# SENTIMENT ANALYSIS NEL SETTORE TURISTICO: RISULTATI E SFIDE FUTURE

Luca Dini*, Massimo Balestrieri**

*Abstract*- In this paper we present an application of sentiment analysis to the domain of tourism. It will be shown that, given the specific aspects of the domain, any sentiment analysis application must be able to deal with very fine-grained features of opinion detection, such as phrase level granularity, intensity of the judgement, capability of identifying potentially dangerous trend. We will show by examples how these features are implemented on the Senti-Miner 1.0 framework, a generic opinion mining system which has been tailored specifically for the target domain.

## 1. Introduzione

Sin dalla fine degli anni '90 (Dini et al. 1997), con la crescente disponibilità di media open source e in particolar modo di contenuto generato dagli utenti, un gran numero di ricercatori si è dedicato alla creazione di programmi che fossero in grado di comprendere le opinioni spontaneamente scritte dagli utilizzatori del web.

Questi programmi sono chiamati "sentiment analysis systems" o "opinion monitoring systems" (il termine in letteratura è considerato equivalente) e sono una parte importante dei sistemi basati su tecnologie di Natural Language Processing (NLP).

Il campo di applicazione delle tecniche di sentiment analysis comprende un gran numero di domini tra cui valutazioni di media, valutazioni di prodotti e commenti , opininioni sui mercati finanziari, etc... Nonostante questo sono molto pochi i lavori dedicati specificatamente al sentiment analysis per il settore turismo, con l'interessante eccezione di Kasper e Vela (2011) e Gräbner et al. (2012), che comunque prendono in esame solamente le valutazioni di Hotel. Per quanto a nostra conoscenza nessun studio è stato condotto sull'analisi delle opinioni nel turismo in un contesto multilingua, e anche in studi monolingua è evidente che le lingue prese maggiormente in considerazione sono Inglese e Tedesco.

In questo lavoro, dopo aver enfatizzato l'importanza del sentiment analysis nel settore turismo (sezione 1), viene presentato *Senti-miner 1.0 for Tourism*, descrivendo le tecnologie di base (sezione 2). Nella sezione 3 si descrivono alcune caratteristiche fondamentali per un rendere un sistema di sentiment analysis realmente efficace nel

* Holmes Semantic Solutions (Grenoble, France), dini @ho2s.com
** Holmes Semantic Solutions (Italia), balestrieri@ho2s.com

dominio specifico e si conclude con una sezione sulle principali sfide future per questo genere di applicazioni.

## 2. L'importanza del Sentiment Analysis nel settore Turismo

Nel settore del turismo si distingue tra due categorie principali di informazioni e tre categorie di attori fondamentali. La prima categoria di informazioni è rappresentata da recensioni su Alberghi e Ristoranti (e simili). Siti web come TripAdvisor e Booking.com adottano da anni sistemi di valutazione delle strutture ricettive basati su indicatori numerici e commenti testuali scritti dagli utenti. In alcuni casi si può avere la certezza che il commento sia stato scritto effettivamente da un utilizzatore della struttura. Questo tipo di fonti di dati sono chiamate "Competition Sources" e l'informazione che può essere estratta "Competition Information". Dall'altra parte si hanno siti come PaesiOnline, 4square, ma anche forum sui viaggi (come Turistipercaso, ed altri) dove gli utenti scrivono i loro commenti a proposito delle mete turistiche spontaneamente o piuttosto sollecitati dalla richiesta di consigli da parte di un altro utente. Questo tipo di fonti sono chiamati "Territory Sources" e le informazioni estratte "Territory Information".

Dal punto di vista dei soggetti interessati, si possono distinguere queste 3 categorie

1. Hotel/Restaurant Managers: seguono una o più strutture con evidenti obiettivi di tipo commerciale.
2. Pubblici decisori: hanno il potere di influenzare l'evoluzione di un territorio o di una meta turistica applicando politiche adeguate (Marketing territoriale);
3. Lobbisti: con questi termini si intendono tutti quei soggetti la cui missione è di promuovere il settore del turismo sia lavorando sul lato hotel / ristorante che sul lato meta di destinazione: Associazioni di Categoria, uffici turistici, centri di studio, ecc

Nella tabella seguente si mostra in che senso ogni tipo di soggetto abbia interesse a comprendere entrambi i tipi di informazioni:

| | Compet. Source | Territory Source |
|---|---|---|
| H/R Managers | Questa è la loro principale fonte di informazioni. Nonostante il fatto che i siti web come Tripadvisor propongano un punteggio numerico, è raro che la griglia di classificazione proposta sia adeguata al gestore che in molti casi deve approfondire aspetti che sono specifici per la | E' di cruciale importanza per impostare il proprio marketing. Il gestore può affinare il proprio marketing e le sue attività proposte , grazie alla comprensione di quali siano le caratteristiche del territorio che realmente sono apprezzate (e criticate). Inoltre, per un |

| | | |
|---|---|---|
| | propria struttura o la propria catena. | manager di catene alberghiere è uno strumento fondamentale per differenziare I propri piani di marketing rispetto al territorio. |
| Public Managers | Essi possono monitorare il livello di ospitalità del territorio, eventualmente proponendo politiche di miglioramento della ricettività. | Possono avere una istantanea in tempo reale della situazione del turismo nel territorio, le sue tendenze e ciò che potrebbe essere migliorato. Il confronto con posti confinanti e con mete con circa le stesse caratteristiche potrebbe essere molto utile. |
| Lobbyst | Come rappresentanti delle categorie hanno sempre bisogno di dati per affrontare e sostenere le azioni di lobbying. | L'Ufficio Turismo utilizzerà i dati del sentiment analysis per promuovere le campagne più efficaci di marketing territoriale |

Tutto questo testimonia l'estrema importanza del sentiment analysis per il settore del turismo, nonché la necessità di adeguati strumenti personalizzati. Nel seguito descriveremo Senti-Miner 1.0 uno strumento che è stato appositamente configurato per rispondere a tali esigenze.

## 3. Senti-Miner 1.0

**Senti-Miner** è una derivazione di **Sybille 2.0**, un sistema per opinion monitoring presentato al DEFT07 (Maurel et al. 2007, vedi anche Maurel & al. 2008, Maurel & al. 2009), dove si è classificato terzo nella competizione fra sistemi e primo tra quelli industriali. Contrariamente a Sybille 2.0, che è basato su XIP (Mokhtar & al. 2001), Senti-Miner è costruito al livello più alto del sistema HOLMES (Hybrid Operable platform for Language Management and Extensible Semantics). L'assuzione di base del sistema HOLMES è che l'ibridazione delle diverse tecnologie è essenziale per ottenere buone prestazioni in opinion mining e per compiti generici di estrazione dati. Esso è basato su un modello di elaborazione flessibile (molto simile alle assunzioni della piattaforma Stanford NLP) dove diversi "annotatori" (sistemi automatici in grado di arricchire il testo; ad esempio: analisi morfologica, disambiguazione lessicale, analisi sintattica, annotazione semantica ...) sono disposti in linea e dove ogni annotatore può beneficiare del trattamento di tutti gli annotatori precedenti.
E' stato adottato il modello generale per cui abbiamo inserito coppie di annotatori con funzionalità comparabili nella linea, uno basato su tecniche statistiche (per lo più

controllate), e uno basato sulla configurazione manuale. Il ruolo del linguista diventa correggere l'uscita del modello statistico sulla base di opportune regole. Ad esempio, HOLMES contiene sia un modulo basato su Conditional Random Field (CRF) per il riconoscimento di named-entity (Lefferty & al. 2001) e un modulo di correzione basato su TokenRegexp (Levy e Galeno 2006), un tagger sintattico stocastico e un componente basato sulla'pplicazione di regole a corrispondenza lineare, un modello basato su MaltParser per l'analisi della dipendenza e un componente basato sulla trasformazione di grafi per individuare e correggere gli errori di analisi.
Nel caso di Senti-Miner, il meccanismo di base HOLMES è stato arricchito con un componente di analisi semantica, descritto nella sezione successiva.

*3.1 Analisi semantica come transformazione di grafi*

Da lungo tempo, un filone di ricerca in semantica computazionale (cf. (Sowa, 2008) per coprire in modo esaustivo la letteratura) ipotizza che una rappresentazione ottimale della semantica del linguaggio naturale possa essere ottenuta utilizzando una rappresentazione a grafi. Il livello semantico di HOLMES (che include il modulo di sentiment analysis) si basa su questo presupposto. Fondamentalmente, I predicati (interpretati secondo la logica dei predicati classica) sono rappresentati da archi che collegano i nodi, i quali a loro volta corrispondono alle entità identificate nel testo e arricchite con specifica informazione semantica.

Per esempio , la frase "*Le **patient** est **réadressé** en **Service d'Orthopédie le 19.02.2012.***" ("Il ***paziente** viene **trasferito** al **Reparto Ortopedia** il **19.10.2012**.") è rappresentato come in Figura 1.

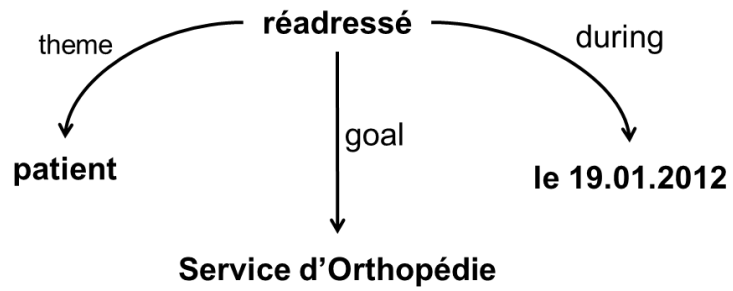



*Figura 1: rappresentazione semantica prodotta da HOLMES.*

Dal momento che l'output del'analisi di base di HOLMES è reppresentato da un grafico di dipendenza, è naturale concepire il successivo processo di arricchimento semantico come un processo di trasformazione di grafi, lungo le linee (Bonfante et al., 2010 e Ribeyre, 2012) . Nel nostro caso è ulteriormente facilitato dal fatto che l'output è conforme al paradigma di dipendenza di Stanford (Manning & Marneffe 2008, Cer & al. 2010, Robin & al., 2013), cioè con rappresentazioni grafiche che sono più vicine a una rappresentazione semantica che a grafi di dipendenza "standard" basati su sintassi .
La vera sfida di concepire l'analisi semantica come trasformazione di grafi sta nel fatto che le regole che disciplinano le varie fasi di trasformazione hanno bisogno di accedere a una grande quantità di informazioni semantiche sintattiche e lessicali,

mentre le piattaforme di trasformazione basate su grafi standard, di solito offrono la possibilità di gestire alfabeti limitati.

Per questo motivo il naturale orientamento è stato verso grafici *con attributi* come descritto in (Fisher & al., 1998), che trova un'implementazione naturale nell'ambiente AGG (Taentzer, 2000). Il software è dotato di un'interfaccia utente grafica che permette all'utente di scrivere le regole per la manipolazione del grafico.

Ogni regola descrive un grafo di ingresso a sinistra, un grafo di uscita a destra e un grafo "host", dove la trasformazione può essere testata. Per quanto sia attraente, l'interfaccia AGG GUI è abbastanza controproducente per compiti intensivi di scrittura regole.
Questo probabilmente perché i linguisti sono più orientati verso la scrittura di regole formali dichiarative piuttosto che disegnare archi tra oggetti. Inoltre, per ragioni di efficienza e manutenzione, vi era la necessità di limitare il potere formale di AGG in modo tale che in fase di applicazione I vincoli non funzionali di efficienza e computabilità fossero soddisfatti. Per queste ragioni, abbiamo progettato un linguaggio dichiarativo per la trasformazione del grafico. Il linguaggio consente operazioni sui grafici, come ad esempio la creazione e la cancellazione di archi e nodi, la dichiarazione e l'assegnazione di attributi, ecc Permette anche chiamate arbitrarie ai metodi Java per testare precondizioni di applicazione e per assegnare i valori funzionali. Una semplice regola nel nostro linguaggio di trasformazione grafico è simile alla seguente:

[1][2 pos_pol=true pos=adj][3 pos=ADV type=invert] {mod(1,2), mod(2,3)} => [1][2]{Negative_opinion(1,2)}

Questa regola stabilisce che se un aggettivo a polarità positiva modifica un sostantivo, ma è esso stesso modificato da un avverbio a polarità negativa, si stabilisce tra il sostantivo e l'aggettivo una relazione di parere negativo (per inciso, la regola inoltre elimina l'avverbio, dato che il suo contributo alla semantica può essere considerato esaurito).
Regole di questo tipo costituiscono il nucleo del modulo che è stato usato per lo studio descritto in questa sede. Le regole sono divise in componenti indipendenti (o livelli), che si applicano nell'ordine scelto dal linguista.

### *3.2 Senti-Miner 1.0 for Tourism*

La grammatica di base di HOLMES per il sentiment analysis contiene circa 50 livelli di trasformazione per lingua (attualmente copre inglese, francese e italiano). Al livello superiore della grammatica italiana per il sentiment analysis abbiamo aggiunto 43 regole per il settore del turismo. Queste regole hanno accesso alle tassonomie e ontologie (luoghi, alloggi, hobby e sport, monumenti, ecc) legati al turismo, al fine di filtrare le informazioni irrilevanti. Inoltre, una parte di queste regole è dedicata a interpretare espressioni linguistiche che hanno una polarità specifica nel settore del turismo, ma non necessariamente nel linguaggio ordinario (e quindi non sono contenute nella grammatica di base). Ad esempio, un'espressione come "mi Sono

divertito molto" può avere una connotazione positiva in un dominio come il turismo o il tempo libero ("A Cortina mi sono divertito molto") ma essere neutrale in altri domini ("mi sono divertito molto a smontare la stampante").

## 4. Cosa dovremmo aspettarci dal Sentiment Analysis nel settore turismo?

Nel seguito descriveremo una serie di caratteristiche che sono fondamentali per effettuare Sentiment Analysis nel settore turismo. Rappresenteremo l'importanza di tali caratteristiche visualizzando i risultati di un sondaggio sentiment analysis condotto con Senti-Miner 1.0 per il turismo nel periodo 01/03/2013 al 2013/01/11, avendo come fonti Tripadvisor e Booking.com per "fonti concorrenza ") e PaesiOnLine (per " fonti territorio ").

### *4.1 La granularità del "sentiment"*

E' una pratica diffusa nelle interazioni del social web assegnare un punteggio globale alle valutazioni di oggetti digitali e fisici (un film, un articolo, un luogo, un hotel, ecc). Questa pratica si scontra con il fatto che i pareri sono complessi e spesso i sentimenti contrastanti e solo raramente possono essere rappresentati come un giudizio "globale". Oltre a questo il giudizio globale può offuscare "gli atomi di giudizio", che sono quelli più utili.

La stessa pratica "dubbia" è stata adottata per un po' nel campo dell'analisi automatica del sentimento: capire se un testo ha un valore positivo o negativo è stato considerato come un problema di classificazione dei documenti (cioè se un documento è stato "naturalmente" classificato nella categoria "documenti negativi "o" documenti positivi "). E' evidente che in questo modo tutte le sfumature di una frase vengono affogate in un punteggio generale che a malapena può contenere informazioni utili: basti pensare che l'87% dei testi di valutazione nel settore del turismo contiene un mix di giudizi positivi e negativi.

Sarebbe quindi naturale, anche se più difficile, passare al livello frase in modo da catturare opinioni più dettagliate. Ma anche le frasi non sono unità di informazione atomiche, e assegnare un punteggio di positività spesso significa sovrascrivere l'intento dell'utente nel descrivere un luogo o un punto di interesse (POI). Ad esempio, in una valutazione al livello della frase, l'espressione

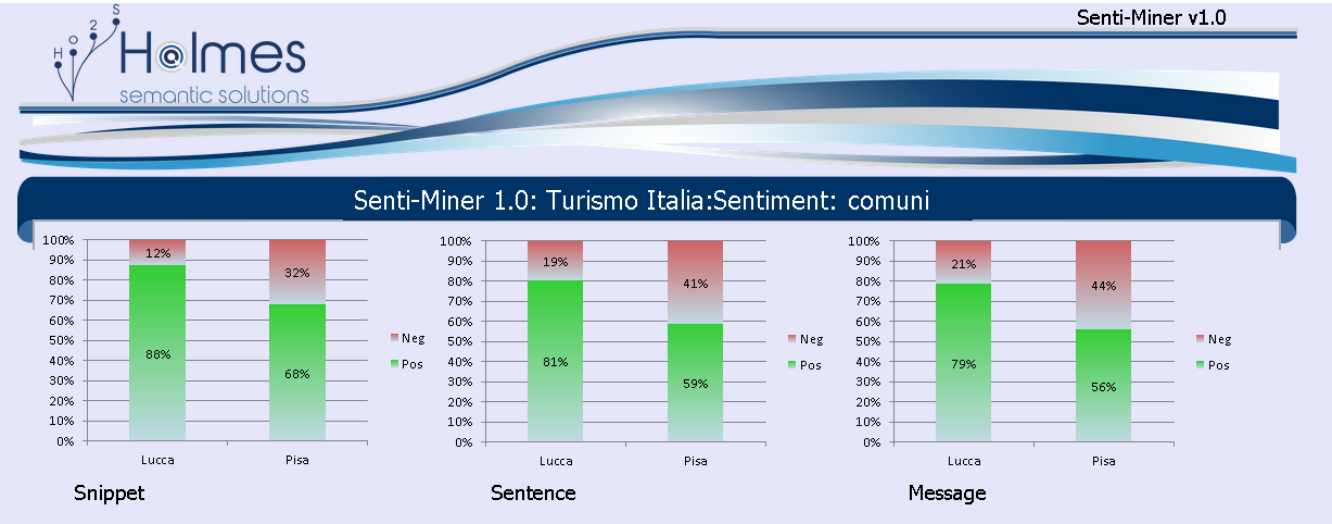
*It is a quite charming place, but accessing it is quite a nightmare*

*E' un luogo molto affascinante, ma arrivarci è un po' un incubo*

otterrebbe un valutazione di neutralità (qualcosa come 0,5 in una scala da 0 a 1), anche se in realtà contiene due espressioni molto chiare. Più precisamente abbiamo osservato che il 46% delle opinioni del nostro campione sul turismo, contiene più di un giudizio e il 23% contiene giudizi di polarità opposte.

Tutto ciò rende chiaro che l'unica granularità affidabile nel sentiment analysis (almeno per il turismo) è sotto le strutture atomiche linguistiche (per lo più sintagmi), che chiameremo "opinioni atomiche" o "snippets". In base a tale concezione un manager del marketing turistico, avrebbe capito, dalla frase di cui sopra che il luogo ha un grande potenziale, ma l'accesso potrebbe essere di gran lunga migliorato.

Per dare un'idea dell'impatto di queste tre diverse granularità nel settore del turismo forniamo qui uno screenshot di Senti-miner con 3 diverse classificazioni del sentiment per le città di Pisa e Lucca. È evidente che un'analisi a livello di documento complessivo rischia di dare valori che sono molto distanti dalla realtà delle impressioni dell'utente.




A livello di intero messaggio (documento) abbiamo un 21% / 44% di giudizi negativi, che si riducono a 19% / 41% se si confrontano le frasi e a 12% / 32% per i singoli snippets.

*4.2 L'oggetto del "sentiment"*

Un punto cruciale per dare senso a milioni di opinioni è naturalmente la capacità di comprenderne l'oggetto. Ad un primo livello l'oggetto di un parere è sempre un entità, ad esempio un hotel, un ristorante, un monumento, una città, ecc.. L'oggetto di un parere viene individuato abbastanza facilmente in siti di recensioni dove I pareri sono naturalmente collegati ad una tale entità. Diventa però più difficile da individuare se si considerano fonti di informazioni come forum e blog. In questi casi solo le tecnologie semantiche quali *Named Entity Recognition* and *Dependency Parsing* possono aiutare a capire quale sia l'entità oggetto del giudizio espresso in una frase come:

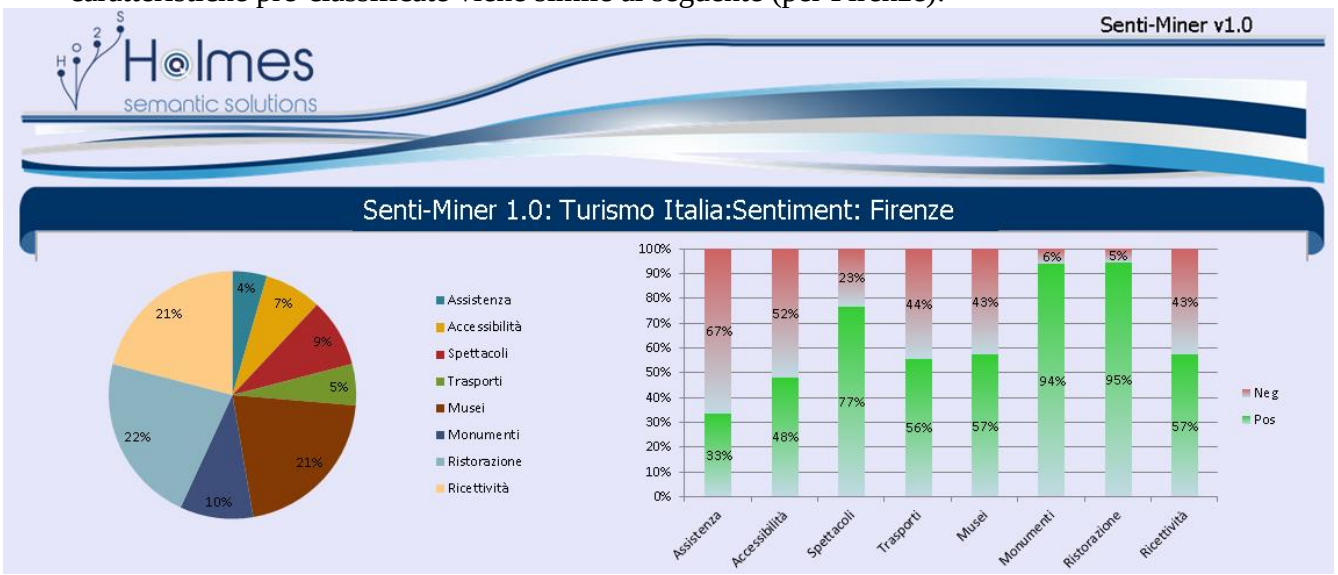
*I have been to Forte dei Marmi and Viareggio: the former is nice and peaceful (but expensive), the latter is chaotic and crowded (but cheaper).*

*Sono stato a Forte dei Marmi e a Viareggio: la prima è bella e tranquilla (ma costosa), la seconda è caotica e affollata (ma meno costosa).*

Identificare l'entità è, naturalmente, la base per un sistema di rilevamento del sentimento: altrimenti sarebbe del tutto inutile. Tuttavia, oltre a comprendere le tendenze generali, la semplice associazione dei giudizi e delle entità è di poco valore per un manager del turismo che ha bisogno di capire quale caratteristica specifica dell'entità sia oggetto di giudizio e di come la caratteristica specifica evolva nel tempo.
In questo senso sono disponibili tre tecnologie, vale a dire l'approccio a caratteristiche predefinite, l'approccio clustering e l'approccio linguistico. Essi sono dettagliati nelle sezioni seguenti.

### *4.2.1. Set di caratteristiche pre-definite*

E' l'approccio più classico per raggiungere una piena comprensione del set di pareri. In generale, l'operatore stabilisce "a priori" una serie di categorie che pensa siano di suo interesse. È poi la fase di configurazione del sistema che produce la mappatura tra l'insieme di tali categorie e la loro espressione linguistica nel contesto delle frasi contenenti le opinioni. Un diagramma di opinione tipico derivante dalle caratteristiche pre-classificate viene simile al seguente (per Firenze):



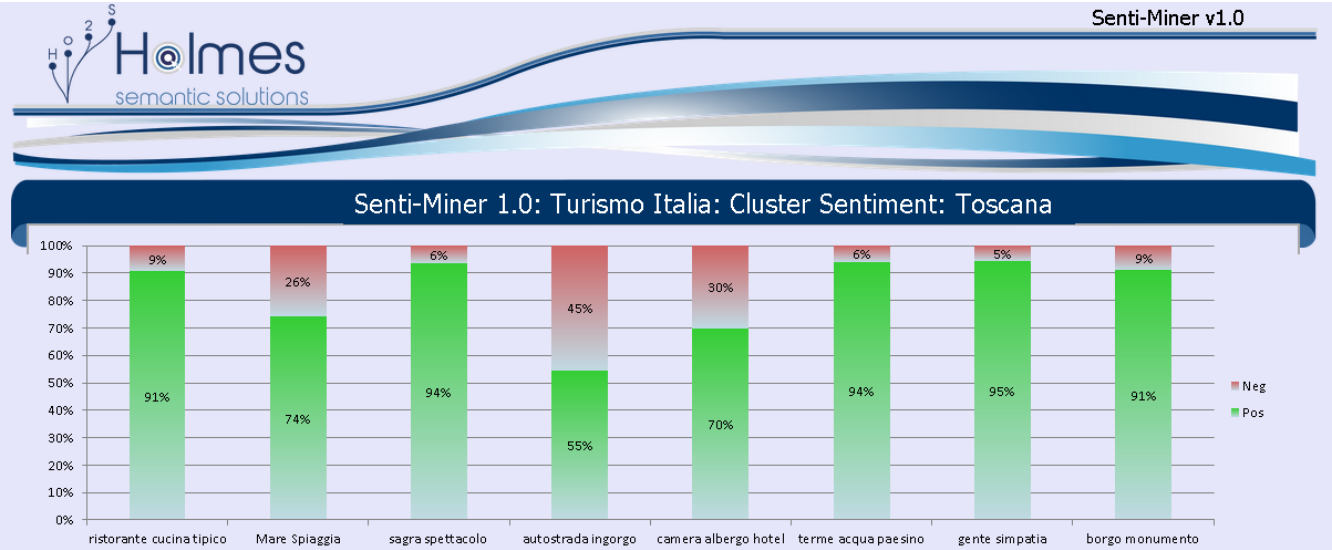

Il vantaggio di questo metodo è, ovviamente, che il manager del turismo, che dovrebbe conoscere il settore in modo molto dettagliato, ha il pieno controllo sui risultati dell'analisi. D'altra parte, vi è un certo rischio di proiettare un modello tradizionale "precostituito" del settore su un insieme di dati che è in continua evoluzione e che potrebbero sfuggire alle categorie di settore / dominio tradizionali.

### *4.2.2. L'approccio Clustering*

"Automatic Clustering" è una tecnologia generale che consente di raggruppare insiemi arbitrari di testi in categorie "naturali" che emergono non in seguito alla sovrapposizione di una struttura predefinita ma come un albero dei temi emergenti direttamente dall'insieme dei testi analizzati. Ad esempio raggruppando tutte le opinioni riguardanti la regione Toscana nel corpus del turismo otteniamo un grafico come segue:


Dove la dimensione delle bolle rappresenta l'importanza di un certo argomento e le parole connesse rappresentano le prime tre parole più rappresentative per quel gruppo. E 'evidente da tale grafico che parte del tema della classificazione predefinito descritto nella sezione precedente non appaia affatto come l'argomento più importante di discussione.

Una volta raggruppati i documenti (o meglio ancora, le frasi o gli snippets), è quindi possibile applicare una metrica ai singoli cluster  per verificare la loro polarità. Per la regione Toscana si ottiene la seguente intersezione cluster / pareri:




Si possono notare temi 'emergenti' come 'sagra spettacolo' , 'terme acqua paesino' , 'gente simpatica'.

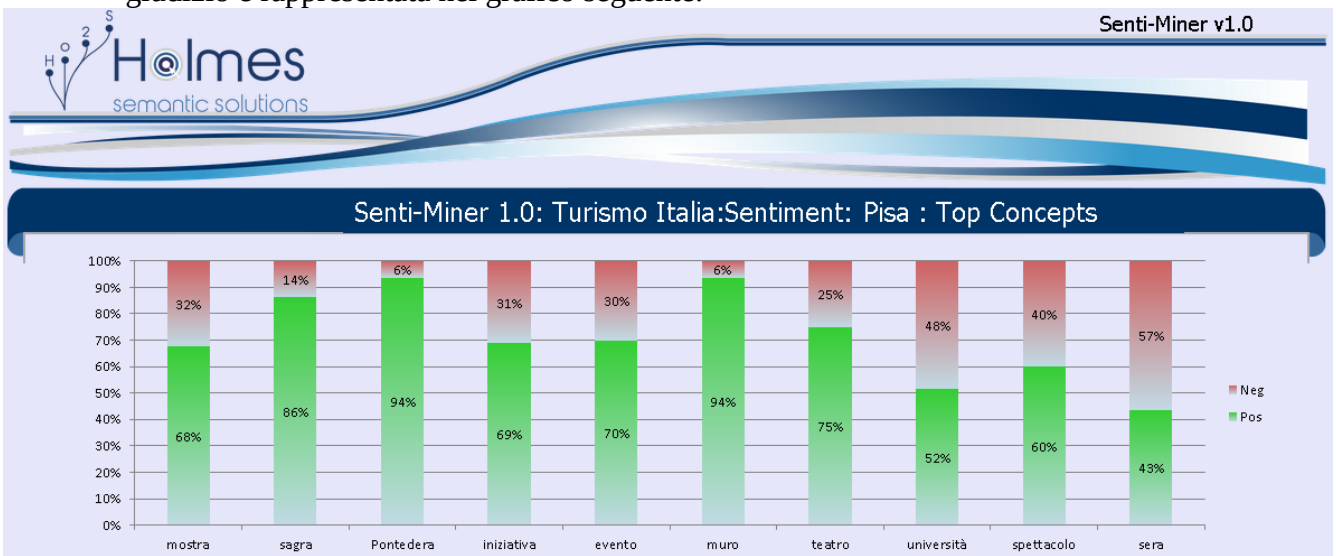
*4.2.3 L'approccio linguistico*

L'approccio linguistico unisce la precisione dell'approccio a caratteristiche predefinite alla flessibilità del clustering. L'idea di base è quella di analizzare l'intera frase in modo da comprendere, da un punto di vista linguistico, qual è il nome della frase oggetto del giudizio e usando la testa lessicale normalizzata (in pratica, la parola più importante) come rappresentazione delle caratteristiche specifiche. Per esempio in una frase come

*That new bridge, I don't like the way they reshaped it!*

*Il nuovo ponte, proprio non mi piace come l'hanno rifatto!*

L'elaborazione linguistica capisce che il sostantivo *ponte* è l'effettivo oggetto diretto del verbo piacere e la considera come una possibile caratteristica di giudizio. Su larga scala, considerando una città come Pisa, la top ten del testo oggetto di giudizio è rappresentata nel grafico seguente:



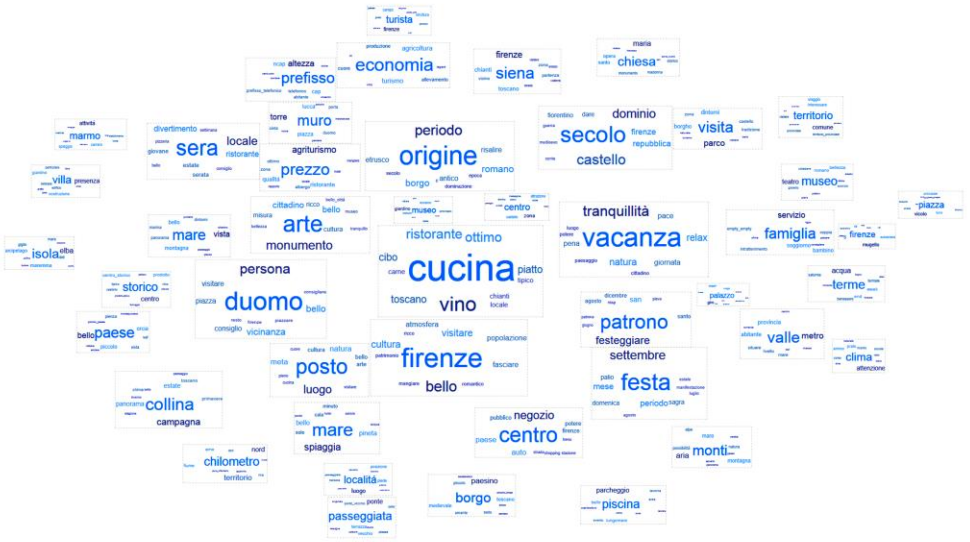
Non ci sorprende se notiamo che la vita notturna ("sera") a Pisa genera un sentimento piuttosto negativo, mentre "muro" (= mura, cinta muraria) ha una valutazione estremamente positiva.

### *4.3 Individuare tendenze inattese*

La Sentiment analysis è sicuramente uno strumento fondamentale per il marketing turistico, sia se orientato alla struttura che al territorio. Tuttavia, ci potrebbero essere aspetti che emergono nelle discussioni degli utenti che non sono necessariamente correlati al sentimento. Il fatto che le persone sui social media abbiano appena cominciato a parlare di un determinato argomento (o, al contrario, che improvvisamente nessuno ne parli piu'), può essere una informazione importante. In genere si tratta di un ottimo strumento per misurare l'impatto delle iniziative di marketing tradizionale e web marketing. E' anche un buon modo per monitorare certi "rumors" che rischiano di diventare nocivi per un territorio specifico.

Comprendere l'imprevisto è un compito tradizionalmente difficile nel settore del Trattamento Automatico del Linguaggio (TAL), quasi per definizione. Tuttavia, alcuni recenti sviluppi tecnologici, come la disponibilità di pacchetti software in grado di calcolare in modo efficiente "Topic Models" secondo l'ipotesi LDA (Latent Dirilecht Allocation), rendere il compito un po' meno difficile. L'idea è quella di calcolare la distribuzione di un argomento latente per un certo periodo e confrontarlo, ad esempio, con la distribuzione per il periodo successivo. Qualsiasi discordanza può essere rappresentata come un grafico in grado di fornire indicazioni utili per l'analista.

Ad esempio, nel seguente raggruppamento notiamo che nel mese di settembre 2013, per la regione di Firenze, gli aspetti legati all'alimentazione hanno la priorità rispetto all'argomento di discussione più tipico, legato ad arte e cultura.




## 5. Sfide e nuove prospettive nella sentiment analysis

In questo lavoro abbiamo stilato brevemente le principali esigenze per eseguire una ragionevole sentiment analysis nel settore del turismo. Tuttavia il requisito a priori, quello che pregiudica tutti gli altri, indipendentemente dalla modalità di rappresentazione delle informazioni è la *qualità*. Il concetto di qualità è abbastanza vago nel campo dell'analisi semantica dal momento che la valutazione quantitativa (ad esempio, precisione e richiamo) è spesso soggettiva e benchmark comunemente riconosciuti non sono disponibili. Pertanto il controllo di qualità deve basarsi su una serie di "condizioni abilitanti" che ogni sistema deve rispettare. Dalla nostra esperienza maturata sul campo emerge che un insieme ragionevolmente completo di indicatori di qualità può essere costruito secondo le seguenti linee:

- *"Linguality"*: il sistema dovrebbe essere esplicitamente sostenuto da intelligenza semantica basata su strutture linguistiche (grammatiche) per le lingue selezionate. Gli approcci a "Sacco di parole" (*bag of words*), come quelli utilizzati, ad esempio, dai motori di ricerca, non saranno mai in grado di comprendere le sfumature dell' espressione di opinioni, in cui l'ordine delle parole, la negazione, la modificazione avverbiale e modale possono cambiare completamente il significato di una frase.

- *"Verificability"*: Grafici a torta e istogrammi sono in grado di fornire un'istantanea della situazione e hanno un altissimo potere descrittivo. Tuttavia è fondamentale che in qualsiasi momento l'utente sia in grado di scendere al livello di frammento / frase / documento, al fine di verificare l'attendibilità

delle informazioni estratte: infatti tutti i sistemi semantici sono soggetti ad errori, ed un'attenta verifica deve sempre essere fatta.

- *"Configurability"*: la configurabilità è come conseguenza della tecnologia sottostante. I Sistemi basati sul apprendimento automatico sono davvero molto veloci da addestrare, a volte molto efficaci, ma non c'è spazio per la configurazione: una volta che il modello è stato appreso, si può solo replicare più e più volte su nuovi dati. Tuttavia, un sistema di sentiment analysis che possa funzionare realmente deve consentire all'utilizzatore (direttamente o tramite un servizio di assistenza) di modificare il comportamento del sistema in qualsiasi momento, ad esempio con l'introduzione di nuovi elenchi di sinonimi, l'aggiunta di nuove entità, l'adattamento di espressioni dipendenti dal dominio, ecc.

Oltre a queste caratteristiche, che caratterizzano lo stato dell'arte dei sistemi di estrazione di opinioni, ci sono caratteristiche che sarebbe auspicabile avere, ma che oggi si trovano ancora al di là della frontiera tecnologica e per le quali sono necessarie ulteriori ricerche. Si tratta in generale fenomeni linguistici e retorici assi complessi, tra i quali:

- Anafora: riferimenti a quanto accennato in precedenza dallo scrivente possono essere compresi con sistemi avanzati, ma il tasso di errore è abbastanza alto, soprattutto quando il riferimento può essere risolto solo grazie ad una conoscenza specifica di dominio.

- Ironia: Alcuni utenti amano esprimersi in termini ironici. L'ironia è un meccanismo retorico talmente complesso che a volte è difficile da capire anche per gli esseri umani. I sistemi di Sentiment Analysis attuali non sono in grado di affrontare questo fenomeno, che per fortuna è abbastanza raro nel dominio del turismo.

- Dialogo: Alcuni media open source (come forum, chat, ecc) sono intrinsecamente dialogici, nel senso che le opinioni sono spesso espresse in relazione alle opinioni espresse in precedenza. La struttura complessa di riferimenti e citazioni contenuta in una discussione tra utenti che interagiscono è qualcosa che in questo momento va al di là della semantica e coinvolge l'analisi del discorso, una disciplina che, da un punto di vista computazionale, è ancora in fase di sviluppo.

## 6. Bibliografia

Gräbner,K., M. Zanker, G. Fliedl and M. Fuchs. Classification of Customer Reviews based on Sentiment Analysis. In M. Fuchs, F. Ricci and L.Cantoni (eds), *Information and Communication Technologies in Tourism 2012*, Vienna: Springer.

Kasper W and M. Vela. Sentiment Analysis for Hotel Reviews. In *Proceedings of the Computational Linguistics-Applications Conference*, Jachranka, Poland, Polskie Towarzystwo Informatyczne, Katowice, 10/2011