Usare i Social Media in Applicazioni Predittive

C. Francalanci, P. Giacomazzi, A. Poli

Sommario. *Sui social media vengono ogni giorno espresse opinioni su fatti, marchi, persone, prodotti e servizi di ogni tipo. Secondo la teoria della "wisdom of the crowd", l'opinione media di un numero sufficientemente elevato di individui rappresenta una buona approssimazione della realtà. Se l'opinione riguarda il futuro, la media delle opinioni può forse essere utilizzata a fini predittivi. La letteratura sul tema e le diverse esperienze fatte in quest'ambito forniscono un interessante quadro delle opportunità e dei limiti dell'uso del parlato sui canali sociali a fini predittivi, mostrando tuttavia molti risultati sorprendentemente positivi.*

Keywords: Predictive analytics, Social media, Twitter

1. Introduzione

Sui social media vengono ogni giorno espresse opinioni su fatti, marchi, persone, prodotti e servizi di ogni tipo. Tali opinioni sono pubblicate in modo spontaneo, rispondendo a un innato bisogno di socializzazione e condivisione del proprio *status*. Diversi canali sociali rispondono a diversi bisogni emotivi o pratici. Ad esempio, Facebook, il canale sociale più diffuso in assoluto, supporta la socializzazione coi propri amici che, in generale, corrispondono a vere relazioni di amicizia nel mondo reale, passate o presenti. Su Twitter gli utenti condividono fatti che ritengono interessanti, rispondendo a un più generale bisogno di informazione. I forum settoriali, ad esempio quelli sulla salute, riuniscono persone accomunate da uno stesso interesse o da uno stesso problema, che è per loro così importante da rappresentare di per sé l'origine della necessità di socializzazione.

Sui canali sociali le opinioni non sono sollecitate da domande fatte da chi ha interesse nel conoscere le risposte. Ad esempio, nel settore turistico le persone condividono esperienze e si danno reciprocamente consigli senza che un operatore turistico abbia loro inviato un questionario per misurare l'indice di gradimento di una destinazione turistica e dei suoi servizi. Le persone parlano dunque di ciò che realmente le appassiona, di quello che ritengono interessante o utile per sé e per altri, di quello che li ha profondamente soddisfatti o chiaramente delusi, con una sincerità spesso facilitata dalla privacy della quale si gode o si può godere nella maggior parte dei canali sociali.

È pur vero che molte delle opinioni espresse sono in realtà di natura pubblicitaria e vengono immesse in rete, a volte in massa, da profili aziendali gestiti con l'obiettivo di sfruttare le capacità relazionali dei canali sociali a fini promozionali. Ad esempio, i profili di leader aziendali, quali Stefano Gabbana (@stefanogabbana) o Valentino Rossi (@ValeYellow46) su Twitter, rappresentano veri e propri canali di comunicazione con i fan, la cui efficacia risiede in parte nella capacità offerta dal canale sociale di relazionarsi con i *follower* in maniera più vivace, meno impersonale, e disinvoltamente innovativa. Lo stile di comunicazione, l'unicità dei contenuti, la velocità con la quale fatti e notizie vengono condivisi, rendono i canali sociali interessanti e inducono alla lettura e alla condivisione. I lettori sono consapevoli che il fine ultimo dei messaggi è spesso promozionale, ma se il contenuto è interessante e originale si lasciano comunque coinvolgere, purché ciò che viene comunicato sia vero e abbia un carattere di autenticità. Negli esempi citati, è chiaro che per Stefano Gabbana comunicare coi fan è parte del mestiere, ma se ciò che viene detto corrisponde alla personalità di Stefano Gabbana dimostrandosi coerente con il suo *stile*, i lettori dimostreranno comunque interesse, in un clima di partecipazione e sincera approvazione.

In sintesi, la gente, cioè la *crowd*, si fida dei canali sociali. La convinzione più o meno espressa, ma largamente condivisa, è che se qualcuno diffondesse un messaggio falso, la massa dei lettori è talmente ampia che arriverebbe la smentita da parte di qualcun altro. Se, viceversa, tutti condividono un particolare contenuto supportandolo, allora quel contenuto è ritenuto *vero*. La letteratura indica due fenomeni come i principali responsabili di questo atteggiamento: la *wisdom of the crowd* e l'assenza sui social media del fenomeno della *sospensione di coscienza* che caratterizza invece i tradizionali canali di marketing.

Secondo la teoria della *wisdom of the crowd*, l'opinione media di un numero sufficientemente elevato di individui rappresenta una buona approssimazione della realtà, in generale migliore di quella che può fornire un singolo esperto. Questa teoria ha avuto origine all'inizio del secolo scorso ed è comunemente associata ad un esperimento condotto dallo statistico Francis Galton nel 1907 [1], ampiamente citato come un interessante aneddoto che ha originato un'ampia letteratura in diversi ambiti, dalla psicologia, all'economia. Nell'esperimento in questione, Galton riporta i risultati di un'indagine empirica condotta a una fiera di paese. Egli mostra come la media di un campione statisticamente significativo di valutazioni del peso di un bue fatte da non esperti era più vicina al peso reale del bue di tutte le stime fatte da esperti prese individualmente. Il suo esperimento è stato largamente discusso e criticato.

Tuttavia, secondo la teoria della *wisdom of the crowd*, la media delle opinioni pubblicate sui canali sociali può costituire un dato di grande interesse e se l'opinione riguarda il futuro, può forse essere utilizzata a fini predittivi.

L'assenza del fenomeno della sospensione di coscienza rafforza ulteriormente l'interesse dell'informazione proveniente dai canali sociali. I media tradizionali sono basati sul paradigma del *broadcasting*, ovvero sono monodirezionali e gli ascoltatori non hanno (quasi) possibilità di replica. Questo ha creato nel tempo il fenomeno della sospensione di coscienza, secondo il quale gli ascoltatori apprezzano un messaggio in broadcasting anche se sanno che non è necessariamente vero. Alcune pubblicità hanno un notevole successo di pubblico, indipendentemente dalle caratteristiche del prodotto o servizio al quale fanno riferimento e con il quale hanno talvolta una relazione molto remota (vedi, ad esempio, la nota pubblicità della Coca Cola [2] o di Mentadent [3]). Queste pubblicità sono talmente piacevoli da guardare che, indipendentemente dalla loro relazione con il prodotto, vera o falsa che sia, inducono il pubblico ad apprezzare la forma senza porla in relazione con la sostanza spesso lontana dalla realtà. In altre parole, il pubblico sospende la propria coscienza, non si concentra sul fatto che il contenuto dello spot pubblicitario è essenzialmente falso (una palese esagerazione della realtà) e apprezza la piacevolezza della storia raccontata.

Questo fenomeno di sospensione di coscienza non vale, in generale, per i social media. Sui canali sociali si pretende che ciò che viene detto sia *vero* e, se non lo è, si utilizza la possibilità di replica lasciando commenti negativi che, nel numero e nella diretta espressione della negatività di opinione sono spesso distruttive dell'immagine del marchio pubblicizzato (vedi [4]). La letteratura riporta molti casi di pubblicità che hanno avuto un buon riscontro in televisione e un pessimo riscontro sui social, dimostrando il rischio di passare da *above the line* (la televisione) a *below the line* (i social).

La fiducia nella veridicità complessiva dell'informazione proveniente dai canali sociali e nella sua capacità di fornire valutazioni medie affidabili rappresenta un presupposto fondamentale delle applicazioni di *predictive analytics*. Se, viceversa, non si ritiene che l'opinione media espressa sui canali sociali rappresenti una buona valutazione della media del fenomeno di interesse, la correlazione fra tale opinione e la futura evoluzione del fenomeno perde il suo fondamento teorico. La robustezza del legame teorico fra opinione media e evoluzione del corrispondente fenomeno di interesse è, a nostro avviso, un fattore determinante del successo o dell'insuccesso dei tentativi pratici di utilizzo dell'informazione sociale a fini predittivi.

2. L'informazione sociale

È naturale domandarsi quali fenomeni sia possibile predire sulla base del parlato sociale. Non esiste una risposta generale a questa domanda, ma è possibile, e relativamente facile, escludere alcuni ambiti. In generale, è possibile utilizzare il parlato sociale a fini predittivi solo se del fenomeno in oggetto si parla a sufficienza. Se invece i volumi sono scarsi, la media delle opinioni non rappresenta un indicatore statisticamente significativo e non può essere ritenuta affidabile, a prescindere da qualunque considerazione teorica.

Quale sia il valore di volumi di parlato necessario dipende poi dalla complessità del fenomeno di interesse. Occorre osservare che i volumi di parlato possono essere elevatissimi. Ad esempio, i volumi di parlato in lingua inglese sul solo canale Twitter sul brand "Nutella" sono circa un milione di tweet al mese, quelli sul brand "Londra" quasi due, i volumi in lingua italiana sul brand "Milano" sono circa 100.000 tweet al mese. Benché Twitter rappresenti il canale con i più elevati volumi di parlato, esistono una miriade di canali sociali concettualmente importanti con volumi singolarmente più bassi, ma complessivamente comunque significativi (ad esempio, TripAdvisor nel settore turistico o Healthboards nel settore salute).

La *wisdom of the crowd* concentra il suo interesse sulle opinioni dei non esperti, ovvero della massa che liberamente condivide il proprio pensiero sui diversi social media. La vasta letteratura sull'ascolto del parlato sociale indica con ampio consenso che le persone tendono a parlare spontaneamente di ciò che li riguarda da vicino e di cui hanno esperienza diretta (vedi, ad esempio, [5]). Essi parleranno, ad esempio, di prodotti e servizi dei quali sono utilizzatori, mentre non parleranno di prodotti e servizi a monte nella filiera produttiva, a meno che questi non siano in qualche modo visibili e sperimentabili. La conseguenza di questo dato di fatto è che il parlato sociale esiste con volumi spesso significativi per i prodotti e servizi classificabili come *B2C* (business to consumer), mentre non esiste o ha volumi molto bassi per prodotti e servizi classificabili come *B2B* (business to business).

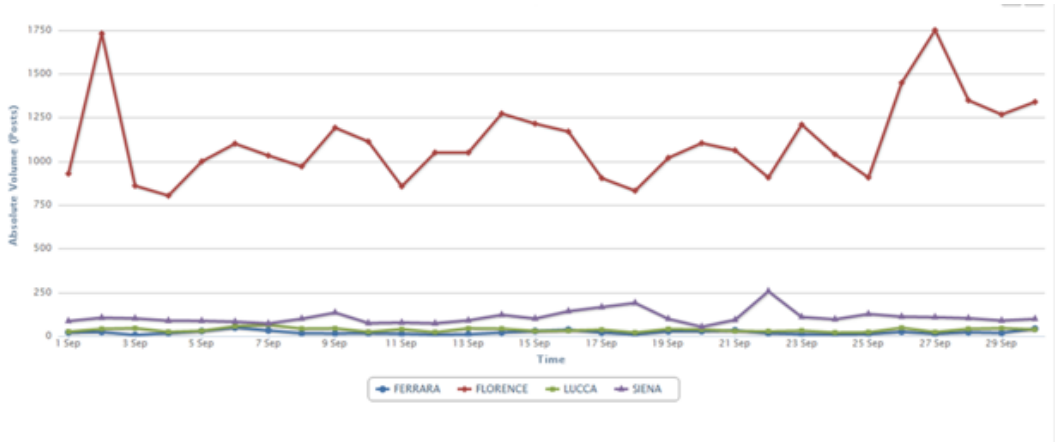

Figura 1
Volumi di parlato su Twitter nel settore turistico (esempio)

Questa suddivisione fra B2C e B2B è criticabile e approssimativa, ma aiuta a farsi un'idea a priori e non crearsi aspettative in ambiti nei quali le persone non forniscono spontaneamente opinioni. È importante notare che anche nel B2B può esserci informazione, ma normalmente non è di tipo sociale. Ad esempio, sugli impianti di refrigerazione industriale è possibile trovare molta informazione sul Web, specialmente sui siti di informazione e istituzionali, mentre è quasi nullo il parlato sociale. L'obiezione secondo la quale su Twitter si può trovare del

parlato su questo argomento non considera la distinzione fra parlato sociale e parlato degli esperti. I tweet di giornalisti che parlano, anche su Twitter, di eventi e notizie che riguardano gli impianti di refrigerazione industriale non possono essere considerati come un campione significativo di opinioni di non-esperti o di generici utenti degli impianti industriali. Viceversa, se esistesse un forum di scambio di opinioni fra persone che lavorano in aziende utenti degli impianti di refrigerazione e se il numero di tali persone fosse sufficientemente elevato, il loro parlato potrebbe costituire il campione di non esperti interessante ai nostri scopi. Tuttavia, tale tipo di forum esiste molto raramente e, se esiste, presenta volumi di parlato che non possono essere considerati statisticamente significativi.

Vogliamo anche osservare che anche nel B2C i volumi possono essere bassi o non essere statisticamente significativi. Ad esempio, il volume di parlato su prodotti elettronici non è sempre elevato. Mentre si parla tanto di smartphone, si parla poco di prodotti di elettronica domestica, quali ad esempio quelli di Bticino. La conseguenza è che, ferma restando la regola generale, i volumi devono essere sempre correttamente valutati e analizzati a priori, con una chiara distinzione fra il parlato degli esperti e quello dei non esperti.

Infine, i volumi possono avere ampia variabilità anche in uno stesso settore. A titolo d'esempio, la Figura 1 riporta i volumi di parlato nel tempo per alcune città italiane, mostrando come essi non siano elevati per tutte le destinazioni turistiche di interesse, ma solo per alcune.

3. Valutazione delle opinioni e sentiment analysis

Una delle principali sfide dell'utilizzo del parlato sociale a fini predittivi è la sua interpretazione. Nella maggior parte dei casi, le opinioni vengono espresse in linguaggio naturale. La difficoltà di interpretazione del linguaggio naturale rende particolarmente preziose le opinioni espresse come voto numerico su un oggetto di interesse, come un hotel o una autovettura, e su alcune sue caratteristiche predefinite, quali la pulizia delle stanze di un hotel o la piacevolezza della vista. I meccanismi di voto sono infatti implementati dalle imprese sui propri siti per ottenere facilmente una valutazione dei propri prodotti e servizi. Le valutazioni numeriche sono facili da interpretare e da aggregare per ottenere valutazioni medie rappresentative dell'opinione di una campione significativo di persone.

Tuttavia, tali valutazioni numeriche aggregate spesso non sono condivise e, quando lo sono, mostrano tutti i limiti dell'approccio a domande chiuse con il quale esse sono ottenute. Gli utenti sono supportati nell'esprimere opinioni su caratteristiche predefinite dei prodotti e servizi delle aziende, che possono non includere gli aspetti dei prodotti e dei servizi sui quali l'utente vorrebbe poter esprimere le proprie opinioni. Questo dato di fatto in alcuni casi è gestito affiancando alla richiesta di valutazioni numeriche la possibilità di esprimere commenti liberi. Ad esempio, TripAdvisor supporta entrambe le modalità e, in effetti, i commenti liberi sugli hotel sono numerosi e contengono indicazioni articolate e precise sull'esperienza di chi li ha scritti. Tali indicazioni sono generalmente ritenute molto utili, ma nella pratica sono poco fruibili a causa del tempo e dello sforzo richiesto per leggerle e farsene una sintesi [6].

Le opinioni espresse in forma testuale rappresentano la gran parte dei contenuti sui social più diffusi, si pensi a Facebook e Twitter. Tali opinioni edvono essere analizzate e strutturate per poter essere utilizzate in modelli matematici predittivi che, ovviamente, hanno bisogno di dati numerici in input. La trasformazione dei dati testuali in dati strutturati facilmente traducibili in dati numerici rappresenta a tutt'oggi una sfida.

Per comprendere la complessità del problema, facciamo riferimento a un esempio pratico. Supponiamo di voler effettuare una previsione sui volumi di vendita della *Ford*. L'assunzione di base che possiamo fare è che se la *crowd* parla tanto e positivamente dei vari modelli di auto Ford, allora è probabile che la Ford abbia in futuro un buon fatturato. Per tradurre questa assunzione in un semplice modello predittivo occorrono:

- Una serie storica di dati di fatturato, ad esempio mensile, della Ford.
- Una serie storica dei volumi di parlato sui principali canali sociali, ad esempio i volumi mensili di Tweet che riguardano la Ford.
- La valutazione dell'opinione espressa sulla Ford da ciascun post.

Supponiamo di avere accesso ai dati di fatturato mensile della Ford. Tali dati sono per natura quantitativi e possono essere facilmente utilizzati in un modello quantitativo. I dati sui volumi mensili di Tweet possono essere ottenuti scaricando tramite le API di Tweeter o di un suo rivenditore autorizzato (ad esempio, Gnip) tutti i Tweet che contengono la parola chiave "Ford" nell'intervallo di tempo considerato. Una volta ottenuti i dati grezzi, occorre eliminare i Tweet che pur contenendo la parola chiave "Ford" non fanno in realtà riferimento alla casa produttrice di automobili, ma ad altri oggetti del mondo reale, ad esempio Harrison Ford, l'attore e qualunque altro omonimo di Ford. Questa operazione si chiama *disambiguazione*. Si tratta di un'operazione molto importante e che non può essere tralasciata, poiché i dati non pertinenti possono rappresentare una percentuale molto rilevante del totale dei volumi, spesso superiore al 50%.

Sui dati rimanenti, occorre poi valutare l'opinione espressa. Questa valutazione prende il nome di *sentiment analysis*. L'obiettivo ultimo della sentiment analysis è comprendere se un testo, tipicamente un post condiviso su un canale sociale, riporta un'opinione positiva, negativa o neutra su un oggetto di interesse. Un testo può esprimere una opinione positiva o negativa esplicita sull'oggetto di interesse. Ad esempio, un post che dichiara che "le auto Ford sono inaffidabili" esprime un'opinione chiaramente negativa sulla sicurezza delle Ford. L'opinione però può essere espressa anche in forma più implicita tramite la condivisione di un fatto che riguarda la Ford. Ad esempio, un post che dichiara che "le auto coinvolte in incidenti autostradali sono per il 15% prodotte dalla Ford" fornisce un fatto negativo sulla sicurezza delle auto Ford, pur non esprimendo esplicitamente un'opinione. Considerare questo tipo di post come neutrali sarebbe evidentemente un errore ai fini predittivi. Questa varietà delle modalità linguistiche con cui le opinioni possono essere espresse è una delle difficoltà tecniche principali della sentiment analysis.

Ulteriore difficoltà è poi creata dal fatto che le opinioni sono espresse in maniera articolata, facendo riferimento a specifiche caratteristiche degli oggetti dei quali si parla. Ad esempio, le persone si lamentano se la batteria della loro auto si rompe prima del tempo, oppure se i consumi sono considerevolmente più elevati di quelli nominali, eccetera. Per utilizzare questo dettaglio a fini predittivi, occorre capire se le lamentele, o l'entusiasmo, sono allineate alla media di mercato, oppure rappresentano una oggettiva debolezza del brand in esame che può riflettersi in una diminuzione del fatturato. Inoltre, occorre selezionare le caratteristiche di interesse del prodotto o servizio che si ritiene possano avere un valore predittivo.

Tra l'altro, visti in maniera relativa, i volumi stessi hanno un valore predittivo. Volumi di parlato in diminuzione possono infatti essere un'indicazione di un minor interesse da parte del mercato in un determinato prodotto o servizio. L'attenzione ai volumi da parte delle numerose agenzie di comunicazione è proprio legata alla correlazione e potenziale impatto economico fra l'attenzione ricevuta dai diversi brand sui canali sociali e il loro successo di mercato.

Figura 2
Tag cloud relativa al brand MSC Crociere (marzo 2013, Twitter)

La Figura 2 riporta, a titolo d'esempio, una tag cloud del parlato su Twitter su MSC Crociere. I colori indicano il sentiment con il quale si parla dei diversi concetti, in rosso e viola i tag molto e blandamente negativi, in verde e giallo quelli molto e blandamente positivi. Il sentiment generale su MSC crociere è molto positivo (in verde), ma si notano alcune negatività, come, ad esempio, la negatività (in viola) su «Costa Fortuna» dovuta nello specifico a sporcizia,

cabine brutte, orari di imbarco impossibili (in relazione soprattutto a crociere in Sud America). In figura, i tag neutri sono in grigio, evidenziando una prevalenza di commenti neutrali che, in effetti, rappresenta una caratteristica generale del parlato sociale. Dal punto di vista predittivo, questo si traduce in una percentuale piuttosto piccola dei post con sentiment e una conseguente necessità di utilizzare intervalli temporali ampi nei modelli quantitativi, per poter contare su volumi di sentiment statisticamente significativi.

4. Applicazioni direzionali: la previsione dei volumi di vendita

L'informazione proveniente dai social media e il suo sentiment possono essere ritenuti indicativi dell'orientamento delle future decisioni di acquisto da parte degli utenti. A tal fine, i volumi del parlato ed il sentiment associato possono essere impiegati per predire i volumi di vendita. La disponibilità di previsioni di vendita può consentire di allocare al meglio i fattori produttivi al fine di rispondere alla variazione della domanda; al contempo, la disponibilità di un modello che correla il parlato sociale alle vendite può suggerire strategie di marketing che agendo sulla leva dei social media cerchino di incrementare i volumi di vendita.

Numerose analisi sono state portate avanti nel settore cinematografico, laddove si è cercato di prevedere il successo del lancio dei film sulla base dei commenti e delle recensioni su forum, blog e altri canali sociali, analizzati sia prima che durante e dopo l'uscita nelle sale cinematografiche. L'ampia disponibilità di informazioni sugli incassi ha reso possibili le analisi da parte di numerosi ricercatori. Come risultato di queste analisi vi è stata la proposta di una serie di sistemi predittivi che usano diversi approcci, ad esempio tecniche di regressione lineare, classificatori k-NN, e reti neurali. I modelli predittivi individuati basano i loro risultati talvolta sui volumi del parlato, talvolta sul sentiment, e talvolta su un uso combinato di sentiment e volumi. Non vi è quindi un accordo in merito alla predominanza della significatività dei volumi piuttosto che del sentiment, ma anzi appare che modelli differenti riconoscano un potere predittivo differente a queste due variabili fondamentali.

Un settore particolarmente interessante è quello della telefonia mobile. Numerosi appassionati discutono quotidianamente, con alti volumi di parlato in particolare su forum e blog specializzati, in merito a telefoni cellulari, smartphone, tablet ed altri dispositivi elettronici mobili. Inoltre, sezioni apposite o interi forum sono dedicati alle discussioni inerenti i servizi e le tariffe degli operatori di telefonia mobile. Gli utenti confrontano e commentano le diverse offerte e promozioni ed inoltre si esprimono in merito alla qualità del servizio offerto, riportando i casi di disservizi subiti. Questi messaggi possono essere anticipatori della volontà degli utenti di cambiare operatore mobile, laddove si siano individuate migliori offerte o si stiano riscontrando disservizi eccessivi. Allo stesso tempo, i messaggi sono in grado di influenzare i lettori, i quali si costruiranno un'opinione sulle diverse offerte e sulla qualità dei servizi disponibili, portando a future decisioni di acquisto. Sulla base di queste ipotesi, appare interessante verificare la correlazione delle vendite con il parlato sui forum tematici. In [7], gli autori prendono in considerazione i dati dei cambi di operatore con portabilità del

numero telefonico (MNP, Mobile Number Portability). In un mercato saturo, queste informazioni sono una buona approssimazione delle acquisizioni/perdite di clienti da parte degli operatori. Sono stati quindi considerati i flussi netti di clienti per ogni singolo operatore e al contempo analizzate le discussioni sui forum in merito all'operatore mobile preso in esame. La Figura 3 mostra un esempio di confronto su un intervallo di 16 settimane tra i flussi netti dei clienti in portabilità di un operatore mobile italiano (asse *sales*) ed il sentiment normalizzato dei messaggi, relativo allo stesso operatore, inseriti sul forum del sito web *telefonino.net*.

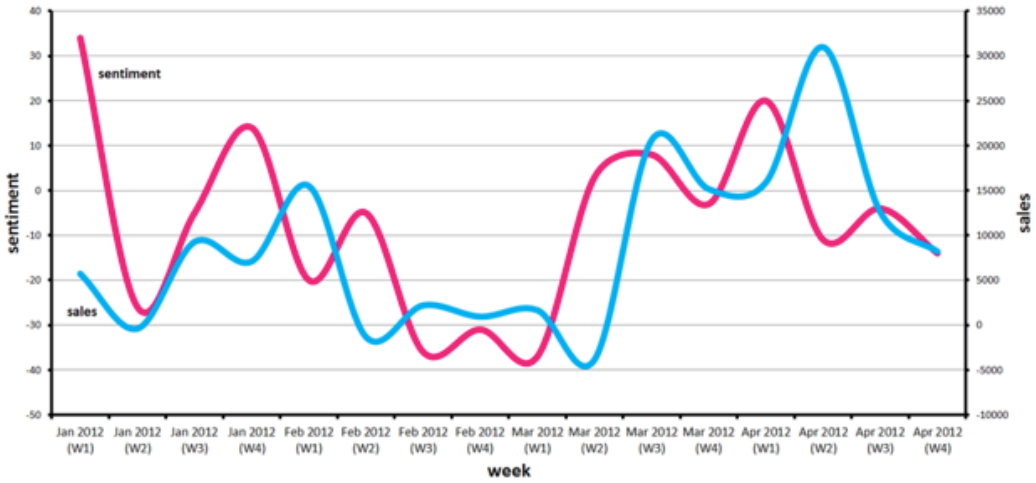

Figura 3
Relazione fra sentiment e vendite (settore telefonico, 2012)

Come si può notare, la curva delle vendite cresce, con un ritardo di circa una settimana, in seguito ad un miglioramento del sentiment. Ciò mostra che il miglioramento del sentiment è indicatore di un futuro incremento delle vendite. La presenza cioè di messaggi favorevoli ad una offerta, tariffa o servizio è un indice del gradimento, che si traduce a breve nello spostamento dei clienti verso un operatore. La decrescita del sentiment, invece, si accompagna anch'essa ad un calo delle vendite, ma è in alcuni casi anticipatoria e in altri casi posticipata rispetto ai passaggi di operatore. Il disagio dei clienti si manifesta infatti a volte in anticipo rispetto alla loro decisione di abbandono, mentre a volte impiega più tempo per essere osservato.

5. Applicazioni nel settore delle scomesse: il caso del calcio

Il calcio e lo sport in generale sono popolari argomenti di discussione su forum e social network da parte di numerosi appassionati. Una parte di questi appassionati scommette – più o meno regolarmente – denaro sull'esito di singoli incontri sportivi o campionati. Esistono inoltre forum pubblici dedicati alla discussione di strategie di scommessa prendendo in esame singoli incontri sportivi, ad esempio *Mondo Scommesse*, *Pronostitalia*, *Pronostigoal* e il forum di

Bet4Win in Italia, e *Punterslounge* e *OLBG* nel Regno Unito. Mentre su questa categoria di forum sono possibili discussioni articolate e motivate delle diverse strategie di scommessa, su Twitter si trovano tipicamente messaggi relativi alle gare e alle squadre, ma con osservazioni generiche e non finalizzate all'individuazione di strategie di scommessa. Inoltre, mentre sui forum specializzati sono prese in considerazione anche scommesse elaborate – vale a dire diverse dalla semplice individuazione del risultato finale "1X2" – i commenti su Twitter in genere forniscono solamente indicazioni di sentiment a favore o a sfavore di una squadra, senza indagare o fornire indicazioni sull'esito di eventi più articolati, come il numero di goal, i risultati del primo tempo, o il risultato esatto. Attraverso una metodologia di classificazione dei messaggi sui forum e l'analisi empirica dell'impatto della lettura di questi messaggi da parte di uno scommettitore, in [8] si osserva come il giocatore sia indotto ad alzare la propria soglia di rischio, ovvero a decidere di scommettere su eventi a bassa probabilità e alta quota di vincita. Ciò si verifica in quanto gli scommettitori esperti che partecipano ai forum tendono a non considerare gli eventi altamente probabili e poco remunerativi, ma al contrario forniscono, per le medesime gare, motivazioni articolate in merito all'opportunità di scommettere su altri eventi più remunerativi.

L'uso di Twitter finalizzato all'individuazione di una strategia di scommessa si è invece focalizzato sull'individuazione dell'esito finale "1X2", prendendo in esame un campione di squadre di calcio (Inter, Juventus, Milan e Napoli per il campionato italiano e Arsenal, Chelsea, Manchester City e Manchester United per quello inglese). Sono stati presi in esame i tweet relativi a queste squadre nei periodi precedenti e seguenti le gare. Da ogni tweet, attraverso un classificatore automatico, sono stati estratti i sentiment – positivi, negativi o neutrali – espressi nei confronti delle squadre citate. Prendendo in considerazione i risultati reali delle partite, e confrontati con le aspettative di risultato sulla base di un indicatore di difficoltà delle gare, sono state condotte delle verifiche sulla significatività statistica di una serie di ipotesi che coinvolgono i volumi dei messaggi osservati su Twitter e il sentiment misurato. Si è dapprima indagato quale impatto i risultati effettivi delle gare hanno avuto sui messaggi di Twitter. I risultati hanno mostrato che l'esito di un evento non ha alcun impatto sul sentiment misurato su Twitter, mentre invece è possibile affermare che l'esito degli incontri influenza sensibilmente i volumi, in particolare un evento negativo comporta la diminuzione dei volumi dei tweet relativi alla squadra, ed eventi positivi comportano talvolta variazioni in positivo e talvolta variazioni in negativo dei volumi dei messaggi. In secondo luogo, si è indagato in merito al potere predittivo dei tweet, confrontando i messaggi antecedenti le gare con i risultati effettivi. I risultati mostrano come a seguito di un aumento della percentuale dei tweet di tipo positivo e una contemporanea diminuzione di quelli di carattere negativo, l'esito dell'evento successivo è in genere positivo. Ciò significa che le buone sensazioni delle persone nei confronti di una squadra sono spesso giustificate. Allo stesso modo, anche un aumento di tweet positivi e un piccolo aumento di post negativi porta statisticamente ad un esito positivo per la squadra. Al contrario, si è dimostrato che a seguito di un aumento della percentuale dei tweet negativi e una contemporanea diminuzione di quelli

positivi, si verifica un evento negativo per la squadra. I volumi, ai fini della predizione, non sono invece stati riconosciuti come significativi. Si è osservato quindi che, in talune situazioni, è possibile trarre indicazioni in merito all'esito delle partite analizzando i messaggi che gli appassionati di calcio postano su Twitter, in particolare analizzando le variazioni del sentiment. E' possibile quindi individuare le gare per le quali si verificano le ipotesi qui descritte, al fine di determinare la puntata con esito più probabile. Tuttavia, le quote offerte dai bookmaker già incorporano, al fine di ridurre il rischio d'impresa, il sentiment degli scommettitori, portando ad offrire quote più basse, confrontate alla loro probabilità, per gli eventi che si ritengono possano ricevere maggiori scommesse. Questo fenomeno potrebbe quindi vanificare la maggiore accuratezza delle predizioni calcolate con i social network, abbassando sensibilmente i ritorni economici derivanti dalle scommesse.

Un analogo esperimento è stato condotto da un blogger americano [10], che ha utilizzato il parlato presente su Facebook, Twitter, blog e forum al fine di prevedere i risultati delle partite del campionato di football NFL (National Football League). L'esperimento ha coinvolto le 32 squadre del campionato e per ogni messaggio è stato assegnato, con tool automatici di analisi del testo, un valore di sentiment. Sono stati successivamente individuati alcuni indicatori che, per ogni squadra, calcolano, ad esempio, la soddisfazione degli utenti, la percentuale di conversazioni relative alla squadra rispetto alle conversazioni totali, o la percentuale di conversazioni con sentiment positivo/negativo rispetto alle conversazioni totali. Questi indicatori vengono quindi utilizzati per confrontare le squadre che stanno per sfidarsi, ed attraverso un algoritmo di decisione viene fornito il pronostico della squadra vincente. Questa metodologia ha consentito di prevedere correttamente 137 risultati a fronte di 111 errori durante la stagione 2012/13.

6. Applicazioni in finanza: il caso del trading

L'idea che si possano fare previsioni dei prezzi di borsa ha origine con le prime critiche all'*ipotesi dei mercati efficienti*. Assumendo che il flusso delle notizie non si interrompa mai, che le notizie abbiano un impatto diretto sul prezzo dei titoli, e che le notizie non siano predicibili, ne deriverebbe infatti che le variazioni di prezzo siano casuali e impredicibili, in quanto i prezzi rifletterebbero pienamente tutta l'informazione conosciuta.

Si è fatta strada inoltre l'idea che i consumatori, gli investitori e i manager siano in qualche modo guidati dall'umore della massa, anche detto *social mood*, il quale influenza le loro decisioni. Infatti, gli indizi forniti da altre persone influenzano le nostre opinioni, facendo nascere e propagare una visione condivisa della realtà (vedi, ad esempio, [11]).

L'uso del sentiment per ottenere informazioni su un titolo quotato è un'attività praticata correntemente ai fini dell'investimento azionario e nel mercato delle valute. Sono stati a tal fine sviluppati prodotti di analisi automatica delle notizie che convertono le news finanziarie in indici di sentiment che possono essere utilizzati dagli analisti finanziari per prendere decisioni di investimento. I leader in questo settore sono Ravenpack, Thomson Reuters e Alexandria. Questo

approccio tuttavia ha due principali difficoltà tecniche: in primo luogo l'estrazione del sentiment da un frammento di testo solleva il problema di adottare tecniche efficienti e precise di elaborazione del linguaggio naturale, ed in secondo luogo le variazioni di sentiment devono poter essere ricondotte alle variazioni di prezzo dei titoli, attraverso un modello appositamente definito.

Prima della diffusione dei social network, i lavori finalizzati all'individuazione del potere predittivo si sono focalizzati su gruppi di discussione online e siti di informazione, mentre negli ultimi sei anni si è iniziato ad investigare in merito all'utilizzo dei social network come fonte di *sentiment universale*, da utilizzare per correlare e predire vari indicatori economici. I risultati ottenuti non sono univoci, ma una larga parte dei ricercatori concorda con il fatto che la presenza di volumi di messaggi particolarmente elevati rispetto alla media su un dato forum o social network, si correli significativamente a variazioni elevate del valore dei titoli, ad un aumento della volatilità, e a un aumento del volume degli scambi. Controversa è invece la possibilità di fornire indicazioni in merito alla direzione della variazione del prezzo, e quindi di poter conseguire un vantaggio economico dall'uso dei modelli predittivi, sebbene alcuni studi affermino di aver raggiunto accuratezze nelle predizioni tra il 60 e l'80%, mediante la classificazione con modelli bayesiani, alberi di decisione e altre tecniche di machine learning [12].

Un tool software comunemente usato per l'estrazione del sentiment dalle news è SentiWordNet [13], il quale è liberamente installabile ed utilizzabile gratuitamente. Il tool, che si presta a numerosi utilizzi, in questo ambito è stato spesso impiegato per calcolare il numero di commenti positivi, negativi e neutrali per i titoli azionari di riferimento, sulla base delle parole presenti nel testo. Queste informazioni, assieme ai prezzi di borsa, sono quindi usate per la costruzione dei modelli predittivi, da utilizzare poi per prevedere gli andamenti futuri.

Un importante studio di Bollen et al. [14] ha coinvolto i messaggi del social network *Twitter*, usati per correlare i valori dell'indice americano *Dow Jones Industrial Average* (DJIA). La collezione dei messaggi giornalieri è stata classificata con tool automatici sulla base di 6 differenti dimensioni di umore: *calma*, *allerta*, *sicurezza*, *importanza vitale*, *gentilezza* e *felicità*. Trasversalmente si è identificata la polarità emozionale nelle due classi *positiva* e *negativa*. Come risultato è stata individuata una significativa relazione di causalità tra la dimensione *calma* e i movimenti di prezzo dell'indice DJIA nei 5 giorni successivi. Inoltre, attraverso un modello a reti neurali, la dimensione *calma* è stata usata al fine di predire l'indice DJIA, usando i dati dei precedenti 3 giorni. Il modello ottenuto ha mostrato una capacità predittiva della direzione dell'indice pari a circa l'87%.

Ad oggi esistono imprese specializzate nella sentiment analysis, come *MarketPsych* o *AlphaGenius*, che utilizzano le numerose teorie e tecniche di finanza comportamentale, data mininig, machine learning e linguistica computazionale per estrarre conoscenza finanziaria rilevante dalle più possibili fonti, come social network, chat, forum, siti di informazione e siti aziendali. Alcune piattaforme di trading inoltre incorporano tecnologie che forniscono agli investitori la possibilità di creare legami sociali e di replicare le strategie di trading altrui.

Molto ampio è inoltre l'interesse dei gestori di hedge fund, i quali hanno iniziato ad affiancare ai tradizionali indicatori finanziari alcuni indici di sentiment osservato sui social media. A titolo di esempio si cita la pionieristica *Derwent Capital Markets* che nel 2011 ha lanciato sul mercato un fondo speculativo che usava Twitter per prendere decisioni di investimento.

In [9], è descritta una metodologia automatizzata per l'individuazione di modelli di correlazione tra i messaggi su Twitter e alcuni indici finanziari settoriali. In particolare, il lavoro si è concentrato sul settore automobilistico coinvolgendo i post relativi ai costruttori di auto e impiegando un indice finanziario che racchiude i relativi titoli (DJTATO, Dow Jones Automobiles & Parts Titans 30 Index). Il sistema utilizza un tool proprietario per l'estrazione del sentiment e la classificazione in varie dimensioni di analisi di rilevanza economica (ad esempio vendite, assunzioni, licenziamenti, prodotti, fabbriche, acquisizioni, fusioni, e management) ed esplora il comportamento di una serie di modelli predittivi messi a confronto per individuare dinamicamente quelli che mostrano una migliore storia di successo nelle predizioni. Il sistema è stato usato per prendere due decisioni giornaliere di acquisto o vendita dell'indice finanziario, mettendolo a confronto con modelli predittivi che usano la sola autocorrelazione dei prezzi e non prendono in esame il parlato sociale.

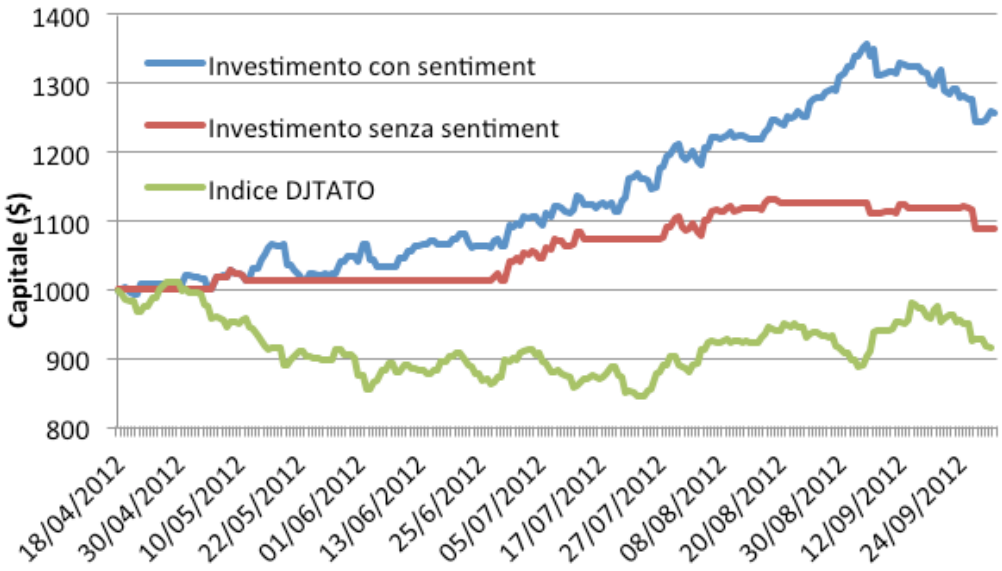

Figura 4
Simulazione di investimento basato su informazione sociale

In Figura 4, è mostrata la simulazione di investimento su un intervallo temporale di circa 5 mesi. A fronte di una perdita di circa l'8% dell'indice considerato, la metodologia di investimento, con l'utilizzo dei dati social, ha consentito di ottenere un ritorno del 26% con una volatilità del 12%. Questo dato va confrontato con la metodologia applicata in assenza dell'informazione sociale, la

quale ha prodotto un ritorno del 9%, mostrando un significativo valore aggiunto ottenibile dal suo sfruttamento.

Vale la pena ribadire che l'affidabilità delle predizioni non può prescindere dalla presenza di volumi statisticamente significativi del parlato sui social media. Perciò, l'efficacia delle predizioni sarà ricercabile solo in settori di mercato laddove vi sia una larga utenza che fornisce il proprio contributo nei social media, escludendo quindi le imprese che vendono prodotti e servizi classificabili come B2B (business to business).

Conclusioni

I contributi scientifici sull'utilizzo dell'informazione sociale a fini predittivi sono numerosi e distribuiti in diversi ambiti. Questo dimostra non solo una naturale curiosità nei confronti dei nuovi media che tanto appassionano i loro utilizzatori, ma anche un sincero interesse per un fenomeno che potrebbe avere interessanti applicazioni di mercato. Tuttavia, l'evidenza fornita dai lavori di ricerca finora svolti non si può considerare conclusiva e mostra ancora alcuni risultati non completamente coerenti e convincenti. Un'applicazione aziendale estensiva dei modelli e delle tecniche qui discussi richiede l'integrazione fra l'informazione sociale e le altre più consolidate fonti informative aziendali. Un'azienda che analizza con interesse i risultati dei modelli predittivi sociali non può non domandarsi se, in ultima analisi, i dati reali di vendita non rappresentino un indicatore più diretto e affidabile del successo dei prodotti aziendali e delle aspettative di vendita future. E' pur vero che i social media forniscono indicazioni sui *trend* emergenti, i cosiddetti segnali deboli che indicano un fenomeno nascente non ancora riscontrabile dai soli dati oggettivi di vendita. Tuttavia, se e come tali segnali deboli si possono integrare nei più consolidati processi di pianificazione aziendale resta oggetto di interessanti futuri lavori di ricerca.

Ringraziamenti

Gli autori desiderano ringraziare Francesco Merlo, collega e designer di alcune delle infografiche mostrate in questo articolo.

Riferimenti

[1] Francis Galton (1907). "Vox Populi," *Nature*, 75(1949), 450-451, March.

[2] https://www.youtube.com/watch?v=yRqUTA6AegA, 16 Dicembre 2013.

[3] https://www.youtube.com/watch?v=BCSwjYsy0QQ, 22 Aprile 2013.

[4] http://vimeo.com/16143908, 24 Ottobre 2010.

[5] L. Bruni, C. Francalanci, P. Giacomazzi, F. Merlo, A. Poli (2013). "The Relationship Among Volumes, Specificity, and Influence of Social Media Information," *International Conference on Information Systems 2013* (full paper), Milano, Dic.

[6] R. Pagano, P. Cremonesi (2014). "Summarization of restaurant reviews," Advanced User Interfaces, Milano.

[7] L. Rasina (2012). “Un’architettura per la predizione delle vendite basata su dati di sentiment,” Politecnico di Milano, Tesi di Laurea Magistrale, Dic.
[8] C. Consolandi (2012). “Una metodologia decisionale per le scommesse sportive basata su social media,” Politecnico di Milano, Tesi di Laurea Magistrale, Dic.
[9] A. Maggioni, L. Mazzoni (2012). “Design and validation of a forecasting trading system based on Twitter,” Politecnico di Milano, Tesi di Laurea Magistrale, Dic.
[10] Esposito, Jeff. Social Media’s NFL Week 1 Picks. [Online] Jeffesposito.com, 8 Settembre 2011. http://jeffesposito.com/2011/09/08/social-medias-nfl-week-picks/.
[11] Nofsinger, J. R. (2005). Social mood and financial economics. Journal of Behavioral Finance, 6 (3): 144-160.
[12] Sehgal, V. and Song, C. (2007). SOPS: Stock prediction using web sentiment. Proceedings of the 7th IEEE. International Conference on Data Mining Workshops, ICDM Workshops ‘07: 21-26.
[13] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the 7th Conference on International Language Resources and Evaluation, LREC ’10: 2200-2204.
[14] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2 (1): 1-8.

Biografie

Chiara Francalanci è Professore di Sistemi Informativi del Politecnico di Milano, dove si è laureata in Ingegneria Elettronica nell'ottobre 1991. Durante gli studi di dottorato è stata ricercatrice ospite per un periodo di due anni presso Harvard Business School. Ha scritto numerosi articoli sulla progettazione dei sistemi informativi e sul valore economico dell'informatica, svolto attività di consulenza nel settore finanziario e manifatturiero, sia in Italia che negli Stati Uniti, è editor del Journal of Information Technology e senior editor delle AIS Transactions on Enterprise Systems.
E-mail: chiara.francalanci@polimi.it

Paolo Giacomazzi è professore di Multimedia Internet al Politecnico di Milano, dove si è laureato con lode in Ingegneria Elettronica. Ha svolto un periodo di ricerca presso la University of Mississippi, presso il National Center for Wireless Communications. Ha svolto attività di ricerca e consulenza nel settore delle reti di telecomunicazione fisse e mobili. Attualmente, si occupa di sistemi previsionali relativi sia al traffico Internet, sia ad altri domini applicativi fra i quali i social media, la grande distribuzione, e il mass market in generale.
E-mail: giacomaz@elet.polimi.it

Alessandro Poli è nato a Cremona nel 1981. Si è laureato al Politecnico di Milano in Ingegneria Informatica e ha ottenuto il titolo di dottore di ricerca in Ingegneria dell'Informazione dal Dipartimento di Elettronica, Informazione e Bioingegneria del Politecnico di Milano, dove attualmente è impiegato come assegnista di ricerca. I suoi interessi accademici sono rivolti alle reti peer-to-peer video streaming, all'ottimizzazione di infrastrutture ICT, all'analisi semantica dei social media per la reputazione dei brand e la predizione, e all'analisi del mercato delle mobile app.
E-mail: alessandro.poli@polimi.it