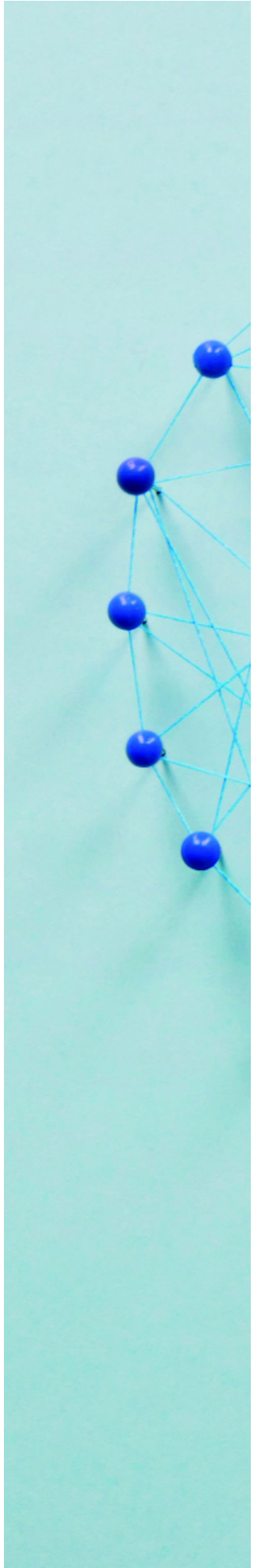
Big Data & Social Media Intelligence

Maurizio Tesconi

I social media sono tra le principali sorgenti di Big Data: producono un'enorme quantità di dati, a ritmi vertiginosi e in diversi formati. Per fare un esempio, un quarto della popolazione mondiale si connette almeno una volta al mese su Facebook. È possibile utilizzare questa massa di informazioni per le attività d'intelligence a tutela della società? Qual è il miglior modello per rappresentarle e quali sono le principali tecniche utilizzate per analizzare i dati provenienti dai social media?

Dalle fonti aperte alla social media intelligence

Sono quasi due miliardi le persone che utilizzano almeno una volta al mese un social media, producendo un flusso continuo d'informazioni che possono essere raccolte e analizzate per avere una sorta di finestra sempre aperta sul mondo e su quanto sta accadendo. L'*Open Source Intelligence*, cioè l'intelligence da fonti aperte, fa uso di qualsiasi documento o informazione di pubblico dominio e accessibile. Benché usati da secoli, i documenti cartacei sono stati rapidamente raggiunti e superati dalla loro controparte digitale in quantità e diffusione.
Con l'espansione di internet, essi si sono affermati in modo prepotente, specialmente nei contenuti prodotti dagli utenti della rete, tant'è che nel giro di pochi anni Wikipedia ha superato il numero di voci dell'Encyclopædia Britannica. Questo ha spostato l'attenzione dell'intelligence verso nuovi tipi di media, anch'essi pubblicamente accessibili in buona parte.

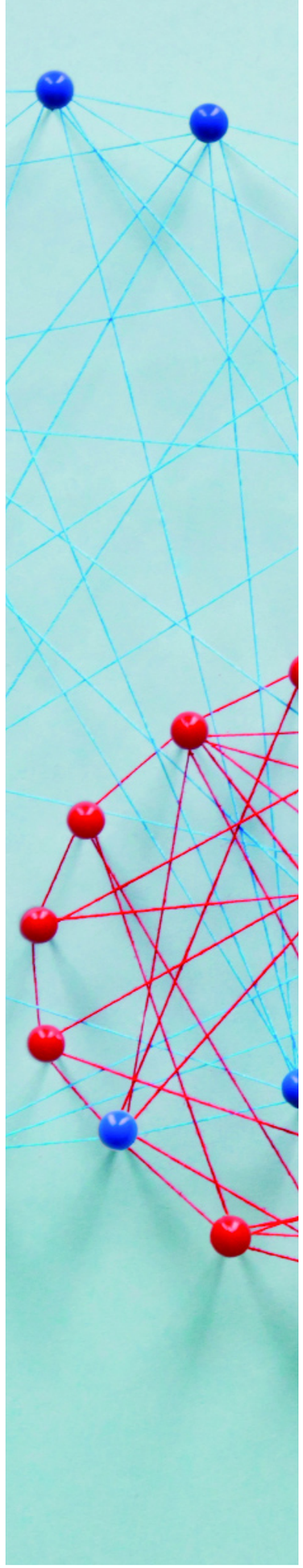


Si tratta di una considerevole mole di dati, esplosa con l'avvento dei social media, un fenomeno in continua estensione che coinvolge il 37%[1] della popolazione mondiale. Dalla ricchezza dei contenuti di queste piattaforme nascono molteplici tipologie di analisi che spaziano su domini diversi da quelli supportati dai documenti testuali.
Esaminando i social media è possibile esplorare dati di dimensioni prima impraticabili e, grazie alla possibilità di collegare tra loro gli utenti, se ne possono mappare i collegamenti e studiare il flusso dei contenuti attraverso la rete sociale delle persone. Le nuove tecnologie offrono spunti importanti all'intelligence che – attraverso processi di analisi sempre più ricchi e complessi denominati *Social Media Intelligence* (Socmint) – può utilizzarli come fonte di acquisizione informativa.
I miliardi di iscritti alle piattaforme di social networking, generando ogni giorno un flusso costante di informazioni di portata inimmaginabile fino a pochi anni fa, costituiscono un 'occhio' sul mondo, un punto di osservazione sui fenomeni che accadono intorno.
Gli utenti si trasformano in ottimi 'sensori' capaci di arricchire la narrazione dei fatti con una molteplicità di dettagli e percezioni riuscendo, nella somma complessiva delle descrizioni, perfino a correggere o filtrare i particolari errati. Questi 'sensori sociali' rappresentano un'incredibile ricchezza per i social network che possono utilizzare i dati raccolti per descrivere in tempo reale eventi o monitorarne l'andamento. Si pensi al contributo che gli utenti forniscono alla descrizione degli accadimenti con il flusso di messaggi di Twitter, con le dirette di Facebook o con i canali di YouTube.

OSSERVARE IL MONDO DA UNO SCHERMO

La Socmint permette di captare informazioni da tutte le aree in cui siano presenti persone dotate di smartphone e connesse alla rete, e rappresenta un valido supporto per fini d'intelligence grazie alla moltitudine d'informazioni fornite, parte delle quali in tempo reale.
Con il flusso d'informazioni condivise durante un determinato evento si possono sviluppare dei sistemi che trasferiscono velocemente un quadro aggiornato di quello che sta accadendo, con la possibilità inoltre d'individuare e contattare gli utenti coinvolti, in modo da aiutare i decisori a gestire un'emergenza come, ad esempio, un attentato terroristico. Il grado di completezza dell'informazione dipende dalla popolazione che vive in un dato territorio e dalla copertura delle reti cellulari o Internet. Quindi il sistema si adatta bene alle aree metropolitane e meno agli ambienti rurali.

1. <https://wearesocial.com/blog/2017/01/digital-in-2017-global-overview> [8-5-2017].



La *Social Network Analysis* è una delle tecniche d'indagine che maggiormente si sposa con questo genere di dati perché consente di ricostruire i legami sociali che intercorrono tra i membri di una comunità, con lo scopo d'individuare le persone chiave all'interno di una rete o scoprire eventuali anomalie e singolarità: con i social network si possono realizzare reti di natura differente, strutturate su legami di parentela, lavoro o amicizia, le quali originano lo sviluppo d'interazioni all'interno di una pagina o di un gruppo. Tali analisi hanno un significato diverso ma permettono di osservare un numero assai elevato di soggetti, al contrario di quello che avveniva in passato. Inoltre, la Socmint permette l'analisi e il tracciamento dei contenuti nell'ambito dei social network e tra social network distinti: ne sono esempi il video YouTube che viene condiviso su Facebook o Twitter, o il selfie Instagram, poi pubblicato anche su Snapchat.
Dal punto di vista grammmaticale, la Socmint si avvale di tutte le tecniche di analisi linguistiche fino a oggi ideate, ma le applica a un contesto assai ampio in cui i contenuti sono inseriti in un ambito social, ricevono tag, commenti e like, e possono rimbalzare da una piattaforma all'altra. Ecco che il contenuto diventa mobile e dinamico, subisce aggiunte o tagli, si evolve. Da non sottovalutare la capacità di identificare determinati soggetti o utenti all'interno dei social network osservando le caratteristiche dei loro profili: un nickname simile o un amico comune possono essere degli indicatori per individuare persone con profili multipli che si celano dietro foto differenti.
La Social Network Analysis e le tecniche di text mining possono aiutare a scovare chi dissimula la propria identità dietro falsi profili, a volte con intenti malevoli (adescamento di minori, diffamazione, diffusione di contenuti falsi ecc.).
Altre volte i social network sono teatro di propaganda razzista e di odio, e possono agevolare la radicalizzazione di estremisti religiosi. In tali scenari, oltre a contenuti testuali che supportano la causa, si trovano anche video e immagini della propaganda. Questa multimedialità diventa il fulcro di nuove verifiche, attraverso le quali si correlano informazioni conosciute con contenuti innovativi. Ecco che le tracce audio diventano testi a loro volta e possono venire analizzate, oppure la voce all'interno di una traccia è confrontata con quella di un database di persone note, oppure i volti che compaiono in un frame video sono comparati con immagini d'archivio.

UN UNIVERSO SOCIAL IN ESPANSIONE

Anche se la Socmint è una disciplina nuova e rappresenta una delle ultime frontiere verso cui il mondo dell'intelligence guarda con interesse crescente, la storia dei social media ha alle spalle due decenni. Da quando il web è nato, all'inizio degli anni Novanta, si sono susseguiti progetti di stampo social, in massima parte naufragati o superati. La svolta arriva con Friendster nel 2002, che conteneva tutte le caratteristiche dei social network moderni, presto superato da Facebook, in un periodo molto prolifico. Nel 2003 esordisce Myspace, utilizzato spesso come vetrina dai gruppi musicali e anche Google prova a creare il suo primo social network, Orkut, che riscuote un discreto successo, soprattutto in India e Brasile. Nel 2005 è la volta della prima grande piattaforma di contenuti video, YouTube, oggi il secondo sito più visitato al mondo dopo Google[2] e nel 2006 Twitter, il microblog da 140 caratteri che spopolerà per l'immediatezza e la novità e che adesso sta attraversando una fase di declino.

Nel 2009 Whatsapp, per la modica cifra di 89 centesimi all'anno, manda in pensione gli sms e nel 2010, in piena era Steve Jobs, sull'iPhone di Apple esordisce l'app di Instagram che, grazie anche alla nascente mania dei selfie, cresce a dismisura in popolarità, tanto da spingere Facebook ad acquisirla per l'incredibile cifra di un miliardo di dollari.

Dal 2011 è iniziato un nuovo trend tra i teenager con l'utilizzo sempre più ampio di chat istantanee, come Snapchat, divenuta famosa grazie ai contenuti a tempo che si cancellano. E nel 2015 con Periscope si affaccia nel mondo social un'altra app, con la possibilità di condividere in tempo reale video live e streaming da smartphone.

Tutto questo senza dimenticare i numerosi social network 'minori' o specifici per determinate categorie di utenti: VKontakte, clone di Facebook per i russi; i social network cinesi QZone, con i relativi sistemi di messaggistica QQ e Sina; Flickr, il sistema di condivisione immagini, e Reddit, la piattaforma per le social news e l'intrattenimento. Anche Google, forte della posizione come motore di ricerca e diffusione video (nel 2006 acquisisce YouTube), ha cercato di ritagliarsi un proprio spazio nella guerra tra social network creando la piattaforma Google Plus, anche se con risultati controversi, e Microsoft che, con la recente acquisizione di LinkedIn, è divenuta un importante player nel settore.

2. <http://www.alexa.com/topsites> [8-5-2017].

Ci sono, dunque, social network per tutte le esigenze, con caratteristiche specifiche che permettono di differenziarli e attirare di volta in volta specifiche tipologie di utenti anziché altre. Dal punto di vista tecnologico, i social network rappresentano delle sfide perché milioni di accessi simultanei ai contenuti rendono difficile la gestione delle risorse, senza contare le ingenti quantità di dati da immagazzinare che arrivano con facilità all'ordine di grandezze dei Petabyte e Zettabyte, tipici dei Big Data.

LA NATURA DI UN SOCIAL NETWORK E IL PROBLEMA DEI BIG DATA

I dati social, per i volumi e la velocità con cui vengono prodotti, hanno spinto le società che li gestiscono a sviluppare nuove tecnologie basate su sistemi distribuiti su tanti server, e abbandonare i modelli relazionali classici che garantivano la consistenza del dato, riducendo al minimo la ridondanza.

Vi sono delle intrinseche difficoltà nel gestire i contenuti dei social network dovute alla grande eterogeneità dei dati. Questa varietà rende complesso maneggiare e comparare dati provenienti da sorgenti diverse, tant'è che la *data fusion*, cioè la convergenza dei dati per ottimizzare le analisi, ha acquistato sempre maggiore importanza.

La coesistenza di formati di dati eterogenei, anche proprietari, codifiche di caratteri diverse, codec audio e video specifici è ormai la norma quando si parla di Big Data e social network. Questi formati devono andare incontro all'immensa diversità di dispositivi mobili e fissi che accedono ai social network da ogni parte del mondo e supportare lingue differenti con simboli disparati, svariate risoluzioni video e qualità audio.

Ma c'è anche qualcosa che accomuna un po' tutti i social ed è la loro struttura esterna. Per quanto siano presentate con nomi diversi, molte funzionalità delle piattaforme sono comuni. Non è inusuale che gli utenti siano collegati tra loro attraverso legami di amicizia o relazioni di following/follower. Anche il concetto di timeline, che organizza i contenuti in ordine cronologico mettendo in alto quelli più recenti, è presente in quasi tutti i social media e può essere utilizzato per rappresentare la wall di un utente, i post di un gruppo o la discussione intorno a un argomento, spesso identificato tramite l'hashtag (#).

Gli utenti sono collegati ai contenuti che producono e a quelli con cui interagiscono anche attraverso vari tipi di interazione implementati all'interno del social network, come i like, i commenti e le condivisioni. Queste interazioni si identificano con il nome di «engagement», ovvero il coinvolgimento prodotto dal contenuto sugli utenti.L'insieme delle caratteristiche forma una base di partenza su cui, in genere, si articolano le analisi utilizzate dalla Socmint per estrapolare conoscenza. Le interazioni sui contenuti e i legami strutturali tra gli utenti possono essere modellati attraverso un gigantesco grafo che connette gli account anche su social media diversi e apre la possibilità ad analisi sempre più sofisticate e interessanti. Le timeline possono essere trasformate in serie temporali e studiate con tecniche prese in prestito dalla teoria dei segnali.
Esiste allora un modello ideale per raffigurare queste informazioni? Sicuramente no, perché esse devono essere rappresentate con tanti modelli diversi in base al tipo di analisi che si vuole effettuare.

LE SFIDE APERTE

La condizione indispensabile per avviare un progetto Socmint di successo è quella di disporre di dati in quantità adeguata; la prima sfida che si pone riguarda la conoscenza e l'attuazione delle tecniche di acquisizione delle informazioni più efficaci e performanti. Innanzitutto, indipendentemente dalle modalità adottate, tutti gli strumenti di raccolta non possono accedere a informazioni che siano protette mediante password o sottoposte a limitazioni di accesso tramite le impostazioni di privacy degli utenti.
Molti social network forniscono agli sviluppatori un insieme di Application Program Interface (API) per esporre alcune funzionalità delle proprie piattaforme in modo da integrare nuovi servizi e applicazioni, un approccio che consente di rendere più ampia e appetibile l'offerta del social in modo da incentivarne l'uso. È il caso del sistema dei pagamenti introdotto da Facebook o dai vari giochi. Le API possono essere pertanto utilizzate per acquisire informazioni dalle piattaforme social.
L'uso di questi strumenti applicativi è sottoposto a politiche d'impiego che limitano l'accesso ai dati in modo da evitare usi impropri da parte degli sviluppatori e, nello stesso tempo, ridurre il carico di lavoro dei server che devono rispondere alle richieste di accesso ai dati formulate con questa modalità.

L'impiego delle Application Program Interface richiede una conoscenza specifica delle tecniche di programmazione e delle modalità con le quali interfacciarsi con le singole piattaforme.
Recentemente i gestori dei principali social network hanno limitato significativamente le informazioni a cui è possibile accedere tramite API per trasmettere il messaggio di un'accresciuta attenzione al tema della tutela della privacy dei propri utenti.
Un'altra modalità per raccogliere i dati è denominata *web scraping* e consiste nell'elaborazione delle pagine web formate in codice Html allo scopo di estrarne automaticamente le informazioni più significative. Una volta enucleati, i dati possono essere elaborati immediatamente oppure possono essere salvati su disco per un 'processamento' successivo. Nel primo caso è necessario disporre di algoritmi ad alte prestazioni, in grado di operare in memoria su una grande quantità di dati; nel secondo, si aggiunge l'esigenza di un'adeguata capacità di conservazione e indicizzazione di imponenti volumi di dati per consentirne la successiva ricerca.
Per quanto le più moderne tecnologie abbiano semplificato l'adozione di ambienti Big Data, implementarne uno che sia in grado di soddisfare tutte le esigenze descritte può rivelarsi un compito molto oneroso, da diversi punti di vista. Innanzitutto è necessario disporre di personale che abbia formazione ed esperienza specifiche sui temi della Socmint. Hardware e software devono essere adeguatamente strutturati per riuscire a mantenere prestazioni accettabili anche nei casi in cui le dimensioni dei dati aumentino oltre misura. Inoltre gli algoritmi che sviluppano i dati devono essere progettati tenendo conto delle performance e delle esigenze di distribuire l'elaborazione su molteplici unità. Non è infatti possibile applicare le stesse tecniche di analisi solitamente utilizzate per piccoli volumi di dati, in quanto la complessità e il numero di variabili sono così elevate che un approccio tradizionale non è in grado di fornire una risposta in tempi ragionevoli.
Chi progetta e implementa ambienti Big Data deve conciliare la capacità di gestire un sistema complesso, che tratta enormi volumi d'informazioni, con l'esigenza degli analisti di disporre di notizie precise, selezionate e di immediata fruibilità.
In conclusione, se oggi si vuol fare Socmint non si può prescindere dalla conoscenza dei Big Data, sia come fenomeno culturale sia dal punto di vista tecnologico, in modo da accrescere la consapevolezza in merito alle potenzialità, ai benefici e ai limiti di questa disciplina