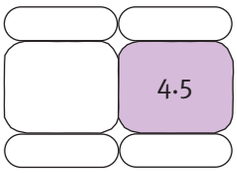
SOCIAL MEDIA INTELLIGENCE
COMPRENDERE IL "POLSO" DI UN TERRITORIO

Paolo Barbesino
Chiara Francalanci
Fiamma Petrovich

Ogni territorio ha una sua identità e reputazione creata dal sedimentarsi delle esperienze di chi lo attraversa. Internet e i Social Network sono disseminati delle tracce di questi passaggi: la social media intelligence consente di leggere questi segni. Lo stato dell'arte nell'analisi della reputazione online per il brand di una città è in continua evoluzione: strumenti semiautomatici basati su tecnologia semantica supportano il monitoraggio in tempo quasi reale dell'ambiente dei social media per il corretto presidio del web 2.0.



1 CHE C'È DI NUOVO? IL MICROBLOGGING E IL TERRITORIO

Nel nuovo millennio l'attrattività di un luogo e in particolare di una città è legata alla storia e alle qualità del posto e delle persone che lo abitano, che costituiscono veri e propri attributi del brand territoriale, così come alla vivacità di ciò che vi accade. La risposta alla domanda "se avessi tempo per una visita o se vivessi lì quanto facilmente vi troveresti cose interessanti o nuove da scoprire?" determina in maniera significativa il successo di una destinazione per il turismo o come territorio d'elezione per viverci[1].
Più di cento milioni di persone scambiano liberamente informazioni in tempo reale attraverso Twitter. La forma testuale corta, che consente messaggi non superiori ai 140 caratteri, ne facilita la produzione e il consumo anche in mobilità e il messaggio è in misura crescente accompagnato da indicazione di localizzazione secondo due modalità: come informazione dichiarata disponibile nel profilo dell'utente e/o come provenienza dichiarata (@); oppure come informazione rilevata attraverso la funzione di geolocalizzazione dell'applicativo utilizzato dall'utente per generarlo.
Ecco perché i 65 milioni di messaggi pubblicati ogni giorno su Twitter in risposta alla domanda "Cosa c'è di nuovo?" costituiscono un enorme serbatoio di intelligenza distribuita (o connettiva, secondo *Derrick de Kerckhove*) che può tradursi in indicazioni per il governo di un territorio, se raccolta ed elaborata attraverso l'analisi dei Social Media.
Negli Stati Uniti, uno studio dell'Università di Harvard registra lo stato emotivo (*polso*) della popolazione su Twitter da costa a costa, nelle diverse ore del giorno, a partire da un *dataset* di 300 milioni di messaggi (Figura 1).

[1] Anholt City Brand Index, 2006 e successivi.
[2] http://www.ccs.neu.edu/home/amislo-ve/twittermood

In Europa, la geografia di Londra è stata mappata in funzione della densità degli scambi su Twitter[3] (Figura 2).
A Milano, l'amministrazione comunale ha promosso e finanziato un progetto realizzato dal Politecnico di Milano in collaborazione con CommStrategy per studiare le dinamiche che presiedono alla costituzione dell'attrattività della città per il popolo di Twitter a partire da un data-set di oltre un milione di mes-

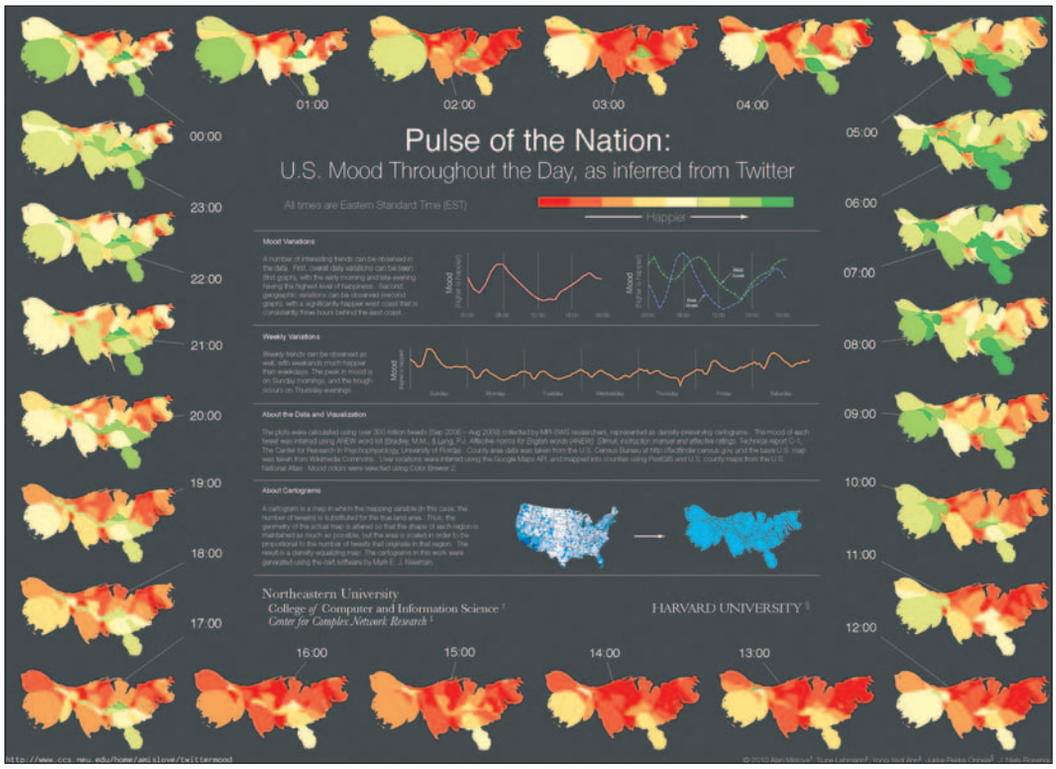


FIGURA 1
Visualizzazione delle variazioni di stato emotivo nel tempo e nello spazio inferite da Twitter

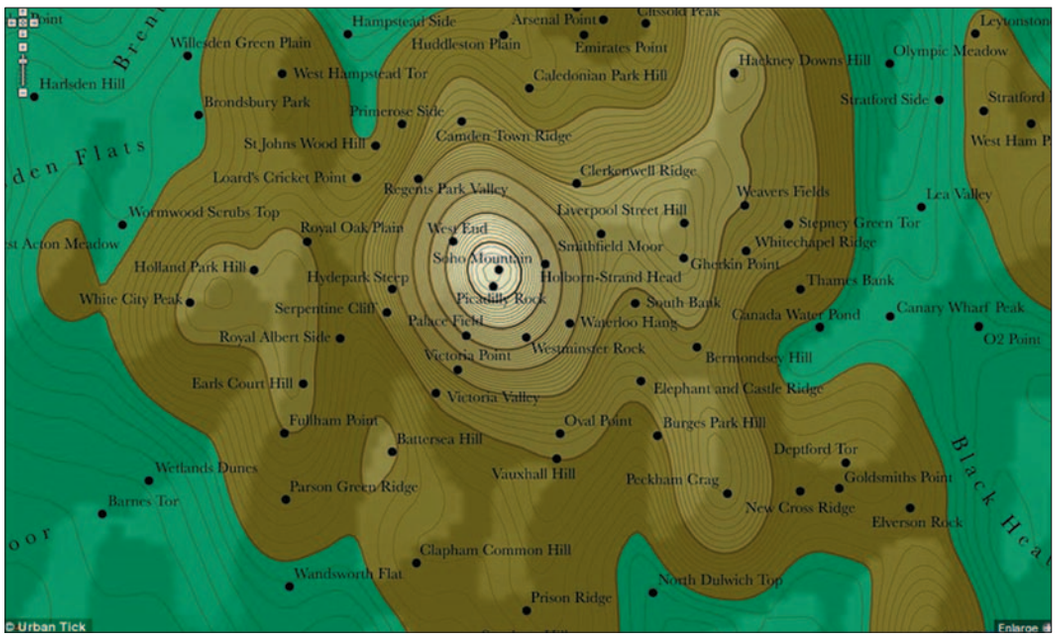


FIGURA 2
Picchi e avvallamenti di comunicazione su Twitter nelle diverse aree di Londra, città e dintorni

[3] Fabian Neuhaus, UCL's Centre for Advanced Spatial Analysis.

FIGURA 3
Dashboard di una listening platform (interfaccia a widget fruita in modalità self-service dall'analista)

saggi in lingua inglese raccolti negli ultimi 6 mesi, su Milano e su un primo gruppo di metropoli europee in diretta competizione quali: Berlino, Londra, Madrid.
Il progetto ha come perimetro di riferimento l'analisi dei Social Media e si estende dal microblogging alle community di viaggiatori (*tripadvisor*) e all'editoria digitale specializzata (*lonely planet*); mentre è in corso di realizzazione l'allargamento alle *fan page* di Facebook. L'obiettivo è la generazione di analisi strutturate e multi-sorgente, capaci di ponderare il gradimento in funzione della reputazione della sorgente informativa, in modo continuo e integrabile.
Il rilascio in versione prototipale di uno strumento di monitoraggio della reputazione online della città è avvenuto nell'ottobre 2010 e alla fase di *testing* attuale seguirà lo sviluppo di un'interfaccia *mash-up* destinata a recepire i requisiti espressi dai responsabili del Dipartimento Marketing Territoriale e Turismo del Comune di Milano.

2. INTRODUZIONE ALLA SOCIAL MEDIA INTELLIGENCE

Un *data-set* molto esteso come quello disponibile nell'ambiente dei social media richiede l'impiego di tecnologie per l'automazione dell'analisi dati: sul mercato della *Business Intelligence & Information Measurement Systems* sono disponibili diverse soluzioni per l'ascolto di Internet. Il processo che viene comunemente adottato prevede la raccolta dati, l'elaborazione e il rilascio, con diversi gradi di automazione.
Le numerose piattaforme oggi disponibili in modalità *software as a service*, dette piattaforme di *listening*, offrono prestazioni apprezzabili in termini di raccolta dati e di rilascio, consentendo la visualizzazione del dataset sia nella forma atomica del singolo messaggio sia in forma aggregata grazie a grafici di sintesi sulla distribuzione dei volumi e delle *keyword* secondo variabili scelte; mentre ancora limitata è la possibilità di attribuzione di un gradimento ai messaggi relativi a un dato dominio (Figura 3).
L'utilizzo di piattaforme di *listening* per l'analisi della reputazione online del brand di una città fornisce alcune risposte di natura quantitativa: ci dice quanto si parla di un territorio nell'ambiente dei social media e se, in coincidenza di alcuni eventi, si verifica una moltiplicazione dei contenuti, come per esempio in occasione di manifestazioni fieristiche o al verificarsi di eventi critici, come nel caso del blocco dei voli a causa dell'eruzione del vulcano islandese la scorsa primavera. Nell'esempio proposto in questa sede, sono state mappate le *keyword* più ricorrenti nel corpus complessivo di conversazioni su Londra nel periodo autunnale fino al 27 ottobre 2010 co-

FIGURA 4

Tagcloud delle top keyword per Londra (con almeno 1% di ricorrenza, su +2,5 mio di messaggi)

stituito da più di 2,5 milioni di messaggi distribuiti su blog, microblogging, Facebook, e forum (Figura 4).

La restituzione di una *listening platform* include nei risultati sostantivi, verbi e aggettivi, agganciati al flusso dati sottostante e quindi dimensionabili in termini di frequenza nel volume complessivo. Da questo tipo di analisi si ottiene immediatezza di riscontro sui temi conversazionali "caldi"; inoltre la presenza rilevante di verbi (*love*) e aggettivi qualificativi (*good, great, best*) a connotazione positiva su circa un quarto dei messaggi conferma l'attitudine verificata in molti studi a utilizzare i social media in primo luogo per esprimere apprezzamento, dando luogo a un vero e proprio electronic *word-of-mouth* (*e-wom*) connotato in modo positivo [1].

Se lo studio dell'ambiente dei social media vuole supportare il governo del territorio, è necessario integrare il contributo di analisti specializzati al fine di organizzare i dati restituiti in informazioni utili per i decisori. Poniamo il caso che l'analisi debba supportare il Marketing Territoriale nella scelta di attività di promozione della destinazione Londra. Un'elaborazione dei dati supportata dalla *listening platform* può giungere a verificare volumi e qualificazione positiva di alcuni elementi di attrattività del luogo, come quelli individuati nel modello di *branding* di Anholt (Figura 5). L'impiego di una struttura descrittiva dell'immagine percepita di un luogo si

FIGURA 5

Organizzazione delle top keyword per Londra sul modello di attrattività di Anholt

rende necessaria per integrare l'analisi dei social media tra gli strumenti già impiegati di governo.

L'analisi testuale finalizzata a filtrare i messaggi per rilevanza rispetto al modello euristico dell'attrattività del brand di un territorio, o secondo indicatori significativi su un *data-set* molto esteso, come è quello proveniente dall'ambiente dei social media, può essere in parte automatizzato come processo attraverso l'impiego del NLP, *Natural Language Processing*, con una diversa complessità tecnologica.

3. ASPETTI TECNOLOGICI DELLA SOCIAL MEDIA INTELLIGENCE

Un applicativo di Social Media Intelligence deve permettere di:

- ❑ monitorare come variano le opinioni nel tempo su un insieme di fattori competitivi;
- ❑ identificare aree di miglioramento e misurare il gap dal *best in class*;
- ❑ avere sotto osservazione i principali attori portatori di opinioni e poter reagire tempestivamente a situazioni pericolose per la propria reputazione.

Un elemento di criticità nei messaggi provenienti dai social network è la bassa qualità dei dati, dovuta all'elevata frequenza, nel già complesso linguaggio naturale, di slang, abbreviazioni e simboli che rendono inutile l'uso di un *parser* di testo tradizionale su flussi di conversazioni come quelli in Twitter: in 140 caratteri l'espressione testuale viene contratta e la tecnologia si trova ad avere poche informazioni di contesto per disambiguare correttamente il significato delle parole.

Si parla di *word sense disambiguation*, in altre parole, l'identificazione di un unico senso per ogni parola del discorso. In inglese, le performance dei vari algoritmi esistenti si attestano su un'accuratezza che raggiunge il 69% su corpora costruiti ad hoc per *SemEval-2007*, *Senseval-2* (*Wikipedia*), ma non c'è da sorprendersi poiché non è sempre facile distinguere di cosa si sta parlando.

«*I have just arrived in Milan. Here food is great!*»
«*I have just read great news about Alyssa Milan.*»
«I *love Milan Kundera...*»
«*Beating AC Milan is going to be a challenge!*»
Si sta parlando della città o di altro?

Nel progetto pilota del Comune di Milano l'applicativo sviluppato è in grado di comprendere testo non strutturato in lingua inglese, estratto da fonti Web eterogenee, attraverso tecniche di elaborazione del linguaggio naturale e l'uso di reti semantiche costruite ad hoc per il segmento di mercato in studio, producendo in output dei report che descrivono: cosa si discute online (gli argomenti d'interesse), chi ne parla (analisi degli *opinion maker*), come (positività e negatività dei giudizi espressi) e quando.

Uno strumento intelligente non si limita a contare le occorrenze di parole, ma integra capacità d'interpretazione semantica per una migliore disambiguazione, tramite l'uso di reti semantiche specifiche per le lingue d'analisi. L'uso di reti semantiche permette di risolvere i conflitti in casi ambigui quali "*milan*" (artista o città) o "*turkey*" (animale o nazione) (Figura 6).

L'accuratezza sulla disambiguazione del senso delle parole ha un elemento in comune con l'analisi delle opinioni espresse dagli utenti. Anche l'analisi di gradimento usa algoritmi complessi, sia con approcci basati su *data mining* sia seguendo metodi di elaborazione del linguaggio naturale (NLP) e non sempre produce come risultato un valore univoco, tanto che è possibile calcolare il gradimento con diverse scale[4]:

1. scala a due gruppi:
   - ***a.*** due possibili valori (0 e 1);
   - ***b.*** positivo/negativo;
   - ***c.*** accordo/disaccordo;
2. scala a più gruppi:
   - ***a.*** più valori possibili (1, 2, 3, 4, 5);
   - ***b.*** positivo/negativo/neutrale;
   - ***c.*** accordo/disaccordo/indecisione;

Noun

- S: (n) **turkey**, Meleagris gallopavo (large gallinaceous bird with fan-shaped tail; widely domesticated for food)
- S: (n) **Turkey**, Republic of Turkey (a Eurasian republic in Asia Minor and the Balkans; on the collapse of the Ottoman Empire in 1918, the Young Turks, led by Kemal Ataturk, established a republic in 1923)
- S: (n) joker, **turkey** (a person who does something thoughtless or annoying) *"some joker is blocking the driveway"*
- S: (n) **turkey** (flesh of large domesticated fowl usually roasted)
- S: (n) **turkey**, bomb, dud (an event that fails badly or is totally ineffectual) *"the first experiment was a real turkey"; "the meeting was a dud as far as new business was concerned"*

FIGURA 6
Esempi di diversi sensi per la parola "turkey" estratti da WordNet

[4] Conversition, 2009.

3. scala a gruppi illimitati:
   a. valori con posizione decimale (1.000, 4.82 ecc.).

Nell'architettura del sistema di analisi della reputazione online, ai moduli tecnologici di raccolta dati e rilascio delle analisi - comuni alle *listening platform* descritte - è necessario aggiungere moduli di *data processing* che operino la disambiguazione, una corretta categorizzazione rispetto al modello euristico prescelto e filtrino correttamente i dati con gradimento positivo o negativo.

4. PRIMA FASE TECNOLOGICA: CRAWLING

La realizzazione del progetto del Comune di Milano ha messo in luce come il modello più efficace ed efficiente per lo sviluppo di una piattaforma di analisi semi-automatica del gradimento sia basato su un'articolazione in quattro fasi distinte, ciascuna delle quali presenta sfide tecnologiche specifiche e richiede l'utilizzo di soluzioni applicative differenti, all'interno di un governo complessivo del progetto finalizzato a rendere fluide le integrazioni tra le varie fasi. In forma sintetica, abbiamo definito queste quattro fasi come: *crawling*, *mapping*, *processing* e *visualization*.
L'acquisizione di dati dal Web avviene grazie all'uso di tecnologie ormai standardizzate. Ciò nonostante, implementare uno strumento di *crawling* multi-sorgente presenta evidenti criticità, dovute all'eterogeneità delle fonti che si vogliono analizzare e ai contenuti cui si desidera accedere.
Si distinguono principalmente due tipologie di *crawler*, quelli basati su *parser* HTML (*HyperText Markup Language*) e quelli basati su API (*Application Program Interface*).
I primi (*crawler* su *parser* HTML) puntano ad acquisire i dati analizzando la struttura di una pagina Web, in modo molto simile al comportamento di un essere umano che naviga su internet: così come l'utente vede una pagina alla volta e naviga tra pagine diverse aprendo dei collegamenti ipertestuali, allo stesso modo il *crawler* analizza una pagina alla volta e sceglie la successiva sulla base degli attuali collegamenti (*link* e indirizzi URL) presenti. L'analisi della pagina Web è fatta attraverso dei *parser*. I *parser* sono algoritmi che identificano i tag HTML presenti e, in base ai tag scelti dall'utente, analizzano il testo presente al loro interno. È questo il contenuto che viene estratto e attraverso opportune connessioni alla base dati viene memorizzato per il riutilizzo dei dati in fase di analisi. Questo primo approccio permette una ricerca molto ampia, completa, a livello di sito Web, ma di contro richiede lunghi tempi di attesa (variabili da uno a venti secondi per pagina scansionata), paragonabili alla velocità di lettura manuale, e una connessione al sito sempre attiva. I tempi di attesa diventano un fattore da tenere in considerazione quando le fonti hanno un aggiornamento elevato, come nel caso dei social network, dove ogni giorno vengono generati contenuti nuovi. La connessione sempre attiva è un elemento non gestibile dalla tecnologia ma dipendente dalle *policy* del sito, perché c'è chi permette l'accesso ai dati e chi li custodisce gelosamente.
I vantaggi e gli svantaggi si invertono con il secondo tipo di *crawler*, basato su API (Tabella 1). Le API sono servizi offerti dal gestore del sito per permettere l'interfacciamento con i dati e/o le funzionalità del sito stesso. L'ottenimento delle informazioni è più veloce, ben gestito ma ristretto, perché l'accesso al sito è limitato a livello spaziale (solo alcune aree) e spesso temporale (numero di chiamate orarie fisse). Se il *crawler* basato su *parser* HTML è sempre utilizzabile, quello basato su API ha performance più elevate, ma richiede che il si-

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

TABELLA 1
Vantaggi e svantaggi per le tipologie di crawler

to Web abbia sviluppato e messo a disposizione interfacce ad hoc non disponibili di default. Gli esempi più noti sono rappresentati da Twitter e Facebook che mettono a disposizione API, in continua evoluzione, attraverso le quali è possibile scaricare i *tweet* o le *fan page*, ma non i messaggi privati o tutti quei dati coperti dalla privacy.

5. SECONDA FASE TECNOLOGICA: *MAPPING*

La fase di *tagging* può essere intesa come una classica categorizzazione di testi. Per *text categorization* si intende l'attività che si occupa di classificare testi digitali scritti in linguaggio naturale, assegnandoli in maniera automatica una o più classi/tag. Esempi di classi possono essere "trasporti", "arte", "sport". Per questa operazione si usano solitamente degli algoritmi con apprendimento automatico supervisionati, sistemi che è necessario addestrare tramite auto-apprendimento a partire da testi taggati manualmente. Altri approcci fanno riferimento ad algoritmi non supervisionati, in cui il sistema non richiede una fase di addestramento preliminare; tuttavia quelli che stanno attualmente ottenendo risultati più concreti sono i semi-supervisionati, dove le decisioni sono lasciate in parte alle macchine e in parte agli esseri umani, sfruttando così i punti di forza di entrambi e limitando gli svantaggi.
Algoritmi di apprendimento supervisionato si basano sull'utilizzo di due insiemi di documenti per elaborare il gradimento finale: il *training set* (insieme che "allena" l'algoritmo a produrre risultati migliori) e il *test set* (insieme per effettuare i test). Molti studi hanno potuto fare riferimento a siti di recensioni (per esempio di un hotel), dove l'utente fornisce già una classificazione, per esempio da 1 a 5 stelle, per il suo messaggio o post.
Le prime tecniche che hanno implementato queste idee sono state i classificatori *bayesiani* e le *Support Vector Machine* (SVM), tutte appartenenti alla branca del *Machine Learning*. Pang et al hanno così classificato le recensioni dei film in due classi, positive e negative, ma il buon risultato ottenuto era in parte dovuto al fatto che non era stata presa in considerazione la possibile esistenza di commenti neutrali. Se i risultati migliori sembrano ottenibili tramite apprendimenti supervisionati, la loro correttezza è limitata a particolari set di documenti, di conseguenza è poco adatta a catalogare tutta l'informazione presente online, che varia dai più comuni fino ai più sconosciuti argomenti, usando termini specifici, abbreviazioni e slang subculturali. Per questo motivo negli ultimi anni hanno ripreso importanza i sistemi non supervisionati, funzionanti in qualsiasi contesto. Questi si basano su uno schema, una struttura di conoscenza costruita a priori, quale una rete semantica o un'ontologia, in cui sono schematizzati tutti i concetti delle realtà (oggetti, azioni ecc.) e le relazioni tra essi ("composto da", "appartiene a" ecc.).
Possedere un'ontologia del linguaggio naturale non è sufficiente a dire di avere a disposizione uno strumento in grado di elaborare qualsiasi informazione scritta su fonti digitali. Diversi lavori hanno dimostrato che sono necessari algoritmi complessi che, basandosi su reti semantiche quali WordNet [2], OntoNotes [3], e SUMO [4], disambiguano il testo rendendolo disponibile a successive elaborazioni. Negli ultimi anni c'è stata una forte crescita tecnologica nel campo, portando risultati molto precisi, ma con costi ancora elevati e limitati contesti applicativi. Le performance ottenute, infatti, variano parecchio, da un'accuratezza del 54.2% ottenuta da Mihalcea [5] fino a valori del 96%, come nello studio di Yarowksy [6]. La minore precisione di Mihalcea non significa però un risultato più negativo, perché tecniche come la sua, applicano reti semantiche indipendenti dal contesto, in cui cioè non c'è una focalizzazione su particolari domini, ma permettono di essere applicati su qualsiasi argomento da analizzare: questo a differenza di tecniche dipendenti dal contesto, spesso basate su algoritmi supervisionati, che ottengono risultati più performanti a scapito dell'applicabilità pratica e del loro costo [7].
L'enorme progresso tecnologico degli ultimi anni viene incontro all'esigenza di analizzare contenuti generati dall'utente, permettendo di ricavare gli elementi essenziali per la ricostruzione di un insieme di senso partendo da analisi soggetto-azione-oggetto; soggettività/oggettività di una frase; analisi dei concetti (persone e luoghi) più importanti presenti nei testi. Queste tre analisi sono abilitate da un *parsing* sintattico e semantico delle

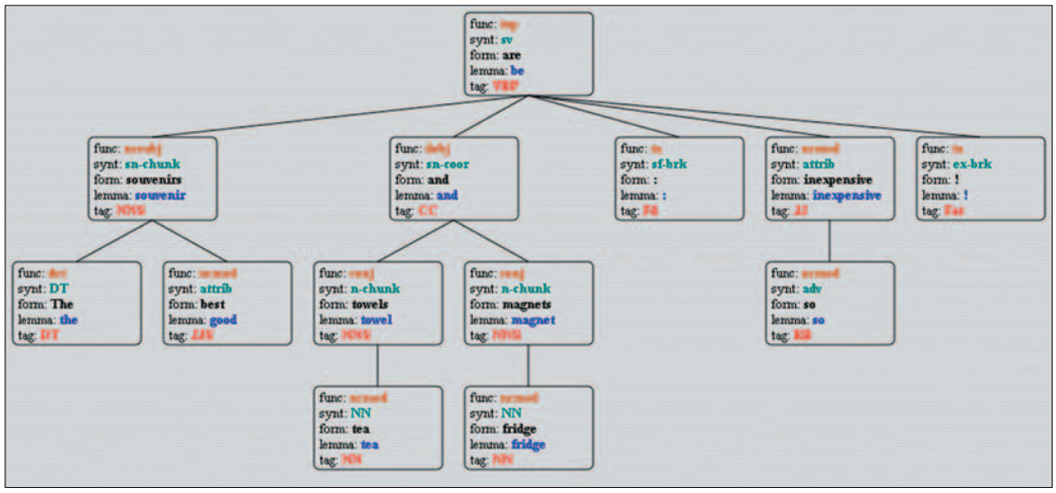


FIGURA 7
Parsing tramite l'analizzatore sintattico - Freeling [1] di una frase presa da TripAdvisor

frasi. Grazie a degli alberi delle dipendenze (*dependency tree*), risultato tipico degli analizzatori sintattici (Figura 7) si può mostrare all'utente i concetti del discorso e le relazioni tra essi con una certa sicurezza.
Pur non esente da errori, questa tecnica garantisce risultati con un'accuratezza superiore al 90-95% ed è di grande aiuto a chi ha il compito di costruire una mappa di concetti. Un parser è indispensabile per ottenere relazioni tra concetti. Se si volessero avere solo le entità citate, quali nomi propri e luoghi, ci si potrebbe basare su moduli NER (*Named Entity Recognition*), che sfruttano pattern, ricorrenze statistiche e matching da dizionari per riconoscere le varie entità.

6. TERZA FASE TECNOLOGICA: *PROCESSING (TAG & SCORING)*

La difficoltà si presenta non solo in fase di assegnazione di un significato alle parole ma anche a livello di assegnazione di un gradimento. Abbiamo già detto che l'*e-wom* tende a essere positivo, ovvero che i contenuti generati dagli utenti esprimono soprattutto entusiasmo per un meccanismo sociale di autorinforzo (l'eccellenza si riverbera dall'oggetto al soggetto) ma le esperienze su cui si esprimono opinioni sono estremamente parcellizzate: *The best Sbagliato in Milan* fa riferimento ad un particolare aperitivo, riconducibile ad un'esperienza di ristorazione.
Grazie a insiemi di sinonimi, parole identificanti concetti equivalenti (Milano, città della Madonnina, capitale lombarda ecc.) e l'identificazione di nomi propri, ogni frase può essere catalogata con precisione all'interno della mappatura adottata, garantendo un solido punto di partenza per il calcolo del gradimento. L'uso della semantica non è l'unica soluzione. Altri due approcci affrontano il problema da un punto di vista di distanza lessicale o da un punto di vista statistico. L'approccio lessicografico si basa su una finestra di n parole considerate per la classificazione. Per esempio, usando una finestra di dimensione due, si prendono in considerazione due parole a sinistra e due a destra dalla parola "cardine" analizzata e su di esse si applicano algoritmi di disambiguazione per contestualizzare e catalogare correttamente la parola. In inglese si chiamano algoritmi *sliding window*, perché la finestra scorre di volta in volta su tutte le parole della frase. Differente è l'approccio statistico, che si basa sulla ricorrenza di particolari relazioni tra lemmi derivate dall'osservazione di grandi basi dati testuali. I suggerimenti nella ricerca di Google ne sono un esempio. La correzione non si basa su un dizionario fisso, ma costruito tramite le ricerche fatte precedentemente da milioni di utenti.
Le diverse tecniche agiscono per individuare parole e per poterle categorizzare con meno errori, ma tra ottenere volumi su un determinato argomento e analizzarne il gradimento c'è un grosso margine di dati inutilizzati. Un vero e proprio imbuto, perché ricerche dimostrano che la percentuale di testi che hanno un gradimento esplicito si attesta su meno di un decimo del totale. Di questo insieme poi, ci sono fattori che introducono distorsione dei dati: la

differenza tra la neutralità e la mancanza di opinione, la tendenza a parlare positivamente di un avvenimento, oggetto o evento.
La maggior parte degli strumenti che tracciano la reputazione online non si preoccupa di differenziare opinioni neutrali e/o non sbilanciate (non sono né a favore né contro questa nuova legge) da frasi completamente senza gradimento (com'è andata ieri in parlamento?). Eppure per valutare correttamente come si parla on-line è necessario prendere queste frasi e analizzarle. Correlato è anche il discorso di considerare frasi soggettive e oggettive: frasi con opinioni personali (mi piace il Duomo) da frasi con evidenze fattuali (il Duomo è alto 150 m).
Le sfumature sono importanti, così come **il silenzio**. Perché non si parla di un certo fatto, evento o brand? L'ascolto del silenzio è possibile impostando analisi di tipo comparativo, per esempio quanto si parla di Food&Drink a Milano rispetto a Londra, ponderando con la base diversa di volumi di partenza.
L'altro aspetto è la **distorsione del gradimento**. Quando una persona pubblica un commento su un sito tende a voler dare un'impressione positiva di sé stesso, come abbiamo già accennato. Descrive quanto bella è stata la vacanza, quanto si è contenti del nuovo acquisto, della bella serata o del fantastico cibo mangiato al ristorante. L'uomo tende a valorizzare ciò che fa e questo comportamento traspare anche nei commenti online. In particolare su Twitter, gli utenti che perseguono il raggiungimento di un maggior numero di *follower* tendono a esprimere eventi positivi. Questo comportamento opportunistico è difficile da catturare con gli strumenti di *Business Intelligence* per l'analisi di gradimento, ma è in realtà molto importante perché introduce un elemento di distorsione che crea un divario notevole tra il numero di frasi positive e quelle con valenza negativa. La differenza tra i valori è notevole, dato che le frasi positive sono un numero quasi quattro volte superiore a quelle negative, considerando una classificazione a tre livelli (positiva, neutra, negativa). Questo risultato è confermato anche da fonti come Steve Kaufer, fondatore di TripAdvisor [8].
Nello scenario precedente si è fatta l'assunzione che la valenza fosse calcolata a livello di frase. In verità l'analisi di gradimento può essere affrontata a più livelli: **documento, frase** e parte di frase a senso compiuto, detto ***snippet***.
Il primo è la cosiddetta classificazione a livello di documento, l'assegnazione di un voto positivo o negativo a un testo. Questo è stato il punto di partenza, da cui poi si è derivata la necessità di estrarre informazioni a granularità più fine rispetto all'intero documento. Da qui è nata la classificazione a livello di frase, che ottiene valori più dettagliati, ma con costi maggiori.
Le tecniche di elaborazione a livello di frase prevedono due passaggi:
1. determinare se una frase ha contenuto soggettivo o oggettivo;
2. se una frase è soggettiva, determinarne la polarità: positiva, negativa, o neutrale.
Molte delle ricerche studiano entrambi i punti, alcune si focalizzano su uno dei due. In entrambi i casi gli algoritmi più utilizzati sono il *bootstrapping*, i classificatori *bayesiani naïve*, SVM e altri metodi statistici. Più precisamente per determinare la soggettività/oggettività sono usate tecniche quali la somiglianza delle frasi, i classificatori *bayesiani* singoli e multipli.
La valenza di una frase soggettiva invece è determinata basandosi su una lista di parole o sensi (in caso di analisi semantiche) aventi associati già un gradimento a priori. La classificazione a livello di *snippet* è la terza granularità (la più fine) per il calcolo del gradimento. A livello di documento si fa un'assunzione, che ci sia solo un valore di polarità per tutto il testo, quindi uno e un solo utente che descrivesse positivamente o negativamente un solo oggetto. Più si scende nei dettagli più si scopre che ciò non è vero: una frase può contenere più opinioni, proposizioni soggettive in contemporanea con proposizioni oggettive, più oggetti descritti. Il livello *snippet* punta a catturare tutti questi contenuti, distinguendo all'interno delle stesse frasi più sezioni, ognuna delle quali è composta da un'unica informazione specifica.
Ancora oggi esiste poco riguardo allo studio degli *snippet*, perché introduce la necessità di avere alta precisione sul contenuto in analisi con una tecnologia matura che non è ancora presente sul mercato.

7. QUARTA FASE TECNOLOGICA: *VISUALIZATION*

Disciplina nel cui ambito si registra attività di ricerca fin dall'inizio delle discipline informatiche, ma applicazione sistematica in ambito aziendale e pubblico solo dall'inizio del millennio, la *data-visualization* è assurta a fattore strategico di governo in occasione dell'impiego pubblico da parte del presidente Obama per mostrare agli elettori l'andamento della disoccupazione nei primi due anni del suo mandato.

I presupposti tecnologici che sostengono uno sviluppo autonomo della fase di visualizzazione risiedono nelle basi del *pattern* architetturale *Model View Controller* (MVC), che mira a tenere separata la parte di modellazione (quale la base di dati) dalla parte di logica (il controller, quale le classi java o C^{++} che chiamano il database ed eseguono algoritmi di analisi) e dalla parte di visualizzazione. Le visualizzazioni sono un mix tra design e tecnologia che devono permettere di far nascere conversazioni intorno ad esse [9], il loro scopo è comunicare e permettere il dialogo; sono il punto di partenza - e non un punto di arrivo - per un dibattito da cui poter estrarre conoscenza. Immagini e video hanno un potere enorme e riassumono in pochi colori e forme dati che spiegati a parole richiederebbero frasi e tabelle piene di valori.

Nel mondo del Web stanno nascendo numerose forme di rappresentazione dei dati, ma non esiste ancora un vero e proprio standard. Una visualizzazione deve essere immediata, sia per ciò che deve trasmettere, sia nella velocità di trasmissione dei dati alla parte grafica. Quest'aspetto tecnologico è importante quando si hanno notevoli quantità di dati, per non perdere il controllo delle informazioni catturate. È utile usare una logica a servizi, in cui solo nel momento del bisogno una rappresentazione fa richiesta di dati a un server con grandi capacità di elaborazione, che permette di alleggerire il carico del client di visualizzazione, che dovrà solo occuparsi di mostrare dati già filtrati e organizzati in base alle necessità.

Applicazioni pratiche che sfruttano l'approccio a servizi sono i *Mash-up* capaci di includere dinamicamente informazioni e contenuti provenienti da più fonti (Figura 8 e Figura 9).

8. CHE ALTRO C'È? PROSPETTIVE PER LA SOCIAL MEDIA INTELLIGENCE

La complessità tecnologica della comprensione dell'ambiente dei Social media e delle sue dinamiche è ben rappresentata dal caso illustrato in questa sede, relativo al progetto promosso e finanziato dal Comune di Milano sull'analisi della reputazione online del brand di una città.

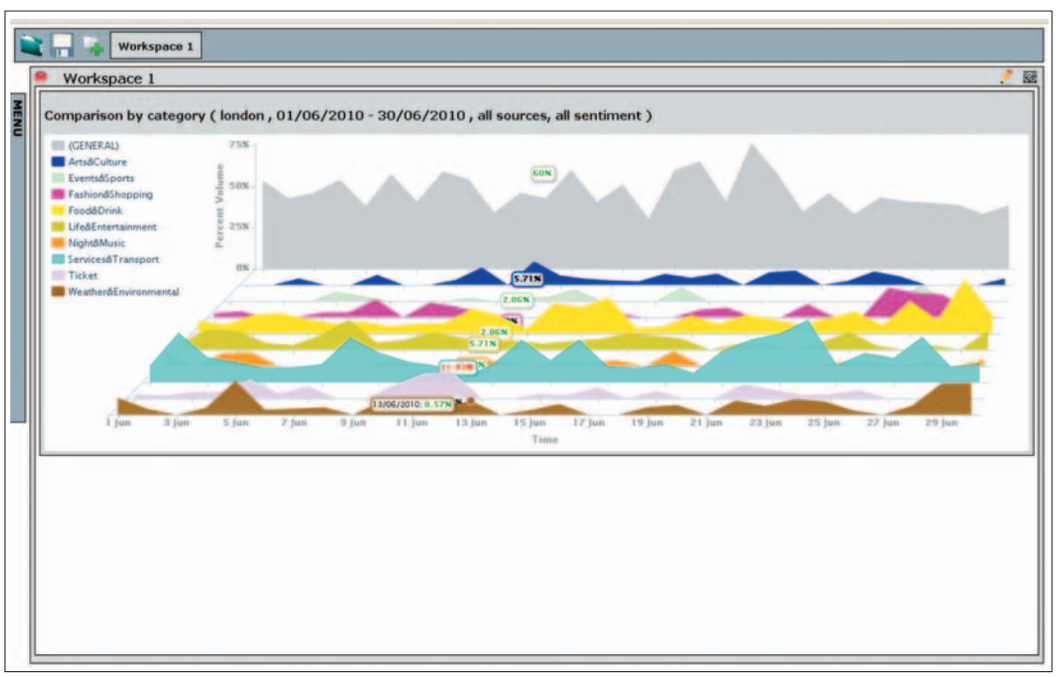


FIGURA 8
Esempio di applicazione mashup: analisi dei volumi per categorie nel settore turismo (Londra, giugno '10)

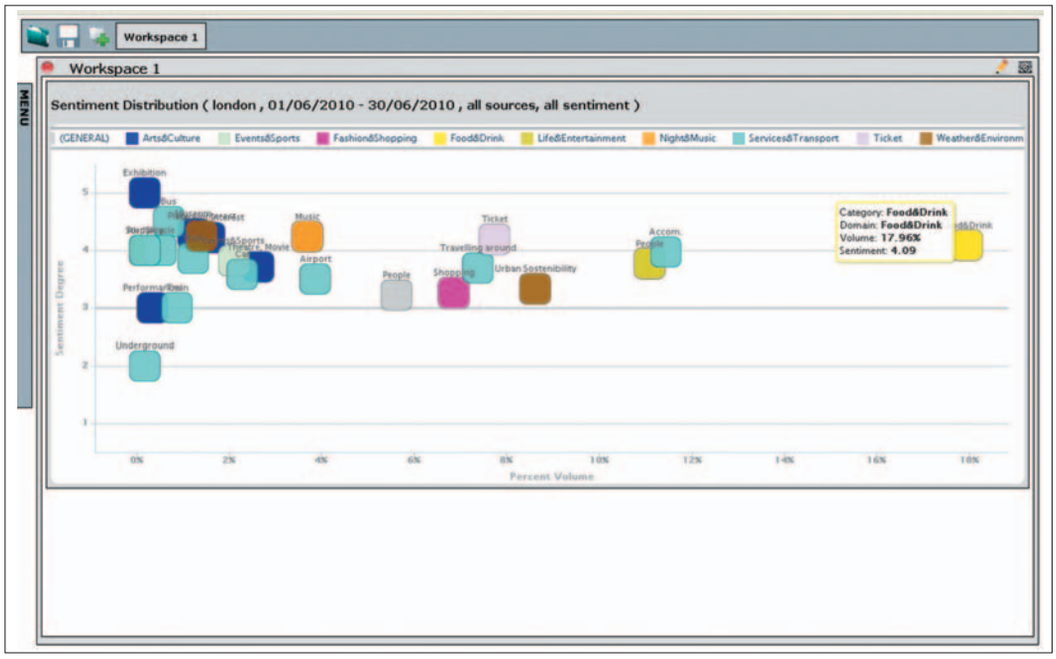


FIGURA 9

Esempio di applicazione mashup: analisi del gradimento per categorie nel settore turismo (Londra, giugno 2010)

Il diffondersi del cosiddetto mobile Internet e l'affermazione di un modello di interazione basato su "app" come quelle rese familiari da iPhone e Android, che da un lato offre un'esperienza utente più ricca e dall'altro integra a livello di metadati la componente di geo-localizzazione, pone indubbiamente nuove sfide: conoscere il "polso" di una città attraversata da persone in transito vuole dire parlare tante lingue, quante quelle usate da chi visita o lavora in una moderna metropoli ed essere molto veloci nell'ascoltare e comprendere un numero sempre maggiore di contenuti, destrutturati, contestuali, spesso espressione di forti appartenenze subculturali, potenzialmente distribuiti su un numero crescente di sorgenti rilevanti, ognuna caratterizzata da una propria architettura specifica a livello logico e tecnologico.

Il Politecnico di Milano in collaborazione con Commstrategy si propone come Centro di Competenza per progetti che vogliano sviluppare sistemi innovativi di Social Media Intelligence, a partire dall'allargamento del progetto in corso con il Comune di Milano. Le frontiere identificate sono in primo luogo quelle dell'allargamento del perimetro di analisi ad altre città, e territori; includendo un numero maggiore di dimensioni e allargando l'insieme delle sorgenti monitorate e delle lingue considerate. In secondo luogo, quelle legate all'estensione dell'architettura tramite moduli aggiuntivi, come quello di individuazione e valutazione degli "orientatori" a livello di notizie, persone e brand. Infine, quelle relative alla mappatura geografica dei nodi di scambio.

Ognuno di questi ambiti di sviluppo presenta sfide tecnologiche e concettuali estremamente rilevanti, ma i presupposti per affrontarle con successo sono da ricercare nei risultati che già oggi sono stati raggiunti.

Bibliografia

[1] Jansen B.J., Zhang M., Sobel K.:Twitter power: Tweet as electronic word-of-mouth. *Journal of the American Society for Information Science and Technology*, Vol. 60 n. 11, 2009, p. 2169-2188.

[2] Fellbaum C.: WordNet: *An electronic lexical database*. Cambridge, MA: MIT Press, 1998.

[3] S.S. Pradhan, E.H.: OntoNotes: *A Unified Relational Semantic. International Conference on Semantic Computing*. Washington DC: IEEE Computer Society, 2007, p. 517-526.

[4] Pease I. N.: *Towards a standard upper ontology*. Formal Ontology in Information Systems, New York: ACM. 2001, p. 2-9.

[5] Mihalcea R.: *Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling*. Human Language Technology Conference, Morristown, NJ: ACL. 2005, p. 411-418.

[6] Yarowsky D.: *Unsupervised word sense disambiguation rivaling supervised methods*. Annual Meeting of the Association For Computational Linguistics, Morristown, NJ: ACL 1995, p. 186-196.

[7] Barbagallo D., Bruni L., Francalanci C.: *Exploiting WordNet glosses to disambiguate nouns through verbs*. The Fourth International Conference on Advances in Semantic Processing, Firenze, 2010.

[8] Kaufer S.: *Social Media Insights from TripAdvisor CEO Steve Kaufer*. E. Qualman, Interviewer, 2009, 4 20.

[9] Wattenberg M.: *IBM Wants Many Eyes on Visualization*. T. O'Reilly, Interviewer, 2007, 01 23.

[10] Atserias J., Casas B., Comelles E., González M., Padró L., & Padró M.: FreeLing 1.3: *Syntactic and semantic services in an open-source NLP library*. International Conference on Language Resources and Evaluation (LREC), Genoa, 2006.

[11] Bartlett F.: *Remembering, a study in experimental and social psychology*. Cambridge University Press, 1932.

[12] *Conversition*: There's Nothing Neutral About Neutral, 2009.

[13] Conway D.: *My Five Rules for Data Visualization,* 2009, 12 03. Retrieved from: http://www.drewconway.com/zia/?p=1582

[14] Friendly M.: *Gallery of Data Visualization*, 2001. Retrieved from York University: http://www.math.yorku.ca//SCS/Gallery/

[15] *Il Sole 24 Ore*, 20 maggio 2010. Retrieved from: http://www.ilsole24ore.com: http://www.ilsole24ore.com/art/SoleOnLine4/Tecnologia%20e%20Business/2010/05/facebook-privacy-modifiche.shtml?uuid=704c84a0-6429-11df-87da-3032239fa3f5&DocRulesView=Libero

[16] Nigam K., Hurst M.: *Towards a Robust Metric of Opinion. Spring Symposium on Exploring Attitude and Affect in Text*. Pittsburgh, 2004.

[17] *Word sense disambiguation*. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/-Word_sense_disambiguation

Paolo Barbesino, PhD in Sociologia, socio fondatore e Managing Director di CommStrategy, New Media Strategic Insight, dal 1997 supporta importanti attori internazionali, pubblici e privati, nella costruzione della strategia e nell'approccio al mercato utilizzando l'ICT come fattore critico di successo. Coinvolto anche come membro del board in nuove imprese tecnologiche, è stato mentor al Qatar Science & Technology Park di Doha. È responsabile della collaborazione con il Politecnico di Milano, sia sul Mobile sia sulla Social Media Intelligence.
E-mail: p.barbesino@commstrategy.com

Chiara Francalanci è professore associato di Sistemi Informativi al Politecnico di Milano. Ha scritto numerosi articoli sulla progettazione e sul valore economico delle tecnologie informatiche, svolto attività di ricerca e consulenza nel settore finanziario e manifatturiero sia in Italia sia presso la Harvard Business School ed è editor del Journal of Information Technology.
E-mail: francala@elet.polimi.it

Fiamma Petrovich, Senior Manager in CommStrategy, New Media Strategic Insight. Master in Marketing (1994) e Master in Internet Business (2001). Appassionata del mondo digitale, da 10 anni coinvolta su progetti di innovazione in ambiente Internet e Mobile. Sta lavorando allo sviluppo di knowledge su Social Media e reputazione online in un progetto di ricerca con Politecnico di Milano, attraverso l'impiego di tecnologia semantica su un brand di territorio. Background professionale in Consulenza Strategica (Servizi, Government, Automotive), e da manager nell'area Marketing/Comunicazione (FMCG).
E-mail: f.petrovich@commstrategy.com

Glossario

App': applicazioni, gratuite o a pagamento, che possono essere scaricate direttamente dal dispositivo mobile o su un computer; a giugno 2010 nell'App Store ne erano disponibili 220 mila, con 5 miliardi di download accumulati.

Branding: progetto strategico e creativo di creazione e gestione dell'identità e dell'immagine di marca.

Business Intelligence: insieme composto dai processi aziendali per raccogliere e analizzare informazioni strategiche, la tecnologia utilizzata per realizzare questi processi e le informazioni ottenute come risultato di questi processi.

Crawling: azione del software che analizza i contenuti di una rete (o di un database) in modo metodico e automatizzato, ad esempio per conto di un motore di ricerca.

E-wom: passaggio di informazioni da persona a persona attraverso la rete digitale.

Follower (in Twitter): utenti che si sono iscritti ai messaggi di un altro utente.

Listening platform: sistema di ascolto della rete digitale con accesso web attraverso password a pagamento.

Mapping: localizzazione di elementi rilevanti all'interno in un sistema organizzato di dati per la creazione e l'organizzazione di una mappa di informazioni.

Mash-up: applicazione web che include dinamicamente informazioni o contenuti provenienti da più fonti.

Natural Language Processing: trattamento automatico delle informazioni in linguaggio umano che ne affronta le ambiguità in un processo di elaborazione con tecniche di analisi lessicale, grammaticale, sintattica e semantica.

Social Media: tecnologie e pratiche online che le persone adottano per condividere contenuti testuali, immagini, video e audio.

Tag: termine associato a un contenuto digitale, anche per facilitarne l'indicizzazione nei motori di ricerca.

Visualization: tecnica per creare immagini, diagrammi, o animazioni per comunicare un messaggio.

Word sense disambiguation: problema aperto del natural language processing, che fa riferimento al processo di identificare il senso di una parola (cioè significato) usato in una frase, quando la parola ha significati multipli (polisemia).