirizzo <http://books.google.it/books?hl=it&lr=&id=YnCcAQAAQBAJ&oi=fnd&pg=PP2&dq=Bryman+e+Bell,+2011&ots=IhtqWrjF0A&sig=Gr3vCNsVfW-iD6cQKMm7CW4nKyk#v=onepage&q&f=false> [ultimo accesso 24 aprile 2014]

Bullen, C. et al., 2013. “Electronic cigarettes for smoking cessation: a randomised controller trial”. *Lancet,* [online] Disponibile all’indirizzo < http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(13)61842-5/fulltext> [ultimo accesso 17 aprile 2014]

Buran, G., 2013. “Battere i concorrenti coi Big data”.*Il sole 24 ore,* [online] Disponibile all’indirizzo <http://www.ilsole24ore.com/art/tecnologie/2013-10-20/battere-concorrenti-data-085252.shtml?uuid=AbPiTzvI> [ultimo accesso 20 febbraio 2014]

Burns, A., et al., 2012. *“#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Flood”*. Brisbane:ARC Centre of Excellence for Creative Industries and Innovation, 2012. [pdf] Disponibile all’indirizzo <http://www.cci.edu.au/floodsreport.pdf> [ultimo accesso 4 marzo 2014]

Butler, D., 2013. “When Google got flu wrong”. *Nature,* [online] Disponibile all’indirizzo <http://www.nature.com/news/when-google-got-flu-wrong-1.12413> [ultimo accesso 28 marzo 2014]

Buzzetti, E., 2014. “Sigarette elettroniche, la tassa fa paura”. *Passione Tecno,* [online] Disponibile all’indirizzo <http://www.passionetecno.com/2014/04/01/sigarette-elettroniche-la-tassa-fa-paura-2789> [ultimo accesso 17 aprile 2014]

Caliandro Aka Ali, A., 2013. “A lezione di Twitter. Reti dalla Rete: brevi spunti sociologici su cosa si può fare e perché farlo”. *Centrostudietnografiadigitale.it,* [online] Disponibile all’indirizzo <http://www.etnografiadigitale.it/2012/05/a-lezione-di-twitter-reti-dalla-rete-brevi-spunti-sociologici-su-cosa-si-puo-fare-e-perche-farlo/> [ultimo accesso 28 marzo 2014]

Cameron, AM. Et al., 2013. “Social Media and Organ Donor Registration: the economic value of user-generated: The Facebook Effect”. *American Journal of Transplantation,* [online] Disponibile all’indirizzo <http://onlinelibrary.wiley.com/doi/10.1111/ajt.12312/pdf> [ultimo accesso 9 marzo 2014]

Caponnetto, P. et al., 2013. “Efficiency and Safety of an electronic cigarette (ECLAT) as Tobacco Cigarettes Substitute: A Prospective 12-Month Randomized Control Design Study”. *PLoS ONE,* [online] Disponibile all’indirizzo <http://www.plosone.org/article/info:doi/10.1371/journal.pone.0066317> [ultimo accesso 17 aprile 2014]

Castillo, C., Mendoza, M., Poblete, B., 2011. “Information Credibility on Twitter”. *Proceedings of the 20th International Conference on World Wide Web,* Pagine 675-684 [pdf] Disponibile all’indirizzo <http://www.ra.ethz.ch/cdstore/www2011/proceedings/p675.pdf> [ultimo accesso 18 marzo 2014]

Cavallo, A., Maggio 2011. “BPP and PriceStats”. *The Billion Prices Project @ MIT,* [online] Disponibile all’indirizzo <http://bpp.mit.edu/bpp-and-pricestats/> [ultimo accesso 20 febbraio 2014]

Center of Disease Control and Prevention, 2014. *“Best Practices for Comprehensive Tobacco Control Programs—2014” [online] Disponibile all’indirizzo <http://www.cdc.gov/tobacco/stateandcommunity/best_practices/index.htm>* [ultimo accesso 17 aprile 2014]

Ceron, A., 2012. “Every tweet counts? How sentiment analysis of social networks can improve our knowledge of citizens’ policy preferences. An application to Italy and France”. *UNIMI-Research Papers in Economics, Business, and Statistics,* [online] Disponibile all’indirizzo <https://www.academia.edu/2865844/Every_tweet_counts_How_sentiment_analysis_of

_social_networks_can_improve_our_knowledge_of_citizens_policy_preferences._An_application_to_Italy_and_France> [ultimo accesso 22 marzo 2014]

Ceron, A., Curini, L. e Iacus, S.M., 2014. *"Social media e Sentiment analysis"*. Milano: Springer per l'Innovazione

Chaudhari, G., 2012. "*Twitter Data Analysis*". M.Tech. Seminar Report. Bombai, 12 Aprile 2012

Chiusi, F., 2014. "Quello che il Big Data (ancora) non dice". *Wired*, [online] Disponibile all'indirizzo < http://www.wired.it/internet/social-network/2014/04/02/quello-che-il-big-data-ancora-non-dice/?utm_source=twitter.com&utm_medium=marketing&utm_campaign=wired> [ultimo accesso 26 marzo 2014]

Chunara, R., Andrews, J., and Brownstein, J., 2012. "Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak". *American Journal of Tropical Medicine and Hygiene*. [online] Disponibile all'indirizzo <http://www.ajtmh.org/content/86/1/39.abstract> [ultimo accesso 20 febbraio 2014]

Chung, J., and Mustafaraj, E., 2011. 'Can collective sentiment expressed on twitter predict political elections?'. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA [pdf] Disponibile all'indirizzo < http://www.christopia.net/data/school/2011/Fall/social-media-mining/project_proposal/sources/chung-2011.pdf> [ultimo accesso 28 marzo 2014]

Ciulla et al., 2012. "Beating the news using Social Media: the case study of American Idol". *EPJ Data Science*, [online] Disponibile all'indirizzo <http://www.epjdatascience.com/content/1/1/8> [ultimo accesso 4 marzo 2014]

Clozel, L., 2014. "E-cigarette firms targeting young people, lawmakers say". *Los Angeles Times*, [online] Disponibile all'indirizzo <http://www.latimes.com/nation/la-na-congress-ecigarettes-study-20140415-story.html#axzz2ywMSzHc0> [ultimo accesso 14 maggio 2014]

Cogburn, D., Espinoza-Vasquez, F., 2011. "From Networked Nominee to Networked Nation: Examining the Impact of Web 2.0 and Social Media on Political Participation and Civic Engagement in the 2008 Obama Campaign". *Journal of Political Marketing*, [online] Disponibile all'indirizzo *<http://www.tandfonline.com/doi/full/10.1080/15377857.2011.540224#.UvpPMfl5PuQ* > [ultimo accesso 28 marzo 2014]

Cook, S. et al., 2011. “Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic”. *PLoS ONE,* [pdf] Disponibile all’indirizzo <http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0023610&representation=PDF> [ultimo accesso 8 marzo 2014]

Cooper, C. P. et al., 2005. “Cancer Internet Search Activity on a Major Search Engine, United States 2001-2003”. *Journal of Medical Internet Research,* [online] Disponibile all’indirizzo <http://www.jmir.org/2005/3/e36/> [ultimo accesso 2 marzo 2014]

Cosenza, V., 2013. “State of the Net”. *Slideshare.net,* [slide] Disponibile all’indirizzo < http://www.slideshare.net/Blogmeter/sotn-2013-le-conversazioni-e-gli-umori-degli-italiani-in-rete?ref=http://vincos.it/2013/06/01/state-of-the-net-2013-parte-1-lo-scenario-italiano/> [ultimo accesso 5 giugno 2014]

Cova, B., Giordano, A., Pallera, M., 2012. “*Marketing non-convenzionale. Viral, guerrilla, tribal, societing e i 10 principi fondamentali del marketing postmoderno”* Milano: Il sole 24 ore libri

Crawford, K., 2009. “Following you: disciplines of listening in social media”. *Continuum,* [online] Disponibile all’indirizzo <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Crawford%20Continuum%20following%20you_0.pdf> [ultimo accesso 18 marzo 2014]

Cressey, D., 2014. “E-cigarettes affect cells”. *Nature* [online] Disponibile all’indirizzo <http://www.nature.com/news/e-cigarettes-affect-cells-1.15015> [ultimo accesso 14 maggio 2014]

Culotta, A., 2010. “Detecting influenza outbreaks by analyzing Twitter messages”. *Knowledge Discovery and Data Mining Workshop on Social Media Analytics,* [pdf] Disponibile all’indirizzo <http://snap.stanford.edu/soma2010/papers/soma2010_16.pdf> [ultimo accesso 4 marzo 2014]

De Choudhary, A. et al., 2012. *“Social Media Evolution of the Egyptian Revolution”*. Communications of the ACM, [pdf] disponibile all’indirizzo <http://cucis.ece.northwestern.edu/publications/pdf/ChoHenLee12.pdf> [ultimo accesso 28 marzo 2014]

De Choudhury, M., Ringel Morris, M., White, R. W., 2014. “*Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media”* Microsoft Research Publication, [pdf] Disponibile all’indirizzo

<http://research.microsoft.com/pubs/208026/health_search_chi2014.pdf> [ultimo accesso 2 marzo 2014]

Dearing, JW., Kreuter, MW., 2010. “Designing for diffusion: how can we increase uptake of cancer communication innovations?”. *Patient Education and Consulting*, [online] Disponibile all’indirizzo <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000559/> [ultimo accesso 12 marzo 2014]

Dedagroup, 2012. “*DEDAGROUP HIGHLIGHTS. BIG DATA: riconoscerli, gestirli, analizzarli*” [pdf] Dedagroup. Disponibile all’indirizzo <http://www.ecos2k.it/allegati/BigData.pdf> [ultimo accesso 20 febbraio 2014]

Dello Iacovo, L., 2013. “Il futuro è già scritto in big data”.*Il sole 24 ore*, [online] Disponibile all’indirizzo <http://www.ilsole24ore.com/art/tecnologie/2012-09-21/futuro-scritto-data-155044.shtml?uuid=AbOQCOhG> [ultimo accesso 18 marzo 2014]

Dello Iacovo, L., 2013. “Social network sono i nuovi termometri del web”.Il sole 24 ore, [online] Disponibile all’indirizzo < http://www.ilsole24ore.com/art/tecnologie/2013-10-18/social-network-sono-nuovi-termometri-web-112828.shtml?uuid=ABVp0bX> [ultimo accesso 28 marzo 2014]

Dello Iacovo, L., 2013. “Startup, ecco chi fa soldi con i big (open) data”.Il sole 24 ore, [online] Disponibile all’indirizzo <http://www.ilsole24ore.com/art/tecnologie/2013-02-15/startup-ecco-soldi-data-175937.shtml?uuid=AbAaWqUH> [ultimo accesso 20 febbraio 2014]

Delmonte, R., 2004. “Strutture Sintattiche dall’Analisi Computazionale di Corpora di Italiano”. Apparso su Anna Cardinaletti, 2004, "Intorno all'Italiano Contemporaneo", Franco Angeli, Milano, pp.187-220. [pdf] Disponibile all’indirizzo <http://project.cgm.unive.it/resource/VIT/Browser-VIT/indices/papers/treeb-it.pdf> [ultimo accesso 16 maggio 2014]

Delmonte, R., 2004. “Text Understanding with GETARUNS for Q/A and Summarization” *Proceeding of the 2nd Workshop on Text Meaning and Interpretation*, Pagine 97-104 [pdf] Disponibile all’indirizzo http://delivery.acm.org/10.1145/1630000/1628288/p97-delmonte.pdf?ip=157.138.189.238&id=1628288&acc=OPEN&key=4DF843340132291B%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=333759396&CFTOKEN=60084380&__acm__=1399384737_d95b44ddb694dd7d6a08f9b5addb108f [ultimo accesso 12 maggio 2014]

Delmonte, R., Bristot, A., Tonelli, S., 2007.“VIT – Venice Italian Treebank: Syntactic and Quantitative Features”. Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories [pdf] Disponibile all’indirizzo <http://tlt07.uib.no/papers/3.pdf> [ultimo accesso 16 maggio 2014]

Efvi, 2014. “Manifesto”. efvi.eu [online] Disponibile all’indirizzo <http://www.efvi.eu/index.it.html#manifest> [ultimo accesso 17 aprile 2014]

Eisenberg, M. J. Et al., 2008. “Pharmacotherapies for smoking cessation: a meta-analysis of randomized controlled trials”, CMAJ [online] Disponibile all’indirizzo <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2443223/?tool=pubmed> [ultimo accesso 17 aprile 2014]

Ferro, E., Tosco, E., Tortone, C., 2009. “Marketing sociale e promozione della salute”, *fact sheet n°1 ASL TO3 Regione Piemonte* [pdf] Disponibile all’indirizzo <http://www.asrem.org/Corsi%20aggiornamento%20e%20convegni/comunicazone%20per%20la%20slute/DorS%20Marketing%20sociale%20e%20prevenzione%20della%20salute.PDF> [ultimo accesso 17 aprile 2014]

Fisher, D. et al., 2012. “Interaction with Big Data Analytics”. *ACM Interaction,* [online] Disponibile all’indirizzo < http://bundlr.com/clips/5005f922eb517d00020000dc> [ultimo accesso 20 febbraio 2014]

Franch, F., 2013. “Wisdom of the Crowds: 2010 UK Election Prediction with Social Media”. *Journal of Information Technology & Politics,* Volume 10, Issue 1

Franzosi, R., 2009. “Quantitative Narrative Analysis.” Capitolo 19 in: Williams, M., Vogt p., V., 2009. “Sage Handbook of Methodological Innovation”. Beverly Hills, CA: Sage.

Franzosi, R., De Fazio, G., Vicari, S., 2012. “Ways of Measuring Agency: An Application of Quantitative Narrative Analysis to Lynchings in Georgia (1875-1930)”. Sociological Methodology, Volume 42, Pagine 1-42

Franzosi, R., 2014. “Analytical Sociology and Quantitative Narrative Analysis: Explaining Lynchings in Georgia (1875-1930)”. In Manzo, G., 2014. “Analytical Sociology: Norms, Actions, and Networks”, Wiley & Son

Franzosi, R., 2004. “Content Analysis”. Capitolo 24 in Hardy, M, Bryman, A., 2004. “Handbook of Data Analysis”. London, Thousand Oaks, New Delhi e Singapore: Sage Publications

Franzosi, R., 1998. “Narrative Analysis-Or Why (And How) Sociologists Should be Interested in Narrative”. Annual Review of Sociology, Volume 24, Pagine 517-554

Franzosi, R., 2010. “Quantitative Narrative Analysis.” (Quantitative Applications in the Social Sciences). Beverly Hills, CA: Sage.

Frediani, C., 2013. “Cosa sono le rivendite autorizzate di tweet”. *Wired*, [online] Disponibile all’indirizzo <http://daily.wired.it/news/internet/2013/09/04/twitter-rivendite-autorizzate-manichetta-564573.html> [ultimo accesso 28 marzo 2014]

Frias-Martinez, V., 2012. *“Enhancing Public Policy Decision Making using Large-scale Cell Phone Data”*. [slide] Madrid: UN Global Pulse. Disponibile all’indirizzo <http://www.slideshare.net/unglobalpulse/unicefgpworkshopvanessa-friasmartinez> [ultimo accesso 20 febbraio 2014]

Fruchterman, J., 2011. “Issues with Crowdsourced Data Part 2”. *Beneblog: Technology Meets Society*, 28

Fu, K., Chan, C., 2013. “Analyzing Online Sentiment to Predict Telephone Poll Results”. *Cyberpsychology, Behavior, and Social Networking*, Volume 16, Numero 9.

Galea, G., 2013. “Social Graph of 8GCHP”. *Blog.gaudengalea.com*, [blog] 30 Giugno 2013. Disponibili all’indirizzo <http://blog.gaudengalea.com/2013/06/social-graph-of-8gchp/#more-100> [ultimo accesso 18 febbraio 2014]

Gayo-Avello, D., 2011. “Don’t turn social media into another ’literary digest’”. *The Communication of ACM*, [pdf] Disponibile all’indirizzo <http://di002.edv.uniovi.es/~dani/downloads/Social-Media-Literary-Digest-authors.pdf> [ultimo accesso 15 marzo 2014]

Gayo-Avello, D., 2012. *“A Balanced Survey on Election Prediction using Twitter Data”*[pdf] Disponibile all’indirizzo <http://arxiv.org/pdf/1204.6441.pdf> [ultimo accesso 15 marzo 2014]

Gelman, A., 2012. “What Too Close to Call Really Means”. *The New York Times*, [online] Disponibile all’indirizzo <http://campaignstops.blogs.nytimes.com/2012/10/30/what-too-close-to-call-really-means/?_php=true&_type=blogs&_r=0> [ultimo accesso 1 marzo 2014]

Gennimata, S. A. et al., 2013. “Acute effect of e-cigarette on pulmonary function in healthy subjects and smokers”. *Proceedings of the 2012 meeting of the European Respiratory Society* [online] Disponibile all’indirizzo

<https://www.ersnetsecure.org/public/prg_congres.abstract?ww_i_presentation=59718> [ultimo accesso 26 aprile 2014]

Ghose, A., Ipeirotis, P., 2008. "The EconoMining project at NYU: Studying the economic value of user-generated content on the internet". *Journal of Revenue and Pricing Management* Volume 8, Pagine 241–246

Ginsberg, J. et al., 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature, International weekly journal of science,* [online] Disponibile all'indirizzo <http://www.nature.com/nature/journal/v457/n7232/abs/nature07634.html> [ultimo accesso 20 febbraio 2014]

Grier, S., Bryant, C. A., 2005. *Social Marketing in Public Health.* Annual Reviews Public Health, Volume 24, Pagine 319-339 [pdf]

Guzzo, S., 2012. "Big Data: cosa sono e come vengono utilizzati in Italia". *blog.artera.it,* [blog] 11 Febbraio 2013. Disponibile all'indirizzo < http://blog.artera.it/curiosita-web/big-data-definizione-e-utilizzo-in-italia> [ultimo accesso 20 febbraio 2014]

Hajek, P., 2013. "Electronic cigarettes for smoking cessation". *The Lancet,*Volume 382, Issue 9905, Pagine 1614-1616 [pdf]

Heavner, K. et al., 2013. "Electronic cigarettes (e-cigarettes) as potential tobacco harm reduction products: Results of an online survey of e-cigarette users". *Tobacco Harm Reduction Working Paper 001* [pdf] Disponibile all'indirizzo <http://www.tobaccoharmreduction.org/wpapers/011v1.pdf> [ultimo accesso 17 aprile 2014]

Helbing, D., Balietti, S., 2011. *"From social data mining to forecasting socio-economic crises"*. The European Physical Journal, [pdf] Disponibile all'indirizzo <http://download.springer.com/static/pdf/144/art%253A10.1140%252Fepjst%252Fe2011-01401-8.pdf?auth66=1391876836_1c833adce7591097cc44eeab7cf4f0f1&ext=.pdf>

Hof, R., 2013. "IPO Filing: Twitter Still Losing Big Money Even As 2012 Revenues Tripled". *Forbes,* [online] Disponibile all'indirizzo < http://www.forbes.com/sites/roberthof/2013/10/03/revealed-in-ipo-filing-twitter-still-losing-big-money-even-as-2012-revenue-tripled-to-317-million/> [ultimo accesso 17 febbraio 2014]

Hopkins,D. J., King, G., 2010. “A Method of Automated Nonparametric Content Analysis for Social Science”. *American Journal of Political Science*, Volume 54, Numero 1, Pagine 229-247, [pdf]

Howard, A., 2013. “Untangling algorithmic illusions from reality in big data”*O’Reilly*, [online] Disponibile all’indirizzo <http://strata.oreilly.com/2013/03/untangling-algorithmic-illusions-from-reality-in-big-data.html> [ultimo accesso 15 marzo 2014]

Huberman, B., Romero, D., Wu, F., 2009. “Social networks that matter: Twitter under the microscope”. *First Monday*, [online] Disponibile all’indirizzo < http://firstmonday.org/article/view/2317/2063> [ultimo accesso 4 marzo 2014]

Il sole 24 ore, 2013. “Le informazioni sociali? Miniera d'oro da esplorare” *Il sole 24 ore*, [online] Disponibile all’indirizzo <http://www.ilsole24ore.com/art/tecnologie/2012-03-02/informazioni-sociali-miniera-esplorare-180412.shtml?uuid=AaxDZ70E> [ultimo accesso 20 febbraio 2014]

Internet World Stats, 2012. *“Internet Users in Europe”*, [online] Disponibile all’indirizzo <http://www.internetworldstats.com/stats4.htm#top> [ultimo accesso 28 febbraio 2014]

Jansen, B. J. et al., 2009 “Twitter Power: Tweets as Electronic Word of Mouth”. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*[pdf] Disponibile all’indirizzo <http://www.cs.rochester.edu/twiki/pub/Main/HarpSeminar/Twitter_power-_Tweets_as_electronic_word_of_mouth.pdf> [ultimo accesso 5 marzo 2014]

Jansen, J. et al., 2009. “Twitter Power: Tweets as Electronic Word of Mouth”. *Journal of the American Society for Information Science and Technology*, [online] Disponibile all’indirizzo <http://www.cs.rochester.edu/twiki/pub/Main/HarpSeminar/Twitter_power-_Tweets_as_electronic_word_of_mouth.pdf> [ultimo accesso 5 marzo 2014]

Johns Hopkins University, 2013. “Using Twitter to Track the Flu”. Notizie pubblicate da Johns Hopkins University Department of Computer Science. [online] Disponibile all’indirizzo <http://www.cs.jhu.edu/2013/01/24/using-twitter-to-track-the-flu/#.UwIrBfl5PuR> [ultimo accesso 15 marzo 2014]

Jungherr, A., Jürgens, P. e Schoen, H., 2011. “Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, t. o., sander, p. g., & welpe, i. m. ‘predicting elections with twitter: What 140 characters

reveal about political sentiment'". *Social Science Computer Review*, Volume 30, Issue 2, Pagine 299-234

Kaplan, A. M., Haenlein, M., 2010. "Users of the world, unite! The challenges and opportunities of social media". *Business Horizons*, Volume 53, Issue 1, Pagine 59-68

Kirkpatrick, R., 2012. "*Big Data for a better world*". O'Reilly Strata Conference, Ottobre 2012. New York: UN Global Pulse [online] Disponibile all'indirizzo <http://www.slideshare.net/unglobalpulse/strata-14934034> [ultimo accesso 20 febbraio 2014]

Kotler, P., 1999. *Il marketing second Kotler*. Milano: IL Sole 24 Ore

Kotler, P., Roberto, E. L., 1991. "*Marketing sociale*". Milano: Edizioni di Comunità

Kumar, P., 2013. "Twitters geography visualized and explained". *Internet Monitor*, [online] Disponibile all'indirizzo <https://blogs.law.harvard.edu/internetmonitor/2013/06/17/twitters-geography-visualized-and-explained/> [ultimo accesso 28 febbraio 2014]

Kwak, H. et al., 2010. "*What is Twitter, a Social Network or a News Media?*". Proceedings of the 19th International World Wide Web [pdf] Disponibile all'indirizzo <http://an.kaist.ac.kr/~hosung/papers/2010-www-twitter.pdf> [ultimo accesso 12 marzo 2014]

Lansdall-Welfare, T., 2012. "Nowcasting the mood of the nation". *Significance*, Volume 9, Isue 4, [online] Disponibile all'indirizzo <http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2012.00588.x/pdf> [ultimo accesso 15 marzo 2014]

Leetaru, K. et al., 2013. "Mapping the global Twitter heartbeat: The geography of Twitter". *First Monday*, [online] Disponibile all'indirizzo <http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654#p6> [ultimo accesso 1 marzo 2014]

Leskovec, J., Adamic, L., Huberman, B., 2007. "*The Dynamics of Viral Marketing*". [pdf] Disponibile all'indirizzo <http://arxiv.org/pdf/physics/0509039.pdf> [ultimo accesso 15 marzo 2014]

Liviu, L., Miahela, T., 2011. "Predicting Product Performance with Social Media". *Informatica Economica*, Volume 15, Issue 2, Pagine 46-56, [pdf] Disponibile

all'indirizzo <http://revistaie.ase.ro/content/58/05%20-%20Lica,%20Tuta.pdf> [ultimo accesso 9 marzo 2014]

Liyakasa, K., 2013. "Big Data and Customer Experience Begin to Converge". *CRM.com*, [online] Disponibile all'indirizzo <http://www.destinationcrm.com/Articles/Columns-Departments/Insight/Big-Data-and-Customer-Experience-Begin-to-Converge-90142.aspx> [ultimo accesso 20 febbraio 2014]

Lohr, S., 2012. "The Age of Big Data". *The New York Times*, [online] Disponibile all'indirizzo <http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=1&> [ultimo accesso 20 febbraio 2014]

Lotan, G., 2011. "#Sidibouzid Twitter Hashtag: an analysis of the people spreading the news". *giladlotan.com*, [blog] 24 Gennaio 2011. Disponibile all'indirizzo <http://giladlotan.com/blog/sidibouzid-twitter-hashtag-an-analysis-of-the-people-spreading-the-news/> [ultimo accesso 27 marzo 2014]

Lotan, G., 2012. "Analyzing UNICEF's #SahelNow Campaign". *giladlotan.com*, [blog] 23 Maggio 2012. Disponibile all'indirizzo <http://giladlotan.com/blog/analyzing-unicefs-sahelnow-campaign/> [ultimo accesso 27 marzo 2014]

Lotan, G., 2012. "ReTweet Revolution". *giladlotan.com*, [blog] Disponibile all'indirizzo <http://giladlotan.com/retweet-revolution/> [ultimo accesso 27 marzo 2014]

Lotan, G., 2013. "The Promise of Realtime". *giladlotan.com*, [blog] 11 Febbraio 2013. Disponibile all'indirizzo <http://giladlotan.com/blog/> [ultimo accesso 20 febbraio 2014]

Lotan, G., Devin, G., Cherie, M., 2011. "An audience analysis of prominent mainstream media news accounts on Twitter". *blog.socialflow.com*, [blog] 2 Agosto 2011. Disponibile all'indirizzo <http://blog.socialflow.com/post/7120243870/audience-study> [ultimo accesso 27 marzo 2014]

Maccaferri, A., 2013. "Il sentiment di Twitter che racconta le città d'arte". *Il Sole 24 ore*, [online] Dsiponibile all'indirizzo <http://www.ilsole24ore.com/art/tecnologie/2013-10-13/sentiment-twitter-racconta-citta-083004.shtml?uuid=AbOtyntI> [ultimo accesso 15 marzo 2014]

Mander, J., 2014. "GlobalWebIndex Social Summary 2014", [slide] Disponibile all'indirizzo <http://www.slideshare.net/globalwebindex/gwi-social-summary-2014> [ultimo accesso 28 febbraio 2014]

Mannella, L., 2013. “Duncan Watts: ‘La risposta alla crisi è nei social network. Ma sarà difficile trovarla’”. *Wired,* [online] Disponibile all’indirizzo < http://daily.wired.it/news/scienza/2013/06/26/watts-premio-lagrange-torino-46278.html> [ultimo accesso 5 marzo 2014]

Mannella, L., 2013. “Studiare Twitter per capire i mercati finanziari”. *Galileo* [online] Disponibile all’indirizzo <http://www.galileonet.it/articles/52554596a5717a06a2000006> [ultimo accesso 3 marzo 2014]

Manyika, J., et al., 2011. *“Big data: The next frontier for innovation, competition, and productivity”*. McKinsey Global Institute Report [pdf] Disponibile all’indirizzo <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation> [ultimo accesso 20 febbraio 2014]

Mar. 2011 [blog] Disponibile all’indirizzo <http://benetech.blogspot.it/2011/03/issues-with-crowdsourced-data-part-2.html> [ultimo accesso 28 febbraio 2014]

Mathiesen, J. et al., 2013. “Excitable human dynamics driven by extrinsic events in massive communities”. *Pnas,* [pdf] Disponibile all’indirizzo <http://www.pnas.org/content/early/2013/10/02/1304179110.full.pdf+html> [ultimo accesso 11 marzo 2014]

Maynard, D., 2012. “Opinion mining for social media”. s*lideshare.net* [slide] Disponibile all’indirizzo <http://www.slideshare.net/dianamaynard/prague2012-opinionmining>

Maynard, D., 2013. “Practical Opinion Mining for Social Media”. s*lideshare.net* [slide] Disponibile all’indirizzo <http://www.slideshare.net/dianamaynard/opinion-miningtutorial>

Maynard, D., 2013. “What do you really mean when you tweet?”. s*lideshare.net* [slide] Disponibile all’indirizzo < http://www.slideshare.net/dianamaynard/opinion-miningkeynote>

McNeil, D., 2011. “Haiti: Cellphone Tracking Helps Groups Set Up More Effective Aid Distribution, Study Says”. *New York Times,* [online] Disponibile all’indirizzo <*http://www.nytimes.com/2011/09/06/health/06global.html?_r=1&*> [ultimo accesso 20 febbraio 2014]

Mendoza, M., Poblete, B., Castillo, C., 2010. “Twitter Under Crisis: Can we trust what we RT?”. *Proceedings of the First Workshop on Social Media Analytics*, pagine 71-79

Metaxas, P.T., Mustafaraj, E., and Gayo-Avello, D., 2011. “How (Not) To Predict Elections”. *Proceedings of PASSAT/SocialCom* [pdf] Disponibile all’indirizzo <http://cs.wellesley.edu/~pmetaxas/How-Not-To-Predict-Elections.pdf> [ultimo accesso 26 marzo 2014]

Microsoft Research, WHO, UNGP, 2013. *“Big Data and Risk Factors of NCDs”*. Tallinn, 10 Settembre 2013 [pdf]

Mislove, A. et al., 2011. “Understanding the Demographics of Twitter Users”. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, [pdf] Disponibile all’indirizzo <http://www.ccs.neu.edu/home/amislove/publications/Twitter-ICWSM.pdf> [ultimo accesso 6 aprile 2014]

Mocanu, D. et al., 2013. “The Twitter of Babel: Mapping World Languages through Microblogging Platforms” [pdf] Plos Magazine. Disponibile all’indirizzo <http://arxiv.org/pdf/1212.5238v1.pdf> [ultimo accesso 20 febbraio 2014]

Mocanu, D. et al., 2013. “The Twitter of Babel: Mapping World Languages through Microblogging Platforms”. *PLoS ONE,* [pdf] Disponibile all’indirizzo <http://www.plosone.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pone.0061981&representation=PDF> [ultimo accesso 15 marzo 2014]

Models for the Next Generation of Software”. *Communications & Strategies,* [pdf] Disponibile all’indirizzo <http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=8&sid=6e3120ad-f796-43f7-a876-abf26ff0da6f%40sessionmgr111&hid=121> [ultimo accesso 11 marzo 2014]

Moramarco, A., 2014. “E-sigarette e fisco”, *Il sole 24 ore,* [online] Disponibile all’indirizzo <http://www.banchedati.ilsole24ore.com/EstrazioneDoc.do?product=BIG&iddoc=SS20140404040BAA&doctype=HTML&uid=1395998526691&tabelle=sole> [ultimo accesso 17 aprile 2014]

Morris, M.R. et al., 2012. “ Tweeting is Believing? Understanding Microblog Credibility Perceptions”. *Microsoft Research Publication,* [online] Disponibile all’indirizzo <http://research.microsoft.com/apps/pubs/default.aspx?id=155374> [ultimo accesso 1 marzo 2014]

Mustafaraj, E. et al., 2011. “Vocal Minority versus Silent Majority: Discovering the Opinions of the Long Tail”. *Proceedings of PASSAT/SocialCom.* [pfd] Disponibile

all'indirizzo <http://cs.wellesley.edu/~pmetaxas/Silent-minority-Vocal-majority.pdf> [ultimo accesso 8 marzo 2014]

My web 2.0, 2013. "Quanti utenti Twitter attivi ci sono nel mondo e in Italia?". *myweb2.0.it* [blog] 21 Novembre 2013. Disponibile all'indirizzo <http://www.myweb20.it/2013/11/quanti-utenti-twitter-attivi-nel-mondo-e-in-italia/> [ultimo accesso 2 aprile 2014]

Nielsen, 2013. "*Tops of 2013: Digital*" [online] Disponibile all'indirizzo <http://www.nielsen.com/us/en/newswire/2013/tops-of-2013-digital.html> [ultimo accesso 28 febbraio 2014]

O'Connor, B. et al., 2010. 'From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series'. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, Washington, DC [pdf] Disponibile all'indirizzo <http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf> [ultimo accesso 19 marzo 2014]

O'Reilly, M., 2012. "Can big data deliver on its promise?". *OECD Observer*, [online] Disponibile all'indirizzo <http://www.oecdobserver.org/m/fullstory.php/aid/3921/Can_big_data_deliver_on_its_promise_.html> [ultimo accesso 20 febbraio 2014]

O'Reilly, T., 2007. "What Is Web 2.0: Design Patterns and Business" [pdf] Disponibile all'indirizzo <http://www.im.ethz.ch/education/HS08/OReilly_What_is_Web2_0.pdf> [ultimo accesso 17 aprile 2014]

Osservatorio Fumo Alcol e Droga, Istituto Superiore della Sanità, 2013. *"Rapporto sul fumo in Italia 2012"* Roma, XV Convegno Tabagismo e Servizio Sanitario Nazionale [pdf] Disponibile all'indirizzo <http://www.iss.it/fumo/index.php?lang=1&anno=2014&tipo=3> [ultimo accesso 17 aprile 2014]

Osservatorio Fumo Alcol e Droga, Istituto Superiore della Sanità, 2014. *"Rapporto sul fumo in Italia 2013"* Roma, XVI Convegno Tabagismo e Servizio Sanitario Nazionale [pdf] Disponibile all'indirizzo <http://www.iss.it/fumo/index.php?lang=1&anno=2014&tipo=3> [ultimo accesso 17 aprile 2014]

Pak, A., Paroubek, P., 2010. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association*

ELRA [pdf] Disponibile all'indirizzo <file:///C:/Users/Federica%20Fedrigo/Downloads/Pak%20and%20Paroubek%20(2010).%20Twitter%20as%20a%20Corpus%20for%20Sentiment%20Analysis%20and%20Opinion%20Mining%20(1).pdf>

Pang, B., Lee, L., 2009. "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval,* Voume. 2, Numero 1-2 Pagine 1–135, [pdf] Disponibile all'indirizzo < http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf>

Paolotti, D. et al., 2013- "Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience". *PubMed,* [online] Disponibile all'indirizzo <http://www.ncbi.nlm.nih.gov/pubmed/24350723> [ultimo accesso 20 febbraio 2014]

Paranyushkin, D., 2011. "Identifying the Pathways for Meaning Circulation using Text Network Analysis". *noduslab.com* [online] Disponibile all'indirizzo < http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/> [ultimo accesso 15 marzo 2014]

Pascale, L., 2013. "Direttiva tabacco, accordo sulla stretta anti-fumo. Ecco le nuove regole". *eunews.it,* [online] Disponibile al'indirizzo <http://www.eunews.it/2013/12/18/direttiva-tabacco-accordo-sulla-nuova-stretta-anti-fumo-ecco-le-nuove-regole/11477> [ultimo accesso 17 aprile 2014]

Paul, J., 2012. "Big Data take center ice". *Marketing Journal* [online] Disponibile all'indirizzo <http://www.marketingmag.ca/news/marketer-news/big-data-takes-centre-ice-66917> [ultimo accesso 7 marzo 2014]

Paul, M. J., Dredze, M., 2011. *"You Are What You Tweet: Analyzing Twitter for Public Health"* Pubblicazioni della 5° International AAAI Conference on Weblogs and Social Media, [pdf] Disponibile all'indirizzo <http://cs.jhu.edu/~mdredze/publications/twitter_health_icwsm_11.pdf> [ultimo accesso 1 marzo 2014]

PeerReach, 2013. "4 ways how Twitter can keep growing". *Blog.peerreach.com* [blog] 7 Novembre 2013. Disponibile all'indirizzo <http://blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/> [ultimo accesso 28 aprile 2014]

Perini, R., 2010. "Reti Sociali (Social Network)", *riccardoperini.com*, [blog] 15 Maggio 2010. Disponibile all'indirizzo <http://www.riccardoperini.com/reti-sociali-social-network.php> [ultimo accesso 20 febbraio 2014]

Persuasion, 2010. “Using Netvizz & Gephi to Analyze a Facebook Network”. *persuasionradio.wordpress.com,* [blog] 6 Maggio 2010. Disponibile all’indirizzo <https://persuasionradio.wordpress.com/2010/05/06/using-netvizz-gephi-to-analyze-a-facebook-network/> [ultimo accesso 9 marzo 2014]

Polosa, R. et al., 2011. “Effect of an electronic nicotine delivery device (*e*-Cigarette) on smoking reduction and cessation: a prospective 6-month pilot study”. *BMC Public Health [online] Disponibile all’indirizzo <http://www.biomedcentral.com/1471-2458/11/786>* [ultimo accesso 17 aprile 2014]

Popova, L., Ling, P. M., “Alternative Tobacco Product Use and Smoking Cessation: A National Study”. *American Journal of Public Health,* [pdf]

Procter, R., Vis, F., Voss, A., 2013. “Reading riots on Twitter: methodological innovation for the analysis of big data”. *International Journal of Social Research Methodology,* Volume 16, Numero 3, Pagine 197-214

Quotidiano.it, 2013. “La Consob indaga su Gad Lerner: su twitter ha "anticipato" l'affare La7”, *Quotidiano.it,* [online] Disponibile all’indirizzo < http://www.liberoquotidiano.it/news/1196458/La-Consob-indaga-su-Gad-Lerner-su-twitter-ha-anticipato-l-affare-La7.html> [ultimo accesso 4 marzo 2014]

R-bloggers, 2010. “Generating graphs of retweets and @-messages on Twitter using R and Gephi”. *r-bloggers.com* [blog] 17 Ottobre 2010, Disponibile all’indirizzo <http://www.r-bloggers.com/generating-graphs-of-retweets-and-messages-on-twitter-using-r-and-gephi/> [ultimo accesso 9 marzo 2014]

Richtel, M., 2014. “Selling a Poison by the Barrel: Liquid Nicotine for E-Cigarettes”. *New York Times* [online] Disponibile all’indirizzo <http://www.nytimes.com/2014/03/24/business/selling-a-poison-by-the-barrel-liquid-nicotine-for-e-cigarettes.html?_r=3> [ultimo accesso 14 maggio 2014]

Rizzoli, D., 2013. “Vespignani: Il pensiero è contagioso. E noi possiamo prevederlo coi Big Data" *Wired,* [online] Disponibile all’indirizzo <http://daily.wired.it/news/scienza/2013/05/30/vespignani-big-data-wired-next-fest-42750.html> [ultimo accesso 16 marzo 2014]

Rusconi, G., 2013. “Tutti i segreti dei big data”. *Il sole 24 ore,* [online] Disponibile all’indirizzo < http://rassegnastampa.unipi.it/rassegna/archivio/2013/12/16SB21093.PDF> [ultimo accesso 20 febbraio 2014]

Sanders, E., van den Bosh, A, 2013. “Relating Political Party Mentions on Twitter with Polls and Election Results”. [pdf] Disponibile all’indirizzo <http://ceur-ws.org/Vol-986/paper_9.pdf> [ultimo accesso 6 marzo 2014]

Sang, E.T.K., Bos, J., 2012. “Predicting the 2011 Dutch Senate Election Results with Twitter”. *Proceedings of SASN 2012, the EACL 2012Workshop on Semantic Analysis in Social Networks, Avignon, France,* [pdf] Disponibile all’indirizzo <http://ifarm.nl/erikt/papers/sasn2012.pdf> [ultimo accesso 9 marzo 2014]

Santori, P. L., 2014. “Strumento di Analisi Linguistica” *datamediahub.it*, [blog]. 5 Febbraio 2014 Disponibile all’indirizzo <http://www.datamediahub.it/2014/02/05/strumento-di-analisi-linguistica/>

Savage, N., 2011 “Twitter as Medium and Message”. *Communication of the ACM, [pdf] Disponibile all’indirizzo <http://mags.acm.org/communications/201103/?pg=20#pg20>*[ultimo accesso 6 marzo 2014]

Scudieri, A., 2013. “Unione Europea firma accordo sulla sigaretta elettronica e sulla lotta al fumo”. *Il Messaggero,* [online] Disponibile all’indirizzo <http://news.young.it/2013/12/20/unione-europea-firma-accordo-sulla-sigaretta-elettronica-e-sulla-lotta-al-fumo/> [ultimo accesso 17 aprile 2014]

Shachtman, N., 2010. “Exclusive: Google, CIA Invest in ‘Future’ of Web Monitoring”. *Wired*, [online] Disponibile all’indirizzo<http://www.wired.com/dangerroom/2010/07/exclusive-google-cia/> [ultimo accesso 18 marzo 2014]

Shneiderman, B., 2008. *“Extreme visualization: Squeezing a billion datapoints into a million pixels” Proc. of the ACM SIGM.OD International Conference on Management of Data.”* [pdf] New York: ACM SIGMOD. Disponibile all’indirizzo <http://www.cs.umd.edu/~ben/papers/Shneiderman2008Extreme.pdf> [ultimo accesso 20 febbraio 2014]

Siegel, M. B., Tanwar, K. L., Wood, K. S., 2013. “Electronic Cigarettes As a Smoking-Cessation Tool. Results from an Online Survey”. *American Journal of Preventive Medicine,* Volume 40, Issue 4, Pagine 472-475

Signorini, A., Segre,A. M., Polgreen, P. M., 2011. “The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic”. *PLoS ONE,* [online] Disponibile all’indirizzo

<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0019467> [ultimo accesso 29 marzo 2014]

Skoric, M. et al., 2012. "Tweets and Votes: A Study of the 2011 Singapore General Election" *Proceedings of the 45th Hawaii International Conference on System Sciences*, [pdf] Disponibile all'indirizzo <http://www.researchgate.net/publication/232631921_Tweets_and_Votes_A_Study_of_the_2011_Singapore_General_Election/file/e0b4952cb91e2616c9.pdf> [ultimo accesso 9 marzo 2014]

Sladden, C., 2012. "Coming Soon: Nielsen Twitter TV Rating". *blog.twitter.com*, [online] Disponibile all'indirizzo <https://blog.twitter.com/2012/coming-soon-nielsen-twitter-tv-rating> [ultimo accesso 5 marzo 2014]

Spierings, K., Jacobs, K., 2013.*"Getting Personal? The Impact of Social Media on Preferential Voting"* [pdf] The Netherlands: Springer. Disponibile all'indirizzo <https://www.academia.edu/3327515/Getting_Personal_The_Impact_of_Social_Media_on_Preferential_Voting> [ultimo accesso 10 marzo 2014]

Sprenger, T.O., Welpe, I.M., 2010. "Tweets and Trades – The Information Content of Stock Microblogs". *Social Science Research Network* [pdf] Disponibile all'indirizzo <http://ssrn.com/abstract=1702854> [ultimo accesso 7 marzo 2014]

Statistic Brain, 2014. *"Twitter Statistics"*, [online] Disponibile all'indirizzo <http://www.statisticbrain.com/twitter-statistics/> [ultimo accesso 28 febbraio 2014]

Stead, L.F. et al., 2008. "Nicotine replacement therapy for smoking cessation". *Cochrane Database System Rev. 2012* [online] Disponibile all'indirizzo <http://www.ncbi.nlm.nih.gov/pubmed/18253970> [ultimo accesso 17 aprile 2014]

Stelter, B., 2008. "Finding Political News Online, the Young Pass It On". *New York Times*, [online] Disponibile all'indirizzo <http://www.nytimes.com/2008/03/27/us/politics/27voters.html> [ultimo accesso 9 marzo 2014]

Stieglitz, S., Dang-Xuan, L., 2013. "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior". *Journal of Management Information Systems,*Volume 29, Numero 4, Pagine 217 - 248

Sutfin, E. L., 2013. "Electronic cigarette use by college students". *Drug and Alcohol Dependence,* Volume 131, Isuue 3, Pagine 214-221

Svapo, 2011. “Che cos’è la sigaretta elettronica”. *svapo.it* [online] Disponibile all’indirizzo <http://www.svapo.it/guide/sigaretta-elettronica.html> [ultimo accesso 17 aprile 2014]

Takhteyev, Y., Gruzdb, A., Wellman, B., 2011. “Geography of Twitter networks”. *Elsevier*, [pdf] Disponibile all’indirizzo <http://homes.chass.utoronto.ca/~wellman/publications/twitter_geo/Geography_of_Twitter_Networks.PDF> [ultimo accesso 13 marzo 2014]

Teevan, J., Ramage, D., Morris, R. M., 2011. “ #TwitterSearch: A Comparison of Microblog Search and Web Search”. *Proceedings of the fourth ACM international conference on Web search and data mining,* [pdf] Disponibile all’indirizzo <http://nlp.stanford.edu/~dramage/papers/twitter-wsdm11.pdf> [ultimo accesso 11 marzo 2014]

The Center for Digital Government, 2013. *“Big Data Big Promise”*. [pdf] Folsom, California: e.Republic. Disponibile all’indirizzo < http://images.erepublic.com/documents/CDG13_SPQ1_V.pdf> [ultimo accesso 20 febbraio 2014]

Tiscali Scienze, 2014. “Cinquecentomila italiani fumano sigarette elettroniche: in calo il consumo di nicotina”. *Tiscali:scienze* [online] Disponibile all’indirizzo <http://notizie.tiscali.it/articoli/scienza/14/02/cinquecentomila-italiani-fumano-sigarette-elettroniche.html> [ultimo accesso 17 aprile 2014]

Tizzoni, M. et al., 2012. “Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm” *BMC Medicine,* [online] Disponibile all’indirizzo <http://www.biomedcentral.com/content/pdf/1741-7015-10-165.pdf> [ultimo accesso 20 febbraio 2014]

Tjong, E., Sang, K., and Bos, J., 2012.” Predicting the 2011 Dutch Senate Election Results with Twitter”. *Proceedings of SASN 2012, the EACL 2012Workshop on Semantic Analysis in Social Network,* [pdf] Disponibile all’indirizzo <http://ifarm.nl/erikt/papers/sasn2012.pdf> [ultimo accesso 9 marzo 2014]

Toledo Bastos, M., Zago, G., 2013. “Tweeting News Articles: Readership and News Sections in Europe and the Americas”. *Sage* [pdf] Disponibile all’indirizzo <http://sgo.sagepub.com/content/3/3/2158244013502496.full-text.pdf+html> [ultimo accesso 18 marzo 2014]

Tremolada, L., 2013. “Chi controlla i Big data? Il mito dei grandi numeri”. *Il sole 24 ore,* [online] Disponibile all’indirizzo

<http://www.ilsole24ore.com/art/tecnologie/2013-10-27/controlla-data-mito-grandi-082147.shtml?uuid=Abd9uvxI> [ultimo accesso 8 febbraio 2014]

Tsotsis, A., 2011. "A Twitter Snapshot of The Tunisian Revolution: Over 196K Mentions of Tunisia, Reaching Over 26M Users". *Techcrunch.com* [online] Disponibile all'indirizzo <http://techcrunch.com/2011/01/16/tunisia-2/> [ultimo accesso 9 marzo 2014]

Tufekci, Z., 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls". *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014* [pdf] Disponibile all'indirizzo <http://arxiv.org/ftp/arxiv/papers/1403/1403.7400.pdf> [ultimo accesso 12 marzo 2014]

Tumasjan, A. et al., 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment". *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* [pdf] Disponibile all'indirizzo <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852> [ultimo accesso 11 marzo 2014]

Tumasjan, A. et al., 2011. "Where There is a Sea There are Pirates: Response to Jungherr, Jurgens, and Schoen". *Social Science Computer Review*, [pdf] Disponibile all'indirizzo <http://ssc.sagepub.com/content/early/2011/05/12/0894439311404123.abstract> [ultimo accesso 7 marzo 2014]

Twitter, 2013. "Normativa sulla privacy" [online] Disponibile all'indirizzo <https://twitter.com/privacy> [ultimo accesso 17 aprile 2014]

UN Global Pulse, Maggio 2012. *Big Data for Development: Challenges & Opportunities* [pdf] UN Global Pulse. Disponibile all'indirizzo <http://unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf> [ultimo accesso 20 febbraio 2014]

UN Global Pulse, World Health Organization, 2013. *"Digital Smoke Signals"*. Tallinn, 10-11 Settembre 2013

Unicef, 2013. "Tracking anti-vaccination sentiment in Eastern European social media networks", Unicef, Aprile 2013 [pdf] Disponibile all'indirizzo <http://www.unicef.org/ceecis/Tracking_anti-vaccine_sentiment_in_Eastern_European_social_media_networks.pdf> [ultimo accesso 29 marzo 2014]

Vance, K, Howe, W., Dellavalle, R., 2009. *"Social Internet Sites as a Source of Public Health Information*". Dermatoligic Clinics, [pdf] Disponibile all'indirizzo <http://www.training-source.org/sites/default/files/comm_value_public_health/socialpdfs/SocialInternetSitesasSourceofPublicHealthInfo.pdf> [ultimo accesso 12 marzo 2014]

Viktor Mayer-Schonberger, 2013. "Big data, big era, big change!" *Il sole 24 ore*, [online] Disponibile all'indirizzo <http://www.ilsole24ore.com/art/notizie/2013-11-03/big-data-big-era-big-change-083808.shtml?uuid=ABPnW8a> [ultimo accesso 20 febbraio 2014]

VINCOSBLOG, 2013. "Cos'è la sentiment analysis?". v*incos.it* [blog] 30 Agosto 2012. Disponibile all'indirizzo < http://vincos.it/2012/08/30/cose-la-sentiment-analysis/> [ultimo accesso 9 marzo 2014]

VINCOSBLOG, 2013. "Ecco a chi vende i nostri dati Twitter". *vincos.it* [blog] 5 Settembre 2013. Disponibile all'indirizzo <http://vincos.it/2013/09/05/ecco-a-chi-vende-i-nostri-dati-twitter-intervista-per-wired/> [ultimo accesso 6 aprile 2014]

VINCOSBLOG, 2013. "I trend dei social media per il 2014". v*incos.it* [blog] 31 Dicembre 2013. Disponibile all'indirizzo <http://vincos.it/2014/01/07/trend-social-media-2014/> [ultimo accesso 28 febbraio 2014]

VINCOSBLOG, 2013. "La mappa dei social network nel mondo-Dicembre 2013". v*incos.it* [blog] 31 Dicembre 2013. Disponibile all'indirizzo <http://vincos.it/2013/12/31/la-mappa-dei-social-network-nel-mondo-dicembre-2013/> [ultimo accesso 28 febbraio 2014]

VINCOSBLOG, 2013. "State of the Net 2013 Parte 1: Lo scenario italiano". v*incos.it* [blog] 1 Giugno 2013. Disponibile all'indirizzo <http://vincos.it/2013/06/01/state-of-the-net-2013-parte-1-lo-scenario-italiano/> [ultimo accesso 28 febbraio 2014]

Vitak, J., et al., 2009. "It's Complicated: Facebook Users' Political Participation in the 2008 Election". *Cyber Psychology, Behavior and Social Networking*, [online] Disponibile all'indirizzo <https://www.academia.edu/412946/Its_Complicated_Facebook_Users_Political_Participation_In_the_2008_Election> [ultimo accesso 16 febbraio 2014]

Voices from the Blogs, 2012. "Campionato di Calcio 2012/13: #TwitterPaul aveva già previsto tutto. L'inizio della "psicostoria"?". *voicefromtheblog.com*, [online] Disponibile all'indirizzo <http://sentimeter.corriere.it/2013/05/03/campionato-di-calcio-201213-twitterpaul-aveva-gia-previsto-tutto-linizio-della-psicostoria/> [ultimo accesso 27 febbraio 2014]

Voices from the Blogs, 2012. “E il vincitore di Sanremo per la rete è…”. *voicefromtheblog.com*, [online] Disponibile all’indirizzo < http://voicesfromtheblogs.com/2012/02/16/cinguettii-da-sanremo/> [ultimo accesso 27 febbraio 2014]

Voices from the Blogs, 2012. “Maturità 2012 e Social Media: cosa twittano gli studenti?”. *voicefromtheblog.com*, [online] Disponibile all’indirizzo < http://voicesfromtheblogs.com/2012/06/19/tracce-maturita-2012-su-twitter/> [ultimo accesso 27 febbraio 2014]

Voices from the Blogs, 2013. “#Sanremo2013: e il Twitter-Oracolo parlò ancora una volta”. *voicefromtheblog.com*, [online] Disponibile all’indirizzo <http://voicesfromtheblogs.com/2013/02/17/sanremo2013-e-il-twitter-oracolo-parlo-ancora-una-volta/> [ultimo accesso 19 febbraio 2014]

Voices from the Blogss, 2011. *voicesfromtheblogs.com*, Disponibile all’indirizzo <http://voicesfromtheblogs.com/> [ultimo accesso 20 maggio 2014]

Voices from the Blogss, 2012. “Elezioni e Social Network: ancora una volta la rete anticipa il voto”. *Sentimeter* [blog] 7 Novembre 2012. Disponibile all’indirizzo <http://sentimeter.corriere.it/2012/11/07/elezioni-e-social-network-ancora-una-volta-la-rete-anticipa-il-voto/> [ultimo accesso 16 marzo 2014]

Voices from the Blogss, 2012. “Twitter BeneComune: previste anche le Primarie!”. *Voicesfromtheblogs.com* [online] Disponibile all’indirizzo <http://voicesfromtheblogs.com/2012/11/27/twitter-benecomune-previste-anche-le-primarie/> [ultimo accesso 11 marzo 2014]

Wagener, T. et al., 2013. “Pilot Investigation of Changes in Readiness and Confidence to Quit Smoking After E-Cigarette Experimentation and 1 Week of Use”. *Nicotine & Tabacco Research*, Volume 16, Numero 1, Pagine 108-114

We are Social, 2014. *“Global Digital Statistics 2014”* [online] Disponibile all’indirizzo <http://etonpreneurs.com/uploads/Global%20Social,%20Digital%20&%20Mobile%20Statistics,%20Jan%202014.pdf> [ultimo accesso 28 febbraio 2014]

Weil, K., 2010. “Measuring Tweets”. *blog.twitter.com*, 22 Febbraio 2010 [blog] Disponibile all’indirizzo <https://blog.twitter.com/2010/measuring-tweets> [ultimo accesso 28 febbraio 2014]

Wired, 2014. “Antimafia: il tuo social network rivela se sei un boss”. *Wired,* [online] Disponibile all’indirizzo <http://www.wired.it/internet/social-network/2014/04/25/incontri-boss-antimafia/> [ultimo accesso 26 aprile 2014]

Wired, 2014. “Twitter lancia la social tv anche in Europa”. *Wired,* [online] Disponibile all’indirizzo <http://www.wired.it/internet/social-network/2014/01/24/twitter-social-tv-europa/?utm_source=facebook.com&utm_medium=marketing&utm_campaign=wired> [ultimo accesso 16 marzo 2014]

Wisniewski, M., 2013. “Reporter's Notebook: 7 Big Data Themes Emerging from Nonbanks”. Bank Technology News, [online] Disponibile all’indirizzo <http://www.americanbanker.com/issues/178_56/seven-big-data-themes-emerging-from-nonbanks-1057759-1.html?ET=americanbanker:e14652:2205959a:&st=email&utm_source=editorial&utm_medium=email&utm_campaign=BTN_Intraday_032113> [ultimo accesso 20 febbraio 2014]

Wood, C., 2013. “Taking responsability” *Public CIO,* [pdf]

Wu, S., 2011. “*Who Says What to Whom on Twitter*”. Proceedings of the 20th international conference on World wide web, Pagine 705-714, [pdf] Disponibile all’indirizzo <https://iriss.stanford.edu/sites/all/files/dsi/Duncan%20Study%201.pdf> [ultimo accesso 16 marzo 2014]

Zaman, T. R., 2010. “Predicting Information Spreading in Twitter”. *Microsoft Research Publication,* [pdf] Disponibile all’indirizzo <http://research.microsoft.com/pubs/141866/NIPS10_Twitter_final.pdf> [ultimo accesso 19 marzo 2014]

Zhang, X., Fuehres, H., Gloor, P. A., 2011. *“Predicting Asset Value Through Twitter Buzz”*. Advances in Intelligent and Soft Computing Volume 113, Pagine 23-34. New York: Springer [pdf] Disponibile all’indirizzo <http://www.ickn.org/documents/Collin2011_Zhang_Fuehres_Gloor.pdf> [ultimo accesso 26 marzo 2014]

Zhang, X., Fuehres, H., Gloor, P. A., 2012. *“Predicting Stock Market Indicators Through Twitter ‘I hope it is not as bad as I fear’”*. Procedia - Social and Behavioral Sciences, Volume 26, Pagine 55–62 [pdf] Disponibile all’indirizzo <http://www.sciencedirect.com/science/article/pii/S1877042811023895> [ultimo accesso 26 marzo 2014]