Università degli Studi di Salerno

Dipartimento di Ingegneria Elettronica ed Ingegneria Informatica

Dottorato di Ricerca in Ingegneria dell'Informazione
XI Ciclo – Nuova Serie

TESI DI DOTTORATO

Un metodo basato su LDA per la Sentiment Analysis

CANDIDATO: **PAOLO ROCCA COMITE MASCAMBRUNO**

TUTOR: **PROF. MASSIMO DE SANTO**

COORDINATORE: **PROF. ANGELO MARCELLI**

Anno Accademico 2011 – 2012

Ringraziamenti

E’ inevitabile che, al termine di questo triennio, i pensieri vadano a ritroso nel percorso che mi ha condotto alla stesura di queste pagine. Ripenso all’alternarsi delle piccole vittorie quotidiane legate alla gioia per una conferma alle mie ipotesi trovata in un articolo, piuttosto che in una nota, e delle sconfitte legate a ogni smentita emersa dalla pagine che stavo consultando. E così via via, giorno dopo giorno, mentre prima una timida ipotesi e poi un progetto di ricerca sempre più articolato cominciavano a prendere forma.

E proprio questo alternarsi di emozioni, la consapevolezza che anche una semplice frase consultata nei lavori di quelli che ci hanno proceduto riesca a scatenare il nostro intuito, riesce a rendere la ricerca in sé estremamente affascinante. E rappresenta, al tempo stesso, una incredibile palestra di vita, dove anche le frustrazioni, ridimensionando le nostre ambizioni a volte troppo esasperate, contribuiscono a formare le nostre coscienze e il nostro modo di essere.

Nello specifico poi, per me è stato ancor più affascinante riuscire a monitorare proprio i moti dell’animo, i sentimenti o meglio il sentiment, ampiamente descritto nelle pagine che seguono. Mi hanno appassionato le possibili implicazioni di questa disciplina e ho percepito quali realmente applicabili al quotidiano i risultati della mia ricerca, piccolo tassello in una moltitudine di scritti sull’argomento.

E anche sentirsi parte di questo insieme, contribuire seppur in modo infinitesimale e nel proprio piccolo, a questo grande patrimonio di scoperte, rappresenta un’indubbia attrazione della ricerca in sé, dalla quale mi sono lasciato affascinare in questi anni.

Ma tutti questi percorsi non sono mai solitari e ci vedono accompagnati da mille figure alle quali è quasi impossibile formulare un ringraziamento esaustivo senza il rischio di dimenticare qualcuno. Eppure in questa moltitudine di persone alle quali in modo unanime va il mio ringraziamento, non posso non evidenziare chi mi è stato

vicino un po’ più degli altri e che più intensamente ha condiviso ogni giornata di questo lungo percorso.

Non posso non iniziare dal mio “prof” De Santo, colui che per primo ha saputo stimolarmi ad approfondire ogni argomento e che non ha mai smesso di chiedermi continuamente il perché di ogni mia affermazione, aiutandomi a capire fino in fondo il senso della “validità” scientifica di ogni elemento della ricerca. E al “prof” va in modo particolare il ringraziamento per l’incondizionata fiducia riposta in me anche nei non pochi momenti di difficoltà. E poi non è facile trovare il giusto modo di dire davvero e “profondamente” grazie a Francesco. Nel quotidiano ha seguito ogni mio passo e con pazienza è riuscito a rendermi ovvie anche le cose che proprio non riuscivo ad assimilare. E soprattutto i sui consigli sono stati utilissimi, e a volte indispensabili, nell’ultimo periodo, per mettere ordine nei pensieri e per costruire nel modo più efficace possibile tutti gli ambienti sperimentali.

Ci sono poi le persone più importanti, i miei genitori e mia moglie.

In modo diverso, ma costantemente, mi hanno supportato e soprattutto “sopportato” in questi anni, ed è a loro che dedico ogni singola frase, ogni immagine e ogni piccola affermazione di queste pagine che seguono.

Indice

Introduzione

La progressiva diffusione dei social network, sia generalisti (quali Twitter, Facebook o Google+, la recente piattaforma messa a punto da Google) sia specializzati (ad esempio le comunità professionali di Linkedin), ha reso disponibili una massiccia e inedita quantità di dati sulle preferenze sulle opinioni degli utenti.

Questi dati, disponibili in quantità significative e in tempo reale, possono essere sia di carattere orizzontale e quindi di caratterizzazione territoriale (grazie anche alla possibilità di geo localizzazione degli smartphone, tablet e integrati nei MID (Mobile Internet Device) di ultima generazione, sia verticale (in termini anagrafici, culturali, professionali e in termini di orientamenti settoriali e capacità di spesa).

La qualità e la contestualità (temporale e spaziale) di questi dati sulle tendenze degli utenti, con l'affermazione (oramai consolidata) del Web 2.0 (oggi Web 3.0), amplia la dimensione partecipativa alle scelte dell'impresa e delle istituzioni. Più che sul giacimento di idee e di opinioni disponibili in rete, imprese e istituzioni dovrebbero puntare sul valore delle comunità virtuali per modificare le proprie strategie di marketing, partendo proprio dal concetto di stakeholder[1], sviluppando, con il monitoraggio degli orientamenti, nuove metodologie di CRM[2].

Lo sviluppo, in questi settori, di applicazioni web e di servizi dedicati al mondo del multimedia in genere deve sempre presupporre

[1] Con il termine stakeholder (o portatore di interesse) si individuano i soggetti influenti nei confronti di un'iniziativa economica, sia essa un'azienda o un progetto. Fanno, ad esempio, parte di questo insieme: i clienti, i fornitori, i finanziatori (banche e azionisti), i collaboratori, ma anche gruppi di interesse esterni, come i residenti di aree limitrofe all'azienda o gruppi di interesse locali. La definizione fu elaborata nel 1963 al Research Institute dell'università di Stanford. Il primo libro sulla teoria degli stakeholder è "Strategic Management: A Stakeholder Approach" di Edward Freeman, che diede anche la prima definizione di stakeholder, come i soggetti senza il cui supporto l'impresa non è in grado di sopravvivere.

[2] Il concetto di Customer relationship management (termine inglese spesso abbreviato in CRM) o Gestione delle Relazioni coi Clienti è legato al concetto di fidelizzazione dei clienti. Per un approfondimento sul concetto si veda Giaccari, F. Giaccari, M., CRM Magazine: Cosa è il CRM, *CRM magazine*, Lecce, 2009

un'attenta fase progettuale basata sull'analisi dei bisogni degli utenti ai quali si vuole rispondere.

L'analisi in questione viene storicamente basata su indagini di soddisfazione, questionari utente e complesse operazioni di rilevamento del mercato basate su tecniche tradizionali, lunghi tempi di realizzazione, elevato costo di implementazione e livello di affidabilità proporzionale solo alla elevata onerosità delle azioni intraprese.

Negli ultimi anni, tuttavia, si sta sviluppando un interessante filone scientifico in risposta a questa tipologia di esigenze che si basa sull'analisi del comportamento degli utenti durante l'esposizione di un messaggio pubblicitario relativo al prodotto o servizio, piuttosto che durante (e questo è proprio il caso degli strumenti web e multimediali) la fruizione dello stesso.

La famiglia di strumenti e metodologie che rientrano in questo scenario è rappresentata dalla Sentiment Analysis (piuttosto che dall'Opinion Mining).

La Sentiment Analysis è alla base delle strategie di marketing e di comunicazione, e ne definisce i rispettivi orientamenti e nuove azioni, come lo stesso marketing virale sui social network.

Oggi queste tecniche di analisi e monitoraggio dei gusti e delle opinioni degli utenti sono in forte crescita, con applicazioni immediate nel settore della business intelligence e del monitoraggio della reputazione dei brand.

L'opportunità è offerta dalla stessa natura dei social network, dove la dimensione cognitiva si coniuga con quella emotiva, valorizzando la condivisione, dunque l'aspetto comunitario piuttosto che quello individuale. Ciò ha riflessi immediati sul consumo di prodotti e servizi, introducendo valori collettivi quali quelli legati all'ambiente, all'etica, al benessere sociale.

Dunque imprese e istituzioni, attraverso la rete, potrebbero acquisire informazioni ed esperienze del pubblico potenziale, definito, come detto, sia in senso orizzontale sia in senso verticale. Si tratta di dati nuovi, che non provengono né dalla rete di prossimità (stakeholder) né dal marketing off line.

Al contrario del marketing, che registra in modo passivo il target potenziale orientando le scelte dell'azienda, la Sentiment Analysis diventa elemento dinamico di interrelazione, alimentando a sua volta la passione (anzi, facendo leva sulle passioni) degli utenti ed il loro attaccamento al brand.

Conoscere, con sufficiente precisione, il "mood[3]" di consumatori e utenti nei confronti di un determinato prodotto o servizio: oggi è possibile grazie a strumenti che analizzano opinioni e commenti, quali i tools usati in ambito di Business/Government Intelligence, CRM e Brand Reputation Management, in grado di individuare, filtrare, organizzare e analizzare i dati, focalizzando sia gli aspetti quantitativi che qualitativi che tracciano la positività o la negatività delle opinioni e il relativo grado di intensità emotiva. Quest'ultima rappresenta un fattore particolarmente importante, in grado di influire su un sentiment negativo rispetto al brand.

La condivisione, infatti, conta più dell'orientamento del commento[4].

Dal punto di vista della ricerca e della tecnologia impiegata si tratta di elementi relativamente giovani, per i quali, tuttavia, esiste già una casistica affermata. Ciò almeno per quanto riguarda l'analisi dei testi e la previsione delle "sensazioni" che gli stessi sono in grado di suscitare nei fruitori. Al contrario, l'analisi delle immagini, del parlato e del multimedia in genere rappresenta uno scenario ancora essenzialmente non molto studiato per il quale non sono numerosi gli esempi applicativi.

Scopo di questo lavoro è quello di fornire innanzitutto un'ampia panoramica dello stato dell'arte delle famiglie di metodologie e strumenti adottati dalla comunità scientifica e dallo stesso mercato nel settore della sentiment analysis.

---

[3] Il *mood* rappresenta lo stato emozionale di un soggetto e differisce dalle emozioni poiché a differenza di esse è meno specifico e meno intenso. Un mood può essere determinato da un evento particolare che ne modifica, appunto, lo stato. Thayer, Robert E. (1998). *The Biopsychology of Mood and Arousal*. New York, NY: Oxford University Press.

[4] Il "mi piace" su Facebook, alla base della strategia di advertising del social network di Mark Zuckerberg

Si partirà proprio da tutto quanto relativo all'analisi dei testi scritti (sia nell'ambito generico che negli strumenti di net society) per poi propagare le ricerche nell'ambito dell'analisi per la sentiment analysis relativa alla multimedialità in generale.

Sarà necessario identificare uno spazio di lavoro utile a sviluppare una mirata fase di sperimentazione che consenta di mettere a confronto questi strumenti con il reale comportamento dei soggetti.

Ciò consentirà, infine, da un lato, di valutarne il livello di affidabilità e, dall'altro, di configurare un ambiente che, a valle di questa sperimentazione, possa essere efficacemente adottato durante la fase di progettazione di qualsivoglia interfaccia o applicazione che si desideri fortemente orientata agli utenti.

Oltre che nel marketing vero e proprio e nel lancio di nuovi prodotti industriali, settori questi dove la ricerca si sta indirizzando in maniera approfondita, questo ambiente consentirebbe un utilizzo in una variegata casistica di scenari che spaziano dalle applicazioni per l'istruzione e l'insegnamento a distanza, le piattaforme per l'erogazione di servizi assistenziali e sanitari, i portali della pubblica amministrazione fortemente interessati a un elevatissimo gradimento nella fruibilità degli utenti e, infine (ma non per importanza) tutto l'insieme di applicazioni web e multimediali dedicati al mondo della diversa abilità. Tali attori, infatti, sono ancora più sensibili ai modelli di fruizione di un messaggio e proprio dall'analisi delle loro reazioni (su prodotti reali o simulate nell'ambito della sperimentazione) si ottengono informazioni molto ricche da utilizzare successivamente nello sviluppo.

Per quanto riguarda l'organizzazione delle pagine che seguono, dopo una prima parte dedicata all'analisi dello stato dell'arte di metodologie, opinioni scientifiche e strumenti, verrà presentato il progetto dell'ambiente sperimentale e individuato il campione di sperimentazione, ottenuto con una serie di adattamenti delle interfacce di utilizzo e di combinazione di strumenti software appositamente dedicati.

Nella terza fase verranno esaminati nel dettaglio tutti i dati sperimentali ottenuti nel corso dell'attività e nella quarta ed ultima parte verranno illustrate le conclusioni scientifiche del lavoro e

presentato l’ambiente definitivo utile ai futuri impieghi sia scientifici che direttamente rivolti al mercato sia del marketing sia dello sviluppo di interfacce utente.

Prima parte

Introduzione alla Sentiment Analysis

1.1 Il sentiment e le emozioni contenute in un testo

Il trattamento automatico del modo di pensare di una classe di individui e della loro opinione, delle loro sensazioni (il loro "sentimento") e della soggettività rispetto a un testo, piuttosto che a un immagine o a un prodotto multimediale in genere sono indicati rispettivamente come "opinion mining", "sentiment analysis" e "subjectivity analysis".

La proliferazione di questi termini riflette la differenza nelle loro connotazioni, e non ha mancato di sollevare confusione.

Il termine "subjectivity" si riferisce a tutti quegli stati emozionali cosiddetti privati, i quali, quindi, non si prestano ad un'osservazione e verifica oggettiva dei dati acquisiti. Sulla base di questa definizione, quindi, tutte le opinioni, le valutazioni e le emozioni ricadono in questa categoria.

Ma con la *subjectivity analysis* si intende il riconoscimento del linguaggio opinion-oriented al fine di distinguerlo dal linguaggio oggettivo.

Per opinion mining si intende, invece, l'elaborazione di un insieme di risultati di ricerca per un determinato elemento, che genera una lista di attributi di prodotto (qualità, caratteristiche, ecc.) e l'aggregazione di opinioni su ciascuno di essi (povero, buono, ecc.).

Con sentiment analysis, infine, si intende l'analisi di un testo (o di un generico multimedia) mediante l'evidenziazione di giudizi e mood previsti che tale oggetto dovrebbe suscitare. E in questo senso quando si adottano i termini "sentiment analysis" e "opinion mining" in senso lato, allora essi denotano lo stesso campo di studio.

Abbiamo utilizzato il termine mood che ricorrerà spesso nelle pagine a seguire; valutiamone quindi un primo significato per poi approfondirne, successivamente tutte le differenti adozioni negli specifici contesti.

Il mood rappresenta una condizione persistente dello stato emotivo di un soggetto. In pratica si può definire come un'espressione di uno stato d'animo di un soggetto nei confronti di un oggetto, una persona

o un avvenimento. Le modalità per rintracciare un mood e caratterizzarlo sono molteplici e potrebbero partire dall'analisi della mimica facciale (e quindi dall'espressione) piuttosto che dalla postura o da taluni movimenti. Ma il mood può essere monitorato anche dal linguaggio, da un post lasciato in una bacheca o gruppo di discussione, piuttosto che dall'utilizzo di particolari termini, dal testo e dal parlato.

E proprio sulle analisi testuali sono numerosi i casi sperimentali presenti in letteratura. Di essi, tuttavia, non verrà fornita una trattazione specifica; la nostra attenzione andrà a soffermarsi sulle combinazioni di questi strumenti e metodologie che consentono di analizzare oggetti multimediali più o meno complessi.

Oltre che lo stesso testo, infatti, si farà riferimento a immagini, video, audio e parlato. Per quest'ultimo, poi, mediante meccanismi di riconoscimento del testo si ricadrà di fatto nelle analisi testuali vere e proprie.

1.2 Le origini della sentiment analysis

Sebbene l'interesse per la Sentiment Analysis rappresenti un settore in rapido sviluppo soprattutto negli anni recenti con l'attivazione di numerosi progetti di ricerca sullo specifico oggetto, l'argomento è stato caratterizzato da un interesse costante già negli anni passati.

E' possibile considerare l'analisi dei "Beliefe Systems[5] come un primo ambito d'indagine di questo settore e quindi gli studiosi interessati[6,7] come veri e propri precursori sull'argomento.

Successivamente a questa prima fase la ricerca si è focalizzata sull'interpretazione delle metafore testuali e in particolare sul riconoscimento e la classificazione dei differenti punti di vista degli autori, degli stati affettivi e le evidenze lessicali, così come per tutti i settori correlati[8,9,10,11,12,13,14,15,16]

---

[5] I "Beliefe Systems" sono sistemi di credenze combinate che si influenzano reciprocamente determinando il comportamento di masse di individui o singoli individui in relazione ad esse.

[6] Jaime Carbonell. *Subjective Understanding: Computer Models of Belief Systems*. PhD thesis, Yale, 1979

[7] Yorick Wilks and Janusz Bien. *Beliefs, points of view and multiple environments*. In Proceedings of the international NATO symposium on artificial and human intelligence, pages 147–171, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

[8] Marti Hearst. *Direction-based text interpretation as an information access refinement*. In Paul Jacobs, editor, Text-Based Intelligent Systems, pages 257–274. Lawrence Erlbaum Associates, 1992.

[9] Alison Huettner and Pero Subasic. *Fuzzy typing for document management*. In ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes, pages 26–27, 2000.

[10] Mark Kantrowitz. *Method and apparatus for analyzing affect and emotion in text*. U.S. Patent 6622140, 2003. Patent filed in November 2000.

[11] Warren Sack. *On the computation of point of view*. In Proceedings of AAAI, page 1488, 1994. Student abstract.

[12] Janyce Wiebe and Rebecca Bruce. *Probabilistic classifiers for tracking point of view*. In Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, pages 181–187, 1995.

[13] Janyce M. Wiebe. *Identifying subjective characters in narrative*. In Proceedings of the International Conference on Computational Linguistics (COLING), pages 401–408, 1990.

[14] Janyce M. Wiebe. *Tracking point of view in narrative. Computational Linguistics*, 20(2):233–287, 1994.

Ma è dall'anno 2011 che si diffonde una consapevolezza diffusa di tutte le criticità relative all'ambito di ricerca della sentiment analysis e di tutte le opportunità che essa può riservare. E' a partire da quella data, infatti, che si rileva la presenza di un numero consistente di articoli[17,18,19,20,21,22,23,24,25,26,27,28] in letteratura sullo specifico argomento.

[15] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. *Development and use of a gold standard data set for subjectivity classifications*. In Proceedings of the Association for Computational Linguistics (ACL), pages 246–253, 1999.

[16] Janyce M. Wiebe and William J. Rapaport. *A computational theory of perspective and reference in narrative*. In Proceedings of the Association for Computational Linguistics (ACL), pages 131–138, 1988.

[17] Claire Cardie, JanyceWiebe, TheresaWilson, and Diane Litman. *Combining low-level and summary representations of opinions for multi-perspective question answering*. In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, pages 20–27, 2003.

[18] Sanjiv Das and Mike Chen. *Yahoo! for Amazon: Extracting market sentiment from stock message boards*. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

[19] Kushal Dave, Steve Lawrence, and David M. Pennock. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In Proceedings of WWW, pages 519–528, 2003.

[20] Luca Dini and Giampaolo Mazzini. *Opinion classification through information extraction*. In Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining), pages 299–310, 2002.

[21] Hugo Liu, Henry Lieberman, and Ted Selker. *A model of textual affect sensing using real-world knowledge*. In Proceedings of Intelligent User Interfaces (IUI), pages 125–132, 2003

[22] Tetsuya Nasukawa and Jeonghee Yi. *Sentiment analysis: Capturing favorability using natural language processing*. In Proceedings of the Conference on Knowledge Capture (K-CAP), 2003.

[23] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. *Thumbs up? Sentiment classification using machine learning techniques*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[24] Kenji Tateishi, Yoshihide Ishiguro, and Toshikazu Fukushima. *Opinion information retrieval from the Internet*. Information Processing Society of Japan (IPSJ) SIG Notes, 2001(69(20010716)):75–82, 2001. Also cited as "A reputation search engine that gathers people's opinions from the Internet", IPSJ Technical Report NL-14411. In Japanese.

[25] Richard M. Tong. *An operational system for detecting and tracking opinions in on-line discussion*. In Proceedings of the Workshop on Operational Text Classification (OTC), 2001

I fattori principali di questa improvviso picco di interesse sono legati a:

- l'aumento delle metodologie e gli strumenti di machine learning methods per il processing del linguaggio naturale e per il recupero delle informazioni;
- la disponibilità di un numero consistente di datasets da utilizzare per l'addestramento degli algoritmi di apprendimento automatico grazie allo sviluppo e alla diffusione del World Wide Web e, in particolare, allo sviluppo di aggregatori e strumenti di sviluppo per i siti web

e, naturalmente

- il fascino della sfide intellettuali proposte dall'argomento e il numero elevato di applicazioni commerciali e di intelligence che l'ambito suggerisce.

1.3 Una panoramica sulle metodologie

Un numero crescente di studiosi stanno affrontando la tematica della sentiment analysis per la rilevazione e classificazione automatica dei mood all'interno di un dominio chiamato "affective computing". Da un lato, l'identificazione e la descrizione delle opinioni come positive, negative e neutrali sono affrontate nell'ottica dell'opinion mining[29]; dall'altro, lo studio sui moods è condotto nello scenario più ampio della sentiment analysis. Entrambe le metodologie ricadono nello scenario globale dell'affective computing in cui I concetti di intelligenza "psicologica" e "computazionale" sono utilizzati, in

[26] Peter Turney. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. In Proceedings of the Association for Computational Linguistics (ACL), pages 417–424, 2002.
[27] Janyce Wiebe, Eric Breck, Christopher Buckley, Claire Cardie, Paul Davis, Bruce Fraser, Diane Litman, David Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. *Recognizing and organizing opinions expressed in the world press*. In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 2003.
[28] Hong Yu and Vasileios Hatzivassiloglou. *Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.
[29] B. Pang and L. Lee, *Opinion mining and sentiment analysis*, Found. Trends Inf. Retr., vol. 2, no. 1-2, pp. 1–135, 2008.

maniera combinata allo scopo di identificare il contenuto affettivo, classificare i sentimenti o simulare agenti affettivi. Applicazioni pratiche in questa nuova area includono, ad esempio, le emozioni nel modo della robotica, classificatori automatici di filmati, interfacce intelligenti per computers o avatars, progettazione di video-games di ultima generazione e indagini di marketing automatiche.

Per affrontare il problema della sentiment analysis bisogna innanzitutto definire cosa siano i sentimenti e questo aspetto può essere affrontato principalmente a partire dai modelli psicologici.

Partendo dai modelli che valutano gli eventi che determinano le regole di mutazione dei sentimenti possono essere utilizzati per modellare il meccanismo affettivo umano.

Nonostante le tecniche di indagine basate sull'osservazione degli eventi siano particolarmente adatte per lo sviluppo degli agenti affettivi artificiali, non costituiscono una base affidabile nell'attività di distinzione dei sentimenti umani e, pertanto, non sono presenti in questa analisi dello stato dell'arte. Tuttavia, va sottolineato che parecchi studi hanno utilizzato queste tecniche utilizzando inferenza e aggregazione fuzzy per lo studio dei meccanismi affettivi che contemplino ambiguità e imprecisione. Ad essi potranno far riferimento i lettori eventualmente interessati.

Oltre al modello della valutazione degli eventi esistono altre due famiglie di modelli psicologici, rispettivamente il modello categorico basato su un set predefinito di stati affettivi e il modello dimensionale che descrive i sentimenti come vettori in uno spazio continuo multidimensionale. Entrambi consentono di descrivere il processo di discriminazione dei sentimenti; essi sono largamente impiegati nel campo dell'addestramento dei computer utilizzati nell'ambito della sentiment analysis.

Ad esempio essi sono adottati in maniera particolare, per l'analisi dei sentimenti per i testi[30], dell'audio[31], dei frames video[32] o delle misure psicologiche[33].

[30] R. Cowie, E. Douglas-Cowie, and A. Romano, *Changing emotional tone in dialogue and its prosodic correlates*, in Proc. of ETRW on Dialogue and Prosody, 1999.

Se da un lato i benefici dell'applicazione di tali metodologie ai casi della vita reale sono certi, molti problemi rimangono irrisolti ivi comprese le metodologie per classificare, identificare e distinguere I sentimenti: ad esempio le espressioni ambigue per il feeling, l'influenza del contest per la loro interpretazione, ma, anche la natura molto personale dei sentimenti che, ad esempio, dipendono dalla cultura, dai linguaggi, dall'età e dalle esperienze.

Partendo dal presupposto che i modelli e i processi adottabili nell'analisi dei sentimenti sono solitamente caratterizzati da ambiguità e imprecisione, l'approccio preferito è quello di considerare delle componenti di gradualità nei processi di analisi automatica, classificazione e identificazione dei sentimenti stessi.

Nella nostra analisi il concetto di gradualità viene affrontato in rappresentazioni che tengano conto delle componenti intrinseche dei modelli psicologici. Al di là di modelli tridimensionali che offrono un quadro naturale graduale si distingue tra tre componenti di gradualità: composizione o mescolamento, intensità ed ereditarietà.

Prendiamo anche in considerazione le componenti estrinseche di gradualità emessi da approcci di intelligenza computazionale. Sebbene la teoria degli insiemi fuzzy offra un'ampia gamma di possibilità di modellare le ambiguità e le imprecisioni degli stati affettivi, sarà possibile analizzare anche i metodi che fanno uso di altre strutture di gradualità basati su un modello di rappresentazione vettoriale.

1.4 La categorizzazione dei modelli affettivi

Un primo approccio per la rappresentazione dei sentimenti è basato sulla visione Darwiniana: essi sono il frutto di un meccanismo di sopravvivenza e costituiscono condizioni necessarie per la

[31] Y.-H. Yang, C.-C. Liu, and H. H. Chen, *Music emotion classification: a fuzzy approach*, in Proc. of the 14th annual ACM international conference on Multimedia, 2006.
[32] A. Hanjalic and L.-Q. Xu, *Affective video content representation and modeling*, IEEE Transactions on Multimedia, vol. 7, no. 1, pp. 143–154, 2005.
[33] R. W. Picard, E. Vyzas, and J. Healey, *Toward machine emotional intelligence: Analysis of affective physiological state*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175–1191, 2001.

conservazione della specie. In accordo a tale visione il sentimento della paura è interpretato come la risposta immediata e istintiva ad un pericolo imminente.

La visione Darwiniana impone che I sentimenti siano organizzati secondo un set di stati affettivi, universalmente comuni all'essere umano.

A seconda degli autori questo set è indicato come "basilare", "primario" o "sentimenti in piena regola" (full blown sentiments). I modelli psicologici che condividono questo punto di vista sono conosciuti come "modelli categorici"[34,35].

A seconda degli autori essi possono variare sia per il numero e la natura degli stati affettivi sia per le strutture nelle quali essi sono organizzati. Ciò spiega perché le componenti di gradualità siano esplicitate in differenti modelli. Sebbene non vi sia un consenso reale sul numero degli stati affettivi di base, gli approcci di intelligenza computazionale fanno spesso riferimento al cosiddetto set dei sei sentimenti di base: paura, rabbia, felicità, tristezza, sorpresa e disgusto. Vi sono differenti metodologie per esplicitare la gradualità intrinseca o estrinseca nell'ambito dei modelli categorici.

All'origine della visione darwiniana, gli psicologi hanno considerato gli stati affettivi come mix tra sentimenti di base. Da questo punto di vista uno specifico insieme di regole di composizione porta ad un insieme finito di maggiore che consente di ordinare gli stati affettivi. Da un altro punto di vista, rilasciando il vincolo sulla natura e sul numero di combinazioni dei sentimenti si può pervenire a stati affettivi più flessibili, caratterizzati comunque dai loro stati costitutivi di base.

1.4.1 I sentimenti complessi

Gli psicologi sostengono che la moderna società abbia aperto la strada ad una nuova forma di sentimenti di natura più complessa[36]. Viene fatta una distinzione tra i sentimenti veri e propri o primari e

[34] R. Plutchik, *The emotions*. University Press of America, 1990.
[35] P. Ekman, *An argument for basic emotions*, Cognition & Emotion, vol. 6(3-4), pp. 169–200, 1992.
[36] P. N. Johnson-Laird and K. Oatley, T*he language of emotions: An analysis of a semantic field*, Cognition and Emotion, vol. 3, pp. 81–123, 1989.

complessi o sentimenti di secondo ordine. Mentre la prima tipologia emerge da riflessi antichi all'istinto di sopravvivenza degli esseri umani, gli altri sono il risultato della la complessità delle nostre interazioni e dalla civiltà moderna.

I sentimenti del secondo ordine possono essere visti come direttamente derivati dal nostro meccanismo primario affettivo in risposta alla trasformazione del nostro ambiente e le stesse esigenze di sopravvivenza.

I sentimenti di secondo ordine sono quindi espressi come specifica combinazione di sentimenti primari (non tutte le combinazioni sono consentite[37,38]. A seconda degli autori i sentimenti del secondo ordine variano per numero e tipologia. Per Plutchik, i sentimenti del secondo ordine sono definiti come combinazioni di coppie di sentimenti primari. Ad esempio, il disprezzo viene considerata come sentimento del secondo ordine ottenuto come combinazione della rabbia e del disgusto. In termini di intelligenza computazionale, in particolare nei tasks volti ad identificare i sentimenti, la teoria della logica fuzzy è ben adatta per la modellazione di sentimenti complessi: gli stati affettivi sono gradi di appartenenza assegnati ai sentimenti base considerati come insiemi fuzzy.

Un classificatore sfocato applicato ad un set di sentimenti del primo ordine è quindi in grado di perfezionarne l'interpretazione in termini di sentimenti secondo ordine[39]. Lo stesso approccio viene adottato ad esempio nella riproduzione dell'espressione facciale dove si parte proprio dalla combinazione di sentimenti di base e sentimenti complessi[40].

---

[37] R. Plutchik, *The emotions*. University Press of America, 1990.

[38] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Prentice-Hall, 1975.

[39] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann, *Real-time facial expression recognition using a fuzzy emotion model*, in Proc of Fuzzy Systems Conference, FUZZ-IEEE., 2007

[40] J. C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, *Multimodal complex emotions: gesture expressivity and blended facial expressions*, International Journal of Humanoid Robotics (IJHR), vol. 3, pp. 269–291, 2006.

Basandosi sugli studi di Ekman sulle espressioni facciali complesse[41], un set di regole fuzzy che specificano la composizione di sentimenti complessi è definito e sfruttato nella rappresentazione di un avatar 3D.

Facendo uso sia della teoria della logica fuzzy sia dei sentimenti di ordine superiore, si riesce a fornire espressioni realistiche dei sentimenti.

1.4.2 Vettorizzazione dei sentimenti:

I sentimenti possono essere considerati sia uno alla volta (ma ciò accade raramente nell'esperienza) sia contemporaneamente[42,43]: a seconda delle circostanze, gli individui mascherano un sentimento esprimendone un altro, mostrando più sentimenti nello stesso momento o sperimentando un rapido cambiamento dello stato affettivo. Alcuni modelli non impongono vincoli sul numero e sulla natura di sentimenti combinati. Secondo una visione più pragmatica, gli stati effettivi possono essere descritti da ogni genere di combinazione possibile di sentimenti. Questi modelli si adattano bene alla vita reale e alle situazioni in cui i differenti sentimenti possono essere vissuti ed espressi in un breve lasso di tempo. Tali strutture, rappresentate come vettori di sentimento possono cogliere bene le diverse componenti utili a descrivere gli stati affettivi.

Sono impiegate diverse semantiche: vettori affettivi binari che definiscono gli stati affettivi combinando sentimenti diversi gradi di certezza, vettori che raffigurano stati affettivi incerti e gradi di appartenenza, vettori che fanno uso della teoria della logica fuzzy per modellare l'ambiguità e la vaghezza di espressione dei sentimenti. Mentre tutti permettono che gli stati affettivi siano composti da più sentimenti di base, i vettori binari affettivi non forniscono ulteriori informazioni sul rapporto tra gli stati affettivi e i loro sentimenti costitutivi.

---

[41] P. Ekman, *An argument for basic emotions*, Cognition & Emotion, vol. 6(3-4), pp. 169–200, 1992

[42] J. C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, *Multimodal complex emotions: gesture expressivity and blended facial expressions*, International Journal of Humanoid Robotics (IJHR), vol. 3, pp. 269–291, 2006.

[43] M. K. Petersen and A. Butkus, *Modeling emotional context from latent semantics*, in UXTV '08: Proceeding of the 1st international conference on Designing interactive user experiences for TV and video. New York, NY, USA: ACM, 2008, pp. 63–66.

D'altra parte i vettori dei gradi di certezza misurano l'affinità tra gli stati affettivi e i loro sentimenti costituenti di base; i vettori dei gradi di appartenenza descrivono gli stati affettivi come composizioni ambigue. Di seguito vengono descritte le tre differenti rappresentazioni.

a) Vettori binari affettivi

In questa rappresentazione gli stati affettivi sono descritti da vettori binari su un insieme eterogeneo di sentimenti di base. Qualsiasi sentimento può partecipare alla realizzazione di uno stato affettivo, in modo da rendere meno rigidi i vincoli sulla numero e la natura dei sentimenti combinati.

A tale scopo, si adottano degli indicatori di presenza per segnalare la partecipazione di sentimenti di base nella composizione degli stati affettivi. In questa rappresentazione, gli stati affettivi sono completamente caratterizzati da vettori binari su l'insieme di sentimenti di base.

Nel processo atto a studiare l'evoluzione degli stati affettivi lungo l'arco temporale questo approccio è stato impiegato per descrivere slot temporali eterogenei a partire dagli stati affettivi e per produrre conseguentemente ricche descrizioni affettive delle strutture narrative[44].

b) Vettori dei gradi di certezza

Le strutture di Richer strutture valutano la misura secondo la quale uno stato affettivo è descritto dai suoi componenti relativi. I tassi di correlazione tra gli stati affettivi e i sentimenti fondamentali vengono poi analizzati in modo da produrre i gradi di certezza. Diverse sono le metodologie proposte. Tra gli altri, il punto di vista di reciproca informazione tra gli stati affettivi e sentimenti di base porta ad una caratterizzazione che tenga conto dell'incertezza

[44] A. Salway and M. Graham, *Extracting information about emotions in films*, in Proc. of the 11th ACM Int. Conf. on Multimedia. New York, NY, USA: ACM, 2003, pp. 299–302.

nell'espressione di sentimenti[45]. Nell'analisi delle espressioni sentimenti nei testi, l'analisi semantica LSA offre una misura di similarità del coseno tra gli stati affettivi e i sentimenti di base e quindi modella i sentimenti latenti nell'espressione[46].

c) Vettori dei gradi di appartenenza

Concentrandosi sulla ambiguità e la vaghezza di espressione dei sentimenti, altri approcci fanno uso di livelli di appartenenza piuttosto che livelli di certezza. Agli stati affettivi è assegnato un livello di appartenenza per ciascuno dei sentimenti base visti come gruppi Fuzzy[47,48,49]. In particolare, per tenere conto dell'ambiguità delle lingue naturali nell'analisi di sentimenti nei testi, i sentimenti di base sono definiti come insiemi fuzzy. L'estrazione di contenuto affettivo dai testi viene eseguita con la ricerca di un elenco predefinito di termini affettivi mappati come sentimenti fuzzy mediante livelli di appartenenza[50]. Il Controllo Fuzzy stato anche proposto al fine di far fronte alla vaghezza di espressione per sentimenti in un discorso[51]: si considerano regole fuzzy i cui ingressi e le cui uscite sono variabili fuzzy, rispettivamente parametri del parlato e dei sentimenti di base. Come accennato in precedenza, un altro approccio sfrutta la teoria della logica fuzzy per affrontare sia gli stati di

---

[45] G. Grefenstette, Y. Qu, D. A. Evans, and J. G. Shanahan, *Computing Attitude and Affect in Text: Theory and Applications*. Springer Netherlands, 2006, ch. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes, pp. 93–107.

[46] M. K. Petersen and A. Butkus, *Modeling emotional context from latent semantics*, in UXTV '08: Proceeding of the 1st international conference on Designing interactive user experiences for TV and video. New York, NY, USA: ACM, 2008, pp. 63–66.

[47] J. C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, *Multimodal complex emotions: gesture expressivity and blended facial expressions*, International Journal of Humanoid Robotics (IJHR), vol. 3, pp. 269–291, 2006.

[48] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.

[49] T. Moriyama and S. Ozawa, *Measurement of human vocal emotion using fuzzy control*, Systems and Computers in Japan, vol. 32, pp. 59–68, 2001.

[50] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.

[51] T. Moriyama and S. Ozawa, *Measurement of human vocal emotion using fuzzy control*, Systems and Computers in Japan, vol. 32, pp. 59–68, 2001.

sentimenti complessi, sia il mascheramento di sentimenti sia i bruschi cambiamenti di sentimenti[52].

d) Componenti di intensità

I modelli psicologici per la rappresentazione dei sentimenti possono anche esprimere la gradualità per mezzo di componenti di intensità. Secondo la visione darwiniana, ogni sentimento di base è specificato all'interno di una gerarchia di intensità gradualmente crescenti. L'insieme dei differenti livelli di intensità può essere rappresentato da ulteriori sentimenti di base: nel modello Plutchik ad esempio, la gioia è specificata mediante la serenità e l'ecstasy[53]. Altri psicologi definiscono anche le scale ordinali, senza far riferimento ad alcuna identificazione di sub-sentimenti[54]. Infatti è risaputo che la naturale ambiguità del linguaggio naturale rende difficile trovare un equilibrio tra le diverse etichette per i sentimenti. Inoltre, le componenti di intensità sono spesso combinate con le componenti di fusione che sono state presentate nelle pagine precedenti. In particolare, Plutchik fa uso sia di fusione che di intensità nel modello che propone.

Da un punto di vista dell'analisi del sentimento, i sistemi basati sull'intensità sfruttano componenti graduali per distinguere gli stati di passione da quelli platonici. In tal modo essi sono in grado di proporre una caratterizzazione molto raffinata dei sentimenti[55],[56]. In particolare, questo approccio è particolarmente comune per analizzare i sentimenti nei testi: un dizionario manuale fa corrispondere alle parole un insieme

---

[52] J. C. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, *Multimodal complex emotions: gesture expressivity and blended facial expressions*, International Journal of Humanoid Robotics (IJHR), vol. 3, pp. 269–291, 2006.
[53] R. Plutchik, *The emotions*. University Press of America, 1990.
[54] K. R. Scherer, *What are emotions? and how can they be measured?* Social Science Information, vol. 44, no. 4, pp. 695–729, December 2005.
[55] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.
[56] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, *Textual affect sensing for sociable and expressive online communication*, in Affective Computing and Intelligent Interaction. Springer Berlin / Heidelberg, 2007, vol. 4738/2007, pp. 218–229.

finito di sentimenti e li associa con una intensità di valore reale da 0 a 1.

Per esempio, euforico, felice e gioioso corrispondono tutti allo stato affettivo gioia, diminuendo rispettivamente l'intensità[57].

Bisogna fare attenzione a distinguere i livelli di intensità e il concetto di prossimità/appartenenza che è già stato presentato.

Infatti, in alcuni approcci[58],[59] si considerano i sentimenti di base come insiemi fuzzy per interpretare i gradi di appartenenza come gradi di intensità, ora, mentre l'intensità rappresenta il livello con il quale si esprime un sentimento, il grado di appartenenza consente di rappresentare la natura ambigua degli stati affettivi. Ad esempio, "molto triste" e "un po'" triste sono entrambi stati affettivi che fanno riferimento al sentimento di base che è la tristezza, e differiscono soltanto in termini di intensità. Altri approcci[60],[61] sottolineano questa differenza e utilizzano i componenti di intensità allo stesso modo delle componenti di fusione.

e) Struttura a "bambola russa"

Mentre la gradualità espressa attraverso componenti di fusione e l'intensità sono aspetti sviluppati principalmente da psicologi, le altre componenti derivano da modelli linguistici per la rappresentazione dei sentimenti. I linguisti, infatti, hanno sviluppato strutture per organizzare i sentimenti basandosi sulle definizioni della semantica del sentimento: i sentimenti di

[57] Ibid
[58] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann, *Real-time facial expression recognition using a fuzzy emotion model*, in Proc of Fuzzy Systems Conference, FUZZ-IEEE., 2007.
[59] M. S. El-Nasr and M. Skubic, *A fuzzy emotional agent for decisionmaking in a mobile robot*, in Proc. of Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., vol. 1, 1998, pp. 135–140.
[60] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, *Textual affect sensing for sociable and expressive online communication*, in Affective Computing and Intelligent Interaction. Springer Berlin / Heidelberg, 2007, vol. 4738/2007, pp. 218–229.
[61] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.

base sono considerati meta-sentimenti sono specificati all'interno di una gerarchia semantica di sub-sentimenti.

Basandosi sullo studio di un corpora testuale, un modello propone una classificazione semantica dei verbi affettivi e nomi soggettivi[62]. Il modello è strutturato come una gerarchia basata su tre componenti: semantica, antinomia e intensità. Le transizioni omogenee tra sentimenti possono quindi essere sfruttate per arricchire l'output di sistemi di analisi dei sentimenti (caratterizzazione dell'intensità) consentendo di trattare l'ambiguità e la vaghezza (caratterizzazione semantica).

Un approccio diverso consiste nel costruire una struttura gerarchica per la rappresentazione dei sentimenti basandosi sulle sole etichette semantiche[63].

Rimandando l'analisi delle definizioni di etichette semantiche dei sentimenti, le etichette vengono prima raggruppate secondo la loro positività o negatività e poi a seconda della loro relazione semantica. In questo modello, i sentimenti sono così ordinati lungo una scala positivo/negativo e caratterizzata lungo livelli di ereditarietà.

1.5 Il modello dimensionale – gli spazi affettivi

Oltre alla visione darwiniana descritta nella sezione precedente, che definisce un insieme finito di sentimenti fondamentali universalmente condivisi da ogni essere umano, esiste un altro punto di vista secondo il quale ogni oggetto di studio (ad esempio parole o stati fisiologici) reca una semantica affettiva il cui valore varia a seconda della cultura, dell' età, e delle esperienze personali.

Come tale, il valore affettivo associato ad un oggetto può essere misurato su scale continue di stati affettivi; e gli stessi sono

[62] Y. Y. Mathieu, *A computational semantic lexicon of french verbs of emotion*, in Computing Attitude and Affect in Text: Theory and Applications. Springer, 2006.
[63] A. Piolat and R. Bannour, *An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing*, Current psychology letters, vol. 25, 2009.

rappresentati come veri e propri vettori considerati in spazi multidimensionali.

Tra le scale più adottate, la valenza rappresenta il piacere procurato da una situazione, l'attivazione misura l'eccitazione fisica causata da una situazione e la potenza ritrae la capacità di un soggetto a rilevare una situazione. Non vi è una vera uniformità per quanto riguarda il numero di scale o la loro natura: alcuni autori sostengono che un modello bidimensionale con valenza e attivazione sia più adeguato[64], mentre per altri la potenza è essenziale, ad esempio per distinguere la paura e la rabbia[65].

In questa sezione presenteremo l'utilizzo della gradualità negli spazi affettivi per analizzare sentimenti. In particolare, mentre gli spazi affettivi possono eventualmente essere segmentati in regioni affettive secondo la prevalenza del corrispondente sentimento, gli approcci che conservano la loro natura continua consentono di esprimere la gradualità attraverso la continuità.

1.5.1 Dagli spazi affettivi agli insiemi affettivi

Gli spazi affettivi sono ampiamente studiati nel campo della psicologia; essi, difatti, sono effettivamente utili per la loro rappresentazione descrittiva ed esauriente dei sentimenti[66].

D'altra parte, gli autori che trattano di intelligenza computazionale preferiscono gli spazi affettivi quando si ritrovano a trattare di input continui.

Tuttavia, al fine di fornire contenuto intelligibile, la maggior parte degli approcci considera gli spazi affettivi affettivi a valle di insieme affettivi segmentandoli in stati affettivi prevalenti. Per fare ciò, i

[64] L. F. Barrett and J. A. Russell, *The structure of current affect: Controversies and emerging consensus*, Current Directions in Psychological Science, vol. 8, pp. 967–984, 1999.
[65] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, *The world of emotions is not two-dimensional*, Psychological science : a journal of the American Psychological Society, vol. 18, pp. 1050–7, 2007.
[66] M. Schr¨oder, *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions*, Lecture notes in computer science, vol. 3068, pp. 209–220, 2004.

sentimenti di base sono mappati sui vettori multidimensionali in base alle rispettive proprietà affettive[67].

Ad esempio, al fine di prevedere gli stati affettivi di individui, un approccio consiste nell'analizzare valori reali di ingresso prodotti da sensori del corpo[68]. A tal fine, si studiano vere e proprie funzioni a valori reali mappando ingressi stimati alle dimensioni di valenza e attivazione. Questi ultimi sono quindi riportati agli insiemi affettivi segmentando lo spazio affettivo bi-dimensionale in regioni prevalentemente affettive.

Contrariamente alla visione darwiniana sui sentimenti, gli insiemi affettivi derivati dalla segmentazione affettiva degli spazi potrebbero essere non più composti da sentimenti di base ortogonali. Infatti, Russel[69] sostiene che le regioni non sono ben delimitate, ma sono piuttosto ambigue e vaghe, caratterizzate da bordi sfocati.

Diviene allora naturale considerare i sentimenti come insiemi fuzzy, stati affettivi definibili su un prodotto cartesiano di insiemi fuzzy. Anche se gli insiemi affettivi prodotti non condividono la visione darwiniana sui sentimenti, condividono le stesso proprietà.

I sentimenti complessi sono espressi considerando le regioni ottenute mediante la segmentazione degli spazi affettivi.

Le composizioni di sentimenti complessi vengono poi definite tra stati emotivi e affettivi e le etichette delle differenti regioni[70],[71]. Quando le regioni affettive sono viste come insiemi fuzzy, agli stati affettivi

---

[67] I. Albrecht, M. Schr¨oder, J. Haber, and H.-P. Seidel, *Mixed feelings: expression of non-basic emotions in a muscle-based talking head*, Virtual Reality, vol. 8, pp. 201–212, 2005.

[68] R. W. Picard, E. Vyzas, and J. Healey, *Toward machine emotional intelligence: Analysis of affective physiological state*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175–1191, 2001.

[69] J. A. Russell, *A circumplex model of affect*, Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178, 1980.

[70] C. H. Chan and G. J. Jones, *Affect-based indexing and retrieval of films*, in MULTIMEDIA '05: Proc. of the 13th ACM Int. Conf. on Multimedia. New York, NY, USA: ACM, 2005, pp. 427–430.

[71] R. Cowie, E. Douglas-Cowie, E. Apolloni, J. Taylor., A. Romano, and W. Fellenz, *What a neural net needs to know about emotion words*, in CSCC'99: Proc. of the 3rd IEEE Int. MulltiConf. on Circuits, Systems, communications and computers, 1999, pp. 5311–5316.

vengono poi assegnati gradi di appartenenza alle regioni fuzzy in modo da produrre sentimenti complessi.

In particolare, sulla base di una rappresentazione bidimensionale di sentimenti definiti da valenza e attivazione, è stato proposto classificatore fuzzy pe un'emozione suscitata dalla musica[72]: esso consente di mappare dei segmenti di canzoni nei quattro quadranti dello spazio affettivo visti come insiemi fuzzy. Lo stesso approccio è stato sfruttato per l'analisi di stati affettivi di giocatori durante l'utilizzo dei video games[73],[74].

L'intensità è espressa da Russel come funzione di due dimensioni nel modello bi-dimensionale composto da valenza e attivazione. L'intensità è definita come un vettore di distanza affettiva dall'origine del piano affettivo.

Gli approcci della sentiment analysis automatica basati sulla rappresentazione degli spazi affettivi permettono di definire l'intensità come una funzione delle dimensioni definite nello spazio affettivo[75,76,77].

---

[72] Y.-H. Yang, C.-C. Liu, and H. H. Chen, *Music emotion classification: a fuzzy approach*, in Proc. of the 14th annual ACM international conference on Multimedia, 2006.

[73] R. L. Mandryk and M. S. Atkins, *A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies*, Int. Journal of Human-Computer Studies, vol. 65, pp. 329–347, 2007.

[74] J. O. Orero, F. Levillain, M. Damez-Fontaine, M. Rifqi, and B. Bouchon-Meunier, *Assessing gameplay emotions from physiological signals : a fuzzy decision trees based model*, in Proc. of the International Conference on Kansei engineering and Emotion Research, 2010, in Press.

[75] R. Cowie, E. Douglas-Cowie, E. Apolloni, J. Taylor., A. Romano, and W. Fellenz, *What a neural net needs to know about emotion words*, in CSCC'99: Proc. of the 3rd IEEE Int. MulltiConf. on Circuits, Systems, communications and computers, 1999, pp. 5311–5316.

[76] J. A. Sanchez, N. P. Hernandez, J. C. Penagos, and Y. Ostrovskaya, *Conveying mood and emotion in instant messaging by using a twodimensional model for affective states*, in IHC '06: Proc. of VII Brazilian Symposium on Human Factors in Computing Systems. New York, NY, USA: ACM, 2006, pp. 66–72.

[77] S. Fitrianie and L. J. Rothkrantz, *An automated online crisis dispatcher*, Int. Journal of Emergency Management, vol. 5, pp. 123–144, 2008.

1.5.2 Continuità come elemento rappresentativo della gradualità

Altri approcci sfruttano gli spazi affettivi nella loro stensione completa: gli stati affettivi sono poi descritti dalla struttura numerica implicita rispetto alle scale di valore affettivo. Questi approcci adottano una rappresentazione dei sentimenti intrinsecamente graduale, naturalmente in grado di gestire l'ambiguo e vago nella natura espressiva dei sentimenti. Tuttavia, l'interpretazione degli stati affettivi in termini di scale affettive risulta più difficile rispetto agli insiemi affettivi. Le etichette affettive sono infatti più esplicite rispetto alle reali misurazioni dei valori affettivi. Un approccio esprime uno spazio bi-dimensionale composta di valenza affettiva e attivazione al fine di analizzare gli stati affettivi di un film[78].

I valori di valenza e attivazione di fotogrammi consecutivi sono memorizzati in una "curva affettiva" che aggrega i diversi stati affettivi del film. Essi vengono interpretati come firme affettive che caratterizzano il contenuto affettivo del film. Inoltre, segmentando lo spazio affettivo in sentimenti prevalenti, le curve affettive descrivono transizioni omogenee tra sentimenti.

Considerando ogni scala affettiva lungo una linea del tempo, le strutture numeriche offrono una rappresentazione naturalmente aggiornabile per studiare i cambiamenti degli stati affettivi nel tempo. Questo approccio è stato adottato per lo studio nei film del contenuto affettivo sulla base di input visivi continui (in parallelo alle "curve affettive" descritte in precedenza), del contenuto affettivo dei film sulla base di input testuali discreti[79], del contenuto affettivo di finestre

[78] A. Hanjalic and L.-Q. Xu, *Affective video content representation and modeling*, IEEE Transactions on Multimedia, vol. 7, no. 1, pp. 143–154, 2005.
[79] F. Dzogang, M.-J. Lesot, M. Rifqi, and B. Bouchon-Meunier, *Analysis of texts' emotional content in a multidimensional space*, in Proceedings of the International Conference on Kansei Engineering and Emotion Research, 2010.

di dialogo in applicazioni text-to-speech [80] o del rilevamento del contenuto affettivo nelle finestre di dialogo audio[81].

1.6 La gradualità utilizzata per lo sfruttamento della ricchezza del linguaggio naturale

Ci sono molti modi diversi per gli esseri umani di esprimere i sentimenti: le principali modalità studiate sono le espressioni del corpo o del viso, le espressioni fisiologiche, le espressioni orali e scritte.

A seconda dello specifico ambito di applicazione della sentiment analysis (video, testi, espressioni audio etc.) molte solo le metodologie sviluppate e gli approcci consolidati. Ad esempio molta ricerca si è prodotta sull'analisi della strutturazione video, il rilevamento degli eventi e la modellazione semantica.

Più di recente, i ricercatori hanno rivelato il significato dell'analisi affettiva da un personale punto di vista mediatico[82,83,84,85]. Ad esempio, molti di questi strumenti favoriscono gli utenti fornendo loro uno strumento più flessibile per selezionare velocemente il capitolo più divertente o più sentimentale di un film, così come le parti più emozionanti di un videogioco sportivo.

Rispetto all'indicizzazione tradizionale di un video, l'analisi del contenuto affettivo pone molta più enfasi sulle reazioni del pubblico e sulle emozioni. Come analisi semantica, l'analisi del contenuto

[80] M. Schr¨oder, *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions*, Lecture notes in computer science, vol. 3068, pp. 209–220, 2004.
[81] R. Cowie, E. Douglas-Cowie, and A. Romano, *Changing emotional tone in dialogue and its prosodic correlates*, in Proc. of ETRW on Dialogue and Prosody, 1999.
[82] A. Hanjalic and L.-Q. Xu, *User-oriented Affective Video Content Analysis*, In Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries' 01.
[83] H.-B. Kang, *Affective Content Detection using HMMs*, In Proc. of ACM Multimedia' 03.
[84] H.-B. Kang, *Emotional Event Detection Using Relevance Feedback*, In Proc. of International Conference on Image Processing' 03.
[85] S. Moncrieff, C. Dorai and S. Venkatesh, *Affect Computing in Film through Sound Energy Dynamics*, In Proc. of ACM Multimedia' 01.

affettivo è anche difficile a causa del divario tra il basso livello delle caratteristiche percettive e l'alto livello di comprensione.

Tuttavia, il contenuto affettivo sottolinea i fattori che influenzano l'attenzione degli utenti, la loro valutazione, e la memoria circa il contenuto di un video. Non richiede una profonda conoscenza del video né dei suoi contenuti audio. E' possibile impiegare alcune tecniche consolidate di indicizzazione per acquisire il contenuto affettivo di un video incorporando la conoscenza dello specifico dominio.

In un certo senso, possiamo considerare l'analisi affettiva come uno strato intermedio tra il livello inferiore relativo alla strutturazione del video e quello più elevato di modellazione semantica.

Ad esempio alcuni scenari di ricerca in questo ambito considerano le caratteristiche di movimento, colore e audio per rappresentare l'eccitazione e la dinamicità[86] di un dato video.

Analoghi approcci e metodologie sono riconoscibili anche nell'analisi del sentiment all'interno di un generico audio o, meglio ancora, nell'analisi del parlato.

Sono comunque molteplici gli ambiti applicativi possibili per un'analisi del sentiment per un generico multimedia, a seconda dello specifico oggetto che si vada ad analizzare.

In ambito medico, ad esempio esistono numerose applicazioni che utilizzano il riconoscimento delle espressioni piuttosto che l'analisi dei video dei pazienti per la classificazione delle sintomatologie per ciascuna patologia. Sempre in ambito medico alcuni dataset estremamente ricchi di immagini statiche o in movimento consentono di evidenziare forme ancora embrionali di disturbi del movimento associati a forme distrofiche o similari.

Ancora l'analisi di video realizzati riprendendo gli utenti durante alcune dimostrazioni di prodotti commerciali consentono di

[86] A. Hanjalic and L.-Q. Xu, *User-oriented Affective Video Content Analysis*, In Proc. of IEEE Workshop on Content-Based Access of Image and Video Libraries' 01

evidenziare la percezione degli utenti o comunque il loro approccio in relazione all'oggetto proposto o alla modalità di fruizione presentata.

Esiste, ancora, una ricca documentazione relativa allo studio (e alla relativa memorizzazione) di alcuni comportamenti umani (e anche animali) che consentono, successivamente, di adeguare le caratteristiche di un prodotto o servizio che si intende fornire.

Tutti gli elementi che possono essere analizzati possono presentare svariate combinazioni delle differenti tipologie di multimedia, a partire dalla immagini (fisse o in movimento) con o senza audio associato, stringhe di testo, direttamente riconoscibili dal video, piuttosto che ricavati dall'audio presente.

Questi media possono successivamente essere analizzati uno alla volta, a gruppi, oppure tutti insieme.

Le complessità dell'analisi è molto elevata quanto più ricco di informazioni è lo specifico oggetto multimediale e, nel caso del video (comprensivo di audio e magari anche testo o immagini statiche), le variabili in gioco sono molteplici (si debbono valutare le espressioni facciali, gli ambienti, i contesti, le implicazioni tonali, le singole frasi). Così come anche un singolo media, se analizzato in dettaglio, può contemplare complessità notevoli. Si pensi ad esempio un audio ascoltato senza poter visualizzare la mimica facciale e quindi la difficoltà di riconoscere un discorso sarcastico da uno serio.

Problematiche complesse vengono identificate, inoltre, per la convergenza delle tecniche in gioco. Da un lato, come evidenziato per i video, si parte dalla stessa organizzazione e dalla sequenzialità delle immagini. Dall'altro, per l'audio vengono considerati le intensità del suono, le tonalità. Questi elementi possono essere fusi ad esempio con il riconoscimento dell'espressività e quindi con l'analisi espressiva a partire dal viso.

Questo ultimo caso può, ad esempio, essere adottato tanto per la valutazione emozionale della percezione di un interfaccia (ad esempio web) quanto piuttosto per l'efficace comprensione all'ascolto (sguardo attonito, rilassato, interrogativo etc.).

Un tipo di indagine combinata può essere condotta solo miscelando opportunamente le tecniche utilizzate per ciascuno degli ambiti

(audio, video, testo etc). e, dal punto di vista dell'analisi multimodale (tale è il termine con il quale si identifica questo approccio) i risultati più significativi risultano a volte sorprendenti.

Ad esempio si evidenzia[87] come, della miriade di informazioni presenti in un video complesso (che combini appunto, immagini in movimento, immagini statiche, audio e testo), poche siano necessarie a caratterizzare l'orientamento emozionale di un soggetto inquadrato.

Più nel dettaglio è possibile assimilare la mimica facciale all'intensità dell'audio e ancora poi lo stesso audio al testo riconosciuto.

Sebbene l'analisi della gradualità del sentimento risenta molto delle sfumature tonali piuttosto che della mimica facciale è possibile considerare un modello molto semplificato di sentiment dove considerare solo un orientamento positivo e uno negativo.

In questo caso è possibile ridurre l'ambito d'indagine al solo parlato (e ancora di più al testo ottenuto da un riconoscimento del parlato).

Questo è il motivo per cui in letteratura grandissimo spazio trova principalmente l'analisi lessicale del testo utilizzata proprio per classificare l'orientamento (positivo o negativo) dell'autore di un testo (o di un messaggio audio).

Altro elemento da prendere in considerazione per giustificare la netta predominanza in letteratura di lavori relativi alla sola analisi del sentiment a partire dai testi rispetto a quelli per immagini e audio è dato dalla grande disponibilità di dataset testuali, molti dei quali già opportunamente classificati per l'addestramento dei sistemi automatici.

Altrimenti non si può dire per dataset video (a meno di qualche esempio relativo ai video su youtube[88]) e audio.

Sono questi i motivi per i quali, seguendo l'orientamento della letteratura consolidata sull'argomento, nelle pagine a seguire

[87] Louis-Philippe Morency, Rada Mihalcea and Payal Doshi, Towards *Multimodal Sentiment Analysis: Harvesting Opinions from The Web*, In Proceedings of the International Conference on Multimodal Interfaces (ICMI 2011), Alicante, Spain, 2011

[88] Ibidem 87

andremmo ad approfondire le tematiche e le tecniche di analisi della sentiment analysis nel solo ambito del testo e del linguaggio naturale. In particolare, così come indicato nei paragrafi seguenti, riprenderemo il concetto di granularità[89] introdotto per l'analisi.

Il linguaggio naturale trasmette molta incertezza e ambiguità che determina la pluralità dei metodi di analisi automatica semantica dei testi. In particolare, la definizione stessa delle parole possono cambiare dai contesti, ad esempio, i termini *mostruoso* e *colpo di testa* sono visti un modo diverso in un contesto di videogames piuttosto che in un contesto di notizie internazionali. Come pure alcuni termini possono riferirsi a concetti vagamente definiti: ad esempio il termine spregevole, che si riferisce al sentimento complesso di "disprezzo" nel modello "Plutchik"[90], esprime rabbia e disgusto. Inoltre, a seconda del contesto, gli elementi di "negazione" possono essere intesi in sensi ulteriori rispetto all'opposizione: per esempio possono essere utilizzati per esprimere ironia o cortesia. Inoltre, alcuni modificatori linguistici che consentono di accentuare o diminuire l'intensità dei termini possono modificare la semantica di complesse costruzioni linguistiche. Nelle espressioni di sentimenti nei testi per ottenere sentimenti complessi, ambigui e sottili come risultato, la gradualità risulta essere una caratteristica importante.

---

[89] Con granularità dell'informazione, che fa riferimento alle dimensioni dell'informazione, si indica il livello di dettaglio delle informazioni stesse. Si parla di granularità "grossolana" o "generica" (tipica di chi occupa i livelli più alti di un'organizzazione) quando si è di fronte ad informazioni descritte in maniera generica, mentre si parla di granularità "fine" o "specifica" (tipica invece di chi occupa i livelli più bassi) quando si hanno informazioni più dettagliate. c.f.r. aa.vv. *ITC e sistemi informativi aziendali*, Milano, McGraw-Hill, 2007

[90] Il punto di partenza preso da Plutchik è di natura evolutiva. Infatti la tesi su cui si fondano le sue ricerche è che le emozioni sono risposte evolutive per consentire alle specie animali di sopravvivere (Plutchik, 1980). Argomenta infatti che ognuna delle emozioni primarie agisce come interruttore per un comportamento con un alto valore di sopravvivenza (es. paura: fight-or-flight response). Secondo Robert Plutchik, vi sono 8 emozioni primarie, definite a coppie: gioia – tristezza, fiducia – disgusto, rabbia – paura, sorpresa – anticipazione. La ruota delle emozioni da lui creata evidenzia gli opposti e l'intensità delle emozioni, via via decrescente verso l'esterno, più i vari stati intermedi (decrescendo di intensità le emozioni si mescolano sempre più facilmente). Plutchik, citando le ricerche di Darwin che hanno ricevuto numerose conferme, sottolinea il ruolo comunicativo delle emozioni. Inoltre ad ogni emozione viene associato uno stimolo esterno ed una risposta dell'animale.

In generale, la sentiment analysis nei testi è divisa in un processo composto da molteplici passi[91,92]. In primo luogo, i termini affettivi sono isolati e raccolti nei dizionari affettivi; sono stati proposti molti approcci per raccogliere automaticamente i termini affettivi e la loro semantica a partire da risorse pubbliche[93,94,95].

A seconda della rappresentazione dei sentimenti, viene definita una mappatura tra i termini e la struttura prevista dal modello. I testi oggetto dell'analisi vengono poi proiettati su il dizionario affettivo e analizzati a partire dai termini affettivi in essi contenuti. Infine, durante il processo di analisi, gli stati affettivi complessivi dei testi sono calcolati sulla base delle loro proiezioni affettive. In questa fase si presenta la possibilità, a partire dalle componenti di gradualità, di offrire risposte ai problemi legati all'analisi automatizzata dei sentimenti nei testi nota come elaborazione del linguaggio naturale. In particolare, vedremo come le componenti di gradualità possono essere integrate nel processo di passaggio multiplo sopra descritto.

A tal fine, sono successivamente prese in considerazione quattro fasi: si parte dall'affrontare il caso in termini inerenti ambiguità e vaghezza, successivamente ci si concentra sull'interpretazione dei modificatori linguistici. Successivamente si affrontano gli aspetti delle negazioni e solo alla fine si affrontano le costruzioni linguistiche complesse.

Una delle particolarità del linguaggio naturale risiede nelle definizioni ambigue dei termini che variano a seconda dei contesti.

---

[91] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.

[92] F. Dzogang, M.-J. Lesot, M. Rifqi, and B. Bouchon-Meunier, *Analysis of texts' emotional content in a multidimensional space*, in Proceedings of the International Conference on Kansei Engineering and Emotion Research, 2010.

[93] A. Esuli and S. Fabrizio, *Sentiwordnet: A publicly available lexical resource for opinion mining*, in LREC'06: 5th Conf. on Language Resources and Evaluation, 2006.

[94] C. Strapparava and A. Valitutti, *Wordnet-affect: an affective extension of wordnet*, in Proc. of the 4th International Conference on Language Resources and Evaluation, 2004.

[95] H. Liu, *Automatic affective feedback in an email browser*, In MIT Media Lab Software Agents Group, Tech. Rep., 2002.

Allo stesso modo, ci si aspetta che i termini espongano effettiva ambiguità semantica. Un approccio per trattare con le ambiguità è quello di codificare la semantica affettiva dei termini e di eseguire una Word Sense Disambiguation (WSD). La risoluzione dell'ambiguità può poi essere rinviata alla fase di elaborazione, cioè quando il contesto affettivo è disponibile. E' opportuno notare che questo approccio è quello del minimo sforzo. Un'altra caratteristica delle lingue naturali è la definizione di termini che riguardano i concetti vagamente definiti. Nel caso dei sentimenti, i termini possono esprimere stati affettivi misti o complessi. Un approccio per l'estrazione di aggettivi portatori di sentimenti considera i gradi di appartenenza ai sentimenti di base che abbiano a che fare con la vaghezza intrinseca dei termini[96].

Un altro approccio è quello consistente nel codificare manualmente la vaghezza da parte di commentatori umani. Per esempio, gli stati affettivi trasmessi dai gradi di appartenenza dei termini ai sentimenti di base sono assegnati manualmente da linguisti[97]. Anche se la soggettività introdotta dai commentatori potrebbe polarizzare i sistemi, la misura delle correlazioni dei tassi di accordo dei differenti commentatori consente di valutare l'obiettività del processo di annotazione. Inoltre, quando sono impiegati spazi dimensionali affettivi, sistemi di analisi di testo sono in grado di sfruttare l'intrinseca continuità implicita nel modello. Infatti, i termini mappati come vettori multidimensionali vengono analizzati come oggetti nella metrica degli spazi e la funzione distanza può essere sfruttata al fine di valutare la vaghezza semantica affettiva dei termini.

Molti avverbi e aggettivi agiscono come modificatori di intensità e hanno dimostrato di contribuire in larga misura al contenuto affettivo dei testi[98,99]. Per perfezionare i risultati dei sistemi, l'azione dei

[96] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, *Analysis of affect expressed through the evolving language of online communication*, in Proc. of Int. the 12th Conf. on Intelligent User Interfaces, 2007, pp. 278–281.
[97] P. Subasic and A. Huettner, *Affect analysis of text using fuzzy semantic typing*, in Proc. of Fuzzy Systems Conference, FUZZ-IEEE, vol. 2, 2000, pp. 647–652 vol.2.
[98] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian, *Sentiment analysis: Adjectives and adverbs are better than 1adjectives alone*, in Proc. of Int. Conf. on Weblogs and Social Media, 2007.

modificatori linguistici sulle espressioni dei sentimenti può essere modellata mediante componenti di intensità. In questo approccio, i modificatori linguistici sono quindi rappresentati come gli operatori che consentono di aumentare o diminuire l'intensità dei testi affettivi[100,101,102].

Un requisito fondamentale per l'analisi dei sentimenti nei testi è l'elaborazione delle negazioni. Poiché l'espressione di un sentimento è molto diversa dalla espressione del suo opposto, analizzando i sentimenti senza prendere in considerazione le negazioni porta a drastiche perdite di pertinenza dei risultati.
Per fare fronte alle negazioni, un approccio semplice consiste nel definire il contrario di un'affermazione affettiva[103,104]. Ad esempio, nel modello Plutchik ciascun sentimento ne possiede uno opposto: la gioia è considerato come un opposto della rabbia[105]. Negli spazi affettivi, gli stati opposti sono espressi come vettori di coordinate opposte lungo le scale continue.
Tuttavia, alcuni gruppi di sentimenti di base non presentano relazioni contrarie: per esempio, considerando l'insieme dei sei sentimenti di base, è difficile assegnare un opposto immediato per la tristezza. Facendo uso di combinazioni di componenti, dell'ambiguità e/o della vaghezza semantica affettiva per riportare la negazione in modo da memorizzare la sensazione e sfruttarla secondo il contesto circostante. Inoltre sembra che le negazioni non solo esprimano le opposizioni, ma condividano i sentimenti multipli. Infatti, esse possono essere

---

[99] V. Hatzivassiloglou and J. M. Wiebe, *Effects of adjective orientation and gradability on sentence subjectivity*, in Proc. of the 18th conference on Computational linguistics, vol. 1, 2000, pp. 299–305.
[100] L. Polanyi and A. Zaenen, *Contextual valence shifters*, in Computing Attitude and Affect in Text. Springer, 2005.
[101] A. C. Boucouvalas, *Real time text-to-emotion engine for expressive internet communications*, in Being There: Concepts, effects and measurement of user presence in synthetic environments. Ios Press,2003.
[102] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, *Analysis of affect expressed through the evolving language of online communication*, in Proc. of Int. the 12th Conf. on Intelligent User Interfaces, 2007, pp. 278–281.
[103] Y. Y. Mathieu, *A computational semantic lexicon of french verbs of emotion*, in Computing Attitude and Affect in Text: Theory and Applications. Springer, 2006.
[104] A. Piolat and R. Bannour, *An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing*, Current psychology letters, vol. 25, 2009.
[105] R. Plutchik, *The emotions*. University Press of America, 1990.

impiegate al fine di esprimere tra gli altri l'ironia, la cortesia, l'ovvietà come pure l'opposizione. Mentre alcuni approcci scelgono di ignorare la partecipazione dei termini di negazione nel trattamento degli stati affettivi di un testo[106], è possibile utilizzare le componenti di intensità per interpretare le loro negazioni semantiche: esse, viste come modificatori linguistici influenzano l'intensità degli stati affettivi.

La ricchezza delle lingue naturali consente costruzioni di linguistiche complesse: modelli come "eccitato" e "ansioso" o "né felice né triste", rappresentano una sfida per l'analisi automatica dei sentimenti da parte dei sistemi automatici. Tuttavia, considerare la combinazione delle componenti nella rappresentazione degli stati affettivi è particolarmente utile per la modellazione tra stati affettivi. Ad esempio, nel quadro della logica fuzzy, i sentimenti di base visti come insiemi fuzzy possono descrivere stati affettivi complessi a partire dai gradi di appartenenza.
In pratica per procedere alla classificazione dei sentimenti presenti, ad esempio, in un testo assegnato, si costruiscono del vettori nei quali ciascun sentimento è indicato come rapportato al grado di appartenenza a un dato insieme (dove 0 indica la non appartenenza all'insieme e 1 definisce l'appartenenza completa all'insieme).

In particolare, considerando la rappresentazione dei sentimenti di Plutchik, la seguente espressione complessa
"sia arrabbiato sia triste"
può essere definita come:

$$\left(\frac{Rabbia}{1{,}0}, \frac{Tristezza}{1{,}0}\right)$$

dove quindi sia rabbia che tristezza sono indicati con massimo grado di appartenenza (1) all'insieme.

Nel quadro delle emozioni complesse, va evidenziato che un'espressione complessa che contiene per esempio, sia un termine che si riferisce interamente alla rabbia, ma contaminato dal disgusto, e

[106] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, *Analysis of affect expressed through the evolving language of online communication*, in Proc. of Int. the 12th Conf. on Intelligent User Interfaces, 2007, pp. 278–281.

un termine che si riferisce interamente alla tristezza può, per esempio essere definita come:

$$\left(\frac{Rabbia}{1{,}0}, \frac{Disgusto}{0{,}3}, \frac{Tristezza}{1{,}0}\right)$$

che differisce da un'espressione complessa contenente tre termini che si riferisca, rispettivamente a rabbia, disgusto e alla tristezza:

$$\left(\frac{Rabbia}{1{,}0}, \frac{Disgusto}{1{,}0}, \frac{Tristezza}{1{,}0}\right)$$

E' opportuno notare come gli indici dei gradi di appartenenza vengono associati con differenti tecniche, non ultima quella di un'analisi soggettiva effettuata da utenti definiti esperti.

1.7 Uno sguardo d'insieme sulle nuove metodologie di ricerca per analizzare emozioni e passioni.

Nell'ambito dell'Opinion Mining l'elemento emotivo diviene cardine di riferimento da combinare con altri fattori: solo a titolo di esempio citiamo il grafico a quattro quadranti contrapposti (Piacere - Dispiacere - Amore - Odio) messo a punto dalla statunitense NetBase per la **"Brand Passion Index".** Ogni brand analizzato viene identificato da un cerchio ampio in base alla quantità di commenti espressi, e posizionato nei quadranti sulla base del rapporto tra intensità emotiva (Passion Intensity) e orientamento (Sentiment Range).

Nuove metodologie prendono il posto di metodi come i focus group e le ricerche di mercato, più costosi e lenti nei feedback. In un mercato dove il tempo assume un vero e propri valore economico, l'immediatezza tra analisi e l'azione di marketing fa la differenza sui sistemi convenzionali. In pratica la Sentiment Analysis eroga in tempo reale le informazioni sugli orientamenti dei target di riferimento.

I tools sono rappresentati dall'incrocio tra analisi statistiche, semantiche e interpretazione del linguaggio scritto, attraverso lo

studio della sintassi, la ripetizione di keywords e la decifrare di modi di dire.

Lo studio del linguaggio naturale è alla base dello sviluppo di piattaforme dedicate, incrociato con sistemi di web crawling e ricerca media. **Le tre fasi** di analisi vanno dall'individuazione delle entità target, individuazione dei relativi documenti sui social network (attraverso, ad esempio, le pagine/azienda di facebook, già predisposte a tal fine ed associate a database anagrafici e territoriali) statistiche di classificazione.

Questa mole di dati estremamente eterogenea (nel caso di analisi incrociata si va da documenti di tipo testuale a database semistrutturati quali XML) viene elaborata incrociando tecnologie di Data Mining e Information Retrieval, associando dati affini tra loro e assegnandoli a categorie predefinite. Questo processo crea una nuova conoscenza, in quanto fornisce dinamicamente nuove interpretazioni dei dati acquisiti, partendo, ad esempio, dalla profilazione dell'utente.

Dalla Sentiment Analysis si sviluppano ulteriori metodologie di ricerca e analisi quali la **Competitive Intelligence** (ricerca di consumer e marketing insight, individuazione di trend di mercato), **Social Network** Analysis (individuazione dei flussi e degli snodi di comunicazione), **Web Reputation** (monitiraggio in tempo reale su ciò che si dive in rete di un'azienda, un marchio, un prodotto, un servizio), **Viral Tracker (**tracking delle campagne di buzz marketing e monitoraggio delle conversazioni on line sul brand).

Spostando l'attenzione dal monitoraggio del prodotto (considerato anche nell'insieme del suo processo produttivo) a quello del servizio, le stesse metodologie di ricerca possono essere applicate in ambito pubblico, pervenendo all'individuazione del rapporto tra orientamento e intensità emotiva come qualificazione delle attività dell'ente (P.A., centrale e locale) come "bene comune". Qui la condivisione assume un significato particolare in quanto l'utente non è mero consumatore, ma al tempo stesso soggetto attivo che potrebbe incidere sulle modalità stesse di erogazione del servizio.

Anche in questo caso siamo all'anno zero, ed i sistemi di Customer Satisfaction in ambito pubblico dovrebbero essere radicalmente rivisti attraverso modalità operative nuove rispetto agli attuali call center e

uffici URP. Questi ultimi sportelli, infatti, rappresentano di per sé altrettanti servizi, soggetti, quindi, alla valutazione di intensità e orientamento da parte degli stessi utenti.

Quali sono i parametri per misurare la Sentiment Analysis sia sotto l'aspetto orientativo che per l'intensità emotiva? Anche se siamo ancora nella fase sperimentale di questo tipo di analisi (due tools differenti, possono portare, su un identico case study, a risultati non sovrapponibili!) vanno verificate le modalità di ricerca, con le disambiguità terminologiche e le varianti semantiche. Difficile poter definire degli standard (come ad esempio quelli sull'accessibilità statuiti dal 3WConsortium), anche perché l'analisi sul brand è giocoforza una ricerca settoriale, e, inoltre, difficilmente oggettivabile in quanto "emozionale".

Altre due difficoltà di non poco conto sono rappresentate dalla disomogeneità dei dati offerti dalle piattaforme sociali (che stanno ulteriormente formattando le pagine ad hoc per agevolare questo tipo di analisi) e dalla criticità per la privacy degli utenti e del trattamento dei loro dati personali. Quest'ultimo aspetto, in particolare, deve essere oggetto di costante attenzione per la definizione di uno studio che non sia invasivo rispetto a contenuti riservati. Per queste ragioni è auspicabile una certificazione della Sentiment Analysis, sia dal punto di vista tecnologico che socio-psicologico, validata anche dal Garante per la Privacy.

Differente, invece, la possibilità di definizione di parametri per l'agire pubblico, che va impostato secondo modalità di trasparenza, tracciabilità, maggiore beneficio per la collettività in termini di rilevanza economica, credibilità, velocità nell'erogazione dei servizi. Da queste connotazioni si può partire per identificare elementi di valutazione. Anche in questo caso, comunque, l'analisi dovrebbe essere affidata a soggetti dotati di certificazione istituzionale ad hoc. Gli scenari, in questo settore, hanno una crescita esponenziale e creano un circuito virtuoso rispetto alle tecnologie che generano le nuove esigenze sociali, delle quali sono, al tempo stesso, figlie. La fase di passaggio ha le caratteristiche di una vera e propria rivoluzione copernicana, dove non sono più il prodotto o il servizio ad essere al centro del mercato, ma l'uomo, il consumatore, l'utente, con i suoi sentimenti. In tal senso vanno aboliti i criteri che lo vedevano un mero dato d'archivio: l'utente stesso diventa un motore di ricerca, con un

approccio basato anche su un dizionario "sentimentale", efficace ed efficiente.

1.8 La scelta delle features

Le metodologie adottate per la sentiment analysis che sono presenti in letteratura si caratterizzano per due criticità principali.

In primo luogo ci si chiede quali "feature" utilizzare più idonee all'analisi di un dato orientamento del sentiment all'interno di testi, in secondo luogo, poi come debbano essere correttamente interpretate le feature che sono state considerate.

Per quanto concerne le scelte di idonee features, infatti, è possibile effettuare scelte differenti quali l'utilizzo di Unigrams (Singole parole) piuttosto che N-grams (Più parole) se non Frasi complete.

All'interno degli ambi enti sperimentali ciascun set di dati raccolti viene utilizzato per estrarre le caratteristiche che saranno impiegate per addestrare il sentiment classifier adottato.

E' possibile utilizzare la generica presenza di n-grammi come caratteristica binaria, al contrario di ambienti dove l'obiettivo è quello di ricercare generiche informazioni (ad esempio per l'ottimizzazione dei motori di ricerca) per i quali la frequenza di occorrenza di una parola chiave è una caratteristica più adatta. E' abbastanza intuitivo infatti che il ripetersi di una parola chiave non possa essere considerato un indicatore dell'orientamento del sentiment generale di un testo (o comunque non rappresenti una condizione sufficiente). Pang et al. hanno dimostrato come, nel campo della sentiment analysis sia possibile ottenere risultati più confortanti considerando l'effettiva presenza di un dato termine in un testo piuttosto che la sua effettiva frequenza[107]. Sulla scelta poi di utilizzare unigrams, bigrammi e trigrammi. Pang et al. ancora hanno riferito che gli unigrams sono più performanti dei bigrammi quando si esegue la sentiment analysis recensioni di film. Dave et al.[108] hanno invece ottenuto risultati contrari: bigrammi e trigrammi hanno funzionato meglio per la

[107] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. *Thumbs up? sentiment classification using machine learning techniques*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86

[108] Kushal Dave, Steve Lawrence, and David M. Pennock.2003. *Mining the peanut gallery: opinion extraction and semantic classification of product reviews*. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pages 519–528, New York, NY, USA. ACM.

classificazione delle polarità (e quindi del sentiment) nel caso dei product-reviews.

Un ambito particolarmente dibattuto nel quale ci si è investigato per la sentiment analysis è quello della classificazione del sentiment all'interno dei Blog. In questo particolare ambito applicativo sono stati sviluppate numerose tecniche tra le quali, ad esempio, quella di Turney[109] il quale ha proposto un semplice algoritmo non supervisionato di apprendimento basato sull'orientamento semantico per la classificazione di recensioni sul web come "pollice levato" e "pollice verso".

In questo caso l'utilizzo proposto è quello di feature relative a intere frasi.

L'orientamento semantico è calcolato come la mutua informazione tra una data frase e la parola "eccellente" meno la mutua informazione tra la frase e la parola data "povero". Come già anticipato, in Pang[110] vengono analizzate differenti tecniche di apprendimento automatico e si è dimostrato che l'adozione di unigram SVM (Support Vector Machines) consente performance ottime per la classificazione di recensioni di film. Tuttavia, ad esempio i blog non sono recensioni di film e dovrebbero essere adottate tecniche diverse. Ad esempio è possibile dimostrare[111] che il Bag of Words (IR) è un approccio "soggetto dipendente".

La scelta di una feature piuttosto che un'altra, ancora una volta potrebbe portare a risultati discordanti. In questo caso l'utilizzo di un classificatore addestrato su recensioni di film potrebbe non fornire risultati ottimali per la classificazione di recensioni di automobili o blog. Oltre ad essere informali, poco strutturati, pieni di errori ortografici e grammaticali, i blog sono potenzialmente argomentati su una molteplicità di settori.

---

[109] P. Turney, 2002. *Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews*. 40th Annual Meeting of the Association for Computational Linguistics (ACL'02) 417–424.

[110] B. Pang, L. Lee and S. Vaithyanathan, 2002. *Thumbs up? sentiment classification using machine learning techniques*.

[111] C. Engstrom, 2004. *Topic dependence in sentiment classification*. Master's thesis, University of Cambridge.

Dal momento che in un blog un determinato post può cambiare significato a seconda della sua posizione un unigram SVM potrebbe essere poco indicato per l'analisi del sentiment allo stesso modo di un analisi del sentiment basata su words (parole) che però siano relative ad altri contesti. Ancora più importante, gli utenti vogliono solo vedere le relative sezioni di un post sul blog nei risultati di ricerca dei motori politici recupero sentimento, non vogliono post del blog interi.

Mullen e Collier[112] hanno introdotto un approccio di classificazione basato su frasi opinionate, incorporando diverse nuove fonti di informazione nel Support Vector Machines. Tale supporto ha consentito a Wiebe, Wilson e Hoffmann[113], un'analisi del sentiment a livello di frase per determinare se un'espressione è neutrale o polare e sulla base di questa analisi, è possibile distingue la polarità dell'intero testo partendo dalle espressioni positive o negative. Il loro approccio comporta in primo luogo l'identificazione della polarità contestuale di frasi sulla base delle parole "indizio" e poi nella disambiguazione delle frasi raccolte.

Wiebe e Riloff[114] esplorano l'idea di analisi della soggettività per migliorare la precisione del sistema di estrazione di informazioni. Yu e Hatzivassilogou[115], invece, si occupano della differenziazione di opinioni sia a livello di documento che a livello di frase. L'approccio adotta un classificatore Bayesiano per discriminare tra i documenti opinionati e descrive tre tecniche per la rilevazione di statistiche non supervisionate a livello di frase.

Con l'approccio dei "minimum cuts", Pang ha diviso i documenti giù in frasi oggettive e soggettive. Le frasi soggettive vengono poi utilizzate da sole come se fossero il documento originale. Nessuna di

---

[112] T. Mullen and N, Collier, 2004. *Sentiment analysis using support vector machines with diverse information sources*. 42nd Meeting of the Association for Computational Linguistics (ACL).

[113] TheresaWilson, Paul Hoffmann, J.W. 2005. *Opinionfinder: a system for subjectivity analysis*. 34–35

[114] Janyce Wiebe, E. R. 2005. *Creating subjective and objective sentence classifiers from unannotated texts*. Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)

[115] Hong Yu, V. H. *Towards answering opinion questions: Separating facts from opinions*

questi approcci utilizza le “parti del discorso basate su n-grammi” e Pang utilizza l'approccio degli unigrammi.

Pang descrive un nuovo modello che limita la caratterizzazione del testo all’analisi delle porzioni soggettive del documento.

Dall’analisi delle differenti tecniche proposte emerge che i risultati più performanti in questo settore si basano sull’utilizzo del metodo di Bayes e sull’utilizzo di SVM.

Dave[116] ha presentato l’analisi di un set completo di recensioni di un prodotto in cui i trigrammi sono più performanti dei bigrams che a loro volta sono più performanti degli uni-grammi nei due test per Classificatori Naive Bayes.

Tuttavia, la valutazione dell’incoerenza, dati sparsi e la distribuzione distorta influiscono sulle prestazioni del sistema.

Consideriamo ora alcun tra i modelli associati alle specifiche features.

1.8.1 Bag of words (IR)

Si tratta di un modello che adotta le singole parole di una frase come caratteristiche, assumendo la loro indipendenza condizionale. Il testo viene rappresentato come un insieme ordinato di parole. Ogni caratteristica del vettore rappresenta l'esistenza di una parola. Si tratta effettivamente di un modello unigram, dove ogni parola è condizionalmente indipendente dalle altri. Tutte le parole (funzioni) nel vettore caratteristica costituiscono il dizionario. La sfida di questo approccio è la scelta di parole che siano appropriate per diventare caratteristiche.

Utilizzando questo modello la frase

This is a great event

può essere rappresentata dal seguente vettore:

$$\dot{\overline{F_0}} = \{'a': 1, 'event': 1, 'great': 1, 'is': 1, 'this': 1\}$$

[116] Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. World Wide Web

(E' presentata vettore funzione come un dizionario Python; NLTK, ad esempio, utilizza questa rappresentazione di un vettore di feature).

Questo sarebbe una rappresentazione soddisfacente se vi fosse una sola frase in tutto il corpus. Se, al contrario, vogliamo essere in grado di rappresentare altre frasi, ad esempio,

It is a great Startrek movie

il precedente vettore non potrebbe essere idoneo ad una efficace rappresentazione.

E' quindi necessario estendere il set di parole, e integrare il vettore caratteristica. L'insieme delle funzionalità descritte in questo caso sarebbe

$$\{'a', 'event', 'great', 'is', 'it', 'movie', 'Startrek', 'this'\}$$

Il vettore di feature, rappresentativo di entrambe le frasi, sarebbe:

$$\dot{F_0} = \{'a' : 1, 'event': 1, 'great' : 1, 'is' : 1, 'it' : 0, 'movie' : 0, 'Startrek': 0, 'this' : 1\}$$

Solo alcune delle parole appaiono in entrambe le frasi, e sono utilizzate per esprimere la somiglianza tra le frasi. Ovviamente, per qualsiasi uso reale, il vettore di caratteristica dovrebbe contenere un numero maggiore di parole.

E possibile registrare o la presenza della comparsa parola in un testo, o la sua frequenza (il numero di volte che la parola è apparsa). La frequenza del vettore caratteristica per la frase

I really really enjoyed the movie

per la parola "*really*" avrebbe valore "2" (numero di occorrenze della parola.) Ciò consente di indicare con un maggior livello di approfondimento la polarità della frase. Tuttavia, dato che si confronta singole frasi, non è molto comune avere una parola che appare più volte. Inoltre, abbiamo già visto come Pang abbia dimostrato che per l'analisi del sentiment, il considerare la semplice presenza o assenza di una parola in una frase determina le stesse prestazioni di

un’informazione più dettagliata contenente anche il numero di occorrenze.

L’ideale vettore delle feature per un Bag-of-Words dovrebbe contenere tutte le parole esistenti in un linguaggio e rappresenta, di fatto, un dizionario della lingua. Tuttavia, questo modello non sarebbe pratico per almeno tre ragioni. Uno dei motivi è la complessità del modello, dal momento che il modello avrebbe catturato più informazioni di quanto effettivamente richiesto. Ciò rappresenterebbe un corpus ideale per il training del testo, ma sarebbe sovradimensionato e porterebbe a pessime prestazione ogni volta che si ha a che fare con nuovi esempi.

Inoltre, la complessità computazionale di apprendimento successivo (per esempio un vettore di lunghezza di un milione di elementi) è enorme. Infine, se pure i due ostacoli precedenti sono stati superati, gestire nuove parole non è ancora possibile.

Le lingue sono molto dinamiche, e le parole nuove sono spesso inventate, soprattutto nella comunità di Internet. Il sentiment dell'autore è spesso espresso attraverso sia parole sia frasi. Per esempio, nella frase

This is a great event

la parola "*great*" è il migliore indicatore del parere dell'autore. Un approccio che può essere immaginato è quello di selezionare manualmente le più importanti parole chiave (come “grande”, “eccellente”, “terribile” quando vogliamo esprimere la polarità di una frase) e usarle come caratteristiche.

Tuttavia, Pang et al. dimostrano che la scelta manuale delle parole chiave è superato da modelli statistici, dove un buon set di parole che rappresentano caratteristiche sono selezionati per la loro presenza nel corpus di formazione esistente. La qualità della selezione dipende delle dimensioni del corpus e la somiglianza di domini di formazione e dati di test.

L’utilizzo di tendenze statistiche da domini differenti, che non hanno le stesse proprietà desiderate come il dominio originale, può portare a risultati imprecisi, ad esempio, se l'analisi di soggettività viene

eseguita su una serie di frasi prelevate da un giornale, in cui la maggior parte delle frasi sono state scritte in stile oggettivo.

E 'importante creare un dizionario completo (vettore di feature) che catturi caratteristiche più importanti sia in set di training sia in esempio inedito.

E' possibile valutare due diversi approcci per la selezione delle funzioni nel modello Bag-of-Words. Entrambi sono basati sulla selezione delle parole più frequenti in un corpus testuale. Un approccio è quello di utilizzare gli elementi provenienti dallo stesso dominio. Si dividere ogni testo in due pezzi. Un pezzo verrà utilizzato come insieme noto, uno per gli scopi di addestramento. Si noti, tuttavia, che la varianza può essere grande a causa del piccolo corpus considerato. Un altro approccio presuppone l'utilizzo di un set di feature basato sulle frequenze di parole nel corpus esistente, che è simile per argomento e per sentiment. Il vantaggio di tale scelta di caratteristiche è la capacità di una migliore comparazione di nuovi messaggi. I messaggi possono appartenere ad altri articoli di un determinato sito web.

La selezione di tutte le parole che appaiono nei corpora può portare a un sovradimensionamento, analogo al caso di selezione di tutte le parole in una lingua, precedentemente descritto. Vi sono anche altri rischi. Devono essere impostati alcuni limiti per vincolare la dimensione del vettore caratteristica.

Ad esempio, è possibile rimuovere alcune delle parole esistenti frequenti tramite una lista nera. Pratica comune per i motori di ricerca è quella di rimuovere le parole come una, la, do,. . . che portano poche informazioni utili.

Questo modello è semplice, ed ha diversi limiti.

Ad esempio è impossibile catturare le relazioni soggettività/polarità tra le parole, distinguendo tra le parti del discorso, di gestire le negazioni, e i diversi significati di una parola. Una estensione del modello che può apparire naturale e semplice sarebbe quella di considerare anche le coppie di parole (bigrammi) come funzioni invece degli unigrammi. Tuttavia la letteratura non mostra il vantaggio di bigrammi rispetto agli unigrammi nell'analisi del sentiment.

1.8.2 Lessici Annotati (WordNet, SentiWordNet)

Le informazioni relative al Sentiment-correlato possono essere codificate lessicalmente all'interno delle parole della frase, sintatticamente mediante proposizioni subordinate, e morfologicamente attraverso i cambiamenti nei toni attitudinali del significato della parola con suffissi (in particolare, in lingue con una ricca sistema di flessioni, come il russo o italiano)[117].

Metodi per l'estrazione e l'annotazione personale di termini includono approcci di apprendimento automatico che esaminano le relazioni di congiunzione tra aggettivi[118], il clustering di aggettivi secondo una similarità distributiva basta su una piccola quantità di parole date[119], l'algoritmo pattern-bootstrapping per estrarre sostantivi[120], l'esame basata sul web di informazioni comuni per estrarre aggettivi[121], e il morphosyllabic tagging sentiment[122].

Un lessico per il sentiment dovrebbe contenere elementi che indichino polarità (positiva o negativa), e anche l'intensità del sentiment o, in alcuni casi, il grado di centralità rispetto alla categoria del sentiment.

Per determinare il livello di intensità del sentiment per una parola, vengono solitamente utilizzati la Latent Semantic Analysis[123], la

---

[117] J Reilly, and L. Seibert. Language and emotion. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), *Handbook of Affective Science*, pp. 535–559, 2003

[118] V. Hatzivassiloglou, and K. R. McKeown. *Predicting the semantic orientation of adjectives*. Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL, pp. 174–181, 1997

[119] J. Wiebe. *Learning subjective adjectives from corpora*. Proceedings of the 17th Conference of the AAAI, 2000.

[120] E. Riloff, J. Wiebe, and T. Wilson. *Learning subjective nouns using extraction pattern bootstrapping*. Proceedings of 7th Conference on Natural Language Learning, pp. 25–32, 2003

[121] M. Baroni, and S. Vegnaduzzo. *Identifying subjective adjectives through web-based mutual information*. Proceedings of the German Conference on NLP, 2004

[122] K. Moilanen, and S. Pulman. *The good, the bad, and the unknown: Morphosyllabic sentiment tagging of unseen words*. Proceedings of ACL-08:HLT, pp. 109–112, 2008

[123] P. D.Turney, and M. L. Littman. *Measuring praise and criticism: Inference of semantic orientation from association*. ACM Transactions on Information Systems, 21(4):315–346, 2003

tecnica di mutua informazione puntuale appuntata[124], e metodi che impiegano le relazioni della struttura WordNet[125,126,127].

La maggior parte dei sistemi basati su lessici per l'analisi del sentiment scontano la difficoltà di assegnare i punteggi del sentiment alle parole che non sono disponibili nei loro database. Per far fronte alla limitazione della copertura del lessico, vengono proposti metodi per creare automaticamente ed espandere il valore di soggettività del lessico rappresentato dalle parole, che sono annotate per polarità del sentiment, punteggio di polarità e pesi.

Il primo passo nella costruzione di un lessico costruito su termini "polarizzanti" implica la raccolta di contenuti rilevanti "part-of-speech" e parole (aggettivi, avverbi, sostantivi e verbi), e l'assegnazione di punteggi predefiniti di polarità (punteggio positività e punteggio negativo) a ogni unità lessicale.

Per "punteggio di polarità di sentiment " si intende la forza o il grado di intensità del sentimento. Per entrambe le valenze opposte, i limiti della punteggio di polarità sono 0.0 (che indica l'assenza di dati di orientamento del sentimento) e 1,0 (il valore massimo).

Come già anticipato lo svantaggio principale di un approccio di sentiment analysis interamente basato su un lessico precostituito con una serie di termini convoglianti per la polarità è la mancanza di scalabilità, poiché il richiamo del metodo lessicale dipende dalla copertura del database utilizzato. Per espandere i lessici precostituiti è possibile sfruttare SentiWordNet[128].

---

[124] J. Read. *Recognising affect in text using pointwise-mutual information*. Thesis. University of Sussex, 2004

[125] S.-M. Kim, and E. Hovy. *Determining the sentiment of opinions*. Proceedings of Conference on Computational Linguistics, pp. 1367–1373, 2004

[126] A. Andreevskaia, and S. Bergler. *Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. Proceedings of the 11th Conference of the European Chapter of the ACL, EACL, 2006

[127] A. Esuli, and F. Sebastiani. *SentiWordNet: a publicly available lexical resource for opinion mining*. Proceedings of the 5th International Conference on Language Resources and Evaluation, pp. 417–422, 2006

[128] Ibid

SentiWordNet è stato sviluppato sulla base di WordNet[129], synset composto da termini sinonimi. Si basa sul presupposto che "*i diversi significati dello stesso termine possono avere diverse proprietà di opinion*e" Vengono impiegati otto classificatori ternari per analizzare quantitativamente la glosse associate al synset. Tre punteggi numerici (Obj (s), Pos (s), e Neg (s), che vanno da 0,0 a 1,0 e in somma pari a 1,0), che caratterizzano in che misura i termini inclusi in un synset sono oggettivi, positivi, e negativi, sono stati determinati automaticamente in base alla percentuale di classificatori assegnati dalla corrispondente etichetta al synset.

1.8.3 Pattern Sintattici

I processi di estrazione delle informazioni (IE) tipicamente utilizzano sistemi lessico-sintattici per identificare le informazioni rilevanti. La concreta rappresentazione di questi modelli varia a seconda dei sistemi, ma la maggior parte dei modelli rappresenta relazioni esistenti nelle frasi tra nome e frasi verbali. Ad esempio, un sistema di IE progettato per estrarre informazioni su dirottamenti potrebbe utilizzare il pattern
dirottatore di <x>

che ricerca il sostantivo dirottamento ed estrae l'oggetto della preposizione "di"relativa al veicolo dirottato. Il pattern

<x> è stato dirottato

estrarrebbe il veicolo sequestrato quando trova il verbo dirottato nella forma passiva, e il pattern

<x> dirottato

estrarrebbe il dirottatore quando trova il verbo dirottato nella voce attiva.

Una delle ipotesi è che i pattern di estrazione siano in grado di rappresentare le espressioni soggettive che hanno significati non composti. Ad esempio, si consideri l'espressione comune

[129] G. A. Miller. WordNet: *An on-line lexical database*. International Journal of Lexicography, Special Issue, 3(4):235–312, 1990

drives (someone) up the wall

che esprime la sensazione di essere arrabbiata con qualcuno.
Il significato di questa espressione è molto diverso dai significati delle sue singole parole (drives, up, wall). Inoltre, questa espressione non è una sequenza di parole fisse che potrebbe essere facilmente catturata da N-grammi. È una costruzione relativamente flessibile che può essere più generalmente rappresentata da

<x> drives <y> up the wall

dove x e y possono essere sintagmi nominali arbitrarie.
Questo pattern potrebbe corrispondere molte frasi diverse, come

"George drives me up the wall",
"She drives the mayor up the wall" or
"The nosy old man drives his quiet neighbors up the wall."

Sono stati sviluppati una varietà di algoritmi per apprendere automaticamente i patterns di estrazione. La maggior parte di questi algoritmi richiede risorse speciali di formazione e apprendimento, come ad esempio testi annotati con tag di dominio-specifici (ad esempio, AutoSlog[130], CRYSTAL[131], RAPIER[132], SRV[133], FRUSTA[134]) o o con parole chiave definite manualmente, frames, o object recognizers (ad esempio, Palka[135] e Liep[136]).

---

[130] E. Riloff. 1996. *Automatically Generating Extraction Patterns from Untagged Text*. In Proceedings of the AAAI-96

[131] S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert. 1995.CRYSTAL: *Inducing a Conceptual Dictionary*. In Proceedings of the IJCAI-95

[132] M. E. Califf. 1998. *Relational Learning Techniques for Natural Language Information Extraction*. Ph.D. thesis, Tech. Rept. AI98-276, Artificial Intelligence Laboratory, The University of Texas at Austin.

[133] Dayne Freitag. 1998. *Toward General-Purpose Learning for Information Extraction*. In Proceedings of the ACL-98

[134] S. Soderland. 1999. *Learning Information Extraction Rules for Semi-Structured and Free Text. Machine Learning*, 34(1-3):233–272

[135] J. Kim and D. Moldovan. 1993. *Acquisition of Semantic Patterns for Information Extraction from Corpora*. In Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications

[136] S. Huffman. 1996. *Learning information extraction patterns from examples*. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, editors, Connectionist,

AutoSlog-TS[137] ha un approccio diverso, richiede solo un corpus di testi non annotati che sono stati divisi in quelli che legati al dominio di riferimento (testi "rilevanti") e quelli che non lo sono (testi "irrilevanti").
Più di recente, sono stati utilizzati due algoritmi di avvio automatico per apprendere pattern di estrazione.
Metabootstrapping[138] apprende sia modelli di estrazione sia un lessico semantico con i testi annotati e considera alcune parole date come input. ExDisco[139] utilizza un meccanismo di bootstrap per trovare nuovi pattern di estrazione che utilizzano testi non annotate e un po'di pattern dati come input iniziale.

1.8.4 Annotazioni a livello di paragrafo

L'uso di annotazioni a livello di paragrafo consente di attribuire etichette ai dati durante il meccanismo di training di un modulo di apprendimento supervisionato in maniera più concisa. Queste annotazioni sono più dettagliate rispetto alle annotazioni tipiche a livello di documento, e comportano un incremento di precisione a livello del classificatore di sentiment, soprattutto per quanto concerne l'analisi dei Blog[140].

A causa della natura in forma libera dei blog negli articoli spesso si discute di argomenti diversi, e anche se le annotazioni a livello di paragrafo dovrebbe essere utile nel trattare il problema della collocazione nel testo dell'argomento, annotazioni a livello di paragrafo possono anche comportare un decadimento delle performance (impiego di più tempo consumo per generare il sentiment).

Statistical, and Symbolic Approaches to Learning for Natural Language Processing, pages 246–260. Springer-Verlag, Berlin

[137] E. Riloff. 1996. *Automatically Generating Extraction Patterns from Untagged Text*. In Proceedings of the AAAI-96

[138] E. Riloff and R. Jones. 1999. *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping*. In Proceedings of the AAAI-99

[139] R. Yangarber, R. Grishman, P. Tapanainen, and S. Huttunen. 2000. *Automatic Acquisiton of Domain Knowledge for Information Extraction*. In Proceedings of COLING 2000

[140] Ferguson, P., O'Hare, N., Davy, M., Bermingham, A., Tattersall, S., Sheridan, P., Gurrin, C., and Smeaton, A. F. (2009). *Exploring the use of paragraph-level annotations for sentiment analysis in financial blogs*. 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA)

1.9 L'interpretazione corretta delle features considerate

Abbiamo visto come la scelta di una determinata feature porti all'adozione di un modello (e di un ambito specifico) da adottare e nel quale attuare la determinata campagna sperimentale.

Oltre alla scelta della feature vera e propria vi è poi un problema di interpretazione associato ad essa. Anche in questo caso sono numerosi gli approcci in letteratura, ciascuno dei quali caratterizzato da specifiche criticità. Di seguito forniremo una rapida rassegna dei principali in modo da fornire una rappresentazione quanto più esaustiva possibile dello scenario presente.

E' possibile classificare due differenti tipologie di approccio.

Da un lato i modelli che si basano sul Machine learning

- Naïve Bayes
- Maximum Entropy Classifier
- SVM
- Markov Blanket Classifier

Dall'altro quello invece dei Metodi non supervisionati

- Adozione di lessici annotati

1.9.1 Naïve Bayes

In termini semplici, un classificatore Naive Bayes assume che la presenza (o l'assenza) di una particolare caratteristica di una classe è correlata alla presenza (o l'assenza) di qualsiasi altra caratteristica, data la variabile di classe.

Ad esempio, un frutto può essere considerata una mela se è rosso, rotondo, e circa 4" di diametro. Anche se queste caratteristiche dipendono l'una dall'altra o anche da altre caratteristiche, un classificatore Naive Bayes considera tutte queste proprietà indipendente l'una dall'altra in relazione alla probabilità la probabilità che questo frutto sia una mela.

A seconda del dettaglio del modello probabilistico, i classificatori Naive Bayes possono essere addestrati in modo molto efficiente in un ambiente di apprendimento supervisionato. In molte applicazioni pratiche, per la stima dei parametri per i modelli Naive Bayes viene utilizzato il metodo della massima verosimiglianza; in altre parole, si può lavorare con il modello Naive Bayes senza adottare la probabilità bayesiana o utilizzare i metodi bayesiani.

A dispetto della loro rappresentazione elementare le ipotesi apparentemente semplicistiche, i classificatori Naive Bayes sono stati adottati (con ottimi risultati) in parecchie situazioni complesse del mondo reale. Nel 2004, l'analisi del problema di classificazione bayesiana ha dimostrato che ci sono alcune ragioni teoriche per l'efficacia apparentemente irragionevole di semplici classificatori di Bayes[141]. Tuttavia, un confronto globale con altri metodi di classificazione nel 2006 ha mostrato che la classificazione di Bayes è superato da più approcci attuali, come ad esempio "boosted trees" o "random forests"[142].

Un vantaggio di un classificatore Naive Bayes è che esso richiede solo una piccola quantità di dati di training per stimare i parametri (medie e varianze delle variabili) necessari per la classificazione. Considerato che le variabili sono assunte indipendenti, devono essere determinate solo le varianze delle variabili per ogni classe e non l'intera matrice di covarianza.

Un approccio per la classificazione di un testo è quello di assegnare a un dato documento *d* la classe

$c^* = \arg\max_c P(c|d)$.

E' possibile ricavare un classificatore Naive Bayes (NB) osservando innanzitutto che dalla regola di Bayes

[141] H. Zhang, *The optimality of naive bayes*, in Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference , (V. Barr and Z. Markov, eds.), Miami Beach, FL: AAAI Press, 2004

[142] Caruana, R.; Niculescu-Mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. Proceedings of the 23rd international conference on Machine learning. CiteSeerX: 10.1.1.122.5901

$$P(c|d) = \frac{P(c)P(d|c)}{p(d)}$$

dove *P (d)* non svolge alcun ruolo nella selezione di c^*.

Per stimare il termine *P(d|c)*, Naive Bayes si decompone assumendo che le f_i sono scorrelate assegnata una classe di *d*:

$$P_{NB}(c|d) := \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

Nonostante la sua semplicità e il fatto che la sua condizionale assunzione di indipendenza chiaramente lo rendono poco rappresentativo di situazioni del mondo reale, il metodo di categorizzazione del testo basato su Naive Bayes funziona ancora sorprendentemente bene[143], anzi, Domingos e Pazzani[144] mostrano che Naive Bayes è ottimale per alcuni problemi di classi con caratteristiche fortemente dipendenti.

1.9.2 SVM

Le Support Vector Machine (SVM) sono tecniche di apprendimento automatice usate per la classificazione che utilizzano una funzione chiamata kernel per mappare uno spazio di punti di dati in cui i dati non sono linearmente separabili su nello spazio sul quale sussistono, con una tecnica di rilevamento per la classificazione errata che tenga conto di punti anomali o comunque della presenza di rumore.

Nel caso, ad esempio, di due classi, si adotta un separatore lineare. Il criterio adottato è quello di spaziarlo massimamente dai punti di entrambi gli insiemi.

La metodologia consiste nei seguenti passi fondamentali:

- Attraverso un dataset di training si definisce il miglior iperpiano di separazione tra le classi.

[143] David D. Lewis. 1998. *Naive (Bayes) at forty: The independence assumption in information retrieval*. In Proc. of the European Conference on Machine Learning (ECML), pages 4–15. Invited talk

[144] Pedro Domingos and Michael J. Pazzani. 1997. *On the optimality of the simple Bayesian classifier under zero-one loss*. Machine Learning, 29(2-3):103–130.

- I dati contenuti nel dataset vengono processati per trovare l'iperpiano mediante una procedura di ottimizzazione quadratica
- Dato un nuovo punto $\vec{x}$ da classificare, la funzione di classificazione $f(\vec{x})$ calcola la proiezione del punto sul iperpiano normale.
- Il segno di questa funzione determina la classe da assegnare al punto (1 o -1).
- Se il punto è entro il margine di classificazione, la classificazione può restituire "Non lo so ", piuttosto che una delle due classi.
- Il valore di $f(\vec{x})$ può essere trasformato in una probabilità di classificazione.

In relazione alla scelta dei differenti patterns con le SVM questi alcuni degli esempi presenti in letteratura

- Pang and Lee [145]tilizzano le SVM con unigrammi, bigrammi e negazioni
- Dave et al.[146] usano le SVM con unigrammi, bigrammi per confrontare i risultati col modello da loro proposto
- Kennedy and Inkpen[147] combinano le SVM con un metodo che conta la parole positive e quelle negative
- Matsumoto et al.[148] usano I "frequent word patterns" e I "dependency tree patterns" sfruttando unigrammi e bigrammi.

[145] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. *Thumbs up?: sentiment classification using machine learning techniques*. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, EMNLP '02, pages 79{86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics

[146] Kushal Dave, Steve Lawrence, and David M Pennock. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*. In Proceedings of the 12th international conference on World Wide Web, 2003

[147] Alistair Kennedy and Diana Inkpen. *Sentiment classification of movie reviews using contextual valence shifters*. Computational Intelligence, 22(2):110{125, May 2006

[148] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. *Sentiment classification using word sub-sequences and dependency sub-trees*. In Tu Ho, David Cheung, and Huan Liu, editors, Advances in Knowledge Discovery and Data Mining, volume 3518 of Lecture Notes in Computer Science, pages 21{32. Springer Berlin / Heidelberg, 2005

- Wu et al.[149] utilizzano "phrase dependency parses" con i "tree kernels"

1.9.3 Markov Blanket Classifier

Nell'ambito della classificazione documentale vengono solitamente utilizzati due metodi: quello a filtro e quello a wrapper[150,151]. Un metodo wrapper valuta iterativamente il classificatore utilizzato per ogni sottoinsieme caratteristica selezionata durante la ricerca, mentre un metodo a filtro trova sottoinsiemi predittivi di caratteristiche indipendentemente dal classificatore finale. Un metodo wrapper richiede dei calcoli estremamente onerosi, che lo rende applicabile solo in un set di dati di grandi dimensioni di interesse.

Il metodo a filtro che utilizza il Markov blanket concept ha suscitato molta attenzione nella letteratura[152,153] così come l'adozione contemporanea dello stesso con altre funzioni di selezione[154].

Negli ambienti di apprendimento automatico del Machine Learning, il Markov Blanket per un nodo A in una rete bayesiana è l'insieme di nodi ∂A composto dai genitori di A, dai suoi figli e da tutti gli altri genitori dei suoi figli. In una rete di Markov, il Markov Blanket di un nodo è il suo insieme di nodi adiacenti. Un Markov Blanket può anche essere indicata con MB (A).

---

[149] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. *Phrase dependency parsing for opinion mining*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3, EMNLP '09, pages 1533{1541, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics

[150] Almuallim, H., Dietterich, T.G.: *Learning with many irrelevant features*. In: In Proceedings of the Ninth National Conference on Artificial Intelligence, AAAI Press (1991) 547–552

[151] Kohavi, R., John, G.H.: *Wrappers for feature subset selection*. Artificial Intelligence 97(1-2) (1997) 273–324

[152] Koller, D., Sahami,M.: *Toward optimal feature selection*. In: Proceedings of the Thirteenth International Conference on Machine Learning,Morgan Kaufmann (1996) 284–292

[153] Yaramakala, S., Margaritis, D.: *Speculative markov blanket discovery for optimal feature selection*. In: ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2005) 809–812

[154] Tsamardinos, I., Aliferis, C.F.: *Towards principled feature selection: Relevancy, filters and wrappers*. In: in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Morgan Kaufmann Publishers (2003)

Il Markov Blanket di un nodo contiene tutte le variabili che proteggono il nodo dal resto della rete. Ciò significa che la coperta Markov Blanket di un nodo è la conoscenza necessaria solo per prevedere il comportamento di tale nodo. In una rete bayesiana, i valori dei genitori e figli di un nodo evidentemente forniscono informazioni su tale nodo, tuttavia debbono essere considerati anche gli altri genitori dei figli del nodo in questione, in quanto possono essere anch'essi utilizzati per spiegare il nodo in questione.

Un Classificatore Markov Blanket può essere in grado di catturare le dipendenze tra le parole, trovare un vocabolario che possa efficacemente estrarre sentimenti e, in ultima analisi, fornire migliori previsioni sul sentimento espresso in un documento di testo, rispetto a diverse altre tecniche di machine learning.

Effettuare una classificazione usando un Markov Blanket (MB) per la variabile sentimento ha due proprietà importanti: si specifica una previsione statisticamente efficiente della distribuzione di probabilità della variabile sentimento dal più piccolo sottoinsieme di predittori, e si lavora con una elevata precisione, evitando sovra-adattamento a causa di un esubero di predittori.

1.9.4 Adozione di lessici annotati

Un lessico annotato per la valutazione del sentiment è più complesso rispetto ad altri lessici basati sul Natural Language Processing (NLP). Due sono le ragioni di questa complessità:

- Ogni voce del lessico riporta le informazioni sulla sua polarità in aggiunta alle sue caratteristiche ortogonali, fonologiche, sintattiche e morfologiche. Queste informazioni di polarità sono di solito rappresentate come positive, o negative o neutre. Per esempio, SentiWordNet[155] utilizza terzine [obiettivi positivi, negativi,], con valore minimo e massimo tra 0,0 e 1,0. La maggior parte delle parole presentano più orientamenti a seconda del loro uso e del dominio di riferimento. Ad esempio consideriamo la frase "Questo danno è permanente". In questa frase, "permanente" è una parola positiva, ma l'orientamento generale del commento è negativo. Inoltre, "imprevedibile" è

[155] Andreevskaia, A. and S. Bergler: Mining *WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses*. In: EACL 2006, Trent, Italy, (2006)

una parola positiva quando viene usata sulla trama di un film, ma diventa negativa per le prestazioni di un forno a microonde.

Il compito di costruzione di un lessico annotato è suddiviso in fasi seguenti:

- Classificare le parole in soggettive e oggettive. Quando l'algoritmo di classificazione è applicato su questi parole, il classificatore ignora semplicemente i termini oggettivi; in questo modo le prestazioni dipendono totalmente dalle parole soggettive.
- Classificare queste parole secondo le regole morfologiche, che lavorano a livello di parola. Queste regole possono cambiare la struttura, il significato e la parte del discorso delle parole. Ad esempio, le regole per la marcatura di un aggettivo con il sostantivo che qualifica, ecc
- Identificare le loro regole grammaticali, che descrivono le possibili strutture di una frase e le posizione del parti del discorso l'una rispetto all'altra
- Individuare le relazioni tra il lessico delle diverse voci. Queste relazioni possono definire sinonimi, contrari e riferimenti incrociati, ecc
- Decidere ed annotare polarità e poi intensità delle voci. In questa fase prima le voci sono classificate come positive o negative e poi ad esse vengono attribuiti i punteggi di intensità. Alcune voci hanno solo orientamenti e alcuni hanno solo intensità (Come modificatori) e alcuni hanno entrambi i valori.

- Si presuppone che il lessico sia composto da voci soggettive o oggettivo. I termini oggettivi sono salvati senza alcun segno di polarità, ma i termini soggettivi sono ulteriormente classificati sulle basi di orientamento e intensità in tre tipi come:
  - I termini solo con orientamento T(O). Questi sono i termini che sono o assoluti positivi o assoluti negativi. Il grado di positività o negatività non è indicato.
  - I termini solo con intensità T(I). Questi sono i termini che non hanno orientamento ma possono intensificare l'orientamento di altre parole nelle frasi.
  - Condizioni sia con orientamento sia con intensità T(O,I). Se un termine contiene sia l'orientamento

(positivo o) negativo e quindi l'intensità si trova in questa categoria ed è contrassegnato con entrambi i valori.

Seconda parte

I Tools di Mercato

Esistono, sul mercato, una serie di applicativi dedicati che integrano strumenti di business intelligence, modelli statistici e analisi semantica, con i quali è possibile esaminare un'ampia copertura di contenuti presenti sul web e sui social media, analizzando migliaia di testi per fornire informazioni su: brand, competitors, reputation.

Bisogna precisare che queste applicazioni non forniscono verità assolute ma solo delle linee guida che devono essere utilizzate nel modo migliore per poter estrarre delle informazioni utili.

Gli strumenti di brand reputation monitoring, possono essere sia gratuiti che a pagamento. Nascono con l'intento di analizzare quello che viene scritto sui social network. Alcuni consentono di effettuare anche analisi più approfondite sui vari trends.

Diverse sono le tecnologie che consentono di reperire le più disparate informazioni sul web. Tutte le tecniche disponibili, a prescindere dalla loro natura, si basano su tre step fondamentali:

1. Data Collection;
2. Data Processing;
3. Deliver;

Data collection – E' un termine utilizzato per descrivere un processo orientato alla preparazione e alla raccolta dei dati. La accolta dei dati si prefigge l'obiettivo della raccolta dei dati che verranno, successivamente, elaborati per estrapolare da essi informazioni su decisioni e questioni importanti. I dati raccolti riguardano, principalmente, la possibilità di fornire informazioni su di uno specifico argomento. La raccolta dei dati avviene di solito nella fase iniziale di un progetto di miglioramento e, spesso viene formalizzato attraverso un piano di raccolta dei dati, che di solito contiene le seguenti attività:

- Pre-collection activity: accordi sugli obiettivi, sui dati di destinazione, sulle definizioni e sui metodi;
- Collection: data collection;
- Present Findings: di solito comporta una qualche forma di ordinamento analisi e/o presentazione;

Quando si effettua l'estrazione delle informazioni si dovrebbe indicare la grandezza delle fonti, considerare il problema del data-overloading e, inoltre, i dati dovrebbero essere tracciati in tempo reale.

Data processing - Per mezzo di varie fonti, i clienti trasmettono le loro opinioni sotto forma di dati. Il sistema informativo riceve come input tali dati e produce in uscita informazioni utili. La conversione dei dati grezzi, in informazioni utili, avviene tramite un'applicazione di elaborazione dei dati.

La necessità della conversione dei dati, in informazioni utilizzabili, ha l'obiettivo di ottimizzare le metriche automatizzate relative alla definizione dei topics, alle informazioni demografiche, all'analisi del sentiment e all'individuazione degli influencer.

Il processo può essere automatizzato ed eseguito su un computer. Esso comporta la registrazione, l'analisi, il calcolo, la diffusione e l'archiviazione dei dati.

Delivery - I dati provenienti dal data processing vengono sottoposti alla fase di presentazione.

La maggior parte delle applicazioni fanno uso di una dashboard utilizzabile dall'utente finale. Queste ultime, solitamente, includono diverse funzionalità e in alcuni casi sono facilmente personalizzabili dall'utente.

2.1 Caratteristiche dei Tool

L'elaborazione del linguaggio naturale (Natural Language Processsing NLP) si occupa dell'interpretazione e dell'estrazione di informazioni da un testo scritto in linguaggio naturale, tramite l'ausilio di calcolatori elettronici. L'aggettivo "naturale" è utilizzato per distinguere il linguaggio umano dai linguaggi formali.

La nascita di applicazioni per il riconoscimento del linguaggio naturale è avvenuta in ambito militare durante la guerra fredda. Nel periodo compreso tra il 1971 e il 1982 nascono i seguenti paradigmi:

- **Simbolico**: studia il linguaggio naturale tramite regole e grammatiche;
- **Logic-based**: unifica le strutture in feature (interconnessioni tra le parti del discorso) analizzabili in modo più potente rispetto alle grammatiche context-free;
- **Natural Language Understanding**: include lo sviluppo del sistema SHRDLU (un programma per calcolatore realizzato a Terry Allen Winograd nel 1972), il quale utilizza un meccanismo di comprensione basato su tre fasi di analisi: sintattica, semantica e deduttiva. Tramite uno schermo grafico l'utente osserva un ambiente virtuale costituito da una superficie piana, una scatola e una serie di oggetti di varie forme. L'utente può interagire in lingua inglese con un immaginario braccio meccanico per spostare gli oggetti. Il programma è in grado di risolvere molte ambiguità della lingua inglese e riuscire a capire, ad esempio, a quale tipo di oggetto ci si riferisce anche se è sottinteso;
- Discourse Modelling: branchia della linguistica che analizza sotto-strutture del linguaggio scritto e parlato;

Un dashboard è una raccolta di analisi e report che offre agli utenti una singola visualizzazione dati che li aiuta a monitorare informazioni associate a un lavoro, prodotto o servizio. Alcuni dashboard offrono un elevato livello di interattività con l'utente, mentre altri visualizzano solo immagini statiche. Il livello e il tipo di interattività dipendono dall'applicazione utilizzata per creare il dashboard.

Molti utilizzano indifferentemente i termini "dashboard" e "scorecard" ma, questi due elementi, sono in realtà molto diversi. Una scorecard è un tipo di report che visualizza un insieme di indicatori di prestazioni chiave (Key Performance Indicators o Key Process Indicators KPI), insieme agli obiettivi di prestazioni per ogni indicatore KPI.

Un dashboard, invece, è un contenitore per un gruppo di scorecard e visualizzazioni di report correlate, organizzate insieme in un sito di SharePoint. In altre parole, un dashboard contiene un insieme di elementi di altro tipo, come scorecard, report e filtri.

Un report è una presentazione dei dati, trasformati in informazioni formattate e organizzate in base a requisiti aziendali specifici. In un dashboard ci si aspetta in genere di trovare griglie e grafici analitici, report di Excel Services, report di tipo pagina Web e report di ProClarity Analytics. I report possono essere costituiti da semplici immagini statiche o da visualizzazioni altamente interattive dei dati.

Una scorecard consente di confrontare le prestazioni con gli obiettivi. Contiene, in genere, indicatori grafici che trasmettono visivamente il successo o l'insuccesso di un'organizzazione nel tentativo di raggiungere un particolare obiettivo.

Ogni web part mantiene una connessione alla relativa origine dati. Le web part possono funzionare indipendentemente l'una dall'altra oppure essere collegate, in modo che facendo clic SU una web part è possibile modificare gli elementi visualizzati in un'altra. Insieme, tali report consentono di ottenere un quadro chiaro e completo delle prestazioni attuali dell'organizzazione.

Un crawler è un software che analizza i contenuti di una rete, in un modo metodico e automatizzato, in genere per conto di un motore di ricerca. I crawler solitamente acquisiscono una copia testuale di tutti i documenti visitati e le inseriscono in un indice. Un uso estremamente comune dei crawler è nel Web. Sul Web, il crawler si basa su una lista di URL da visitare fornita dal motore di ricerca (il quale, inizialmente, si basa sugli indirizzi suggeriti dagli utenti o su una lista precompilata dai programmatori stessi). Durante l'analisi di un URL, identifica tutti gli hyperlink presenti nel documento e li aggiunge alla lista di URL da visitare. Il processo può essere concluso manualmente o dopo che un determinato numero di collegamenti è stato seguito.

Inoltre i crawler attivi su Internet hanno la facoltà di essere indirizzati da quanto indicato nel file "robots.txt" posto nella root del sito. All'interno di questo file, è possibile indicare quali pagine non dovrebbero essere analizzate. Il crawler ha la Facoltà di seguire i consigli (ma non l'obbligo).

Viene svolta, ora, una ricerca dei principali tool di Sentiment Analysis, di Brand Reputation Monitoring e Management, presenti sul mercato.

2.2 NM Incite

NM Incite, nato da un'iniziativa della Nielsen McKinsey Company, è un software di social media, di ricerca e consulenza per soluzioni di marketing che hanno, come obiettivo, quello di aiutare le aziende nel mondo della concorrenza contando sulla competitività del loro brand.

Rappresenta un supporto per tutte quelle aziende che vogliono sfruttare, in modo approfondito, le potenzialità dei social media al fine di ottenere le migliori performance, in ambito economico, attraverso la loro organizzazione. Attualmente è uno dei leader mondiali nel campo delle applicazioni di social media rivolti alla soluzione di problemi di marketing. NM Incite, infatti, opera su oltre 30 mercati tra cui gli Stati Uniti, Canada, Regno Unito, Germania, India, Giappone, Cina e Australia.

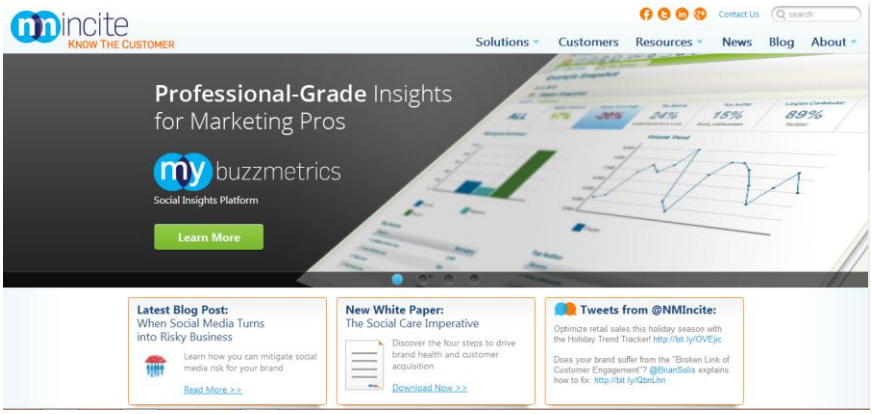


Figura 1 - NM Incite

Il Sentiment Customer Service (CSS) score è un servizio sviluppato da NM Incite. Offre, alle aziende, un punteggio che riflette il livello del sentimento positivo che i clienti nutrono o dimostrano nei confronti di prodotti e servizi offerti da un'azienda/società. Più alto è il punteggio, più alta risulta essere la soddisfazione dei clienti.

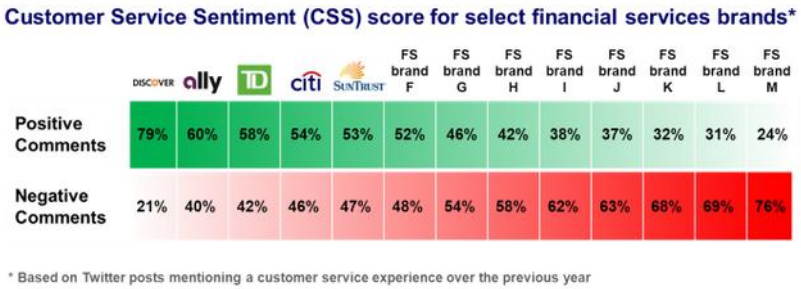


Figura 2 - Customer Service Sentiment (CSS) score

La socialsphere rende possibile conoscere in dettaglio i diversi segmenti della propria clientela. La sfida sta nel riuscire a domare questa immensa fonte di dati al fine di individuare i giudizi dei diversi segmenti di consumatori. Buzzmetrics (MBM) è una web-based Social Insights Platform progettata per il marketing Fortune 1000 e le loro agenzie per organizzare, segmentare e analizzare in modo affidabile, intuizioni globali specifiche del settore, in tempo reale.

Non importa quando o dove si svolgono le conversazioni sociali quali possono essere Facebook, Twitter, Ameblo e Weibo ; MBM analizza i dati provenienti da oltre 30 mercati in 15 lingue, applicando lo stesso approccio di Nielsen Measurement Science.

Con il rilascio di BuzzMetrics Exchange, NM Incite è entrato su un nuovo territorio: offre ai clienti una piattaforma di marketing per la gestione del proprio brand, fidelizzazione dei clienti e impegno diretto in tempo reale.

L'idea di aggiungere un prodotto come Exchange per la dashboard BuzzMetrics è nata quando è cambiato il modo in cui i clienti interagivano con le piattaforme di social media vautando la possibilità di passare dal semplice ascolto all'impegno diretto con i propri consumatori.

Il Marketing di brand, hanno bisogno di una piattaforma che vada al di la dell'ascolto passivo e che poissa gestire in modo attivo la salute del proprio brand con l'avvio di nuove conversazioni con i propri clienti sul qualsiasi piattaforma di social networking come Facebook, Twitter e YouTube.

Caratteristiche di Exchange a BuzzMetrics:

- Precisione. Sfruttando la stessa tecnologia proprietaria di raccolta dei dati utilizzata BuzzMetrics, è stato dimostrato che Exchange si basa su dati precisi e affidabili. BuzzMetrics Exchange, infatti include feed in tempo reale dai siti di social networking come Twitter, Facebook, YouTube, Digg, Reddit, e Flickr, così come oltre 160 milioni di blog e bacheche.

Figura 3 - NM Incite e BuzzMetrics

- Efficienza. La capacità di filtrare i messaggi consente un impatto diretto sui temi più rilevanti in riferimento al proprio brand e consente quindi, di agire di conseguenza. Ogni giorno

si ottengono migliaia di menzioni su brand, ciò risulta di notevole importanza per dare priorità alle conversazioni in base ai criteri come il sentimento, la lingua, la categoria e il numero di seguaci.

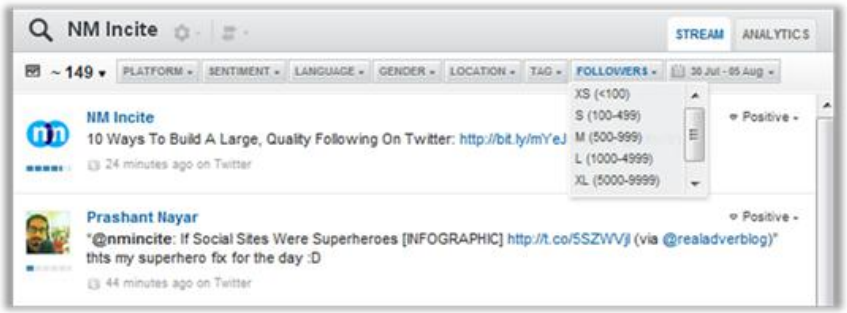


Figura 4 - NM Incite - Mensioni sui brand

- Tempestività. Con i feed in tempo reale, è possibile rispondere facilmente e rapidamente ai clienti attraverso i canali più appropriati finanziare iniziative prioritarie, oltre a programmare nuovi posti da pubblicare in un secondo momento;

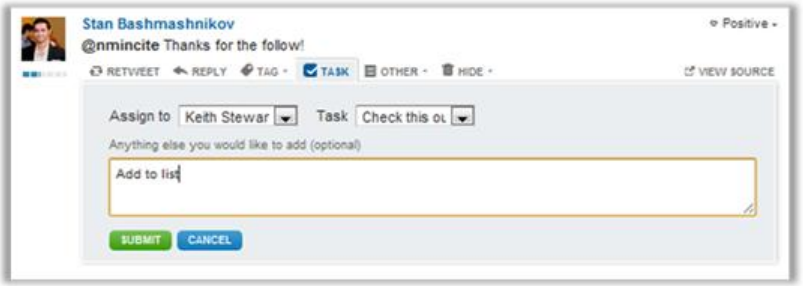


Figura 5 - NM Incite - Feed in tempo reale

2.3 Radian6

(Tipologia: reputation monitoring. - Tipo di licenza: a pagamento.)

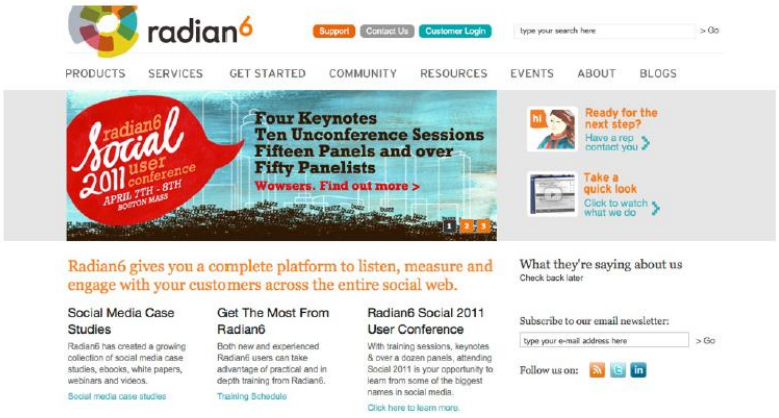


Figura 6 - Radian6

Radian6[156] è lo strumento prodotto dall'omonima azienda canadese rivolto alle aziende che vogliono conoscere la loro Web reputation. È in grado non solo di analizzare un ampio numero di fonti ma anche di svolgere analisi, sia in tempo reale che su serie storiche, anche di tipo comparativo fra diversi brand.

Tutte queste funzionalità non sono ottenute tramite un vero e proprio approccio semantico, ad esempio le ricerche sono svolte solo tramite keywords. Solo nell'ultimo anno Radian6 ha introdotto anche un modulo di sentiment analysis automatizzata e al momento supporta solo la lingua inglese. L'analisi è svolta a livello di frase, per poi aggregare un unico giudizio (positivo, negativo o neutro) sul post; se un particolare documento (o post) tocca più argomenti, il sentiment può essere determinato per ciascuno di essi separatamente.

[156] http://www.vincenzorisi.com/blog/?p=159 Last visited 25/02/2013

È un programma completo in quanto monitora diverse entità informatiche:

- blog;
- forum;
- video;
- micro blogging;
- mainstream;
- foto;

Per un singolo profilo accessibile ad un singolo utente, si devono pagare circa $600 al mese. Sono inclusi 10.000 risultati con un media di 30 giorni. Un anno di dati storici costa altri $600, (una tantum) per tutti i profili già aperti al momento dell'upgrade.

Radian6 risulta piuttosto intuitivo. In alto, sempre visibile, c'è la barra dei menu con 4 voci: Dashboard, Configuration, Help e Video Tutorials.

All'interno della sezione Configuration si trova la sezione Topic Profilies. Prima di poter ottenere dei risultati, è necessario impostare un profilo. I profili sono il core di Radian6. Per profilo si intende una campagna, ovvero un insieme di filtri e keywords pensati per un singolo cliente. All'interno della sezione Details, è possibile configurare il profilo attraverso i seguenti parametri:

- lingua (compreso l'Italiano);
- tipologia di media: video, immagini, forum, commenti, etc…;
- regione;
- filtro delle fonti: è possibile filtrare per sorgenti, ovvero liste URL da non prendere in considerazione nella restituzione dei risultati;

Per quanto riguarda la lingua italiana, Radian6 non è preciso. Per quanto riguarda, invece, la tipologia dei media, si possono riscontrare errori di classificazione.

Un altro punto importante da considerare è la sezione Estimated Monthly Volume. Infatti, se la media su 30 giorni del suo valore,

supera la soglia di 10.000, si viene inviati a restringere il campo di ricerca o a comprare il pacchetto per ulteriori 10.000 risultati.

Il secondo passaggio nel setup di un topic profile è ovviamente la creazione di query. Infatti, in Radian6, se non vi sono particolari esigenze, e se brand e prodotti hanno nomi non ambigui, si possono usare semplici espressioni di keywords. Se invece c'è la necessità di monitorare più prodotti dai nomi ambigui, più brand concorrenti o più tematiche inerenti ad un cliente, è opportuno settare gruppi di keywords. Radian6 permette di associare ciascun gruppo a una categoria tra "brands", "competitors" e "industry". In questo modo si semplificherà l'analisi dei risultati. Infine, è possibile configurare il mix di ingredienti che alimentano l'algoritmo che identifica i siti/blog o persone più influenti.

2.4 Social Mention

(Tipologia: search and analysis. - Tipo di licenza: gratuita.)

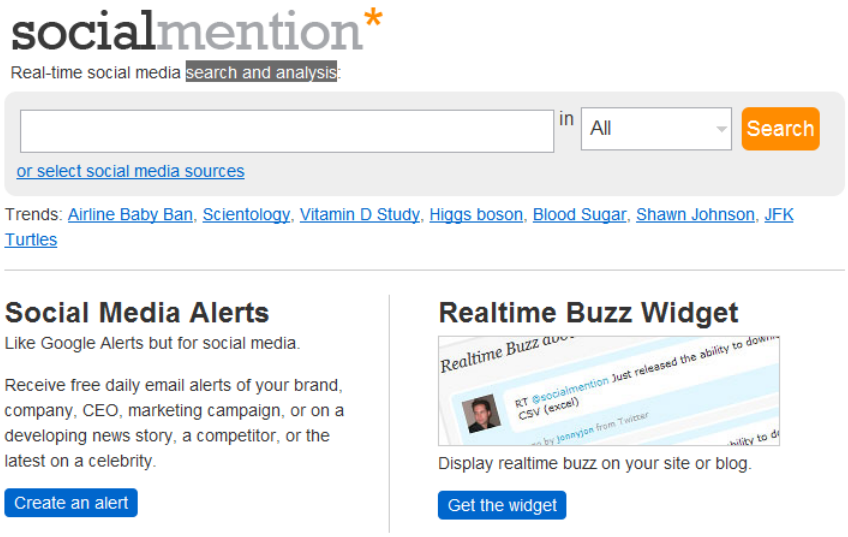


Figura 7 - SocialMention

SocialMantion[157] è un motore di ricerca dove il sociale è la base del suo database. È un semplice strumento per il monitoraggio dei social media. Aiuta a tenere sotto controllo chi sta facendo riferimento ad una specifica azienda, prodotto o materia. Esso aggrega i contenuti generati dagli utenti da diversi social network.

SocialMention si offre al consumatore finale, è facile da usare e soprattutto gratuito. Come altri strumenti per il monitoraggio dei social media, Social Mention offre sia una versione gratuita che un servizio a pagamento e aggiunge funzionalità extra.

Una volta terminata la ricerca, l'utente si trova davanti una pagina divisa in 3 sezioni:

1. al centro vengono disposti i risultati ottenuti dalla ricerca nel database, con piccole icone di riferimento poste vicino al titolo della pagina utile per identificarne la fonte;

[157] http://www.socialmention.com/ Last visited 25/02/2013

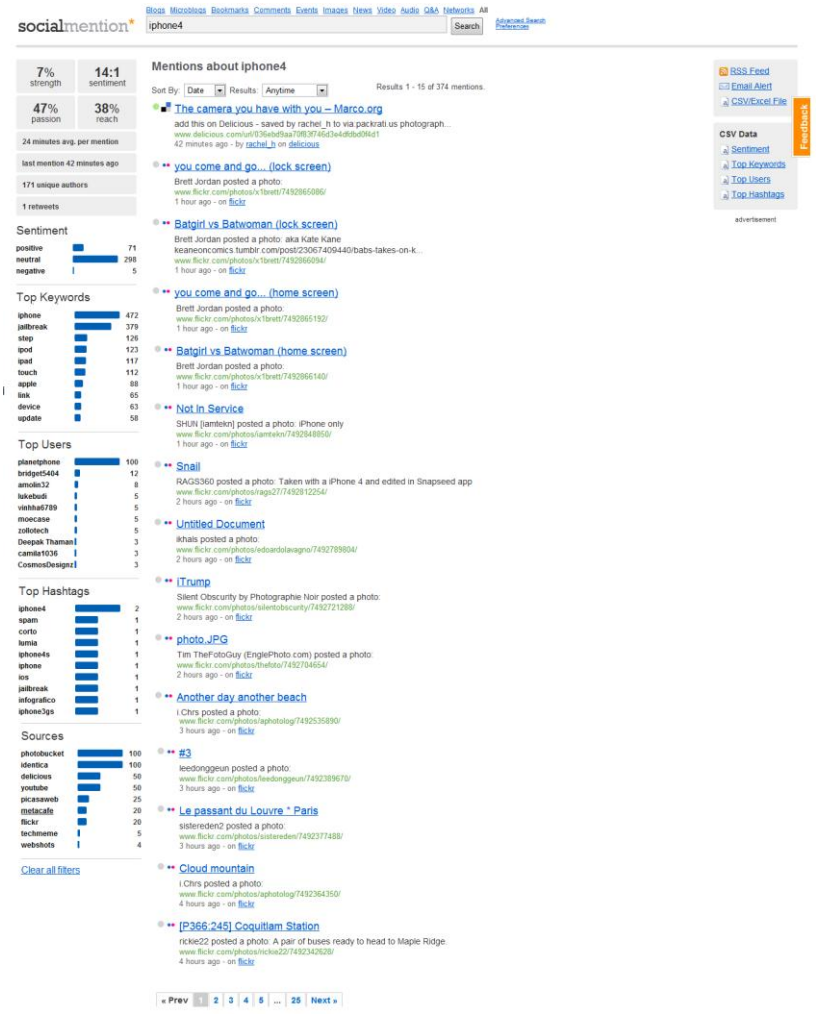


Figura 8 - Esempio ricerca

2. a sinistra si individuano strumenti di Social Mention Explained contenenti informazioni interessanti come:
    a. Sentiment: indica il sentimento positivo, negativo o neutro. Non è un risultato scientificamente preciso, ma da una idea completa di ciò che si pensa e si dice nei social media in riferimento alla query inserita;
    b. Top Utenti: indica gli utenti più attivi;
    c. Top Hash Tags: indica i tag più popolari in associazione con la query di ricerca;

3. a destra è possibile aggiungere gli RSS generati al proprio reader per essere avvisati in caso di nuovi risultati. Non solo, è anche possibile scaricarli nel formato CSV/Excel.

2.5 Alterian

(Tipologia: reputation monitoring. Tipo di licenza: a pagamento.)

Figura 9 - Alterian

Alterian[158] è una società multinazionale con sede principale a Bristol (UK) e alte 22 sedi distaccate in altri paesi del mondo. Alterian SM2 è la piattaforma di web monitoring basata su tecniche di business intelligence orientate ai social media. In particolare permette di eseguire monitoraggio automatico della Sntiment Analysis.

Grazie all'utilizzo di Alterian SM2 è possibile:

- cogliere ed analizzare facilmente i dati provenienti dai canali Social Media;
- controllare i Brand e i competitor;
- identificare community e influencer chiave;
- indirizzare correttamente i clienti con servizi ad hoc;
- condurre ricerche obbiettive;
- generare nuovi sales-leaders;

Contiene più di 9 miliardi di citazioni provenienti da conversazioni online. Tutte le sorgenti delle conversazioni sono tracciate e catalogate in modo da dare una panoramica completa di chi si sta parlando in rete, della categoria e del brand di interesse.

[158] http://www.alterian.com/ Last visited 25/02/2013

La piattaforma d'ascolto consente di accedere a tutti i dettagli, fino al singolo "clip", alla data di pubblicazione e alla posizione geografica dell'autore. Le conversazioni online vengono continuamente esaminate, indicizzate e aggiunte al database che è in continua crescita e sviluppo.

Ogni giorno vengono raccolti oltre 40 milioni di nuovi risultati, analizzando tutte le fonti disponibili online: l'analisi della serie storica può fornire utili informazioni sulla tipologia di eventi capaci di incrementare l'interesse su una categoria o un brand. La piattaforma d'ascolto, inoltre, classifica automaticamente tutti i risultati in base al sentimento (positivo, negativo, neutro) utilizzando parole chiave.

2.6 Alexa

(Tipologia: motore di ricerca con un servizio di web directory.- Tipo di licenza: gratuita.)

Figura 10 - Alexa

Alexa Internet Inc[159]. è un'azienda americana che si occupa di statistiche sul traffico internet. È anche definita come motore di ricerca con un servizio di web directory, vale a dire un elenco di siti suddivisi in maniera gerarchica. Una Web directory non è un motore di ricerca, ne tantomeno archivia i siti attraverso tag, ma, li presenta attraverso categorie e sottocategorie tematiche. È stata fondata nel 1996 da Brewster Kahle e da Bruce Gilliat. Inizialmente offriva una

[159] http://it.wikipedia.org/wiki/Alexa Last visited 25/02/2013

barra di navigazione che guidava gli utenti in base alle strutture del traffico della maggior parte degli utenti; in seguito ha fornito informazioni riguardanti il contenuto di ciascuno dei siti visitati: il proprietario, il numero di pagine componenti il sito, il numero di collegamenti che puntavano al sito e la frequenza degli aggiornamenti.

Inoltre fornisce, web crawls all'Internet Archive. Web crawls è un software che analizza i contenuti di una rete o di un database, in un modo metodico e automatizzato, in genere per conto di un motore di ricerca. Il crawler si basa su una lista di URL da visitare fornita dal motore di ricerca il quale, inizialmente, si basa sugli indirizzi suggeriti dagli utenti o su una lista precompilata dai programmatori stessi. Durante l'analisi di un URL, identifica tutti gli hyperlink presenti nel documento e li aggiunge alla lista di URL da visitare. Il processo può essere concluso manualmente o dopo che un determinato numero di collegamenti è stato seguito. In collaborazione con l'Internet Archive, gli ingegneri informatici di Alexa hanno ideato Internet Archive's Wavback Machine, una biblioteca digitale non-profit che ha lo scopo dichiarato di consentire un "accesso universale" alla conoscenza. Offre uno spazio digitale permanente per l'accesso a collezioni di materiale digitale che include, tra l'altro, siti web, audio, immagini in movimento ovvero video e libri. Nel 1999 l'azienda è stata acquistata per 250 milioni di dollari da Amazon.

Alexa classifica i siti basandosi sulle visite effettuate dagli utenti della Alexa Toolbar, per Internet Explorer, e dalle barre degli strumenti integrate in Mozilla e Netscape. Oltre alle estensioni della barra di stato, esistono svariate estensioni fornite da terze parti per Mozilla Firefox:

- SearchStatus (che mostra il PageRank di Google ed il TraffickRank di Alexa);
- About This Site Plug-in di Firefox, (che mostra il TraffickRank di Alexa;

Capire se la base di utenti di Alexa sia rappresentativa del comportamento degli utenti di internet è controversa. Infatti, se la base di utenti di Alexa è un campione statistico rappresentativo della popolazione degli utenti internet, il ranking di Alexa dovrebbe essere accurato. In realtà, poco si conosce riguardo le caratteristiche del

campione, e delle possibili distorsioni. Una fonte di distorsione dei dati è data dal fatto che l'installazione del software per il monitoraggio del traffico, ha luogo a discrezione degli utenti di Alexa. L'incidenza di queste scelte sulla modalità di elaborazione del ranking di Alexa non è conosciuta.

Una seconda preoccupazione riguarda la possibilità di manipolare il ranking di Alexa. Alcuni webmaster sostengono che possono migliorare in maniera significativa il ranking piuttosto basso di alcuni siti impostandoli come pagina iniziale, scambiando traffico web con altri webmaster e, chiedendo ai propri utenti di installare la barra degli strumenti di Alexa. Tali asserzioni sono basate su aneddoti e non è possibile verificarle mediante dati statistici o altre evidenze. Esisterebbero, inoltre, altri metodi grazie ai quali semplici siti web, che non ricevono molto traffico, hanno ottenuto un ranking alto su Alexa utilizzando un semplice script che, tuttavia, non indica il vero traffico del sito. Alexa non fornisce dati di traffico per i propri servizi.

2.7 SentiMetrix

(Tipologia: monitoraggio delle opinioni. - Tipo di licenza: a pagamento.)

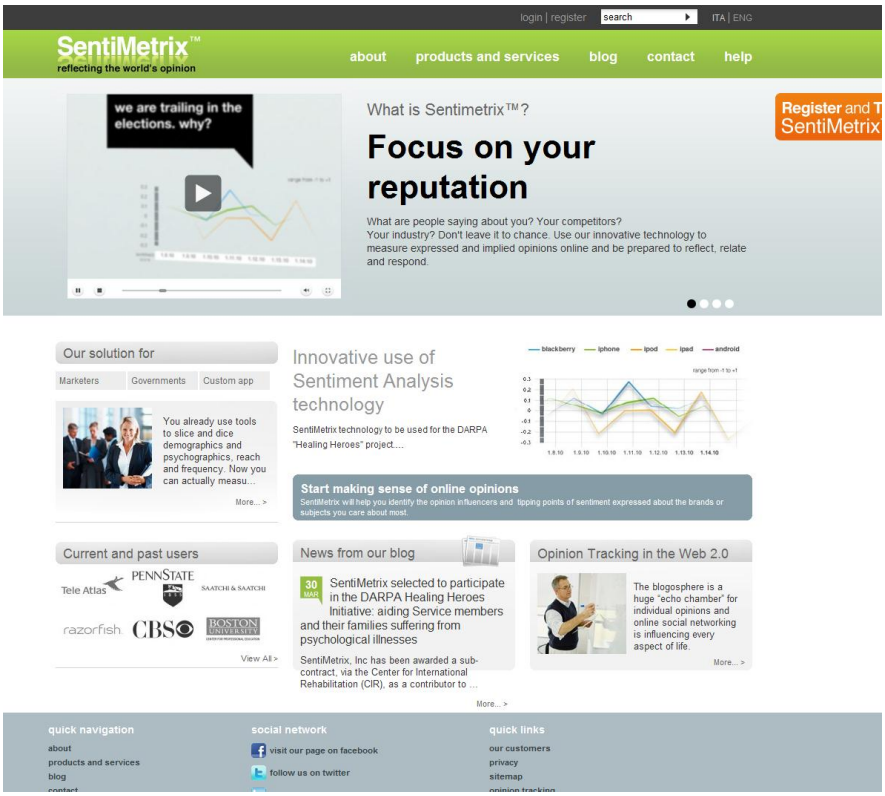


Figura 11 - SentiMetrix

SentiMetrix[160] stata fondata nel 2007 da un Team di professionisti di Internet e da ricercatori universitari. Il software è basato sulla premiata tecnologia OASYS, sviluppata presso l'Università del Maryland Institute for Advanced Computer Sciences (UMIACS).

L'azienda offre un quadro innovativo della tecnologia ai fini di un'accurata valutazione dei sentimenti e delle opinioni espresse nei social media (come news, blog, newsgroup) in tutto il mondo. I dati raccolti vengono combinati e su di essi viene effettuata un'accurata

160 http://www.sentimetrix.com/newsite/sentimetrix/products-and-services.html e http://www.sentimetrix.com/newsite/sentimetrix/ Last visited 25/02/2013

analisi in real-time, in modo tale da monitorare le opinioni espresse su determinati argomenti di interesse.

Il software SentiMetrix viene paragonato, in termini di precisione, al più avanzato calcolatore al mondo: l'uomo. È il frutto di una collaborazione di Natural Language Processing, analisi statistica e intelligenza artificiale, un mix che consente di determinare anche la più piccola variazione nelle opinioni. L'informazione rappresenta il bene più prezioso della nostra società. SentiMetrix è nato con l'obiettivo di conoscere il sentimento delle persone, positivo o negativo che sia, nell'intera blogosfera, ma soprattutto cerca anche di capire il perché, ad un brand, viene associata una buona o cattiva reputazione.

Fornisce un punteggio preciso sul sentimento grazie ad un'analisi dell'opinione di altissima qualità in ben nove lingue. È possibile personalizzare la ricerca modificando le fonti dati, la provenienza, il tipo oppure utilizzando ulteriori lingue.

SentiMetrix offre la famiglia di servizi SentiGrade per il tracciamento delle opinioni, su qualsiasi argomento, espresse sui mezzi di informazione online, come blog, forum, recensioni degli utenti e anche nei database forniti dai clienti. La tecnologia SentiGrade, può essere utilizzata in due modi:

- SentiGrade Dashboard: un'interfaccia web-based semplice da usare che permette di tracciare fino a 10 argomenti contemporaneamente, confrontare i cambiamenti di opinione nel tempo in base all'argomento, alla lingua e alla fonte dell'informazione. I dati estratti dalla SentiGrade Dashboard possono essere esportati in forma grafica o in file CSV;
- SentiGrade Data Service: una suite di servizi web per i clienti che già possiedono le proprie soluzioni di visualizzazione e desiderano aggiungere ai propri dati le informazioni sulle opinioni. I SentiGrade Data Service consentono anche di analizzare i database di proprietà dei clienti;

SentiGrade fornisce l'accesso ai dati raccolti da una moltitudine di sorgenti da tutto il mondo, inclusi più di 50.000 news/media outlets ed un milione di top blog. Fonti aggiuntive personalizzate possono essere

aggiunte su richiesta. I servizi SentiGrade forniscono strumenti all'avanguardia con la loro combinazione unica di funzionalità:

- un'ampia lista di sorgenti di informazioni;
- una scelta contigua per misurare le opinioni;
- misurazione dell'intensità dell'opinione in base al modo con il quale le persone la esprimono;
- algoritmi proprietari, in attesa di brevetto, per il calcolo real-time del "sentiment score";
- comprovata accuratezza della misurazione del sentiment;
- supporto multilingue;
- architettura efficiente e scalabile;

La suite di prodotti di SentiMetrix offre una raccolta di informazioni su un'intera nazione, o comunità seguendo in real-time l'evoluzione delle informazioni multilingue sotto forma di social media.

Caratteristiche del prodotto:

- costoso;
- i risultati richiedono settimane o mesi dopo la valutazione della domanda;
- ogni volta che l'analista ha bisogno di nuove domande bisogna rispettare i medesimi tempi per la nuova risposta;
- generalmente è previsto un elenco di domande frequenti ottenute da sondaggi;
- economico: il servizio di base SentiGrade ha un costo pari a $ 399 al mese.
- real-time. L'analista accede al sito web SentiGrade e ottiene le risposte di sentimento in pochi secondi dalla richiesta della query.
- una volta ottenuto l'account , l'utente può effettuare qualsiasi tipo di query, infatti SentiMetrix tiene traccia di qualsiasi argomento disponibile, su milioni di fonti in tutto il mondo;
- gli analisti che usano SentiGrade otterranno, sicuramente risposte soddisfacenti;

Questi attributi rendono SentiMetrix un ideale partner per le agenzie governative statunitensi e dei loro subappaltatori di fiducia.

Descrizione del servizio:

- Continuous Scoring SentiMetrix's: il punteggio che SentiMetrix attribuisce ad un sentimento avviene in real-time ed è facile da interpretare: -1 corrisponde ad un sentimento negativo, 1 corrisponde ad un sentimento positivo;
- Internet Scale Breadth After: dopo aver ricevuto istruzioni da parte di un analista, SentiMetrix identifica gli articoli e le notizie correlati ad un determinato argomento, li mette a confronto con l'universo conosciuto e determina l'intensità del sentimento espresso sull'argomento della query calcolandone il punteggio complessivo;
- Time Series Analysis Analysts: è in grado di monitorare gli oggetti nel tempo e di agire di conseguenza se si riscontra un sentiment inaspettato. Quando l'analista viene avvertito da tali cambiamenti, deve semplicemente individuare il documento responsabile del cambiamento del parere e presentarlo nella lingua e contesto di origine. Ciò permette all'analista di capire non solo il punteggio del sentimento, ma anche la ragione che c'è dietro;
- Source based Analysis Analysts: SentiGrade può essere utilizzato per conoscere come, il sentimento su di un argomento, cambia al variare delle fonti. Per esempio si può quantificare quali notizie sono più negative o positive di altre;

2.8 Tweetfeel

(Tipologia: search and analysis. - Tipo di licenza: gratuito.)

Figura 12 - Tweetfeel

TweetFeel è un tool di Sentiment Analysis specializzato sulla fonte Twitter, permettendo di stimare la polarità e i tweet in modo automatico, senza bisogno di training. Fa una ricerca in base a termini che ha nel database ai quali assegna valore Vpositivo o negativo.

Il software consente anche di riconoscere i trend; tuttavia non si limita a mostrare delle word cloud ma, tenta di ricostruire la storia che si nasconde dietro dei burst. Tramite l'interfaccia semplice e flessibile, è possibile confrontare più brand. La ricerca del topic avviene per mezzo di una ricerca libera tramite keyword ma con funzionalità avanzate, come la possibilità di escludere altre keywords dai risultati.

2.9 KISSmetrics

(Tipologia: search and analysis. Tipo di licenza: a pagamento.)

Qualsiasi azienda, possiede diverse categorie di utenze: users e customers o entrambe; KISSmetrics è un tool con caratteristiche molto differenti da quelli già citati; serve come supporto alla gestione e da un aiuto alle aziende che hanno la necessità di gestire e soddisfare al meglio, le esigenze della propria clientela.

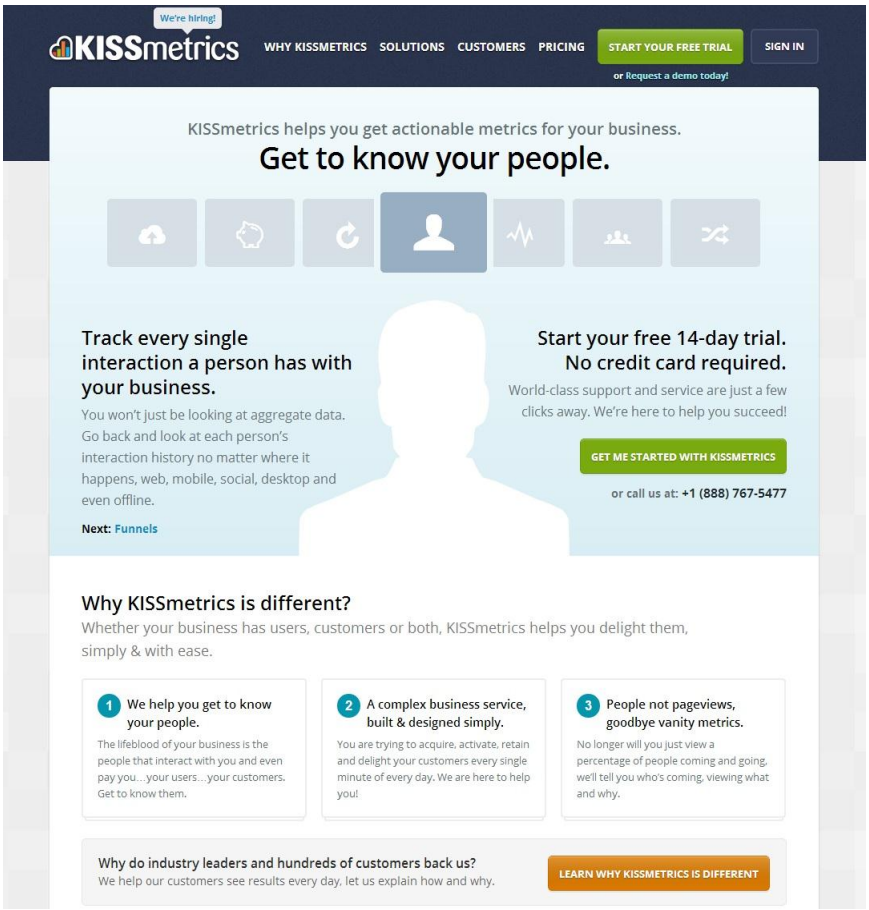


Figura 13 - KISSMetrics

È possibile tenere traccia di ogni singola interazione tra una persona e un'azienda. Il suo obiettivo non riguarda solo la ricerca di nuove aggregazioni di dati, ma è sempre possibile tenere traccia della "storia passata" di ogni cliente a prescindere dal tipo di tecnologia che

utilizza: web, mobile, social, desktop e anche offline. È un servizio di business complesso, costruito e progettato in un modo estremamente semplice che cerca di . Si sta tentando di acquisire, attivare, trattenere e motivare clienti ogni singolo minuto. KISSmetrics ha ricevuto attenzione da Techcrunch, Forbes, Mashable, Fast Company, The Wall Street Journal, The Atlantic ed è regolarmente nominato come una delle migliori soluzioni di analisi sul mercato.

2.10 Technorati

(Tipologia: search and analysis. - Tipo di licenza: gratuita.)

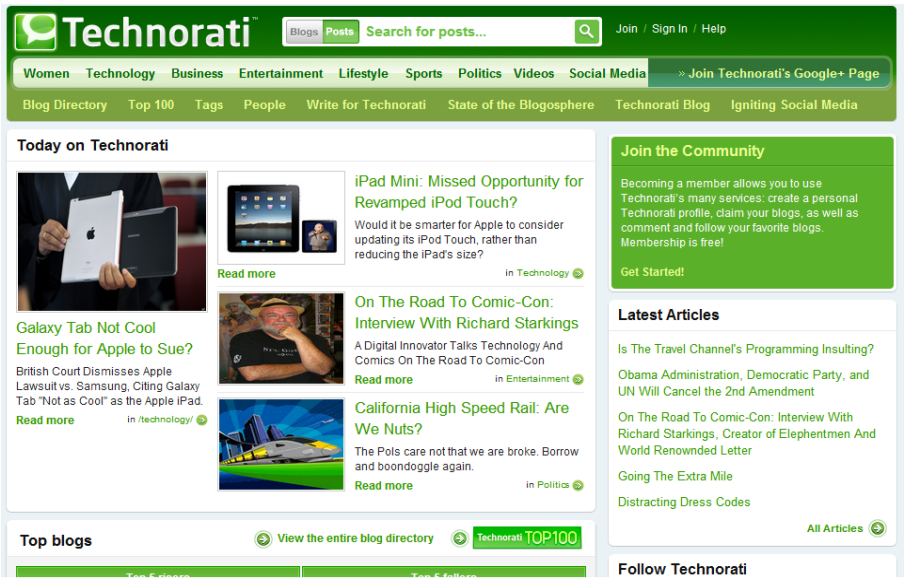


Figura 14 - Technorati

Technorati[161] è un motore di ricerca dedicato alla blogosfera ovvero al mondo dei blog. Dal dicembre 2005 technorati indicizza più di 20 milioni di blog. È stato fondato da Dave Sifry e la sede è situata a San Francisco (California, USA).

Il termine technorati è una crasi, cioè un termine nato dall'unione di due parole: Technological e literati, (traducibile in italiano come intellettuali tecnologici). Il rank di Technorati, conosciuto anche come Autority, è un valore di grande importanza per ogni blog, perchè rappresenta l'indice di gradimento del pubblico del web nei suoi confronti, e da quanto si evince, è tenuto in grande considerazione dai motori di ricerca (Google in testa).

Dopo avere avviato un Blog, il passo successivo da fare è quello di renderlo visibile. Ci vuole un po' di tempo ed il suo successo dipende da molti fattori, primo tra tutti la qualità del suo contenuto e la sua capacità di fare engagement.

[161] http://www.gallito.eu/2009/01/16/technorati-cose/ Last visited 25/02/2013

Un primo step per presentare un nuovo blog al mondo del web è quello di iscriverlo e farlo indicizzare da Technorati.

2.11 BlogScope

(Tipologia: Links, tendenze, conversazioni della blogosfera. - Tipo di licenza: gratuita.)

Figura 15 - BlogScope

Blog Scope è un sistema per l'analisi e la visualizzazione online di stream di dati testuali che compongono la blogosfera. È frutto di un progetto di ricerca in corso presso l'Università di Toronto. Oggi tiene traccia di oltre 725 milioni di posts su blog.

Tramite la raccolta, il monitoraggio e l'analisi delle informazioni sui blog, BlogScope è in grado di fornire informazioni chiave sull' "opinione pubblica" e su una varietà di temi, quali prodotti, opinioni politiche o di intrattenimento.

BlogScope è un sistema per l'analisi on-line e la scoperta dei flussi di informazioni di testo. Un flusso di testo in questo caso è definita come un insieme ordinato temporalmente di documenti di testo.

BlogScope è scritto in Java e funziona su quattro computer Sun V40z server machine with RedHat Linux AS4. I principali componenti sono:

- un multi-threaded crawler con analizzatore di spam;
- indicizzazione e ricerca del modulo;
- raccolta e panoramica delle statistiche di accesso;
- curva del generatore popolarità;
- statistics collection;
- access framework;
- natural language processor;
- interfaccia web.

La figura sottostante riassume l'architettura del sistema ad alto livello.

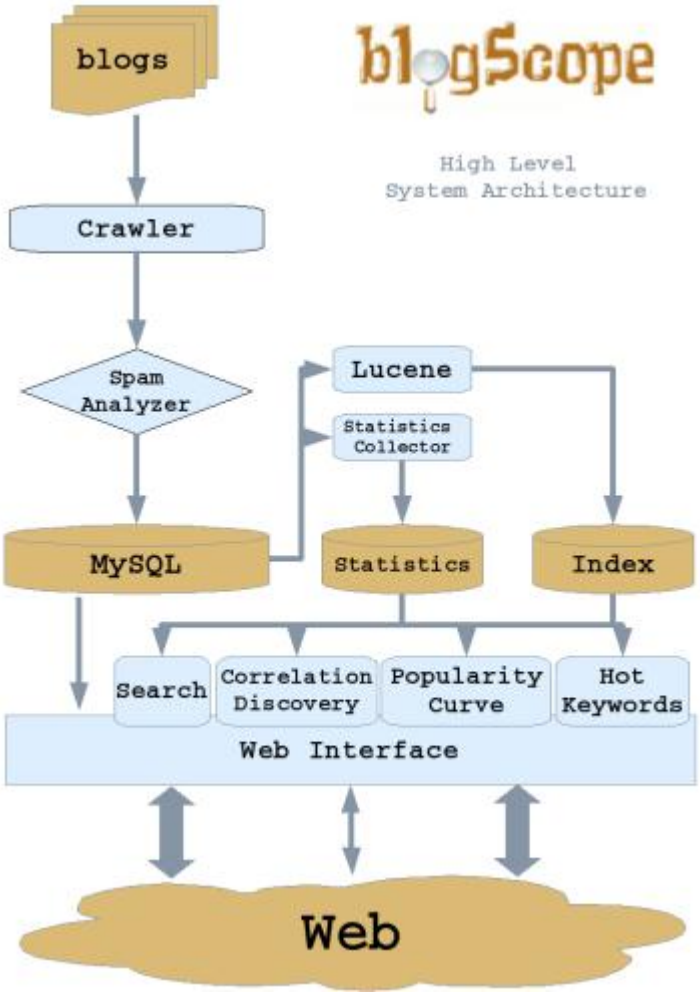


Figura 16 - Architettura del sistema ad alto livello[162]

BlogScope si propone di analizzare la blogsfera al fine di identificare che cosa è interessante, quando è stato interessante, perché è interessante e dove è interessante. Per rispondere alla prima domanda, BlogScope mette a disposizione sia un tag cloud che visualizza le hot keyword del giorno, sia una tradizionale funzionalità di query basata su keyword. Sui risultati delle query, BlogScope non solo costruisce

[162] http://www.blogscope.net/about/index.html Last accessed 25/02/2013

una classifica dei post rispetto ad una metrica di rilevanza, ma svolge anche un'analisi temporale per costruire una curva di popolarità delle keyword ricercate, evidenziando la presenza di eventuali burst. Il terzo step dell'analisi riguarda il perché una keyword è interessante.

Per aiutare l'utente nella comprensione di questo tipo di informazione, viene mostrato l'elenco di parole che sono maggiormente correlate alle keywords della ricerca nell'intervallo temporale di riferimento e, in particolar modo, durante i burst; queste parole dovrebbero aiutare nella disambiguazione del termine in quel particolare frangente. Infine per il problema del "dove la keyword è interessante", BlogScope sfrutta le informazioni di geolocalizzazione che caratterizzano i posts, al fine di identificare da quali regioni del globo provengono i risultati proposti dal motore di ricerca. Riassumendo, le principali funzionalità offerte sono:

- analisi spazio-temporale dei post;
- navigazione flessibile della blogsfera tramite burst;
- analisi della correlazione tra keywords;
- burst synopsis, ovvero l'identificazione delle parole correlate alle keyword di ricerca e che presentano un burst nello stesso intervallo dove lo manifestano le keyword.

2.12 Lithium

(Tipologia: search-specific mentions and sentiment in social media. - Tipo di licenza: a pagamento.)

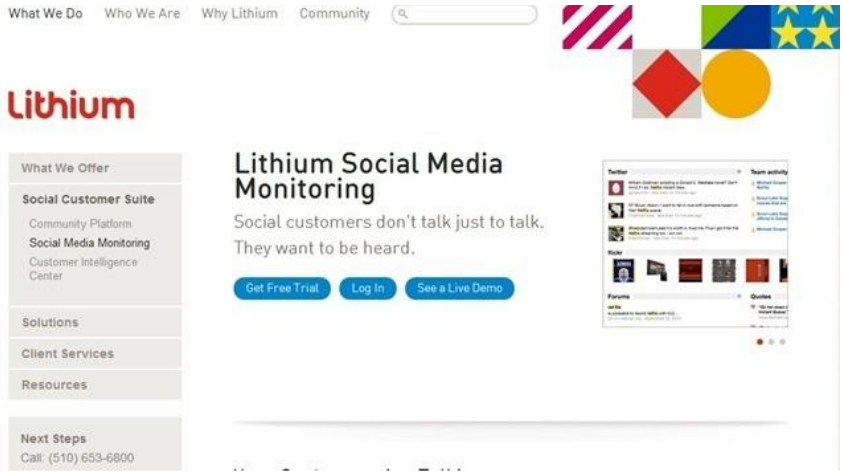


Figura 17 - Lithium

Lithium Social Media Monitoring (noto come Labs Scout) riporta importanti informazioni sulle conversazioni dei clienti da Facebook, Twitter e milioni di altre fonti. E li trasforma in modo tale da facilitarne la lettura tramite grafici.

Lithium aiuta le aziende a capire le passioni dei propri clienti, per aumentare le vendite, ridurre i costi di servizio e accelerare l'innovazione.

2.13 Global Terrorism Database Explorer

(Tipologia: open-source database information. - Tipo di licenza: gratuita.)

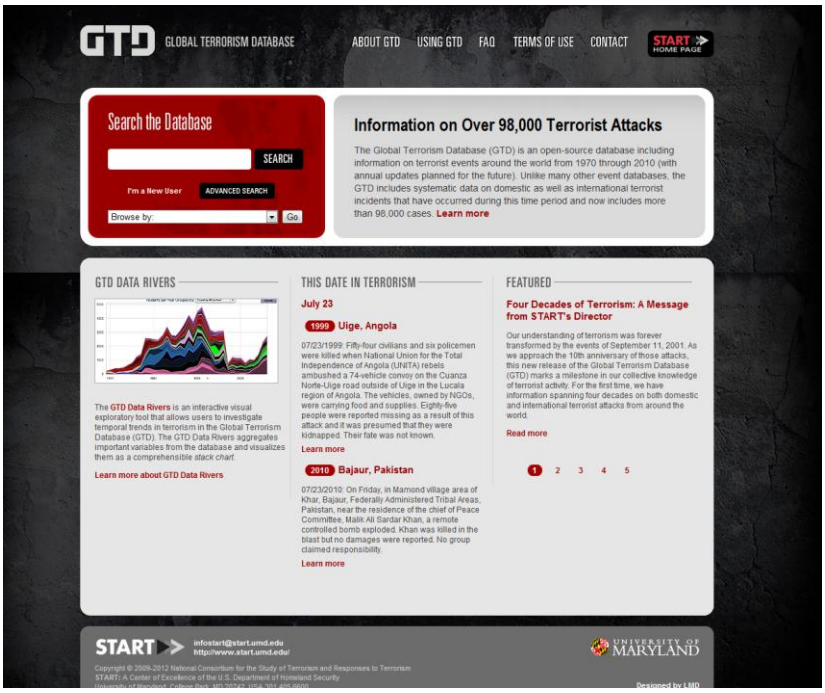


Figura 18 - GTDexplorer

Global Terrorism Database Explorer[163] (GTD) è un database open-source che raccoglie informazioni sugli eventi terroristici nel mondo dal 1970 al 2010 (con ulteriori aggiornamenti annuali previsti per il futuro). A differenza di molti altri database degli eventi, GTD include dati sistematici sugli incidenti terroristici transnazionali e internazionali, che si sono verificati durante questo periodo di tempo ed ora include più di 98.000 casi. Per ogni incidente GTD, sono disponibili informazioni sulla data e sul luogo dell'incidente, le armi utilizzate e la natura del bersaglio, il numero delle vittime e, laddove sia identificabile, il gruppo o persona responsabile.
Il Consorzio Nazionale per lo Studio del terrorismo e Responses to Terrorism (START), rende la GTD disponibile tramite questa interfaccia on-line nel tentativo di aumentare la comprensione della

[163] http://www.start.umd.edu/gtd/ Last accessed 25/02/2013

violenza terroristica, in modo che possa essere più facilmente studiati e sconfitta.

Caratteristiche del GTD:

- contiene informazioni su oltre 98.000 attacchi terroristici;
- include informazioni su più di 43.000 bombardamenti, 14.000 omicidi, 4.700 sequestri dal 1970;
- include informazioni su almeno 45 variabili per ciascun caso, con gli avvenimenti più recenti, comprese le informazioni su più di 120 variabili;
- supervizionata da un comitato consultivo di 12 esperti in ricerca in ricerca di terrorismo;
- per raccoglire i dati relativi agli attacchi sono state riviste oltre 3.500.000 articoli e 25.000 fonti;

Il tool considera i vari incidenti terroristici e per ciascuno misura lo stream di dati, rappresentando i vari stream in pila, in modo da mettere in evidenza sia l'andamento individuale sia quello cumulato (figura 18). L'ampiezza del singolo stream rappresenta variabili come il numero di attacchi e il numero di armi utilizzate. Lo stream in sé può, invece, rappresentare la nazione colpita, l'organizzazione terroristica artefice dell'attacco, il tipo di target, il tipo di attacco, etc… Gli stream possono essere tanti, quindi è data la possibilità all'utente di filtrarli tramite una query libera.

2.14 General Sentiment

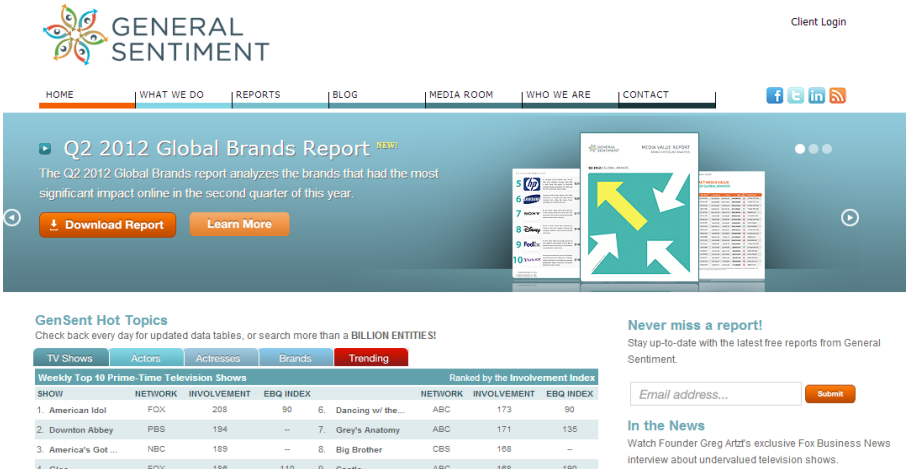


Figura 19 - General Sentiment

Attraverso Internet, news, blog, tweet e molti altri media online, milioni di persone si scambiano diverse tipologie di opinioni. Gli argomenti possono riguardare il lancio di un prodotto, campagne pubblicitarie, eventi di pubbliche relazioni, rapporti sugli utili, l'esperienza che un consumatore ha avuto con un prodotto o un servizio e molti altri fattori scatenanti. General Sentiment, analizza oltre 30 milioni di fonti: , "ascoltando" in tempo reale, le opinioni espresse su brands, prodotti, politica, celebrità, aziende etc..

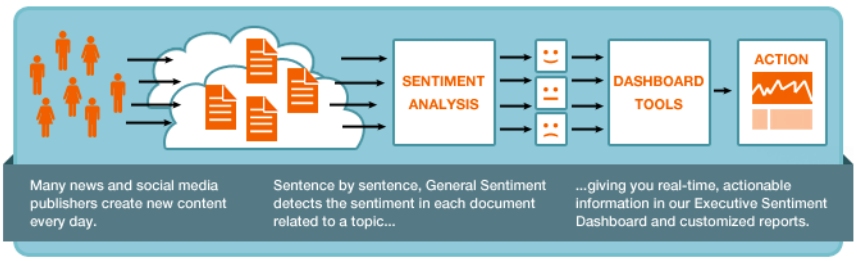


Figura 20 - General Sentiment - Sentiment Analysis

La Media Measurement Dashboard utilizza strumenti innovativi che presentano, un'istantanea sull' opinione pubblica in merito a marchi o

persone importanti in riferimento alla propria azienda. Si tratta di un utile quadro a livello decisionale.

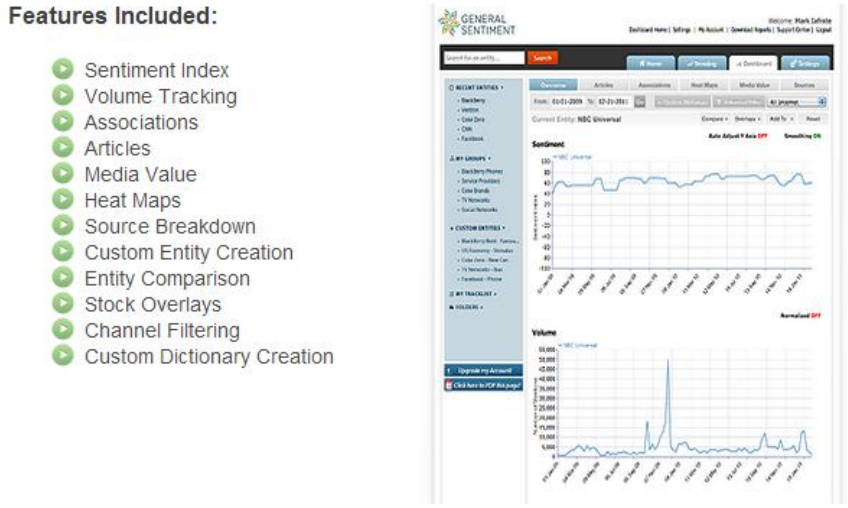


Figura 20 - General Sentiment - Features

È possibile ottenere reporting personalizzati per conoscere l'andamento (positivo o negati) della campagna pubblicitaria, valutare un possibile aumento di volume del proprio brand, le nuove tendenze di mercato etc.

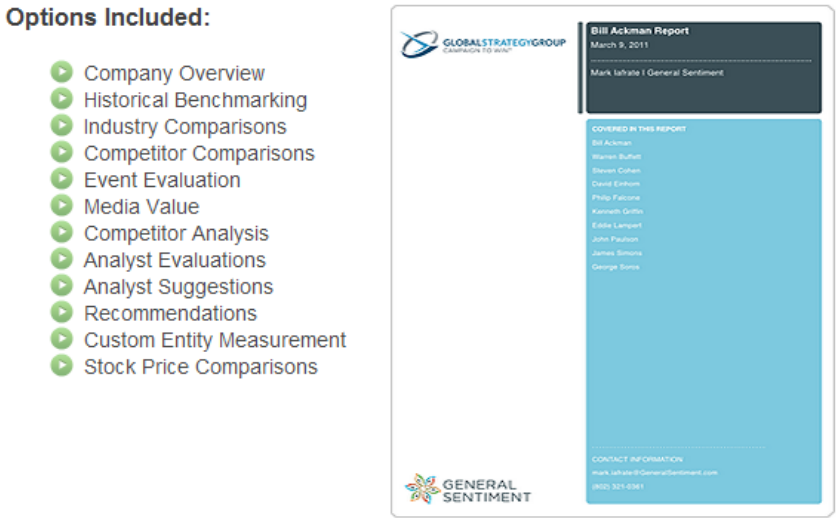


Figura 21 - General Sentiment - Options

GeneralSentiment offre diversi tipi di subscriptions per i diversi industry reports, tra cui tra prime-time television reports, the Top 20 Global Brands series ,e Fast Food Industry Report. A breve sarà disponibile anche un Reposrt per il settore dell'industria automobilistica.

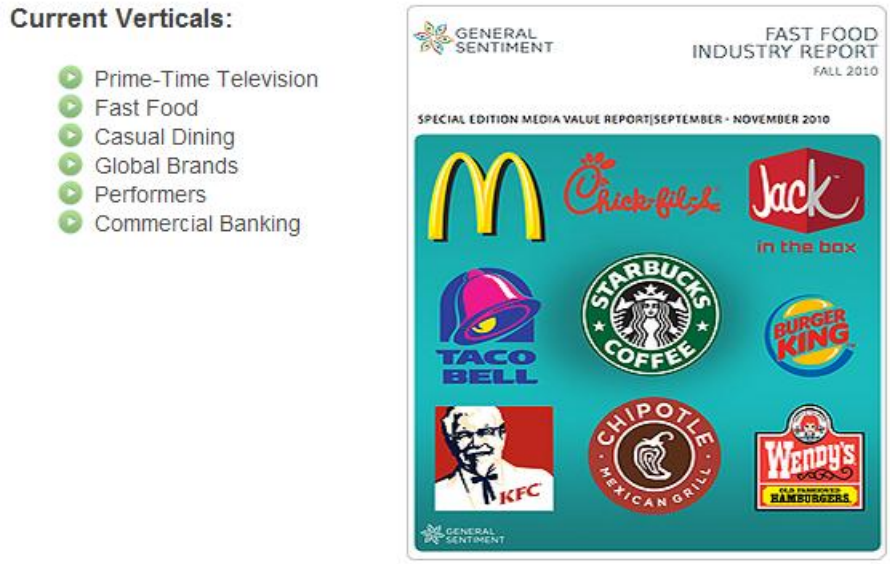


Figura 22 - General Sentiment - Verticals

L'API di GeneralSentiment consente l'utilizzo di media measurement indicators , in modo tale da soddisfare le necessità della propria azienda.

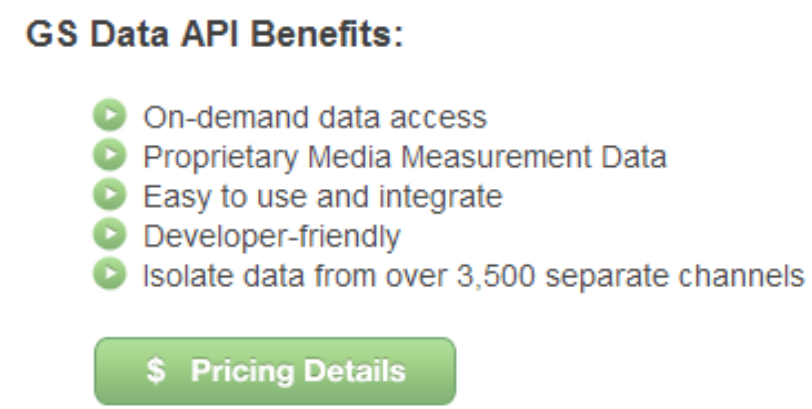


Figura 23 - General Sentiment - API

2.15 Viralheat

(Tipologia: monitoraggio social. - Tipo di licenza: a pagamento.)

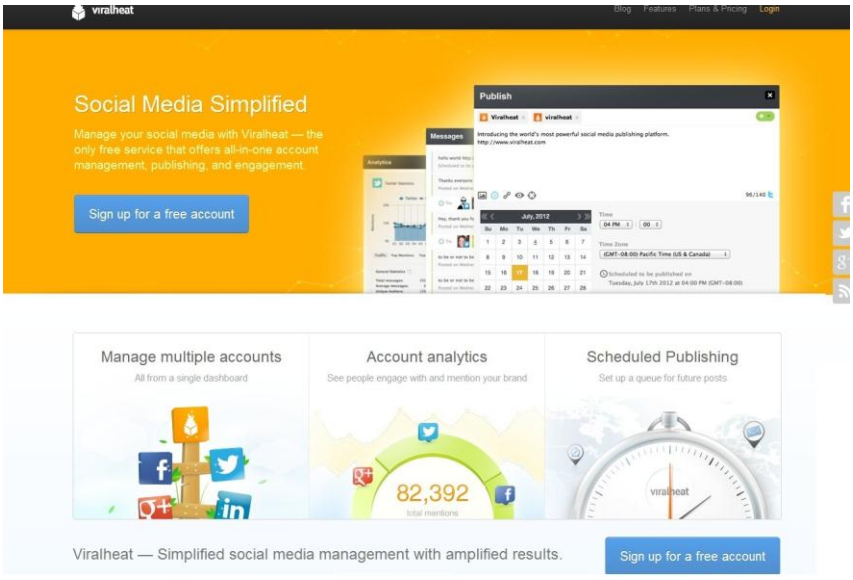


Figura 24 - Viralheat

Viralheat permette di tenere sotto controllo le attività di social media e social marketing. Fornisce un'interfaccia grafica molto semplice ma professionale, con numerose utilità che permettono di tenere sotto controllo i flussi di movimento delle attività legate ai social networks.

Impostando le proprie esigenze, legate al marchio, al prodotto o all'interesse, il sito monitorizza automaticamente i siti social per offrire un report aggiornato comprensivo delle volte in cui la parola chiave legata viene citata.
Si possono creare profili riguardanti un singolo nome o una determinata azienda, perlustrando circa 30 siti video, il Web in generale e Twitter.
Numerose sono le funzioni messe a disposizione: ordinare i risultati di una data settimana o di un giorno specifico, il paese dove è più vivace la discussione su un certo tema, il numero medio di citazioni per giorno, etc... Per quanto riguarda Twitter, è possibile monitorare:

- il numero totale di citazioni per settimana e per giorno;
- l'utente più attivo che ha parlato di un certo marchio;
- il linguaggio più comune dei messaggi; la percentuale di retweets riguardanti un certo nome;
- il giorno più attivo della settimana per la segnalazione di una certa marca, etc…;

Tutte queste misurazioni possono essere fatte naturalmente anche per le piattaforme video. Per esempio, sapere qual è il video più popolare, oppure registrarne la caduta di interesse, il numero medio di download al giorno, e quante volte è stato visto. Si può, inoltre, inviare direttamente (da Viralheat) una e-mail o un messaggino Twitter riguardante un certo video.

La peculiarità di Viralheat consiste nei prezzi:

- 10$ al mese per 10 profili;
- 40$ al mese per 50 profili;

Inoltre, prima di acquistare un qualsiasi piano, è possibile provarlo gratuitamente per 7 giorni.

2.16 SentiStrength

SentiStrength è un lessico basato su un classificatore che utilizza informazioni e regole sulla linguistica per rilevare la forza del sentimento espresso, in brevi testi informali in lingua inglese.

SentiStrength riporta due punti di forza del sentiment:

- -1 (not negative) to -5 (extremely negative)
- 1 (not positive) to 5 (extremely positive)

Questi punteggi variano da 1 a 5 per quanto riguarda l'analisi de sentiment positivo e da -1 a -5 per l'analisi del sentiment negativo.

Al valore 1 non corrisponde nessun sentimento mentre al 5 corrisponde il massimo del sentimento esprimibile. Risulta essere abbastanza preciso se la ricerca si limita a brevi testi di social web in lingua inglese, ad eccezione dei testi politici.

Ad esempio, un testo con un punteggio di 3, 5 corrisponde un sentiment moderatamente positivo e un forte sentiment negativo.

È possibile scegliete anche diversi tipi di risultati:

- binari (positivo / negativo);
- ternari (positivo / negativo / neutro)
- o su singola scala (da -4 a +4).

SentiStrength è stato originariamente sviluppato per la lingua inglese, ma può essere configurato per altre lingue e contesti modificando i file di input.

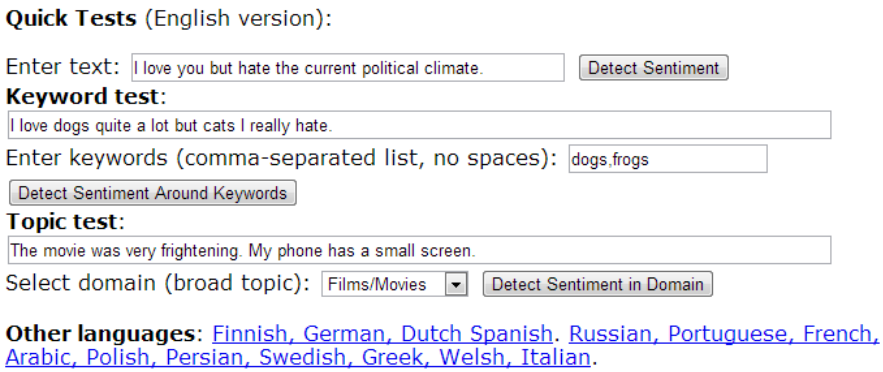


Figura 26 - SentiStrength

La versione gratuita funziona solo sotto Windows e viene fornita senza alcuna responsabilità o garanzia per qualsiasi uso. Per quanto riguarda la licenza commerciale, SentiStrength è disponibile per £ 1000. La versione Java di SentiStrength viene normalmente utilizzata a fini commerciali.

Per i testi che contengono sentimenti sia positivi che negativi, vengono utilizzate due scale di valutazione. L'obiettivo è quello di rilevare il sentiment espresso piuttosto che la polarità complessiva (Thelwall, Buckley et al., 2010). Di seguito viene riportato un elenco delle caratteristiche principali di SentiStrength (Thelwall, Buckley et al., 2010). Quelli contrassegnati con ^ sono stati sostituiti nella versione 2.

- A sentiment word list with human polarity and strength judgements^.
  La parola "miss" è un caso particolare che possiede una forza positiva e negativa pari a 2. È spesso usato per esprimere sia un senso di tristezza che di affetto;
- Spelling correction algorithm: cancella le lettere che si ripetono in modo frequente rispetto al normale. Queste vengono corrette e se non sono presenti nel dizionario (in questo caso parliamo della lingua inglese), vengono aggiunte come nuove. (esempio, hellp -> help).
- A booster word list: utilizzato per rafforzare o indebolire l'emozione espresso nelle parole.

- An idiom list^: viene utilizzata per identificare il sentimento di alcune frasi comuni.
- A negating word list^: viene utilizzaao per invertire termini consecutivi (saltando le parole che si ripetono).
- At least two repeated letters: aggiunge, un punto in più a tutte quelle parole che esprimono un sentiment maggiore ad esempio: la parola haaaappy risulta essere molto più positiva della parola happy. Invece, alle parole neutrali viene assegnata un punteggio positivo di 2.
- An emoticon list with polarities: questa lista viene utilizzata per identificare ulteriori sentiment.
- Sentences with exclamation marks: frasi che esprimono un minimo di sentiment (meno di 2);
- Repeated punctuation: la presenza di uno o più punti esclamatvi, aumenta la forza del sentimento.
- Negative sentiment is ignored in questions^.

Ci sono due versioni di SentiStrength: supervised e unsupervised. La versione supervisionata ha il seguente componente aggiuntivo e cioè un algoritmo di training in grado di ottimizzare i punti di forza del sentimento di una parola. L'algoritmo controlla la forza del sentimento di ogni termine per vedere se un aumento o una diminuzione di 1 aumenterebbe la precisione della classificazione su un insieme di testi classificati manualmente. È un algoritmo iterativo che si ripete fino a quando non sono state controllate tutte le parole senza necessità di apportare alcuna modifica.

La versione originale di SentiStrength è stata testata solo su sui brevi messaggi di amicizia informali di MySpace SNS. La nuova versione è stata sviluppata per una varietà di testi. Il cambiamento principale risiede in una significativa estensione del lessico dei termini negativi.

SentiStrength è stato testato sui seguenti insiemi di dati:

- BBC Forum posts (notizie pubbliche): discussioni su vari argomenti come notizie nazionali e mondiali su argomenti quali politica e religione;
- Digg.com posts (commenti pubblici);
- MySpace comments;
- Runners World forum posts;

- Twitter posts;
- YouTube comments;
- Etc;

I risultati dei test dimostrano alcune limitazioni importanti di SentiStrength:

1. I risultati degli esperimenti condotti non sono esaustivi in quanto, in alcuni ambienti di social web environment, SentiStrength non funziona. Questo perché si basa su un linguaggio standard che non prevede l'uso di ironia e sarcasmo, tipico dei blog.
2. non garantisce un utilizzo diretto dei termini affettivi, molte volte, l'uso di tali termini risulta essere ambiguo come nel caso della parola, "shocking" che da solo assume un significato diverso in base al contesto di riferimento come nella frase "colour shocking pink".

SentiStrength è un algoritmo robusto per il rilevamento della forza del sentiment sui social web data. Ma, in alcuni ambienti SentiStrength non funziona, così come alcune tecniche di machine learning in particolare quelle di logistic regression.

In conclusione, SentiStrength sembra essere adatto per il rilevamento della forza del sentimento sui social web, è consigliato per tutte quelle applicazioni in cui è importante lo sfruttamento dei termini affettivi diretti.

2.17 SentiRate

SentiRate[164] effettua un'analisi attendibile sul sentimento e permette l'estrazione di emozioni, da qualsiasi contenuto di testo in digitale, come e-mail, moduli di feedback e documenti in generale. Garantisce un ordine di priorità di comunicazione.

È possibile utilizzare una versione gratuita di SentiRate per poterlo testare personalmente.

Ecco un elenco di possibili settori in cui risulta utile SentiRate:

[164] http://sentirate.com Last accessed 25/02/2013

- in organizzazioni che ricevono, ogni giorno, un grande quantitativo di corrispondenza da parte dei loro clienti; SentiRate è in grado di identificare gli elementi di maggiore interesse in modo tale da potergli attribuire delle priorità;
- In una organizzazione che invia, ogni giorno, un gran numero di corrispondenza ai propri clienti; in questo caso, SentiRate è in grado di garantire che i messaggi siano sempre positivi;
- Consente di pubblicare automaticamente commenti positivi su un blog. È anche possibile porre delle restrizioni sui commenti negativi;
- L'uso è praticamente illimutato;

Terza parte

Obiettivi e progetto dell'ambiente sperimentale

Come abbiamo avuto modo di approfondire nelle pagine precedenti la Sentiment Analysis, laddove adottata nella classificazione e nell'estrazione di polarità, rappresenta una tecnica molto utile per fornire una approfondita analisi circa l'orientamento di informazioni contenute all'interno di un testo.

Il nostro obiettivo è quello di automatizzare il processo di identificazione del sentiment all'interno di un testo o di un insieme di testi.

Come abbiamo avuto modo di approfondire nelle pagine precedenti la classificazione di un sentimento per un testo presuppone una serie di passi fondamentali che utilizzano la classificazione dei testi stessi.

A seconda del metodo utilizzato, poi, si procede in questa fase considerando i singoli termini, piuttosto che le coppie di termini o intere frasi. Per quanto riguarda la classificazione dei testi utilizzando i singoli termini uno dei metodi maggiormente utilizzati in letteratura è quello della LDA (Latent Dirichlet Allocation), modello statistico, cosiddetto generativo (per i motivi che andremo ad approfondire in seguito) in grado di estrapolare alcuni termini caratterizzanti un dato argomento di un documento[165].

In pratica LDA, in statistica, è un modello generativo che permette di effettuare una serie di osservazioni per spiegare la correlazione fra le parole chiave e topic (argomenti) simili fra loro. Infatti si presppone che un documento sia una miscela di un piccolo numero di argomenti

[165] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John. ed. *Latent Dirichlet allocation*. Journal of Machine Learning Research 3 (4–5): pp. 993–1022.

e che l'utilizzo di ogni parola sia attribuibile a uno dei temi del documento.

LDA e tutti i modelli presenti in letteratura che poggiano su esso sono in grado di classificare efficacemente un testo identificando i termini che ne caratterizzano lo specifico ambito (ad esempio definendo se i testi analizzati sono relativi alla filmografia, piuttosto che a degli sport e così via).

L'ipotesi di lavoro sulla quale ci siamo mossi nella nostra attività di ricerca è che i modelli ch utilizzano LDA, oltre a classificare il dominio di determinati insiemi di testi possano, di fatto, essere utilizzati anche per caratterizzare il sentiment (e quindi l'orientamento positivo o negativo del redattore di quel testo nei confronti dell'oggetto), del servizio o argomento generico trattato nello stesso testo. In pratica in questo modo siamo in grado di caratterizzare un dato dominio con termini specifici e siamo anche in grado di evidenziare quali sono i termini, all'interno di quello specifico dominio, in grado di caratterizzare un orientamento positivo piuttosto che un orientamento negativo.

Si capisce quanta utilità potrebbe avere una simile aggiunta innovativa alle numerose applicazioni presenti in letteratura per LDA. Tramite LDA quindi potremmo, per un quasiasi dominio specifico, riuscire a capire quali sono i termini utilizzati per esprimere un sentimento negativo o positivo.

Trattandosi di un'idea di fatto innovativa per questo specifico aspetto, rispetto a quanto analizzato in letteratura, abbiamo inteso, in una prima fase della nostra ricerca sperimentale, valutare innanzitutto la fondatezza della nostra ipotesi e l'effettiva possibilità di utilizzare anche un modello basato su LDA per classificare il sentiment oltre che lo stesso dominio dei documenti.

Abbiamo quindi costruito un sottoinsieme di un dataset (quello relativo alle recensioni di films Movie Reviews) utile solo a questo specifico task.

Durante questa prima fase della nostra sperimentazione abbiamo elaborato un processo (basato proprio su LDA) utile ad estrarre il lessico di uno specifico dominio attraverso l'identificazione di parole

(termini) semanticamente rilevanti appunto sulla base di analisi di ontologie.

Una volta ottenuto il vettore dei termini rappresentativi di uno specifico dominio, sulla base della nostra ipotesi di partenza, abbiamo reiterato lo stesso processo prima sul dataset dei commenti già etichettati come positivi e successivamente su quello dei commenti già classificati come negativi.

Abbiamo in questo modo ottenuto anche i vettori rappresentativi tanto la polarizzazione positiva quanto la polarizzazione negtiva.

Successivamente abbiamo utilizzato questi vettori (dei termini caratterizzanti l'orientamento positivo e di quelli caratterizzanti l'orientamento negativo) ricercandone le ricorrenze all'interno del commento che volevamo classificare.

Mediante una funzione di score specifica, poi abbiamo confrontato i due valori ottenuti a partire, da questi confronti, del valore di sentiment positivo di quello di sentiment negativo per il commento del quale calcolare il sentiment.

La funzione di score ci ha permesso di definire la positività o la negatività del sentiment del commento considerato.

I risultati di classificazione del sentiment per il nostro ridotto insieme di questa prima fase sperimentale sono stati confrontati con quelli di tecniche differenti presenti in letteratura applicate allo stesso insieme

Va specificato che, come anticipato, questa prima sperimentazione ha utilizzato solo un piccolo sottoinsieme del dataset considerato.

Pertanto, poiché la nostra sperimentazione e quelle presenti in letteratura che abbiamo utilizzato come riferimento trattano insiemi sperimentali diversi (appunto il nostro non è completo) i risultati ottenuti non sono di fatto comparabili.

Tuttavia i risultati raggiunti sono stati in grado di fornirci una valutazione almeno quantitativa della bontà del metodo e quindi della possibilità di attuare una campagna sperimentale completa sull'intero dataset.

Abbiamo quindi costruito una nuova campagna sperimentale, questa volta considerando lo stesso dataset (ma questa volta nella sua interezza) ampliamente adottato in letteratura in modo che i risultati ottenuti dal modello da noi proposto potessero essere confrontati con tutti quelli ottenuti da altre metodologie presenti e consolidate.

Abbiamo in questa seconda campagna formulato anche una ulteriore ipotesi di lavoro, volta ad arricchire il nostro modello. Oltre alla caratterizzazione derivante dai vettori del sentiment positivo e quelli per il sentiment negativo abbiamo considerato i valori che tali termini hanno all'interno del lessico annotato SentiWordNet nella versione 3.0. Inoltre, dato il vettore di termini caratterizzanti il dominio abbiamo osservato che i termini caratterizzanti ciascun gruppo di testi e commenti, indipendentemente dal sentiment positivo o negativo già classificato (si trattava di dataset di training già opportunamente classificati) presentavano una serie (molto alta) di termini che si ripetevano.

Abbiamo quindi ipotizzato che tali termini comuni fossero sì caratterizzanti il dominio dello specifico insieme di documenti, ma non utili a classificare lo specifico sentiment (positivo o negativo) degli stessi documenti.

Allora abbiamo ottimizzato i vettori andando ad eliminare dagli stessi tutti i termini reciprocamente presenti (e quindi considerati non significativi ai fini della specifica polarizzazione).

Questa seconda campagna sperimentale ha consentito di dimostrare la validità di entrambe le nostre ipotesi (quella di poter utilizzare LDA anche per classificare il sentiment oltre che lo specifico dominio e quella che fosse possibile eliminare dai nostri calcoli anche tutti quei termini del vettore caratterizzanti il dominio ma non utili ai fini della caratterizzazione dello specifico sentiment.

L'utilizzo poi di SentiWordNet per pesare opportunamente i termini dei vettori caratterizzanti, così come per alcuni elementi era stato anticipato da Ohana e Tierney[166], ci ha consentito di incrementare notevolmente i parametri di performance della metodologia e

[166] Ohana B., Tierney B.: *Sentiment Classification of Reviews Using SentiWordNet*, 9th. IT & T Conference School of Computing - Dublin Institute of Technology 2009

dell'ambiente sperimentale appositamente realizzato. I risultati sperimentali hanno così rivelato che il nostro approccio ha consentito di incrementare le performances dei metodi già presentati che non utilizzano le ontologie. Inoltre è stata proposta una nuova metodologia per la valutazione di recensioni per servizi e prodotti diversi su uno stesso insieme e una proposta per la progettazione di un metodo per astrarre il lessico generico da ambiti specifici.

Come anticipato, nel nostro metodo sono stati utilizzati dei datasets contenenti valutazioni su prodotti e servizi e in particolare relative a recensioni sulla filmografia.

Tale scelta è motivata dal fatto che il dataset scelto per la sperimentazione è presente in numerosi articoli che sono già stati presentati nelle pagine precedenti. Numerose poi sono le metodologie sperimentali che hanno visto coinvolti questi stessi dati per i quali sono ben definiti i sottoinsieme di training e quelli di test.

Ma, prima di entrare nel merito della descrizione della metodologia proposta è opportuno fornire una descrizione dell'LDA e una panoramica delle metodologie basate su essa oltre a una panoramica dei dataset utilizzati in letteratura con un breve elenco delle caratteristiche fondamentali.

3.1 LDA

In statistica, la Latent Dirichlet Allocation (LDA) è un modello generativo[167] che permette ad una serie di campioni osservati di essere descritti grazie a una serie di gruppi, non osservati, che consentono di giustificare le similitudini tra alcune sezioni dei dati inizialmente disponibili.

Più in dettaglio, se ad esempio i campioni osservati sono parole raccolte in un insieme di documenti, si postula che ogni documento sia una miscela di un numero limitato di topics e che la creazione di ogni parola sia attribuibile a uno degli specifici topics del documento.

LDA è un esempio di topics model presentato per la prima volta come modello grafico per il tema del rilevamento da David Blei, Andrew Ng, e Michael Jordan nel 2002.

[167] In probabilità e statistica, un modello generativo è un modello per la generazione casuale dei dati osservabili, generalmente assegnati alcuni parametri nascosti. Specifica una distribuzione di probabilità congiunta su osservazione e sequenze di etichette. I modelli generativi sono utilizzati generalmente nell'ambito del machine learning direttamente per i dati di modellazione, o come fase intermedia per costruire un funzione di densità di probabilità condizionale. Una distribuzione condizionale può essere costruita da un modello generativo attraverso l'uso di regola di Bayes. I modelli generativi contrastano con i modelli discriminatori, in quanto un modello generativo è un modello probabilistico completo di tutte le variabili, mentre un modello discriminante fornisce un modello solo per la variabile o le variabili di destinazione condizionali per le variabili osservate. Per questo un modello generativo può essere utilizzato, per esempio, per simulare (ossia generare) valori di qualsiasi variabile nel modello, mentre un modello discriminante consente solo il campionamento delle variabili target condizionali sulle quantità osservate. Se i dati osservati vengono campionati a partire dal modello generativo, successivamente si utilizzano i parametri del modello generativo per massimizzare la probabilità di dati. Tuttavia, poiché la maggior parte dei modelli statistici forniscono solo approssimazioni alla distribuzione reale, se l'applicazione del modello è quella di inferire su un sottoinsieme di variabili condizionali su valori noti di altri, allora si può affermare che l'approssimazione fornisce più ipotesi di quelle necessarie per risolvere il problema. In questi casi, un modello discriminante può essere più accurato per modellare direttamente le funzioni di densità condizionali, anche se saranno i dettagli dell'applicazione specifica in definitiva a consigliare l'approccio più adatto al caso particolare. Nonostante il fatto che i modelli discriminativi non abbiano bisogno di modellare la distribuzione delle variabili osservate, non possono generalmente esprimere relazioni più complesse tra le variabili osservate e quelle di destinazione. Essi generalmente non ottengono prestazioni migliori rispetto ai modelli generativi nei processi di classificazione e di regressione.

Le idee alla base di un probabilistic topic model sono piuttosto semplici e sono basate su un modello probabilistico per ciascun documento in una raccolta. Un topic o argomento è una distribuzione di probabilità multinomiale sulle parole V uniche nel vocabolario del corpus,

in sostanza, un dado a V-facce da cui si può generare (senza memoria) "un Bag of Words[168]" o una serie di un certo numero di parole di un documento. Così, ogni argomento è il un vettore di probabilità

$p\left(\frac{w}{t}\right) = [p\left(\frac{w_1}{t}\right), \ldots, p\left(\frac{w_V}{t}\right)]$,

dove

$$\sum_{v} p\left(\frac{w_v}{t}\right) = 1$$

e ci sono **T** argomenti in totale, $\mathbf{1 \leq t \leq T}$.

Un documento è rappresentato come una miscela finita dei topics T.

Si suppone che ogni documento **d**, $\mathbf{1 \leq d \leq N}$, abbia come propria serie di coefficienti la miscela di topics,

$$\left[\boldsymbol{p}\left(\boldsymbol{t} = \frac{\mathbf{1}}{\boldsymbol{d}}\right), \ldots \boldsymbol{p}\left(\boldsymbol{t} = \frac{\boldsymbol{T}}{\boldsymbol{d}}\right)\right.$$

vettore probabilità multinomiale tale che

$$\sum_{t} p\left(\frac{t}{d}\right) = 1$$

[168] Il modello *Bag of Words* è una rappresentazione semplificata adottata nell'elaborazione del linguaggio naturale per il recupero delle informazioni (IR). In questo modello, un testo (ad esempio, una frase o un documento) è rappresentato come una collezione non ordinata di parole, trascurando la grammatica e perfino l'ordine delle parole. Il modello Bag of Words è usato comunemente nei metodi di classificazione di documenti, in cui vengono considerate le occorrenze (frequenza) di ogni parola quale elemento per la costruzione di un classificatore. C.f.r.Girolami, Mark; Kaban, A. (2003). *On an Equivalence between PLSI and LDA*. Proceedings of SIGIR 2003. New York: Association for Computing Machinery.

Così, una parola selezionata casualmente dal documento d ha una distribuzione condizionata $\boldsymbol{p}(\frac{\boldsymbol{w}}{\boldsymbol{d}})$ che è una miscela di più argomenti, in cui ogni argomento è un multinomiale di più parole:

$$\mathrm{p}\left(\frac{\mathrm{w}}{\mathrm{d}}\right) = \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{p}(\frac{\mathrm{w}}{\mathrm{t}})\mathrm{p}(\frac{\mathrm{t}}{\mathrm{d}})$$

Se dovessimo simulare parole **W** per il documento **d** che utilizzano questo modello ci sarebbe ripetere la seguente coppia di operazioni **W** volte:: prima estrarre un argomento **t** secondo la distribuzione $\mathbf{p}\left(\frac{\mathbf{t}}{\mathbf{d}}\right)$ e quindi selezionare una parola **W** secondo la distribuzione $\mathbf{p}\left(\frac{\mathbf{w}}{\mathbf{d}}\right)$,

Dato questo modello generativo per di una serie di documenti, il passo successivo è quello di apprendere le distribuzioni topic-parola e documento-topic sulla base dei dati osservati.

Vi sono stati progressi notevoli su algoritmi di apprendimento per questi tipi di modelli negli ultimi anni. Hofmann[169] ha proposto un algoritmo per l'apprendimento in questo contesto, denominato "probabilistic LSI" o "pLSI". Blei, Ng e Giordania, invece hanno affrontato alcune delle limitazioni dell'approccio pLSI (come la problematica dell'Overfit) e hanno riformulato il modello e di apprendimento in un quadro più generale di impostazione bayesiana.

Si parla così di Latent Dirichlet Allocation (LDA), essenzialmente una versione Bayesiana del modello appena descritto e l'algoritmo di apprendimento è basato su una tecnica nota come approssimazione variazionale apprendimento.

Un algoritmo alternativo di stima, molto efficiente, basato sul campionamento di Gibbs è stato proposto da Griffiths e Steyvers[170], una tecnica che è strettamente correlata alle idee precedenti derivate in modo indipendente per i modelli a miscela statistica genetica[171]. Dal

[169] Hofmann, T.: *Probabilistic Latent Semantic Indexing*. 22nd Int'l. Conference on Research and Development in Information Retrieval (1999)
[170] Griffiths, T.L., and Steyvers, M.: *Finding Scientific Topics*. National Academy of Sciences, 101 (suppl. 1) (2004) 5228–5235
[171] Pritchard, J.K., Stephens,M., Donnelly, P.: *Inference of Population Structure using Multilocus Genotype Data*. Genetics 155 (2000) 945–959

momento che le Griffiths e la carta Steyvers è stato pubblicato nel 2004, un certo numero di gruppi hanno applicato con successo il modello ad una varietà di grandi corpora, compresi grandi collezioni di documenti Web[172], una collezione di 250.000 email Enron[173], 160.000 estratti dalla raccolta di informatica CiteSeer[174], e 80.000 articoli relativi a notizie del XVIII secolo (Pennsylvania Gazette)[175].

Come abbiamo visto, in LDA ogni documento può essere visto come una miscela di vari topics. LDA rappresenta un'evoluzione di modelli statistici utilizzati in precedenza quali come l'analisi semantica latente probabilistica (pLSA).

Considerando le osservazioni in forma di co-occorrenze

(w, d)

delle parole e dei documenti, pLSA modella la probabilità di ciascuna co-occorrenza come un miscuglio di distribuzioni multinomiali condizionatamente indipendenti:

$$P(w,d) = \sum_c P(c)\, P\left(\frac{d}{c}\right) P\left(\frac{w}{c}\right) = P(d) \sum_c P\left(\frac{d}{c}\right) P\left(\frac{w}{c}\right)$$

La prima formulazione è la formulazione simmetrica, dove ***w*** e ***d*** sono entrambi generati dalla classe ***c*** latente in modi simili (usando le probabilità condizionate $P\left(\frac{d}{c}\right) e\, P\left(\frac{w}{c}\right)$), mentre la seconda formulazione è la formulazione asimmetrica, in cui, per ciascun documento ***d***, viene scelta una classe latente condizionatamente al documento secondo $P\left(\frac{c}{d}\right)$, e una parola viene quindi generata da tale classe secondo $P\left(\frac{w}{c}\right)$.

[172] Buntine, W. , Perttu, S. , Tuulos, V.: *Using Discrete PCA on Web Pages*. Proceedings of the Workshop W1 on Statistical Approaches for Web Mining (SAWM).Italy (2004) 99-110
[173] McCallum, A., Corrada-Emmanuel, A., Wang, X.: *Topic and Role Discovery in Social Networks*. 19th Joint Conference on Artificial Intelligence (2005)
[174] Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: *Probabilistic Author-Topic Models for Information Discovery*. 10th ACM SIGKDD (2004)
[175] Ibid

Anche se in questo esempio abbiamo usato parole e documenti, la co-occorrenza di ogni coppia di variabili discrete può essere modellata con lo stesso processo.

Quindi, il numero di parametri è uguale a ***cd*** **+** ***wc***. Il numero di parametri cresce linearmente con il numero di documenti. Inoltre, sebbene pLSA sia un modello generativo per i documenti di una dataa collezione, esso non è un modello generativo per nuovi documenti.

Rispetto al pLSA in LDA la distribuzione dei Topics si assume avere una precedente Dirichlet. In pratica, questo si traduce in una miscela di topics più ragionevoli in un documento. Tuttavia, il modello pLSA è di fatto equivalente al modello LDA per distribuzioni di Dirichlet uniformi.

In questo modello un documento è dato dagli argomenti (dai topics). Questa è un'assunzione standard anche nel modello Bag of Words, e rende le singole parole sostituibili.

3.1.1 Il funzionamento di LDA

Abbiamo visto come LDA consente di rappresentare i documenti come miscele di argomenti (i topics) che evidenziano i termini con relative probabilità determinate.

Si assume che i documenti vengono prodotti nel modo seguente: quando viene prodotto ciascun documento si procede con i seguenti passi elementari:

- Si decidono il numero di parole N che avrà il documento (ad esempio secondo una distribuzione di Poisson).
- Si sceglie una miscela di topics per il documento (secondo una distribuzione Dirichlet su un insieme fisso di temi K).
- Nel documento si genera ciascuna parola w_i:
    - In primo luogo scegliendo uno dei topics della miscela (in base alla distribuzione multinomiale del campionamento campionato di cui sopra).
    - Utilizzando l'argomento per generare la parola stessa (secondo la distribuzione multinomiale del tema).

Dato per assodato questo modello generativo per un insieme di documenti, LDA poi cerca di procedere a ritroso dai documenti per rintracciare i topics che possono aver generato la collezione.

Consideriamo un esempio pratico

Secondo il processo appena presentato, quando si genera qualche particolare set di documenti D, si potrebbe:

- Scegliere per il particolare set D un numero di parole pari a 5 (per ciascun documento)
- Decidere che D sarà composto per ½ da termini relativi a films (I° Topic) e per ½ da termini relativi alla pittura (II° Topic).
- Si sceglie una prima parola relativa al topic film e si sceglie ad esempio “regista”
- Si sceglie una seconda parola relativa al topic pittura e si sceglie ad esempio “colore”
- Si sceglie una terza parola relativa al topic pittura e si sceglie ad esempio “tavolozza”
- Si sceglie una quarta parola relativa al topic film e si sceglie ad esempio “musica”
- Si sceglie una quinta parola relativa al topic film e si sceglie ad esempio “attore”

Così il documento generato con il modello di LDA sarà "regista, colore, tavolozza, musica, attore" (si noti che LDA è una modello a Bag-of-Words).

Vediamo ora le modalità di apprendimento.

Si fornisce ora uno schema estremamente semplificato del funzionamento del modello LDA valido al solo scopo di rappresentarne il processo elementare.

Si supponga di avere un insieme di documenti. Si sceglie un numero fisso di temi K da scoprire, e si vuole usare LDA per apprendere la rappresentazione degli argomenti di ciascun documento e le parole associate a ciascun argomento.

Una delle modalità (noto come "campionamento collassato di Gibbs[176]") è il seguente:

- Si entra nel documento e in modo casuale si assegna una parola dello stesso ad uno degli argomenti **K**.
- Si noti che questa assegnazione casuale dà già entrambe le rappresentazioni tematiche di tutti i documenti e le distribuzioni delle parole per tutti gli argomenti (anche se non molto buoni).
- Quindi, per migliorare questa rappresentazione, per ogni documento **d**
- Si passa attraverso ogni parola w in d e:
- E per ogni argomento **t**, si calcolano due valori: 1) $\boldsymbol{p}(\frac{\boldsymbol{t\ topic}}{\boldsymbol{d\ document}})$ che rappresenta la percentuale di parole nel documento **d** che sono attualmente assegnate all'argomento t, e 2) $\boldsymbol{p}(\frac{\boldsymbol{t\ topic}}{\boldsymbol{d\ document}})$ che rappresenta la proporzione di assegnazione all'argomento **t** di tutti i documenti che provengono da questa parola **w**. Si assegna poi la parola **w** a un nuovo argomento **t** dove abbiamo scelto l'argomento **t** con probabilità $\boldsymbol{p}\left(\frac{\boldsymbol{t\ topic}}{\boldsymbol{d\ document}}\right) * \boldsymbol{p}(\frac{\boldsymbol{w\ parola}}{\boldsymbol{t\ topic}})$ (secondo il nostro modello generativo, questa è essenzialmente la probabilità che il topic **t** abbia generato la parola **w**, così ha senso che noi ricampioniamo l'argomento della parola corrente con questa probabilità).
- In questa fase, stiamo assumendo che tutte le assegnazioni di argomento (tranne per la parola corrente) in questione siano corrette, e possiamo aggiornare l'assegnazione della parola corrente utilizzando il nostro modello partendo dalla considerazione di come i documenti vengono generati.
- Dopo aver ripetuto il passaggio precedente un gran numero di volte, è possibile ottenere o meno uno stato stazionario dove le assegnazioni siano soddisfacenti.

[176] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. *Fast collapsed gibbs sampling for latent dirichlet allocation*. In Proceedings of SIGKDD.

Si utilizzano poi queste assegnazioni per stimare le miscele di argomenti di ciascun documento (contando la proporzione di parole assegnate a ciascun argomento in tale documento) e le parole associate a ciascun argomento (contando la proporzione di parole assegnate a ciascun argomento complessivamente).

3.2 iSoS

Nell'ambito della nostra campagna sperimentale e per testare opportunamente il nostro modello avevamo necessità di un frame work che, basandosi su LDA, ci consentisse in maniera effeicace di rclassificare opportunamente i commenti e riconoscere i relativi domini.

La nostra scelta è caduta su iSoS Lite[177], un frawework sviluppato in Java e Java Server Pages, che include una versione personalizzata dell'API open source di Apache Lucene, già utilizzato dal gruppo di ricerca.

In particolare abbiamo utilizzato iSoSLite[178] come LDA per costruire i nostri grafi elementari.

Utilizzando iSoSlite possiamo modellizzare semanticamente il nostro insieme focalizzandosi sulla struttura delle relazioni associative tra parole e termini del linguaggio naturale e relazioni tra termini aggregatori e termini non aggregatori.[179,180,181].

Con iSoS definiamo un "Graph of Terms (concepts) come

$$G_C = \{N;E;C\}$$

dove N fa riferimento a un set definite di nodi, E come un set di archi pesati su N con degli indici $\psi_{i,j}$ così che {N;E}è un grafico diretto e C definisce un set di concetti così che, per ogni nodo è tale che

4. $n \epsilon N$
5. C'è un solo termine aggregatore
6. $c \epsilon C$

[177] iSoSLite è stato sviluppato dal DIEI dell'Università di Salerno nel 2010

[178] iSoSlite è un'applicazione web based, scritta in Java e Java Server Pages che include version personalizzate delle API Open Source di Apache Lucene, un motore di ricerca Full text.

[179] W. Kintsch. T*he role of knowledge in discourse comprehension: A construction-integration model*. Psychological Review, 95:163{182, 1988

[180] M. C. Potter. *Very short term conceptual memory*. Memory & Cognition, (21):156{161, 1993

[181] K. A. Ericsson and W. Kintsch. *Long-term working memory*. Psychological Review., 102:211-245, 1995

7. I pesi $\psi_{i,j}$ possono essere considerati come livelli di correlazione semantic tra due termini aggregatori.
8. c_i _*is-related* $_{\psi i,j}$*-to* c_j
9. e possono essere calcolati come probabilità
10. $\psi_{i,j} = P(c_i; c_j)$

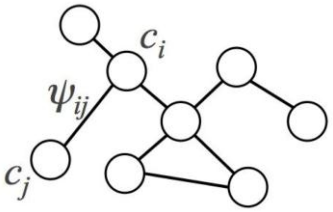


Figura 25 - Grafo di concetti

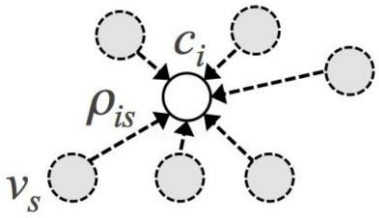


Figura 26 - Graph of a Concept

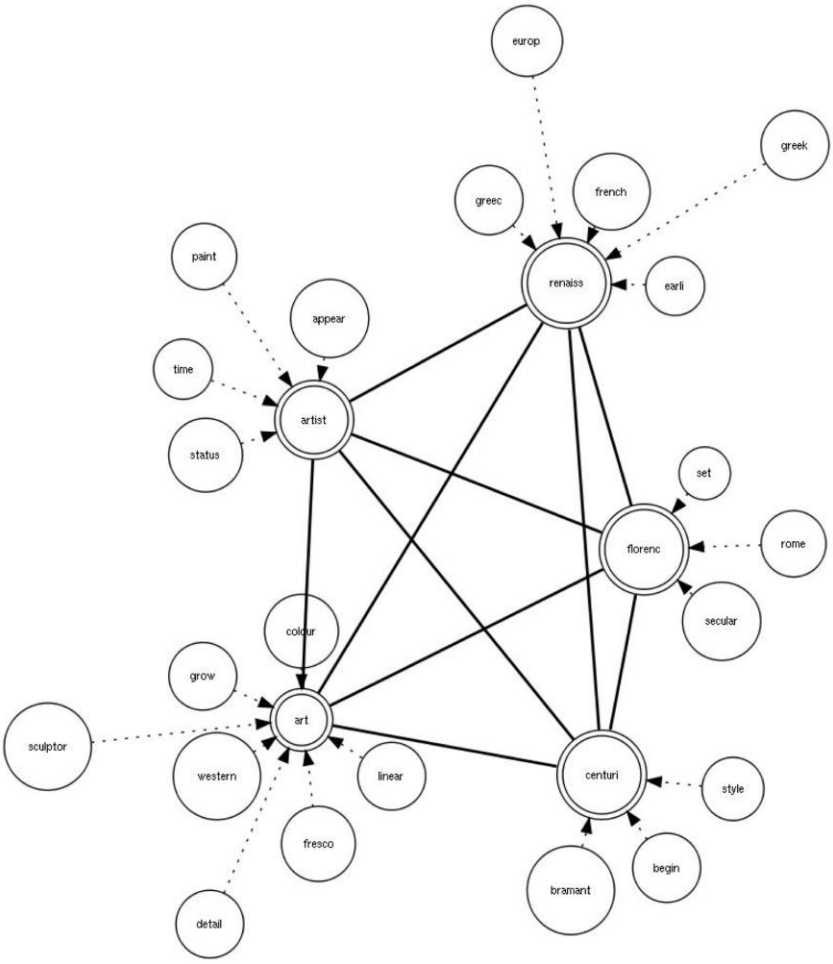


Figura 27 - Graph of Concepts

Rappresentazione della conoscenza:

(a) Grafo di concetti: I pesi $\psi_{i,j}$ mostrano la probabilità per due concetti di essere semanticamente correlati.

(b) Graph of a Concept: I pesi ρ_i mostrano la probabilità per un termine e un termine aggregante di essere semanticamente correlati.

(c) Graph of Concepts: rappresenta una informal Lightweight Ontology estratta da un set di documnti.

Per un dataset iSoSlite estrae automaticamente un Graph of Concepts (GC). Abbiamo due fasi iniziali: la prima dove focalizziamo la nostra

attenzione ai termini aggregatori del nostro vocabolario (apprendendo le relazioni tra termini e termini aggregatori) e nell'altra calcoliamo le relazioni tra I termini aggregatori (graph learning).

E' importante rimarcare che i nostri Graph of Terms non mostrano una matrice di co-occorrenza. Difatti, così come mostrato in letteratura da Efthimiadis e Manning[182,183] le tecniche basate su queste matrici sono limitate e comunque non utili per gli obiettivi che intendiamo perseguire.

Alla base del nostro grafo vi è la probabilità $P(v_i; v_j)$ calcolata dal Topic Model probabilistico e dalla word association:

Sulla base di queste premesse la nostra rappresentazione è più vicina a quella di una grafo piuttosto che all'idea di una matrice.

3.3 I datasets sperimentali

3.3.1 HetRec 2011 Delicious Bookmarks Data Set

Questo dataset contiene informazioni di social networking, bookmarking e tagging relativi a un insieme di 2000 utenti (Provenienza: sistema di social bookmarking Delicious) con i seguenti dati: 1867 utenti, 69226 URLs, 38581 URLs principali, 7668 relazioni bidirezionali tra gli utenti, 53388 tags, 437593 tag assignments (tas), 104799 bookmarks.

Tutti I dati vengono sono formattati con un record per riga (separati da tabulazioni, "\ t"):

Di seguito l'elenco dei file contenuti nel dataset.

User_contacts.dat - user_contacts-timestamps.dat

[182] Efthimis N. Efthimiadis. *Query expansion*. In Martha E. Williams, editor, Annual Review of Information Systems and Technology, pages 121{187. 1996

[183] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch□utze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008. ISBN 978-0-521-86571-5. URL http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

Questi file contengono le relazioni di contatto tra gli utenti del database. Una relazione di contatto interviene quando per due utenti reciprocamente l'uno è fan dell'altro in Deliciuous.

I file contengono anche il timestamp relativo alla creazione delle relazioni di contatto.

User_contacts.dat							
userID	contactID	date_ day	date_ month	date_ year	date_ hour	date_ minute	date_ second
8	28371	4	10	2010	2	14	19

user_contacts-timestamps.dat		
userID	contactID	timestamp
8	28371	1286151259000

Bookmarks.dat

Questo file contiene informazioni sugli URL impostasti come segnalibri.

id	md5	title
1	ab4954b633 ddaf5b5bba 6e9b71aa6b 70	IFLA - Il sito ufficiale della Federazione Internazionale delle Associazioni e delle Istituzioni Bibliotecarie

url	urlPrinc ipal	md5Principal
http://www.ifla. org/~~V	www.ifl a.org	7f431306c428457bc4e1 2b15634484f

Tags.dat
Questo file contiene l'insieme di tag disponibili nel dataset.

id	valore
1	collection_development

User_taggedbookmarks.dat - user_taggedbookmarks-timestamps.dat

Questi file contengono le assegnazioni tag URL dei segnalibro forniti da ciascun utente.

Essi contengono anche il timestamp relativi a questa associazione.

User_taggedbookmarks.dat

userID	bookmarkID	TAGID	giorno	mese	anno	ore	minuti	secondi
8	1	1	8	11	2010	23	29	22

User_taggedbookmarks-timestamps.dat

userID	bookmarkID	TAGID	timestamp
8	1	1	1289255362000

Bookmark_tags.dat

Questo file contiene i tag assegnati agli URL, e il numero di volte in cui sono stati assegnati all'URL.

bookmarkID	TAGID	tagWeight
1	2	276

3.3.2 HetRec 2011 Last.FM Data Set

Questo dataset contiene le informazioni relative all'ascolto musicale, al social networking, al tagging, e agli artisti per un set di 2000 utenti del sistema musicale Last.fm.

Questi i dati: 1892 utenti. 17632 artisti, 12717 relazioni di amicizia bi-direzionali, 92834 ascolti di artista per utenti(utente, artista, numero di ascolti), 11946 tag, 186479 tag attribuiti agli artisti.

Artists.dat

Questo file contiene informazioni sui musicisti ascoltati e taggati dagli utenti.

id	nome	url	pictureURL
707	Metallica	http://www.last.fm/music/Metallica	http://userserve-ak.last.fm/serve/252/7560709.jpg

Tags.dat

Questo file contiene l'insieme di tag disponibili nel dataset.

TAGID	tagValue
1	Metal

User_artists.dat

Questo file contiene gli artisti ascoltati da ciascun utente e fornisce anche un conteggio di ascolto per ogni coppia [user, artist].

userID	ArtistlD	weight
2	51	13883

User_taggedartists.dat - user_taggedartists-timestamps.dat

Questi file contengono i tag attribuiti agli artisti da ciascun utente e il timestamp relativo a questa attribuzione.

User_taggedartists.dat					
userID	artistID	tagID	day	month	year
2	52	13	1	4	2009

User_taggedartists-timestamps.dat			
userID	artistID	tagID	timestamp
2	52	13	1238536800000

User_friends.dat

Questo file contengono le relazioni di amicizia tra gli utenti del database.

userID	friendID
2	275

3.3.3 HetRec 2011 MovieLens Data Set

Questo dataset è un'estensione del dataset MovieLens10M, pubblicato da GroupLeans. Esso collega i film del dataset MovieLens con le loro pagine web corrispondenti nei sistemi di recensioni cinematografici Internet Movie Database (IMDb) e Rotten Tomatoes.

Rispetto al dataset originale, sono stati mantenuti solo gli utenti sia con rating sia con informazioni di tagging.

Questi i dati: 2113 utenti, 10197 film, 20 generi cinematografici, 20809 assegnazioni di generi cinematografici, 4060 registi, 95321 attori, 72 paesi, 10197 assegnazioni di nazionalità, 47899 assegnazioni delle località, 13222 tag, 855598 voti.

Movies.dat

Questo file contiene informazioni sui film del database.

(Le informazioni originali di titolo e anno per i film disponibili nel dataset MovieLens10M sono state ampliate con dati pubblici forniti in IMDb e i siti web Rotten Tomatoes)

Campi	Valori di esempio
1. id 2. title 3. imdbID 4. spanishTitle 5. imdbPictureURL 6. year 7. rtID 8. rtAllCriticsRating 9. rtAllCriticsNumReviews 10. rtAllCriticsNumFresh 11. rtAllCriticsNumRotten	1. 1 2. Toy story 3. 0114709 4. Toy story (juguetes) 5. http://ia.media-imdb.com/images/M/MV5BMTMwNDU0NTY2Nl5BMl5BanBnXkFtZTcwOTUxOTM5Mw@@._V1._SX214_CR0,0,214,314_.jpg 6. 1995 7. toy_story 8. 9 9. 73 10. 73 11. 0

12. rtAllCriticsScore 13. rtTopCriticsRating 14. rtTopCriticsNumReviews 15. rtTopCriticsNumFresh 16. rtTopCriticsNumRotten 17. rtTopCriticsScore 18. rtAudienceRating 19. rtAudienceNumRatings 20. rtAudienceScore 21. rtPictureURL	12. 100 13. 8.5 14. 17 15. 17 16. 0 17. 100 18. 3.7 19. 102338 20. 81 21. http://content7.flixster.com/movie/10/93/63/10936393_det.jpg

Movie_genres.dat

Questo file contiene i generi dei film.

MovieID	genere
1	Adventure

Movie_directors.dat

Questo file contiene i registi dei film.

MovieID	directorID	directorName
1	john_lasseter	John Lasseter

Movie_actors.dat

Questo file contiene i principali attori e attrici dei film.

Viene presentata una classifica agli attori di ogni film secondo l'ordine in cui appaiono sulla pagina web cast IMDb del film.

MovieID	actorID	actorName	ranking
1	annie_potts	Annie Potts	10

Movie_countries.dat

Questo file contiene i paesi di origine dei film.

MovieID	location1	Location2	Location3	Location4
2	Canada	British	Columbia	Vancouver

Movie_locations.dat

Questo file contiene i luoghi delle riprese dei film.

id	valore
1	terra

Tags.dat

Questo file contiene l'insieme di tag disponibili nel dataset.

MovieID	TAGID	tagWeight
1	13	3

User_taggedmovies.dat - user_taggedmovies-timestamps.dat

Questi file contengono le assegnazioni dei tag ai film forniti da parte di ciascun utente. Essi contengono anche i timestamp, relativi a ciascuna di queste assegnazioni.

User_taggedmovies-timestamps.dat			
userID	MovieID	TAGID	timestamp
75	353	5290	1162160415000

User_taggedmovies.dat								
userID	MovieID	TAGID	date_day	date_month	date_year	date_hour	date_minute	date_second
75	353	5290	29	10	2006	23	20	15

Movie_tags.dat

Questo file contiene i tag assegnati ai film, e il numero di volte con le quali tag sono stati assegnati a ogni film.

userID	MovieID	Voto	timestamp
75	3	1	1162160236000

User_ratedmovies.dat - user_ratedmovies-timestamps.dat

Questi file contengono le valutazioni dei film forniti da ciascun utente. Esse contengono anche i timestamp relativi a ciascuna votazione.

userID	MovieID	Voto	date_day	date_month	date_year	date_hour	date_minute	date_second
75	3	1	29	10	2006	23	17	16

3.3.4 Anonymous Ratings from the Jester Online Joke Recommender System

Dataset 1: oltre 4,1 milioni di voti continui (da -10,00 a +10,00) di 100 barzellette da parte di 73,421 utenti: raccolti tra aprile 1999 - maggio 2003.

- jester_dataset_1_1.zip: (3.9MB) dati da 24,983 utenti che hanno dato un punteggio a 36 o più barzellette, matrice di dimensioni 24.983 X 101.
- jester_dataset_1_2.zip: (3.6MB) I dati di 23.500 utenti che hanno dato un punteggio a 36 o più barzellette, matrice di dimensioni 23.500 X 101.
- jester_dataset_1_3.zip: (2.1MB) I dati di 24,938 utenti che hanno dato un punteggio a un numero di barzellette tra 15 e 35, matrice di dimensioni 24.938 X 101.

Formato:

Una riga per utente.

La prima colonna indica il numero di battute per le quali è stato dato un punteggio dall'utente e le successivi 100 colonne contengono i voti per 01 - 100.

La sub-matrice di colonne compreso solo {5, 7, 8, 13, 15, 16, 17, 18, 19, 20} è densa. Quasi tutti gli utenti hanno valutato quelle battute.

Il testo delle battute può essere sono contenute nel file

- jester_dataset_1_joke_texts.zip (92KB)

Formato:
100 files

Ogni file ha titolo init_.html, dove _ è 1-100

I titoli corrispondono a quello del l'ID delle barzellette nei file di Excel di cui sopra

Dataset 2: oltre 1,7 milioni di voti continui (da -10,00 a +10,00) di 150 barzellette da parte di 63,974 utenti: raccolti tra novembre 2006 - maggio 2009.

Formato:

- jester_ratings.dat: Ogni riga è formattata come [User ID] [ID articolo] [Punteggio]
- jester_items.dat: Fornisce una mappa di tutti gli ID e le barzellette

Si noti che i voti sono veri valori che vanno da -10,00 a 10,00. A partire dal maggio 2009, le barzellette {7, 8, 13, 15, 16, 17, 18, 19} sono "set di riferimento" mentre le barzellette {1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 20, 27, 31, 43, 51, 52, 61, 73, 80, 100, 116} sono state rimosse (esse infatti non erano mai state valutate).

3.3.5 MovieLens Data Sets

- MovieLens 100k
  - Composto da 100.000 voti (1-5) di 943 utenti su 1682 films.
  - Ogni utente ha votato al massimo per 20 films.
  - Contiene informazioni demografiche per gli utenti (età, genere, occupazione,CAP)
- MovieLens 1M
  - Consiste di 1 milione di voti di 6040 utenti che hanno scelto Movie Lens nel 2000 su 3900 films
- 10M MovieLens
  - Composto da 10000054 milioni di voti e 100.000 applicazioni dei 95580 tag applicati a 10.681 film da parte di 71567 utenti.

Tutti questi dataset non sono preventivamente classificati in positivi e negativi.

3.3.6 Book-Crossing Data Set

Il BookCrossing (BX) dataset contiene 278,858 utenti (anonimi ma con informazioni demografiche) che forniscono 1,149,780 voti (espliciti / impliciti) su circa 271,379 libri.

Il Book-Crossing dataset è composto da 3 tabelle.

BX-utenti
Contiene gli utenti. Si noti che gli ID utente (User-ID ``) sono stati resi anonimi e mappati per interi. I dati demografici sono forniti (`Località`, `` Età), se disponibili. In caso contrario, questi campi contengono valori NULL.

BX-Books

I libri sono identificati dai loro rispettivi ISBN. I codici ISBN validi sono già stati rimossi dal dataset. Inoltre, vengono fornite alcuni contenuti basati su informazioni (`Book-title`, `Book-Author`, `anno-di-Pubblicazione`, Publisher), ottenuto da Amazon Web Services. Si noti che in caso di diversi autori, viene fornito solo il primo. Sono anche fornite le URL delle immagini, che appaiono in tre formati diversi (`immagine-URL-S`, `immagine-URL-M`, `immagine-URL-

L`), vale a dire, piccolo, medio, grande. Questi URL puntano al sito web Amazon.

BX-book-Ratings

Contiene le informazioni sul punteggio di ciascun libro. Le valutazioni (`Libro-Voto`) sono esplicite, espresse su una scala da 1-10 (valori più alti denotano l'apprezzamento più alto), o implicite, espresse da 0.

3.3.7 Movie Reviews

Le informazioni del dataset sono relative alla valutazione di films.

Ne esistono differenti versioni, tutte disponibili a partire dal sito web della Cornell University – Department of Computer Science[184] o del Laboratorio di intelligenza artificiale dell'Università di Standford[185].

In particolare la versione "Sentiment Analysis Datasets with Latent Explanation Initializations" per i Movie Reviews contiene 3 directory:

- **CvList_short** cartella che contiene i nomi dei file che si trovano in ogni set (training/valifation/test)
- **Pos** cartella che contiene i documenti positivi (1000)
- **Neg** cartella che contiene i documenti negativi (1000)

La divisione delle frasi è già stata effettuata utilizzando OpinionFinder.

Relativamente al CvList_Short ciascun file di testo contiene l'elenco dei commenti che sono stati selezionati per costruire gli insiemi di training e di test[186].

[184] http://www.cs.cornell.edu/people/pabo/movie-review-data/ Last accessed 25/02/2013

[185] http://ai.stanford.edu/~amaas/data/sentiment/ Last accessed 25/02/2013

[186] Omar F. Zaidan, Jason Eisner, and Christine Piatko. 2007. *Using "annotator rationales" to improve machine learning for text categorization*. In NAACLHLT 2007; Proceedings of the Main Conference, pages 260–267, April

3.3.8 U.S. Congressional floor debates

Il dataset, relativo a un dibattito congressuale del 2006 (interamente trascritto) è già pronto per la sperimentazione con SVM (Support Vector Machine).

Il dataset contiene tre file già opportunamente formattati per SVM, rispettivamente per l'addestramento la validazione e il test.

3.3.9 Customer Review Data

Questa cartella contiene le recensioni dei clienti annotate su 5 prodotti.

1. Fotocamera digitale: Canon G3
2. Fotocamera digitale: Nikon Coolpix 4300
3. Telefono cellulare: Nokia 6610
4. Lettore MP3: Creative Labs Nomad Jukebox Zen Xtra 40GB
5. dvd player: Apex AD2600 Progressive-scan DVD

Tutte le recensioni provengono da Amazon.com.

Simboli utilizzati nelle recensioni annotate:

- [T]: il titolo della recensione: Ogni tag [t] segna l'inizio di una recensione.
- xxxx [+|-n]: xxxx è una caratteristica del prodotto.
- [+ N]: Parere positivo, n è l'entità del parere: 3 più forte, e ai più deboli 1. Poiché la valutazione dell'entità è abbastanza soggettiva si consiglia di ignorarla, considerando solo + e –
- [-N]: parere negativo
- # #: Inizio di ogni frase. Ogni riga è una frase.
- [U]: funzione non presente nella frase.
- [P]: funzione non presente nella frase. E' necessaria la risoluzione di un pronome.
- [S]: suggerimento o raccomandazione-
- [Cc]: confronto con un prodotto concorrente di una marca diversa.
- [Cs]: confronto con un prodotto concorrente della stessa marca.

3.3.10 Riepilogo dei datasets

Per alcuni dei datasets precedenti e partendo dall'analisi dello stato dell'arte dei capitoli precedenti, è stato ricavato una tabella di

riepilogo contenente, per ciascun articolo, tutte le tipologie dei datasets utilizzati.

Le tabella di riepilogo è riportata di seguito.

Articolo	anno	Polarity mining techniques used	Text granularity	Features	Data sources/Domains	Performance (accuracy)
Hatzivassiloglou and McKeown (1997)	1997	log-linear regression model	document	conjunctions, part-of-speech	Wall Street Journal corpus	adjectives: precision: >90%
Das and Chen (2001b)	2001	lexicons and grammar rules	document	words	financial news	62%
Pang and Lee (2002)	2002	Nave Bayes, maximum entropy classi-fication support vector machines	document	unigram, bi-gram, contextual ef-fect of negation, feature presence or frequency, position	IMDb (Movie review)	82.9%
Turney (2002)	2002	pointwise mutual information	document	bi-grams	Known positive terms such as excellent and negative terms such as poor movies, cars, banks	66–84%
Morinaga et al. (2002)	2002	decision tree induction	document	characteristic words, co-occurrence words, and phrases	cellular phones, PDAs and internet service providers	N/A
Dave et al. (2003)	2003	support vector machines	document	semantic features based on substi-tutions and proximity	Amazon Cnn.Net	88.9%
Yi et al. (2003)	2003	sentiment lexicon and semantic pattern	subject-spot terms	feature lexical semantics	digital cameras, music al-bums	85.6%
Turney and Littman (2003)	2003	SO-LSA(Latent Semantic Analysis), SO-PMI (Pointwise Mutual informa-tion) General Inquirer	document	words and phrases	TASA-ALLcorpus (from sources such as novels and newspaper articles)	65.27% (SO-LSA) 61.26% (SO-PMI)
Pang and Lee (2004)	2004	Nave Bayes support vector machines	document	sentence-level subjectivity sum-marization based on minimum cuts	IMDb	86.4%
Hu and Liu (2005)	2005	Opinion word extraction and aggrega-tion enhanced with WordNet	product features	opinion words opinion sentences	Amazon Cnn.Net	cameras: 93.6% DVD plyr: 73% MP3 plyr: 84.2% cellphone: 76.4%
Nigam and Hurst (2004)	2004	syntactic rules based chunking	sentence	a lexicon of polar phrases and their parts-of-speech, syntactic patterns	online resources (e.g., Usenet, online message boards) in a particular domain	general polarity analysis: precision: 77% (positive), 84% (neg-ative); recall: 43% (positive), 16% (nega-tive)
Hiroshi et al. (2004)	2004	transfer-based machine translation, principal patterns auxiliary/nominal, patterns polarity lexicon	sentiment unit	full parsing semantic analysis	bulletin boards forums on digital cameras	precision: 89% recall 43%
Bai et al. (2005)	2005	two-stage Markov Blanket Classier	document	dependence among words, mini-mal vocabulary	IMDb, infonic	movie: 87.5%, news: 89-96%
Gamon et al. (2005)	2005	Nave Bayes classier	sentence	stemmed terms, their frequency and weights, go list (salient words in a domain)	car reviews	recall: 96% (positive) 5-24% (negative and other)
Popescu and Etzioni (2005)	2005	relaxation labeling clustering	phrase	Syntactic dependency templates, conjunctions and disjunctions, WordNet	Amazon Cnn.Net	Opinion phrase polarity: precision: 86% recall: 97% relation-ships
Wilson et al. (2005)	2005	AdaBoost	phrase	subjectivity lexicon	multiperspective Question,Answering Opinion Corpus	contextual polarity: 65.7%
Kennedy and Inkpen(2006)	2006	support vector machines, term-counting method, a combination ofthe two	document	term frequencies	General Inquirer dictionary,CTRWdictionary & Adj, IMDb (Movie review)	enhanced combinedmethod: 86.2%
Chesley et al. (2006)	2006	Support Vector Machines Wiktionary	document	textual features (e.g., exclamationpoints and question marks) andlexical semantics	Web sites of CNN, NPR, At-lanta Journal and Constitu-tion, newspaper columns, re-views, political blogs, etc.	positive: 84.2% negative: 80.3%objective: 72.4%

Thomas and B. Pang(2006)	2006	support vector machines	speechsegment	reference classification	2005 U.S.floor debate in theHouse of Representatives	with same-speakerlinks and agreementlinks: 71.16%
Kaji and Kitsuregawa(2007)	2007	phrase trees and word co-occurrence,Pointwise Mutual Information	phrase	lexical relationships, word co-occurrence	HTML documents	62.7–92.9%
Blitzer et al. (2007)	2007	Structural Correspondence Learning	document	word frequencies and co-occurrences, part-of-speech	book, DVD, electronics andkitchen appliance product re-views	66.1–86.6%
Godbole et al. (2007)	2007	lexical (WordNet)	word	graph distance measurementsbetween words based on rela-tionships of synonymity and anonymity, commonality of a words	newspapers, blog posts	82.7–95.7%
Annett and Kondrak (2008)	2008	lexical (WordNet) & Support VectorMachines	document	number of positive/negative ad-jectives/adverbs, presence, ab-sence or frequency of words, min-imum distance from pivot wordsin WordNet	movie reviews, blog posts	65.4–77.5%
Zhou and Chaovalit (2008)	2008	ontology-supported polarity mining	document	n-grams, words, word senses	movie reviews	72.2%
Hou and Li (2008)	2008	Conditional Random Fields	sentence	POS tags, comparative sentenceelements	product reviews, forum dis-cussions; labeled manually and automatically	precision: man.: 89% aut.: 75% recall: man.: 81%, aut.: 71%
Ferguson et al. (2009)	2009	Multinomial Naive Bayes (MNB)	phrase	binary word feature vectors	financial blog articles	75.25%
Tan et al. (2009)	2009	Naive Bayes Classifier with feature adaptation using Frequently Co-occuring Entropy	document	words	Education reviews, stock re-views, and computer reviews	F1 score: 69–91%
Wilson et al. (2009)	2009	boosting, memory-based learning, rulelearning, and support vector learning	phrase	words, negation, polarity modification features	MPQA Corpus	83.6%
Melville et al. (2009)	2009	Bayesian classification with lexiconsand training documents	document	words	Blog posts reviewing soft-ware, political blogs, moviereviews	Blogs: 91.21% Political: 63.61% movies: 81.42%
Yessalina et al. (2010)	2010	support vector machines, annotator rationales	document	words and phrases	Movie Review Polarity Dataset v2.0 (modified by Zaidan et al. (2007))	norationales (SVM Lite) 88.56 % HUMANR (tecnica di Zaidan pura) 91.61% HUMANR@SENTENCE (proposta degli autori) 91.33%
He et al. (2011)	2011	LDA, JST JST (Joint sentiment-topic)	document	words and phrases	Movie Review Polarity Dataset v2.0	LDA 83.76% JST 94.98%
Taboada et al. (2011)	2011	Lexicon Based Semantic Orientation CALculator (SO-CAL)	document	words	Movie Review Polarity Dataset v2.0 (solo 1800 e non 2000 ma con dichiarazione esplicita che le differenze sono irrilevanti)	Con i dizionari SentiWordNet-Full 61.89% - SentiWordNet-Basic 62.89% - SO-CAL-Full 76.37% - SO-CAL-Basic 68.05%
Bonzanini et al. (2012)	2012	support vector machines , naive bayes classifier	document	words	Movie Review Polarity Dataset v2.0	NB 83.31% SVM 87.10%
He et al. (2012)	2012	LDA-GE (LDA using using generalized expectation criteria)	document	words and phrases	Movie Review Polarity Dataset v2.0	LDA-DP (random init) 64.00% LDA-DP (init with prior) 71.15% LDA-GE (random init) 65.87% LDA-GE (init with prior) 72.67%
Dinu and Iuga (2012)	2012	Naive Bayes Classifier	document	words	Movie Review Polarity Dataset v2.0	Accuracy 86.07% Neg precision 97.37% Neg recall 74.14% Pos precision 79.12% Pos recall 98.00%

3.4 Il modello proposto

L'obiettivo principale del nostro lavoro è quello, dato un commento o un insieme di commenti relativi a un prodotto, un servizio, un oggetto o una persona, di ottenerne, in maniera automatica, la polarizzazione.

Stabilire, cioè, in maniera automatica, se il commento è positivo o negativo .

Come già anticipato nelle pagine precedenti il modello proposto presuppone l'utilizzo del metodo LDA per caratterizzare il nostro insieme di testi come già frequentemente fatto in letteratura.

Solo questa prima fase (dalla quale ottenere un vettore e una rappresentazione grafica del nostro dominio) è comune alle nostre due fasi sperimentali che si basano su modelli differenti.

Nel primo modello proposto, infatti, si utilizza il vettore caratterizzante nella sua interezza per calcolare la polarizzazione del nostro set di commenti.

Nel secondo modello proposto, grazie ad una serie di ipotesi, successivamente dimostrate durante la fase di sperimentazione, si opera su una versione estremamente ridotta del vettore rappresentativo per il calcolo della polarizzazione e si interviene anche con l'aggiunta di una componente semantica nell'analisi del testo utile a migliorare ulteriormente il processo.

3.4.1 Prima sperimentazione - Modello iniziale semplificato

In questa sperimentazione, la nostra metodologia è caratterizzata dai passi seguenti:

Step 1

Dato l'insieme iniziale di commenti per i training ricaviamo da essi alcuni sottoinsiemi elementari e, per ciascuno di essi, costruiamo dei grafici rappresentative con il componente iSoS Lite basato sul modello LDA.

Lavorando sugli insiemi "di addestramento" del nostro dataset (già costruiti e classificati) costruiamo uno specifico lessico relativo al nostro dominio utile a identificare i termini che si presentano con la frequenza più elevate all'interno delle espressioni e caratterizzano i commenti positivi e quelli negativi.

Durante questo passaggio adottiamo una sorta di lite ontologies (dei semplici grafi all'interno dei quali identifichiamo solo relazioni gerarchiche e tassonomie) che ci consentono di identificare i termini più frequenti o le più comuni coppie di termini all'interno delle espressioni.

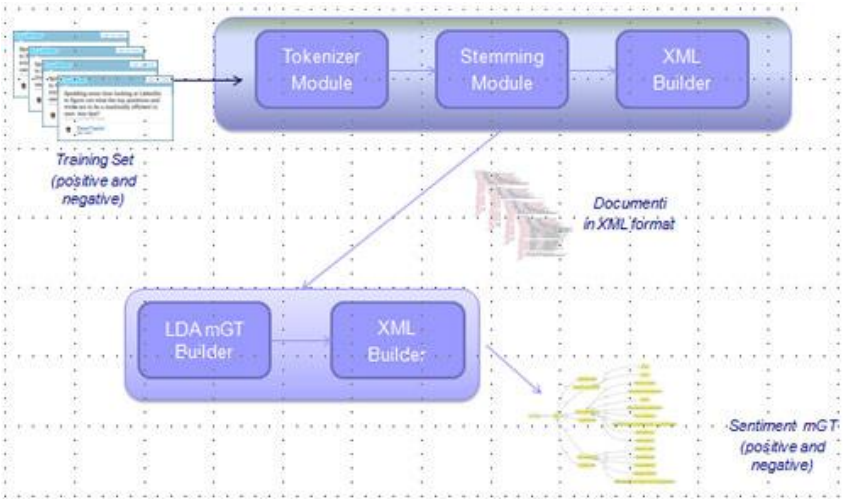


Figura 28 - Costruzione dei vettori caratterizzanti

Costruiamo i nostri grafi elementari a partire da iSoS Lite. Per costruire la nostra tabella utilizziamo la *TF-IDF*[187] (che mette a disposizione un set predefinito di funzionalità utili allo scambio di informazioni.

Possiamo dividere lo Step 1 in alcune sottofasi.

a) Dividiamo il nostro dataset di training in più dataset elementary di commenti;
b) Generiamo un grafo per ciascun dataset elmentare (con iSoS);

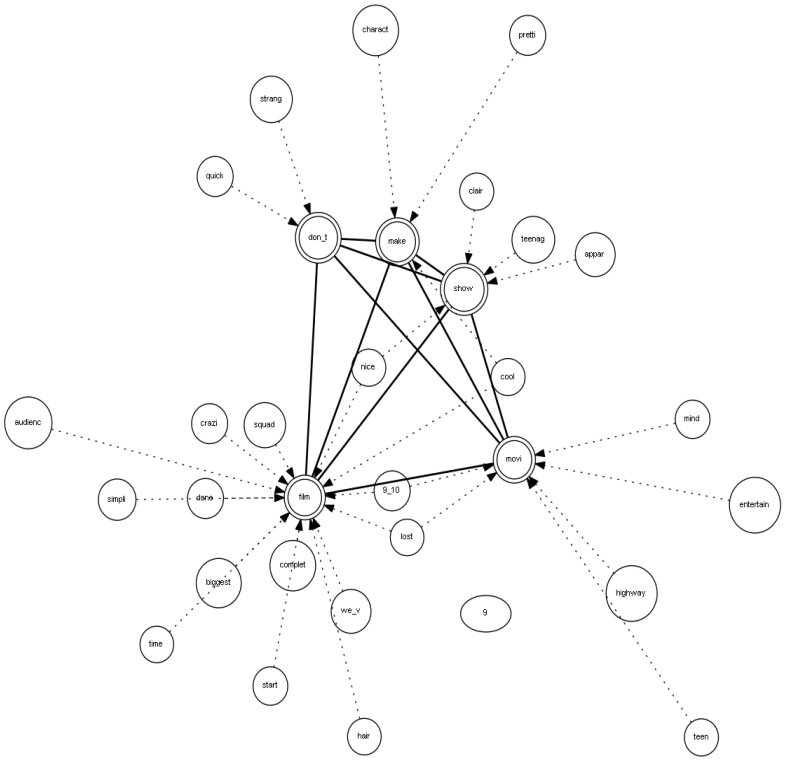


Figura 29 - Un esempio di grafo elementare di iSoS

[187] Y. Yang & J.O. Pederson (1997) *A Comparative Study on Features selection in Text Categorization*. Proceedings of the 14th international conference on Machine Learning (ICML). pp 412-420. Nashville, Tennessee.

Da ogni dataset elementare (utilizzando il grafo) ricaviamo i termini con frequenza più elevata e costruiamo un lessico relativo allo specifico dominio utilizzando solo questi termini più frequenti sia per i commenti positivi, sia per i commenti negativi.

	Movie Reviews Dataset (Positivi)	**Movie Reviews Dataset (Negativi)**
1	Man	Film
2	Play	Time
3	Perform	Play
4	Time	Thing
5	Film	Movi
6	Movie	Year
7	Good	Bad

Figura 30 - Vettori rappresentativi degli orientamenti

Step 2.

In questo passo operiamo invece con il test set. Preleviamo un commento e di volta in volta, sempre con iSoS ne calcoliamo lo specifico vettore rappresentativo e ne calcoliamo la distanza da quello caratterizzante la polarizzazione negativa e quello caratterizzante la polarizzazione positiva.

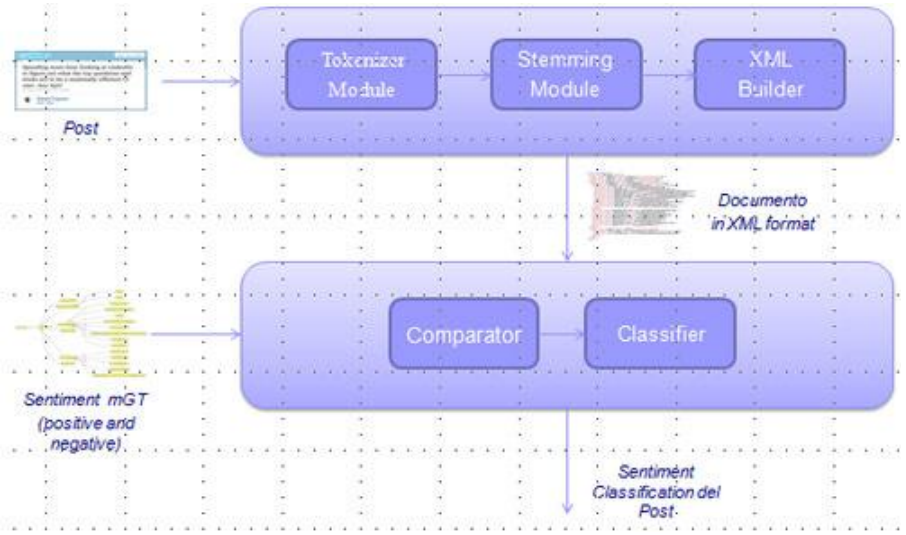


Figura 31 - Classificazione del commento

Per questa operazione utilizziamo una specifica funzione di score di seguito definita.

- Positive : if ($Score_+ > Score_-$)
- Negative: if ($Score_- > Score_+$)
- Neutral: if ($Score_- = Score_+$)

Stiamo cioè calcolando la distanza tra il vettore rappresentativo del nostro commento (prelevato dal test set) con la metodologia proposta. Confrontando i valori ottenuti siamo in grado di ottenere una polarizzazione positiva o negativa. Può anche accadere che, coincidendo i due valori di score sia nel confronto del vettore positivo che di quello negativo, la polarizzazione potrebbe essere non valutabile.

La funzione di scoring utilizzata è:

- Per i documenti positivi:

$$Score_+ = Per_{Agg+} + Per_{Wor+} + Occ_{AggAgg+} + Occ_{AggWor+}$$

- Per i documenti negativi:

  $Score_{-} = Per_{Agg-} + Per_{Wor-} + Occ_{AggAgg-} + Occ_{AggWor-}$

Dove con le abbreviazioni utilizzate indichiamo:

Per_{Agg+}	Percentuale di termini aggregatori positivi
Per_{Agg-}	Percentuale di termini aggregatori negativi
Per_{Wor+}	Percentuale di termini non aggregatori positivi
Per_{Wor-}	Percentuale di termini non aggregatori negativi
$Occ_{AggAgg+}$	Percentuale di mutua occorrenze tra termini aggregatori e termini aggregatori positivi
$Occ_{AggAgg-}$	Percentuale di mutua occorrenze tra termini aggregatori e termini aggregatori negativi
$Occ_{AggWor+}$	Percentuale di mutua occorrenze tra termini aggregatori e termini non aggregatori positivi
$Occ_{AggWor-}$	Percentuale di mutua occorrenze tra termini aggregatori e termini non aggregatori negativi

In dettaglio, il calcolo degli elementi specifici viene ottenuto secondo le formule che seguono:

$$Per_{Agg_+} = \frac{N_{Agg_+}}{N_{Agg_+} + N_{Agg_-}}$$

$$Per_{Agg_-} = \frac{N_{Agg_-}}{N_{Agg_+} + N_{Agg_-}}$$

$$Per_{W_+} = \frac{N_{W_+}}{N_{W_+} + N_{W_-}}$$

$$Per_{W_-} = \frac{N_{W_-}}{N_{W_+} + N_{W_-}}$$

$$Occ_{AggAgg_+} = \frac{Per_{AggAgg_+}}{Per_{AggAgg_+} + Per_{AggAgg_-}}$$

$$Occ_{AggAgg_-} = \frac{Per_{AggAgg_-}}{Per_{AggAgg_+} + Per_{AggAgg_-}}$$

$$Occ_{AggWor_+} = \frac{Per_{AggWor_+}}{Per_{AggWor_+} + Per_{AggWor_-}}$$

$$Occ_{AggWor_-} = \frac{Per_{AggWor_-}}{Per_{AggWor_+} + Per_{AggWor_-}}$$

Dove con l'abbreviazione "Agg" abbiamo considerato i termini aggregatori e con "Wor" i termini non aggregatori. Dal punto visto della LDA e della rappresentazione ottenuta mediante iSoS i termini aggregatori sono quei termini che all'interno del vettore rappresentativo si presentano con una mutua probabilità più elevata con tutti gli altri. I termini non aggregatori sono invece quelli che hanno una bassa percentuale di mutua occorrenza con gli altri.

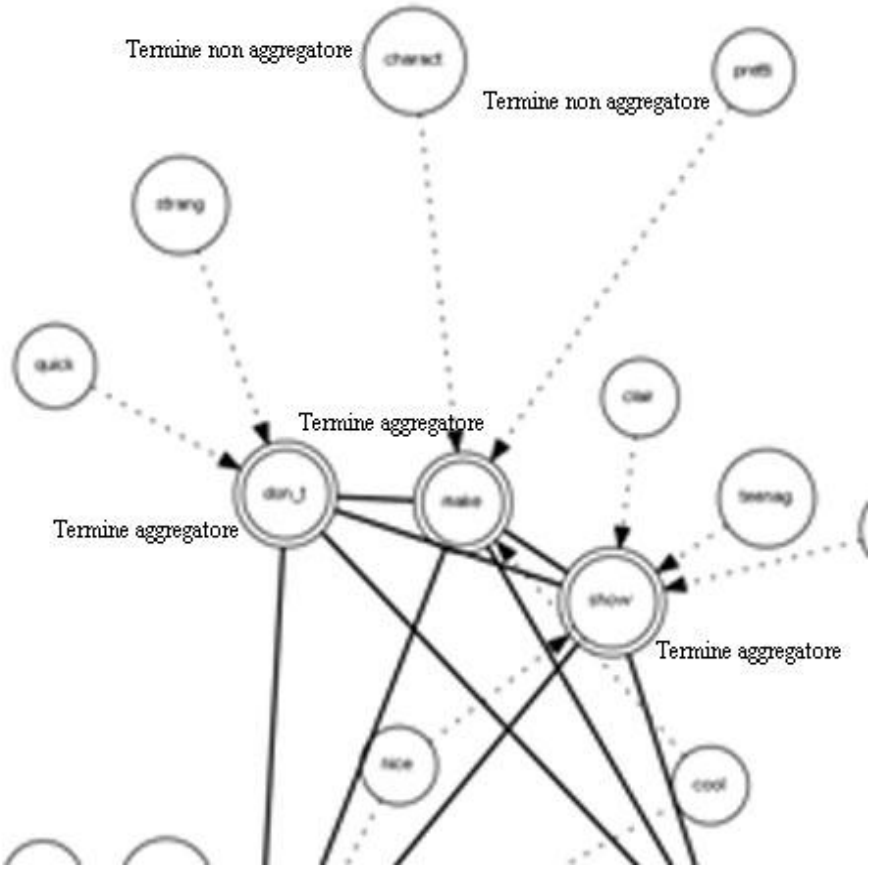


Figura 32 - Termini aggregatori e non aggregatori

3.4.2 Seconda sperimentazione - Il modello avanzato ottimizzato

Step 1

Per quanto riguarda il primo passo si replica esattamente quanto visto per il modello semplificato. Dato, cioè, un insieme iniziale di commenti ricaviamo da essi alcuni sottoinsiemi elementari e, per ciascuno di essi, costruiamo dei grafici rappresentative con il componente iSoS Lite basato sul modello LDA.

Step 2.

Dal passo precedente abbiamo ottenuto due tabelle (con relativi grafi elementari e files xml associati) contenenti i termini aggregatori (e quelli non aggregatori) caratterizzanti lo specifico dominio.

A questo punto si osserva che alcuni termini aggregatori si presentano con una elevata frequenza tanto nei vettori positivi, quanto in quelli negativi.

Formuliamo a questo punto l'ipotesi di lavoro già accennata. Partiamo dal presupposto che LDA (e quindi iSoS che abbiamo utilizzato) sia particolarmente idoneo al'individuazione dei termini caratterizzanti il dominio. Si ottiene una tabella contenente i temini comuni per tutti i grafi (caratterizzanti quindi lo specifico dominio come un dataset di recensioni di films o altro) e una lista di termini aggregatori caratterizzanti la polarizzazione. L'idea è che all'interno delle tabelle solo alcuni termini siano effettivamente utili a caratterizzare successivamente la polarizzazione perché quelli comuni tra positivi e negativi non portino informazioni di polarizzazione e servano solo a caratterizzare il dominio.

Possiamo, sulla base di questa ipotesi, peraltro com visto già affrontata in letteratura in altri contesti, eliminare dai nostri vettori rappresentativi tutti questi temini comuni ottenendo un notevole vantaggio computazionale poiché rimangono pochi termini rispetto ai quali calcolare la funzione di scoring.

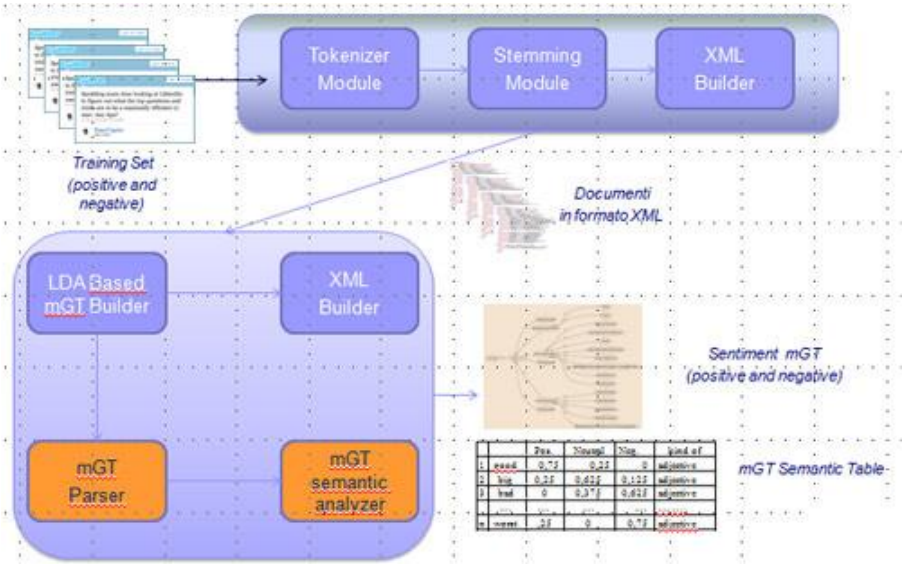


Figura 33 - Generazione dei vettori caratterizzanti - 2

Step 3

Allo scopo di ottimizzare ulteriormente il modello aggiungiamo un nuovo livello, questa volta semantico, che tenga conto anche dei valori di "positività" e negatività" di ciascuno dei termini rimanenti dallo step precedente.

Solo per questi ultimi, quindi, assegniamo un valore numerico per ciascun termine che va da -1 a +1.

Category	WNT Number	pos	neg	Synonyms
A	01123148	0.875	0	good#1
A	00106020	0	0	good#2 full#6
A	01125429	0	0.625	bad#1
A	01510444	0.25	0.25	big#3 bad#2
N	03076708	0	0	trade_good#1 good#4 commodity#1
N	05144079	0	0.875	badness#1 bad#1

Figura 34 - SentiWordNet Fragments

Category	Average Pos Score	Average Neg. Score	Word
Noun	0.531	0	good
Adjective	0.5	0	good
Adverb	0.188	0	good

Figura 35 - A word good average positive and negative score

A questo scopo utilizziamo il motore web "SentiWordNet"[188] grazie al quale possiamo attribuire questi valori ottenendo una tabella del tipo di quella riportata qui sotto:

Movie Reviews Database					
		Pos.	Neutral	Neg.	kind of
1	good	0,75	0,25	0	adjective
2	big	0,25	0,625	0,125	adjective
3	bad	0	0,375	0,625	adjective
4	enjoy	0,375	0,625	0	verb
5	great	0	1	0	adjective
6	worst	,25	0	0,75	adjective

Figura 36 - SentiWordNet Fragments

Il contesto semantico nel quale ci muoviamo ci consente di identificare anche i termini presenti (nei vettori) sulla base del loro contesto semantico (aggettivi, avverbi, nomi, pronomi).

Step 4

[188] A.H.Rohaim: *Reviews ClassificationUsing SentiWordNet Lexicon* in "The Online Journal on Computer Science and Information Technology (OJCSIT)" Vol.(2)No.(1)

Al termine di questo processo (nel quale abbiamo utilizzato il training set) utilizziamo tutti i test sets. Anche per essi consideriamo dei sottoinsiemi per i quali andiamo a generare grafi elementari.

Dai vettori ottenuti si opera l'opportuna semplificazione andando ad eliminare tutti itermini individuati nei passaggi precedenti come caratterizzanti il dominio ma non la polarizzazione.

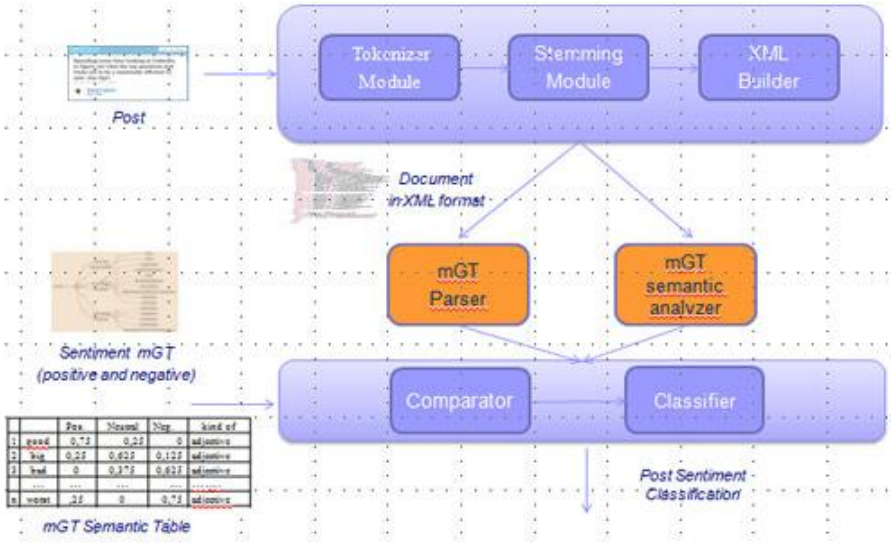


Figura 37 - Classificazione del post - 2

Questa volta per il calcolo della funzione di scoring non si considera solo la distanza tra i vettori rappresentativi secondo la vecchia fuznione di scoring.

La nostra funzione di score, come detto, si arricchisce di una componente semnatica che tiene conto della valorizzazione dei termini presenti econdo i valori desunti da SentiWordNet.

Anche in questo caso la valutazione della funzione può portarci a un riconoscimento di una polarizzazione positiva, negativa o neutra.

Funzione di Score

Valutiamo ora la nuova funzione di score che tiene conto anche della componente semantica.

Abbiamo sempre la definizione:

- Positive : if ($Score_+ > Score_-$)
- Negative: if ($Score_- > Score_+$)
- Neutral: if ($Score_- = Score_+$)

Questa volt, però la funzione di scoring utilizzata è:

- Per i documenti positivi:

$$Score_+ = Val_{SWN+} + Per_{Agg+} + Per_{Wor+} + Occ_{AggAgg+} + Occ_{AggWor+}$$

- Per i documenti negativi:

$$Score_- = |Val_{SWN-}| + Per_{Agg-} + Per_{Wor-} + Occ_{AggAgg-} + Occ_{AggWor-}$$

Dove, appunto il ValSWN+ rappresenta la sommatoria dei vaolri ricavati per i termini positivi rispetto ai vaolri estrapolati da SentiWordNet.

3.4.3 Un esempio del processo avanzato

Possiamo rappresentare il nostro processo secondo lo schema seguente.

STEP 1

Si ottengono dei datasets elementari

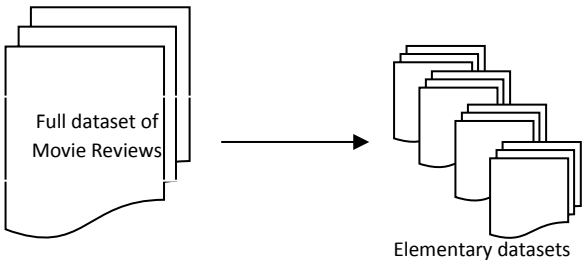


Figura 38 - Datasets elementari

Si costruiscono dei grafici elementari

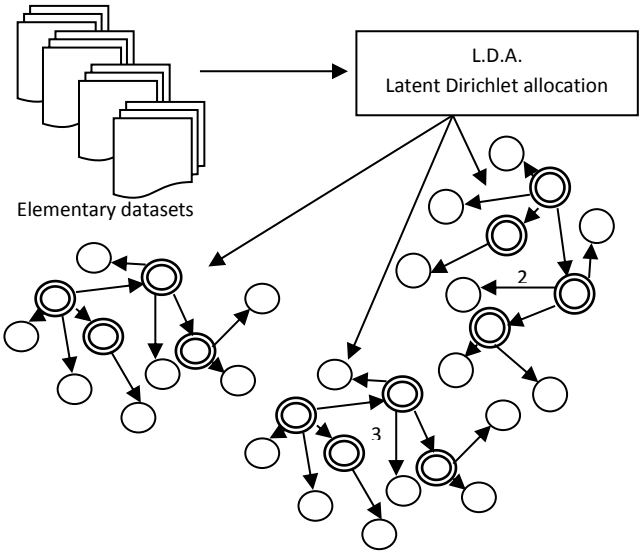


Figura 39 - Grafi elementari

Si utilizzano i grafi elementari per estrarre una tabella di termini aggregatori comuni tra i grafici caratterizzati dalla frequenza più elevata.

positivi		negativi	
1	**film**	1	**film**
2	play	2	time
3	good	3	sleep
…	…	…	…
n	nice	n	bad

Figura 40 - Tabella di aggregatori

Si evidenziano i termini aggregatori comuni all'interno di tutti i grafi (sia positivi che negativi)

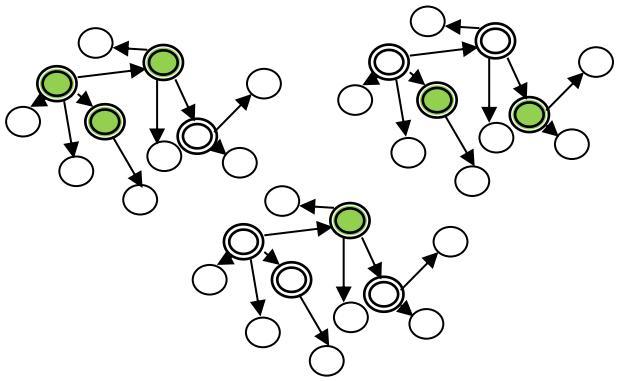


Figura 41 - Evidenza di termini aggregatori comuni

Si rimuovono i termini aggregatori comuni (facendo riferimento al lessico ottenuto a partire dal training set)

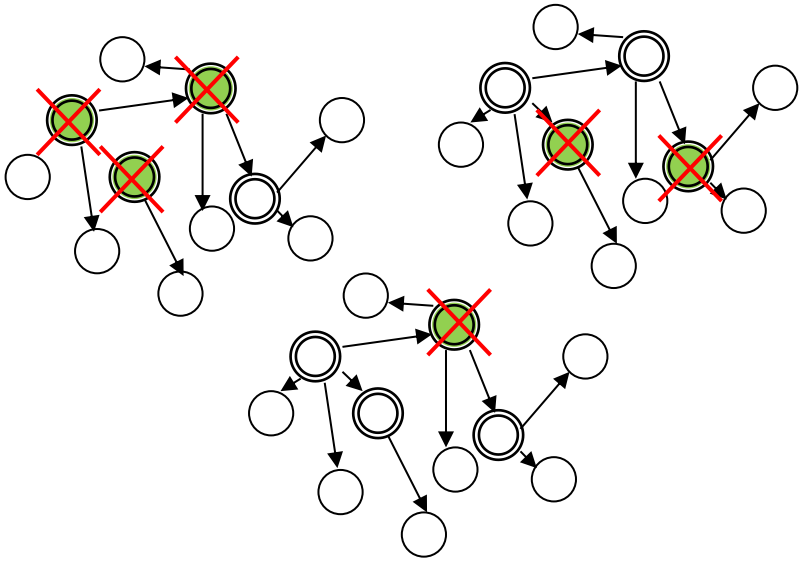


Figura 42 - Eliminazione termini aggregatori comuni

STEP 2

Abbiamo ottenuto un “modello” di grafi elementari caratterizzanti uno specifico dominio (recensioni, prodotti o altro).

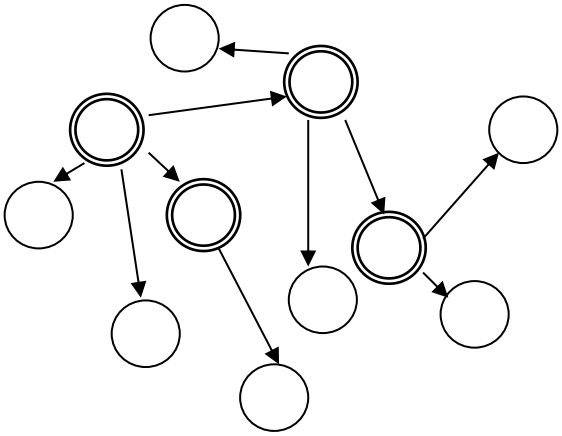


Figura 43 - Modello di grafo elementare

E una tabella contenente questi termini aggregatori

A questo punto passiamo al test set e costruiamo una nuova tabella per ogni commento (con termini aggregatori differenti) provvedendo a polarizzarne anche i temini mediante SentiWordNet (associando a ciascun termine i valori tra -1 e +1). (Abbiamo in precedenza effettuato la stessa operazione anche er i vettori rappresentativi semplificati dei documenti positivi e dei negativi ottenuti dal training set).

Figura 44 - SentiWordNet - Adjective

Movie Reviews Database					
		Pos.	Neutral	Neg.	kind of
1	good	0,75	0,25	0	adjective
2	big	0,25	0,625	0,125	adjective
…	…	…	…	…	…
n	enjoy	0,375	0,625	0	verb

Figura 45 - Tabella di aggregatori con analisi semantica

STEP 4

Calcoliamo i valori della funzione di score e valutiamo la polarizzazione. Qualora ciò non sia stato possibile poiché il valore è pari a zero (e non riusciamo quindi a ricavare la polarizzazione) applichiamo il processo iterativamente in modo da polarizzare gli altri termini (laddove necessario) presenti nel grafo elementare (tutti i temini non aggreganti).

	Level of the Concept		Value of Polarization
1	good		+0,75
2	big		+0,25
3		Bad	-0,625
4		enjoy	+0,625
5		great	0
6		worst	-0,25

Figura 46 - Valorizzazione semnatica degli aggregatori

Quarta parte

La sperimentazione e i risultati

4.1 Prima sperimentazione (Sottoinsieme del dataset)

Per questa prima fase sperimentale abbiamo scelto di utilizzare il dataset Movie Reviews provenienti dalla pagina ”Sentiment Analysis Datasets with Latent Explanation Initializations” della Cornwell University.

L’intenzione era quella di testare quantitativamente e non qualitativamente la’applicabilità del modello e abbiamo quindi considerato un numero ridotto di elementi del dataset di riferimento.

Del dataset iniziale contenente 1000 commenti positivi e 1000 commenti negativi già etichettati è stato considerato un training set di 100 documenti (50 già etichettati come positivi e 50 già etichettati come negativi) e un test set di 1000 documenti (500 positivi e 500 negativi) organizzati secondo lo schema seguente:

- Training Set
  - 50 Commenti positivi
  - 50 Commenti negativi
- Test Set
  - 500 Commenti positivi
  - 500 Commenti negativi

I risultati della prima fase sperimentale sono in linea con quelli ottenuti con mtodi analoghi in letteratura, sebbene i valori non siano comparabili proprio per la nostra scelta iniziale di considerare un dataset ridotto i cui componenti sono stati selezionati arbitrariamente.

Abbiamo affrontato un primo step costituito da una suddivisione dell'insieme di training iniziale in sottoinsiemi di 10 commenti ciascuno (tanto per l'insieme dei commenti classificati come positivi che per quelli classificati come negativi).

Successivamente abbiamo analizzato i nostri sottoinsiemi utilizzando iSoS Lite[189], un frawework sviluppato in Java e Java Server Pages, che include una versione personalizzata dell'API open source di Apache Lucene.

Nel primo step abbiamo utilizzato iSoSLite[190] come LDA per costruire i nostri grafi elementari.

Utilizzando iSoSlite possiamo modellare semanticamente il nostro insieme focalizzandosi sulla struttura delle relazioni associative tra parole e termini del linguaggio naturale e relazioni tra termini aggregatori e termini non aggregatori.[191,192,193].

Come riportato nelle pagine precedenti per i nostri esperimenti abbiamo adottato i Datasets Movie Reviews Sentiment Analysis with Latent Explanation Initializations in un format preprocessato per il funzionamento con SVM^{sle} package con inizializzazione di latent explanations.

Lo abbiamo scaricato al link:

http://www.cs.cornell.edu/~ainur/sle-data.html

I dati di Movie Reviews hanno le latent explanation initializations per utilizzare le OpinionFinder e le Annotator initializations. Le

[189] iSoSLite was developed by DIEI of University of Salerno in 2010
[190] iSoSlite is a web based application written in Java and Java Server Pages, iSoSlite, which includes a customized version of the API Open Source Apache Lucene, a full text retrieval engine. iSoSlite was developed in 2010 by DIEI – Engineering and Electronic Department of University of Study of Salerno.
[191] W. Kintsch. *The role of knowledge in discourse comprehension: A construction-integration model*. Psychological Review, 95:163{182, 1988
[192] M. C. Potter. *Very short term conceptual memory*. Memory & Cognition, (21):156{161, 1993
[193] K. A. Ericsson and W. Kintsch. *Long-term working memory*. Psychological Review., 102:211-245, 1995

OpinionFinder initializations sono state derivate utilizzando il word-level polarity classifier ” Opinion Finder”.

Abbiamo due cartelle contenenti ciascuna 1000 commenti relativi a film (in un caso negativi, nell’altro positivi).

Abbiamo cominciato la nostra sperimentazione costruendo i nostri datasets elementary a partire dai database classificati originali.

Abbiamo considerato i datasets elementari per 50 recensioni positive e altrettanti per recensioni negative.

Ciascuno dei datasets elementari contiene commenti secondo lo schema seguente:

- PD1 (Positive dataset 1) = pos. reviews 01-10
- PD2 = reviews 11-20
- PD3 = reviews 21-30
- PD4 = reviews 31-40
- PD5 = reviews 41-50

e

- ND1 (Negative dataset 1) = neg. reviews 01-10
- ND2 = reviews 11-20
- ND3 = reviews 21-30
- ND4 = reviews 31-40
- ND5 = reviews 41-50

Per utilizzare iSoS Lite dobbiamo inizializzarlo e per farlo abbiao utilizzato i seguenti parametri:

NUM_CONCEPTS=7

MAX_PAIRS=70

LDA_ALPHA=0.5

LDA_BETA=0.001

LDA_TOPICS=30

LDA_ITERATIONS=10000

K_MEANS_CP=5

K_MEANS_PP=5

K_MEANS_ITERATIONS=100

La struttura di ciascun commento è libera e non tokenizzata. Ad esempio:

> *films adapted from comic books have had plenty of success , whether they're about superheroes (batman , superman , spawn) , or geared toward kids (casper) or the arthouse crowd (ghost world) , but there's never really been a comic book like from hell before . for starters , it was created by alan moore (and eddie campbell) , who brought the medium to a whole new level in the mid '80s with a 12-part series called the watchmen . to say moore and campbell thoroughly researched the subject of jack the ripper would be like saying michael jackson is starting to look a little odd . the book (or " graphic novel , " if you will) is over 500 pages long and includes nearly 30 more that consist of nothing but footnotes . in other words , don't dismiss this film because of its source . if you can get past the whole comic book thing , you might find another stumbling block in from hell's directors , albert and allen hughes . getting the hughes brothers to direct this seems almost as ludicrous as casting carrot top in , well , anything ,*

but riddle me this : who better to direct a film that's set in the ghetto and features really violent street crime than the mad geniuses behind menace ii society ? the ghetto in question is , of course , whitechapel in 1888 london's east end . it's a filthy , sooty place where the whores (called " unfortunates ") are starting to get a little nervous about this mysterious psychopath who has been carving through their profession with surgical precision . when the first stiff turns up , copper peter godley (robbie coltrane , the world is not enough) calls in inspector frederick abberline (johnny depp , blow) to crack the case . abberline , a widower , has prophetic dreams he unsuccessfully tries to quell with copious amounts of absinthe and opium . upon arriving in whitechapel , he befriends an unfortunate named mary kelly (heather graham , say it isn't so) and proceeds to investigate the horribly gruesome crimes that even the police surgeon can't stomach . i don't think anyone needs to be briefed on jack the ripper , so i won't go into the particulars here , other than to say moore and campbell have a unique and interesting theory about both the identity of the killer and the reasons he chooses to slay . in the comic , they don't bother cloaking the identity of the ripper , but screenwriters terry hayes (vertical limit) and rafael yglesias (les mis ? rables) do a good job of keeping him hidden from viewers until the very end . it's funny to watch the locals blindly point the finger of blame at jews and indians because , after all , an englishman could never be capable of committing such ghastly acts . and from hell's ending had me whistling the stonecutters song from the simpsons for days (" who holds back the electric car/who made steve guttenberg a star ? ") .

don't worry - it'll all make sense when you see it . now onto from hell's appearance : it's certainly dark and bleak enough , and it's surprising to see how much more it looks like a tim burton film than planet of the apes did (at times , it seems like sleepy hollow 2) . the print i saw wasn't completely finished (both color and music had not been finalized , so no comments about marilyn manson) , but cinematographer peter deming (don't say a word) ably captures the dreariness of victorian-era london and helped make the flashy killing scenes remind me of the crazy flashbacks in twin peaks , even though the violence in

the film pales in comparison to that in the black-and-white comic . oscar winner martin childs' (shakespeare in love) production design turns the original prague surroundings into one creepy place . even the acting in from hell is solid , with the dreamy depp turning in a typically strong performance and deftly handling a british accent . ians holm (joe gould's secret) and richardson (102 dalmatians) log in great supporting roles , but the big surprise here is graham . i cringed the first time she opened her mouth , imagining her attempt at an irish accent , but it actually wasn't half bad . the film , however , is all good . 2 : 00 - r for strong violence/gore , sexuality , language and drug content

Analizziamo così ciascun dataset mediante la piattaforma iSoS ottenendo 2 set di 5 grafi ciascuno e corrispondenti due set di 50 commenti ciascuno.

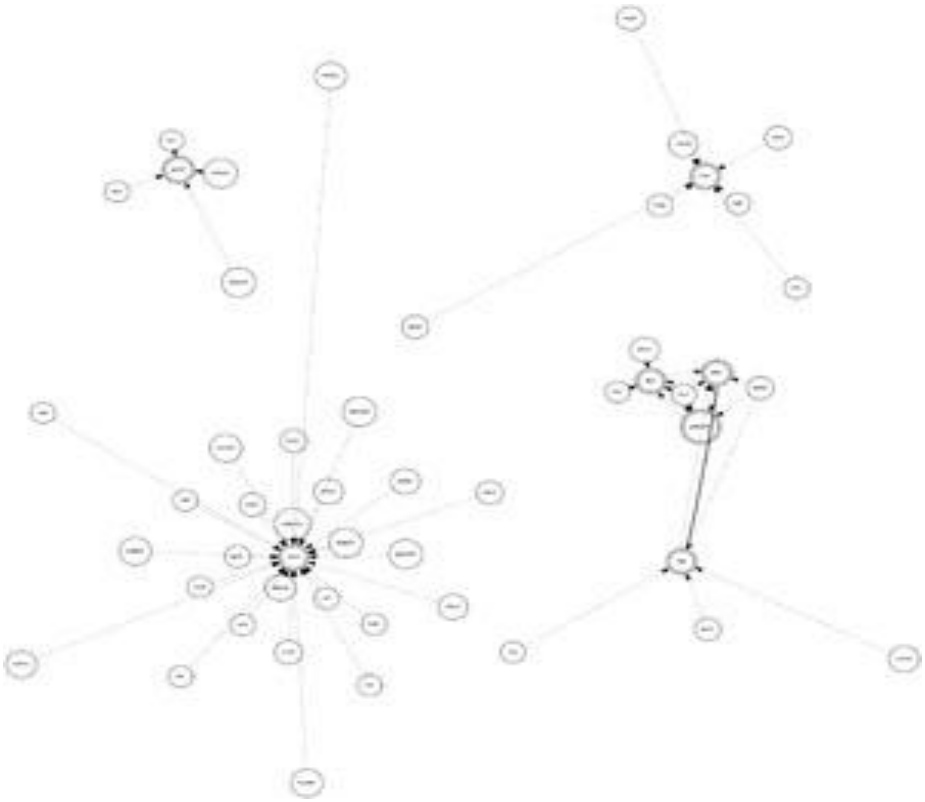
Un esempio di PD1 ottenuto è quello della tabella che segue

film	play	0.13701893
perform	time	0.062393293
make	perform	0.041859515
reminisc	movi	0.004769098
suppos	movi	0.008956215
show	time	0.014972816
eye	time	0.010344413
rest	movi	0.015507774
studio	movi	0.017573185
clear	movi	0.010109969
time	perform	0.062393293
lead	movi	0.01498134
play	perform	0.057584893
eye	perform	0.009889391
film	play	0.13701893
screen	movi	0.03281216
end	good	0.033437777
detroit	movi	0.0049439394
cost	movi	0.013013876
perform	play	0.057584893
didn't	movi	0.020316245
great	man	0.029829014
charact	good	0.06915402
live	good	0.017889373
star	time	0.022243071
watch	movi	0.036862325
make	film	0.34210178
men	man	0.013176498
genr	movi	0.008908458
time	play	0.05790719
money	movi	0.07638378
play	time	0.05790719
element	movi	0.021293838
virtual	man	0.0045812796
role	film	0.13378885
plot	movi	0.037044898

head	man	0.0073024607
fun	movi	0.021807557
comedi	movi	0.01886484
eye	play	0.009713686
town	movi	0.016405856
run	movi	0.020627879
scene	film	0.16153648
begin	man	0.017419962
general	movi	0.014847046
age	man	0.004115405

make	play	0.045755077
hand	movi	0.024839537
point	film	0.071417995
come	movi	0.014406486
babi	movi	0.012616211
interest	good	0.02185373
roll	movi	0.009424985
imag	man	0.0051560826
person	movi	0.017030358
effect	movi	0.010745475

Con i corrispondenti grafi elementari.



Nel corso del secondo step,poi, abbiamo classificato i termini più comuni ottenuti tra tutte le tabelle ottenute.

Un esempio delle tabelle rappresentative i sottoinsiemi è il seguente:

Di seguito un esempio con 5 tabelle positive

PD1	PD2	PD3	PD4	PD5
Man	Film	Movi	Film	Make
Play	Charact	Time	Movi	Movi

Perform	Make	Story	Make	Life
Time	Time	Scene	Time	Charact
Film	Movi	Make	Scene	Film
Movie	Play	Film	Effect	End
Good	Enjoy	Great	Great	Good

.

e 5 negative.

ND1	ND2	ND3	ND4	ND5
Film	Film	Movi	Film	Charact
Time	Play	Big	Thing	Work
Play	Make	World	Time	Time
Thing	Scene	Film	Charact	Movi
Movi	Charact	Charact	Star	Doesn.t
Year	Movi	Watch	Movi	Film
Bad	End	Work	Bad	Make

Al termine della sperimentazione sul training test abbiamo ottenuto una tabella contenente il lessico caratterizzante il nostro dominio sotto forma di termini aggregatori.

1	2	3	4	5	6	7	8
Film	Movi	Make	Make	Scene	Play	Charact	Thing

E possiamo utilizzare questa tabella con tuti i commenti del test set.

Consideriamo ora un primo esempio di commento del nostro test set

...with his successful books and movies , michael crichton is doing well .

with early successes with westworld (1973) and coma (1978) , and recent films such as jurassic park (1993) , his films have been entertaining .

however , he seems to taken a wrong turn somewhere with sphere .

this $100 million mess by good director barry levison (disclosure) is dull , long winded , and a huge disappointment .

considering the huge budget , the all star cast , and a story by crichton , ***sphere is majorly disappointing*** *.*

the film opens with norman goodman (hoffman) , a psychologist who thinks he is visiting an airplane crash to console the survivors .

however , when he arrives , he his told by supervisor barnes (peter coyote) that he is actually investigating an spacecraft .

along with goodman is mathematician harry adams (jackson) , biologist beth halperin (stone) and ted fielding (liev schrieber) they investigate the spaceship , find a massive sphere inside , meet an alien intelligence called jerry , and basically weird crap happens.

unfortunately , ***something went wrong along the way with sphere . the film starts off entertaining enough , but throughout this very long movie , it just gets sillier and sillier . the film jaunts along from scene to scene , never fully explaining what is going on . the actors and directing don't help*** *, either .*

hoffman is on autopilot (and almost seems embarrassed) throughout the movie , churning out dull lines , and probably wondering what the hell he is doing in this movie .

stone is useless , displaying no emotion , and fails to convince the audience that she has any feelings for hoffman *. the only person who seems to be having fun in this movie is jackson , who's funny as the mathematician who slowly goes crazy and entering the sphere .*

but he's hardly in it , and by the end of the film he is just as dull as hoffman and stone . the same goes for peter coyote , who hams it up as the officer , but is then killed off halfway through .

the director , barry levinson , who directed the better crichton adaptation disclosure (1994) messes up with the drama and the action . ***the drama scenes are , quite frankly , boring , and the action scenes suffer from overkill , with levison throwing the camera all over the place (much like the godawful speed 2 , 1997) the writing doesn't help much , either .***

although crichton is great with plots , he's terrible with dialogue , and practically every line in sphere is a dud . the speech is too simple *, i was hoping it would be a bit more intelligent .*

practically every line is just stating the obvious .

none of it is smart . also , where the hell did the budget go ? the sphere itself is impressive , and there's a few nice special effect shots , but where the $100 million went is anyone's guess . there's a giant squid attack in the picture , but not once does the audience see the squid , even though the film has a massive budget.

i assume the picture was trying to build up tension by not showing the squid , and if handled correctly it probably would .

but the whole scene is done badly , and i was just hoping we could see the stupid squid .

finally , the film has no idea what genre to be .

levison can't handle his own plot .

it leaps from hokey sci-fi , to horror , and finally the shining/event horizon psychological thriller .

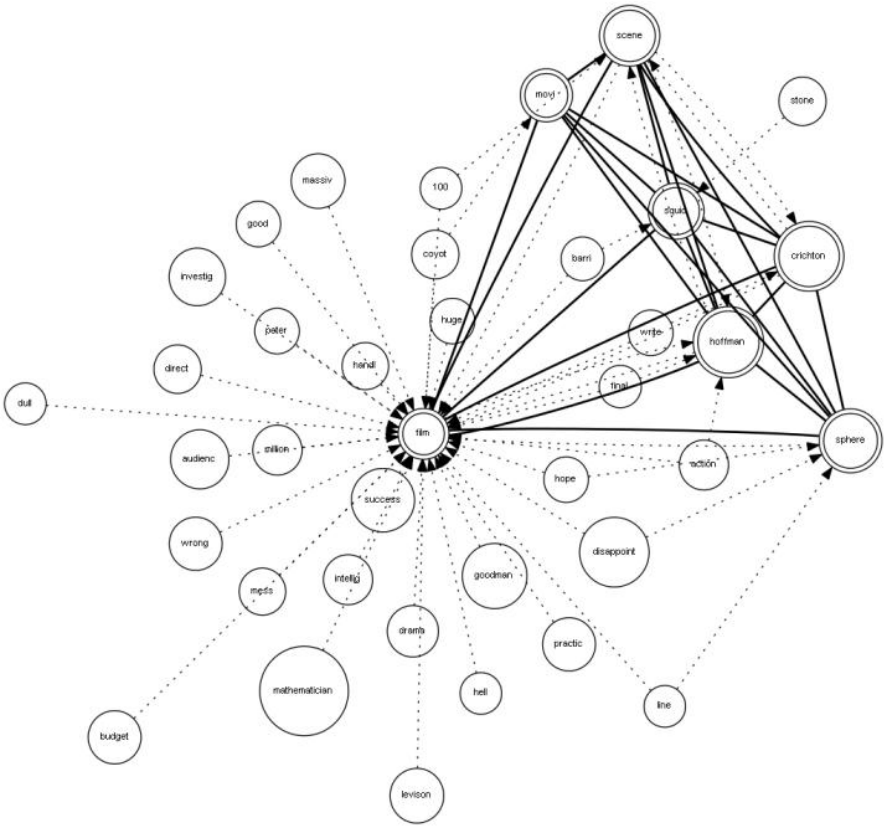
and , of course , the film is very much like the abyss (1988) , although in it's defense , crichton did write sphere before the abyss was released (and is far superior to this rubbish) it's not all that bad though . the plot is all right , there's a few jump scenes (although nothing very scary) and there's the occasionally interesting bit .

but overall , sphere is a big waste of some fine talent , a lot of money , and a potentially good movie .

not really worth seeing .

per il quale abbiamo provveduto anche ad evidenziare il testo chiaramente caratterizzante la polarizzazione negativa di questo comment.

Se trattiamo questo commento con iSoS otteniamo un grafo elementare di questo tipo.



Con questa tavola di termini aggregatori

Film
Movi
Squid
Sphere
Hoffman
Crichton

Applicando ricorsivamente lo stesso processo di costruzione dei grafi elementari e degli specifici vettori caratterizzanti a ciascuno dei

commenti del test set, poi, confrontiamo questi vettori a quelli ricavati dal training set.

Caiscun commento sarà quindi caratterizzato da uno score positivo (la la funzione di score calcolata per il vettore positivo) e da uno negativo.

Il confronto tra questi due valori di score ci consente di identificare lo score del commento.

Calcolo del sentiment per i commenti negativi

Num.	Nome del file	Valore di score positivo	Valore di score negativo	Sentiment del file
1	520-539.txt	0	0	NON VALUTABILE
2	500-519.txt	0	2	NEGATIVO
3	540-559.txt	0	1	NEGATIVO
4	560-579.txt	0	0	NON VALUTABILE
5	580-599.txt	1	1	NON VALUTABILE
6	600-619.txt	1	0	POSITIVO
7	620-639.txt	0	2	NEGATIVO
8	640-659.txt	0	0	NON VALUTABILE
9	660-679.txt	1	1	NON VALUTABILE
10	680-699.txt	2	0	POSITIVO

Figura 47 - Generazione dei vettori caratterizzanti

L’ambiente sperimentale è stato costruito su piattaforma PHP - MySQL. Si tratta, quindi di un ambiente web che può essere fruito da dispositivi che operano su diverse piattaforme.

In particolar modo si sono costruite le tabelle su DB MySql al solo scopo di contenere opportunamente i vettori caratterizzanti i domini e caratterizzanti mentre sono stati predisposti script PHP che consentono il caricamento del dataset per la generazione di detti vettori e lo specifico commento del quale valutare di volta in volta il sentiment.

Figura 48 - Primo ambiente sperimentale

I risultati sperimentali di questa prima fase hanno evidenziato dei valori soddisfacenti così come riportato nella tabella che segue.

	Positive	Negative	Non classificati	Totali
Positive Test set	365	120	15	500
Negative Test set	85	410	5	500
Precision	0,7923			
Recall	0,7904			

Figura 49 - Risultati prima sperimentazione

Si tratta di risultati in linea con quelli presenti in letteratura per sperimentazioni analoghe.

Autore	Metodo	Dataset	Performance
Pang and Lee '02	Naive Bayes, maximum entropy class. support vector machines	IMDb	82,90%
Kennedy and Inkpen '06	support vector machines, termcounting method, a combination of the two	IMDb	enhanced combined method: 86.2%
Zhou and Chaovalit '08	ontology-supported polarity mining	IMDb	72,2
Melville et al. '09	Bayesian classification with lexicons and training documents	IMDb	81,42
LDA for SA (2012)	LDA	IMDb	Precision 79,23% Recall 79,04% Accuracy 77,50%

Figura 50 - Confronto con altri metodi

Come già accennato, a comunque evidenziato che nella tabella presentata nella figura sovrastante i valori sono utili per valutare il solo andamento quantitativo e non qualitativo. Infatti i risultati sperimentali relativi alle altre metodologie della letteratura prendono in considerazione l'intero dataset e non un sottoinsieme come nel nostro caso.

Per questo motivo è stato opportuno passare alla successiva sperimentazione che tenesse conto diquesto aspetto.

4.2 Seconda sperimentazione (Dataset completo e ipotesi SWN3)

Allo scopo di superare i limiti della prima sperimentazione derivanti dall'utilizzo di un sottoinsieme limitato del dataset sperimentale, in questa seconda fase della nostra attività abbiamo utilizzato il dataset considerato nella sua interezza.

In questo modo abbiamo potuto confrontare le prestazioni e tutta l'operatività del modello.

In particolare sono stati considerati un training set costituito da 100 commenti etichettati come positivi e altrettanti etichettati come negativi e un test set che ha considerato tutti quelli disponibili (1000 etichettati come positivi e 1000 etichettati come negativi).

Come nella prima sperimentazione abbiamo utilizzato iSoS Lite applicandolo al training set in modo da ottenere i vettori caratterizzanti tanto il dominio nella sua interezza quanto i commenti polarizzati quelli negativamente e quelli polarizzati positivamente.

Il primo passaggio della seconda sperimentazione, quindi, è di fatto identico a quello della prima, a meno della differenza di dimensione del training set stesso.

Un volta ottenuti i tre vettori rappresentativi si sono però applicate le due nuove ipotesi sperimentali.

Come visto nell'esempio di progetto avanzato dei paragrafi precedenti, abbiamo, in primo luogo, ipotizzato che non tutti i termini contenuti all'interno dei vettori fossero realmente rappresentativi o polarizzanti.

In questo modo abbiamo proceduto a una semplificazione degli stessi andando ad eliminare, ad esempio, i termini contenuti tanto nella polarizzazione positiva, quanto nella polarizzazione negativa.

Come nella prima sperimentazione anche in questo caso abbiamo ottenuto le mGT rappresentative di ciascun sottoinsieme di commenti del training set.

L'unica differenza, in questo caso, è che i sottoinsieme considerati non contengono più 10 commenti ciascuno, ma 20 (per un totale, appunto, di 5 gruppi da 20 commenti e quindi 100 commmenti complessivi tanto nel caso positivo, quanto nel caso negativo).

Di seguito un esempio con 5 tabelle positive e 5 negative.

PD1	PD2	PD3	PD4	PD5
Man	Film	Movi	Film	Make
Play	Charact	Time	Movi	Movi
Perform	Make	Story	Make	Life
Time	Time	Scene	Time	Charact
Film	Movi	Make	Scene	Film
Movie	Play	Film	Effect	End
Good	Enjoy	Great	Great	Good

.

ND1	ND2	ND3	ND4	ND5
Film	Film	Movi	Film	Charact
Time	Play	Big	Thing	Work
Play	Make	World	Time	Time
Thing	Scene	Film	Charact	Movi
Movi	Charact	Charact	Star	Doesn.t
Year	Movi	Watch	Movi	Film
Bad	End	Work	Bad	Make

In questo caso abbiamo già provveduto a identificare i termini di polarizzazione, evidenziandoli (sono quelli comuni alle tabelle di analoga polarizzazione).

Per questi valori noi consideriamo i termini comuni come caratterizzanti il nostro specifico dominio (delle recensioni di films).

1	2	3	4	5	6	7	8
Film	Movi	Make	Make	Scene	Play	Charact	Thing

Le due tabelle "alleggerite" contenenti i termini caratterizzanti la polarizzazione positive e quella negative (riportiamo sempre l'esempio per sole 5 tabelle)

PD1	PD2	PD3	PD4	PD5
Man	Enjoy	Story	Effect	Life
Perform		Great	Great	End
Good				Good

ND1	ND2	ND3	ND4	ND5
Year	End	Big	Star	Doesn.t
Bad		World	Bad	
		Watch		

Quale seconda ipotesi di lavoro abbiamo introdotto un livello semantico, andando ad attribuire il relativo valore ricavato da Sentiwordnet per ciascuno dei termini caratterizzanti ricavati.

Adottando il nostro metodo,poi, possiamo innanzitutto cancellare i termini aggregatori caratterizzanti il nostro dominio e successivamente possiamo assegnare il valore numerico a ciascun termine aggregante (a partire da SentiWordNet) e ottenere (con una somma algebrica) la polarizzazione di questo commento.

Effettuiamo una verifica di funzionalità elementare sugli stessi sottoinsiemi di commento utilizzati per il training set.

Film	deleted
Movi	deleted
Squid	0
Scene	deleted
Sphere	0
Hoffman	Not value
Crichton	Not value

Ma solo per questi termini aggregatori la nostra somma è pari a zero ed è quindi necessario considerare anche i termini non aggregatori (sempre al netto dei termini caratterizzanti il dominio).

Questi i risultati:

action	-0,125
audienc	0,125
barri	0

budget	0
coyot	0
direct	-0,25

disappoint	-0,25
drama	0,5
dull	-0,375
final	-0,625
good	0,75
goodman	0
handl	-0,25
hell	-0,375
hope	0,25
huge	-0,125
intellig	0,25
investig	0
levison	0
line	-0,5

massiv	0
mathematician	0
mess	-0,125
million	-0,125
peter	-0,125
practic	0
stone	0,125
success	0,125
write	0
wrong	0,75
TOTAL	-0,375<0

Anche in questo caso (per un singolo comment del nostro test test) la polarizzazione , ottenuta mediante il nostro metodo, è verificata.

PD1	PD2	PD3	PD4	PD5
0	0	0	-0,25	0
0		0,75	0,75	0
0,75				0,75
TOT=0,75>	TOT=0,375	TOT=0,75>	TOT=0,50>	TOT=0,75>

0	>0	0	0	0
POSITIVE	POSITIVE	POSITIVE	POSITIVE	POSITIVE

Nel caso dei datasets positivi tutte le polarizzazioni sono verificate.

ND1	ND2	ND3	ND4	ND5
0	0	0,25	0,125	Doesn.t= does+not
0,625		-0,375	-0,625	0-0,625
		0		
TOT=0,625 <0	TOT=0	TOT= -0,125 <0	TOT= -0,50 <0	TOT= -0,625 <0
NEGATIV.	NOT VAL	NEGATIV.	NEGATIV.	NEGATIV.

Per quelle negative nell'esempio di sole 5 tabelle considerate sono verificate 4 su 5 polarizzazioni.

Per il dataset ND2 abbiamo la necessità di considerare anche gli altri termini oltre a quelli aggreganti.

Lo schema dei valori corrispondenti per la tabella ND2 è il seguente:

film	play	0.001401037
charact	make	7,58E+03
charact	time	0.15335715

film	time	0.005458816
make	time	0.0015305853

film	good	0.049190193
film	charact	0.0014529874
charact	movi	0.006188647
make	good	4,80E+03
movi	time	3,45E+03
play	movi	0.0022324596
film	movi	8,27E+03
film	make	0.00903131
charact	good	0.0047535864
play	make	6,99E+01
movi	good	0.06213597
play	time	3,48E+02
play	good	0.0544198
charact	play	0.019798456
good	time	4,71E+03
movi	make	0.0034514382
life	good	0.015870364
star	play	0.030370068
play	good	0.0544198
actor	play	0.03658688

pretti	make	0.03067402
flick	movi	0.027436636
partner	play	0.012254573
suppos	good	0.012045436
act	movi	0.046772372
night	play	0.033367075
talk	make	0.029678483
turn	play	0.028924495
piec	make	0.042884707
perform	film	0.16651271
director	good	0.015930064
redempt	make	0.023858277
can't	time	0.017943539
we'r	good	0.004580008
end	charact	0.07686126
movi	good	0.06213597
happen	good	0.011375908
stori	charact	0.13029324
problem	play	0.027792422

night	good	0.008300829
impress	play	0.010024479
time	charac t	0.15335715
guy	movi	0.15996572
good	play	0.0544198
script	charac t	0.045967206

can't	charac t	0.044420853
charact	time	0.15335715
actor	good	0.009731259
peopl	make	0.07749928

Dobbiamo cancellare i termini comuni ricavati dalla tabella rappresentativa dello specifico dominio

Term	Value from SentiWord Net
act	0
actor	0
can't	-0,625
director	0
end	0
flick	0
good	0,75
guy	0
happen	-0,375
impress	0
life	0
night	-0,375
partner	0
peopl	0
perform	0
piec	0
pretti	-0,25
problem	-0,625
redempt	0
script	0
star	0,125
stori	-0,375
suppos	0,125
talk	-0,125
turn	0
we'r	0
SUM	-1,75

E anche in questo caso (negative) la polarizzazione è verificata

Verificata la funzionalità di massima del metodo, abbiamo passato in rassegna l'intero test test con risultati ottimali e un notevole incremento di prestazioni rispetto al caso della prima sperimentazione.

L'ambiente di sperimentazione è stato realizzato ampliando le funzionalità di quello della prima sperimentazione. Anche in questo caso, quindi, si tratta di un ambiente web based PHP-MySQL.
Sono stati modificati gli script per il calcolo della funzione di score (essendo la stessa modificata rispetto al caso precedente) e sono state inserite le nuove funzionalità di interrogazione al motore SentiWordNet per ricavare i parametri di polarizzazione di ciascuno dei temini contenuti nei vettori rappresentativi.

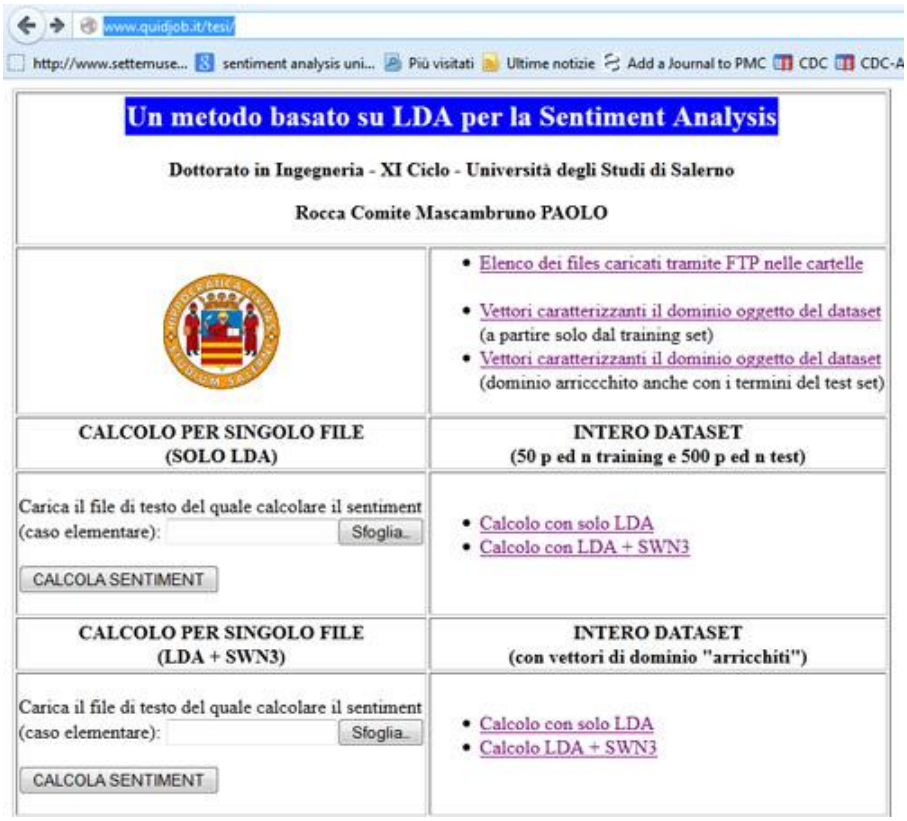


Figura 51 - Ambiente della seconda sperimentazione

Inoltre, allo scopo, di monitorare anche i tempi di esecuzione e valutare, quindi, le performances anche in questo ambito, abbiamo

inserito anche degli script utili a mstrare i dati relativi a questi tempi sulle pagine di ciascuna attività.

Effettuando una stima dei parametri di perfomrances sull'intero test set otteniamo i seguwnti valori

Il caricamento della pagina ha impiegato:10.553 secondi.

Sulla base di queste definizioni e sulla sperimentazione effettuata i nostri valori di precision e recall sono:

- **Precision = 0.98076923076923**
- **Recall = 0.98**

Una ulteriore miglioria che abbiamo applicato in questa seconda sperimentazione è relativa al fatto che i vettori rappresentativi el dominio possono essere progressivamente arricchiti a partire da i termini caratterizzanti la polarizzazione che si ricavano progressivamente dalla stessa sperimentazione.

Facendo riferimento al solo training set questi sono i valori rappresentativi del dominio

Rappresentazione del vettore caratterizzante il dominio oggetto dell'analisi.

Vettore di dominio: Vettore dei termini (comuni alle mGT positive e a quelle negative) che caratterizzano il dominio ma non la polarizzazione

1	audienc	**2**	charact	**3**	end	**4**	film	**5**	good
6	life	**7**	love	**8**	make	**9**	man	**10**	movi
11	perform	**12**	play	**13**	scene	**14**	stori	**15**	thing
16	time	**17**	work	**18**	year				

Vettore caratterizzante POSITIVO

1	action	**2**	american	**3**	day	**4**	director	**5**	enjoy
6	feel	**7**	find	**8**	great	**9**	live	**10**	people
11	show	**12**	success	**13**	world				

Vettore caratterizzante NEGATIVO

1	act	**2**	actor	**3**	bad	**4**	big	**5**	dont
6	line	**7**	plot	**8**	problem	**9**	surpris	**10**	real
11	run	**12**	star	**13**	watch				

Se invece andiamo a considerare anche i termini ricavati dalla sperimentazion sul test set il nostro dominio si caratterizzerà con nuovi vettori.

Alcuni dei termini che in precedenza caratterizzavano un orientamento (positivo o negativo) sono adesso caratterizzanti il solo dominio (come ad esempio il termine "**dont=don't**")-

Vettore di dominio: Vettore dei termini (comuni alle mGT positive e a quelle negative) che caratterizzano il dominio ma non la polarizzazione

1	back	**2**	charact	**3**	director	**4**	dont	**5**	end
6	film	**7**	find	**8**	good	**9**	love	**10**	make
11	movi	**12**	people	**13**	play	**14**	scene	**15**	stori

16	thing	**17**	time	**18**	watch	**19**	work		

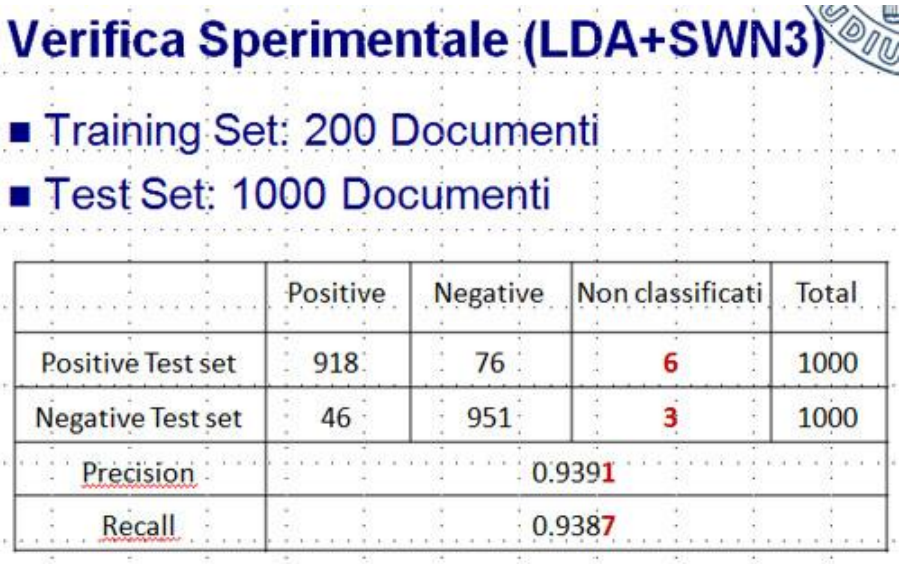
Vettore caratterizzante POSITIVO

1	action	**2**	american	**3**	day	**4**	enjoy	**5**	feel
6	great	**7**	life	**8**	live	**9**	made	**10**	perform
11	show	**12**	success	**13**	world				

Vettore caratterizzante NEGATIVO

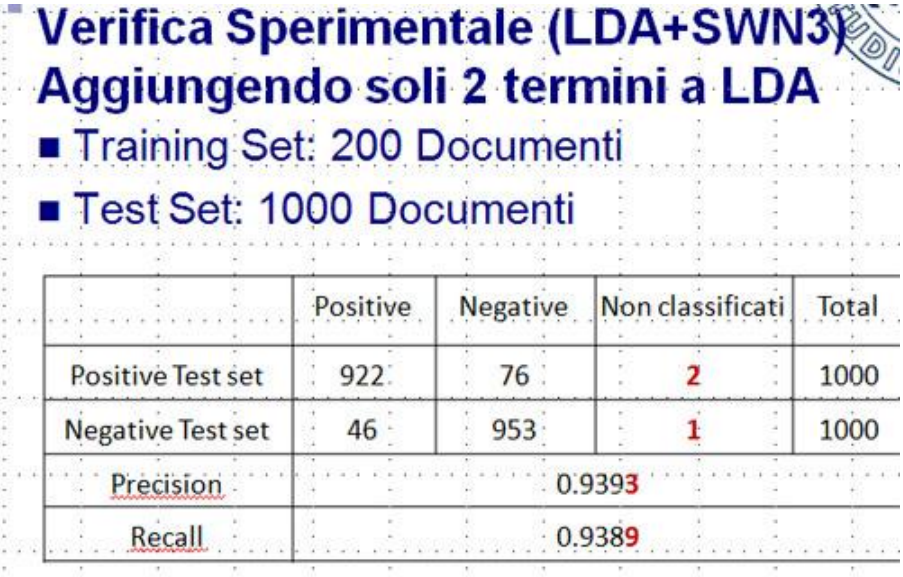
1	act	**2**	actor	**3**	bad	**4**	big	**5**	cast
6	doesnt	**7**	line	**8**	minut	**9**	plot	**10**	point
11	problem	**12**	real	**13**	role	**14**	run	**15**	set
16	star	**17**	surpris	**18**	turn	**19**	year		

Apportando questa ulteriore miglioria le prestazioni del nostro ambiente sperimentale subiscono un ulteriore incremento di performances andando a superare quasi tutti i risultati delle differenti metodologie presenti in letteratura secondo le due immagini seguenti:



	Positive	Negative	Non classificati	Total
Positive Test set	918	76	6	1000
Negative Test set	46	951	3	1000
Precision	0.9391			
Recall	0.9387			

Figura 52 - Risultati sperimentazione vettore "puro"



	Positive	Negative	Non classificati	Total
Positive Test set	922	76	2	1000
Negative Test set	46	953	1	1000
Precision	0.9393			
Recall	0.9389			

Figura 53 - Risultati sperimentazione vettore con aggiunta termini

Andando inoltre a caratterizzare anche i tempi di computazione e i relativi ordini di grandezza si ottengono dei risultati nche al di sopra di

quanto ipiotizzato circa la qualità della metoologia nella prima sperimentazione.

Abbiamo quindi potuto procedere a un'analisi dei risultati ottenuti e un confronto con tutti gli altri metodi secondo la tabella di seguito riportata.

Autore	**Metodo**	**Dataset**	**Performance**
Pang and Lee '02	Naive Bayes, maximum entropy class. support vector machines	IMDb	82,90%
Kennedy and Inkpen '06	support vector machines, termcounting method, a combination of the two	IMDb	enhanced combined method: 86.2%
Zhou and Chaovalit '08	ontology-supported polarity mining	IMDb	72,2
Melville et al. '09	Bayesian classification with lexicons and training documents	IMDb	81,42
LDA for SA (2012)	LDA+SWN 3.0	IMDb+MR	Precision 93,93% Recall 93,89% Accuracy 93,75%

Per quello che riguara i tempi di esecuzione, il calcolo completo di tutti i parametri di performances sull'intero tet set (ricordiamo essere composto da 2000 commenti (1000 positivi e 1000 negativi) impiega **8.721 secondi.**

4.2.1 Validazione dei risultati sperimentali

Le misure utilizzate per il calcolo delle performances del nostro ambiente sperimentale sono quelle solitamente utilizzate per la valutazione delle performances degli ambient sperimentali.

Si tratta di *precision* (*P*), *recall* (*R*), ed *F-measure* (*F*).

Precisione e Recall sono due comuni classificazioni statistiche, utilizzate in diversi ambiti del sapere, come per es. l'information retrieval. La precisione può essere vista come una misura di esattezza o fedeltà, mentre la recall è una misura di completezza.

Nell'Information Retrieval, la precision è definita come il numero di documenti attinenti recuperati da una ricerca diviso il numero totale di documenti recuperati dalla stessa ricerca, e la recall è definita come il numero di documenti attinenti recuperati da una ricerca diviso il numero totale di documenti attinenti esistenti (che dovrebbe essere stato recuperato)[194].

$$Precision = \frac{\#Classes\ found\ and\ correct}{Total\ class\ found}$$

$$Recall = \frac{\#Classes\ found\ and\ correct}{Total\ class\ correct}$$

Altrimenti detta, la precision è la frazione di documenti attinenti che sono stati trovati, mentre la recall è la frazione di di documenti trovati che sono attinenti. Dalla definizione, è possibile intuire che precisione e recall sono grandezze inversamente proporzionali: maggiore è la precision in una ricerca, minore sarà la recall, e viceversa. Ne consegue dunque che motori di ricerca "perfetti", cioè che ritrovino tutti e soli documenti pertinenti ad una particolare ricerca, non sono possibili.

[194] Powers, David M W (2007/2011). "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies* **2** (1): 37–63

La $F_{measure}$[195] è una media armonica e rappresenta la combinazione di precision e recall. Può essere calcolata secondo la formula seguente:

$$F_{measure} = \frac{2 \times P \times R}{P + R}$$

Infine l'accuratezza può essere calcolata come segue

$$Accuracy = \frac{TP \ + TN}{TP \ + \ FP \ + FN \ + TN}$$

[195] K. Nigram, A. K. Maccallum, S. Thrun & T. Mitchell (2000) *Transductive Text Classification from Labeled and Unlabeled Document using EM*. In: Machine Learning. 39(2/3). pp. 103-134

4.3 Terza sperimentazione (Altri datasets)

Quale elemento di ulteriore completezza per la nostra attività sperimentale, avendo esaurito la nostra analisi sullo spcifico dataset delle recensioni cinematografiche Movie Reviews della Cornwell University abbiamo esteso la nostra ttività sperimentale anche ad alcuni degli altri datasets utilizzati in letteratura.

In particolare per quest'utlima camapagna sperimentale abbiamo applicato le nostre metodologie e il nostro ambiente ai seguenti datasets:

- **Subjectivity dataset v1.0**[196] utilizzato in Bo Pang and Lillian Lee, "*A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*", Proceedings of the ACL, 2004..
- **Blog author gender classification data set**[197] utilizzato in Arjun Mukherjee and Bing Liu. "Improving Gender Classification of Blog Authors." *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*. Oct. 9-11, 2010, MIT, Massachusetts, USA..

Il **Subjective dataset v 1.0** si caratterizza per la presenza di 5000 frasi etichettate come soggettive ed altrettante etichettate come oggettive. In particolare nel loro lavoro Bo Pang e Lillian Lee sono stati in grado di costruire un grande corpus di frasi automaticamente etichettate come soggettive e oggettive.

Per raccogliere frasi soggettive hanno raccolto 5000 frammenti di MovieReview (ad esempio, "audace, fantasioso, e impossibile da resistere") da http://www.rottentomatoes.com.

Per ottenere invece i dati oggettivi, hanno considerato 5.000 frasi provenienti dai riassunti delle trame dei films disponibili dall'Internet Movie Database (http://www.imdb.com).

[196] Il dataset può essere scaricato al link: http://www.cs.cornell.edu/people/pabo/movie-review-data/ Last accessed 25/02/2013

[197] Il dataset può essere scaricato al link: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html Last accessed 25/02/2013

Sono stati considerati solo frasi selezionati o frammenti lunghi almeno una decina di parole e tratti da recensioni o sintesi della trama di film usciti dopo il 2001, allo scopo di evitare sovrapposizioni con gli altri dataset di polarità.

Nell'ambito del **Blog author gender classification data set**, invece, per mantenere il problema della classificazione di genere relativa al testo informale il più generale possibile, Arjun Mukherjee and Bing Liu hanno raccolto post sul blog da molti blog hosting e motori di ricerca per blog, ad esempio, http://www.blogger.com , http://www.technorati.com etc.

Il dataset è composto da 3100 blog. Ogni blog è etichettato con il genere del suo autore. Il genere dell'autore è stato determinato visitando il profilo dell'autore. Laddove disponibili sono stati utilizzati foto, profilo o avatar associato per individuare il genere quando il sesso non era disponibile in modo esplicito.

Per garantire la qualità dell'operazione di etichettatura, un gruppo di studenti ha raccolti i blog e ha fatto l'etichettatura iniziale e l'altro gruppo ha controllato due volte le etichette visitando le pagine del blog effettivi.

Dei 3100 post individuati, 1588 (51,2%) sono stati scritti da uomini e 1512 (48,8%) sono stati scritti da donne. La durata media di ciascun messaggio è di 250 parole per gli uomini e 330 parole per le donne.

Al fine di attivare la sperimentazione per ciascuno di questi datasets sono stati costruiti, secondo gli schemi utilizzati nell sperimentazioni precedenti il training set e il test set.

In particolare, per il **Subjectivity dataset v1.0** il training set è stato costruito prelevando in maniera aleatoria 1000 commenti soggettivi e 1000 commenti oggettivi.

Il test set, invece ha considerato tutti i rimanenti commenti (4000 soggettivi e 4000 oggettivi).

Per quanto riguarda i vettori caratterizzanti dominio e polarizzazione "soggettiva" e "oggettiva" abbiamo ottenuto i seguenti risultaiti:

1. **Vettore di dominio**: Vettore dei termini (comuni alle mGT soggettive e a quelle oggettive) che caratterizzano il dominio ma non la polarizzazione

1	film	2	life	3	love	4	stori

2. **Vettore caratterizzante SOGGETTIVITA'**

1	charact	2	end	3	film	4	life	5	love
6	make	7	movi	8	stori	9	time		

3. **Vettore caratterizzante OGGETTIVITA'**

1	film	2	find	3	friend	4	life	5	live
6	love	7	stori	8	year				

N.B.: In giallo sono evidenziati i termini non caratterizzanti che sono comunque presenti nei vettori caratterizzanti. Nell'ambito della sperimentazione, allo scopo di ottimizzare la stessa, tali termini non vengono considerati.

I risultati di questa sperimentazione, relativamente ai parametri di performance e ai tempi di esecuzione sono riportati nella tabella seguente:

I singoli valori calcolati sono pari a

- P+ = Soggettivi correttamente identificati = 3862
- P- = Soggettivi presi per oggettivi = 92
- N+ = Oggettivi presi per soggettivi = 128
- N- = Oggettivi correttamente identificati = 3831
- NVP = Non valutabili (soggettivi) = 46
- NVN = Non valutabili (oggettivi) = 41

Ricordiamo ora i valori di Precision e Recall:

- Precision = (P+/(P+ + N+) + N-/(N- + P-)) / 2

- Recall = (P+/(P+ + P-) + N-/(N- + N+)) / 2

Sulla base di queste definizioni e sulla sperimentazione effettuata i nostri valori di precision e recall sono:

- **Precision = 0,972234**
- **Recall = 0,972201**

Il tempo necessario alla calcolo delle polarità per tutti i commenti positivi del test set è stato di **6,712 secondi** mentre quello per i commenti negativi è stato di **8,013 secondi.**

Nel caso della sperimentazione applicata al **Blog author gender classification data set**, invece, si è proceduto innanzitutto a rendere lo stesso simmetrico considerando un numero pari a 1500 di commenti per gli uomini e altrettanti per le donne. I commenti che non sono stati considerati nella sperimentazione perché eccedenti il numero di 1500 (88 per gli uomini e 12 per le donne) sono stati scelti in maniera del tutto aleatoria.

Il training set è stato considerato costituito da 200 commenti per gli uomini e altrettanti per le donne, mentre il test set è stato costruito con i restanti 1300 commenti per gli uomini e altrettanti per le donne.

Per quanto riguarda i vettori caratterizzanti dominio e polarizzazione "soggettiva" e "oggettiva" abbiamo ottenuto i seguenti risultaiti:

4. **Vettore di dominio**: Vettore dei termini (comuni alle mGT soggettive e a quelle oggettive) che caratterizzano il dominio ma non la polarizzazione

1	**make**	**2**	**work**	**3**	**time**

5. **Vettore caratterizzante i commenti di UOMINI**

1	**day**	**2**	**make**	**3**	**people**	**4**	**separ**	**5**	**time**	**6**	**work**

6. **Vettore caratterizzante OGGETTIVITA'**

1	city	2	island	3	make	4	time	5	women	6	work

N.B.: In giallo sono evidenziati i termini non caratterizzanti che sono comunque presenti nei vettori caratterizzanti. Nell'ambito della sperimentazione, allo scopo di ottimizzare la stessa, tali termini non vengono considerati.

I risultati di questa sperimentazione, relativamente ai parametri di performance e ai tempi di esecuzione sono riportati nella tabella seguente:

I singoli valori calcolati sono pari a

- P+ = Maschili correttamente identificati = 1216
- P- = Maschili presi per femminili = 63
- N+ = Femminili presi per maschili = 132
- N- = Femminili correttamente identificati = 1088
- NVP = Non valutabili (maschili) = 21
- NVN = Non valutabili (femminilo) = 80

Ricordiamo ora i valori di Precision e Recall:

- Precision = (P+/(P+ + N+) + N-/(N- + P-)) / 2
- Recall = (P+/(P+ + P-) + N-/(N- + N+)) / 2

Sulla base di queste definizioni e sulla sperimentazione effettuata i nostri valori di precision e recall sono:

- **Precision = 0,923671**
- **Recall = 0,921273**

Il tempo necessario alla calcolo delle polarità per tutti i commenti positivi del test set è stato di **5,913 secondi** mentre quello per i commenti negativi è stato di **6,019 secondi.**

Quinta parte

Conclusioni, scenari di utilizzo e prospettive

La progressiva diffusione dei social network, sia generalisti (Twitter, Facebook, Google+) che specializzati (Linkedin, Viadeo), ha reso disponibili una massiccia e inedita quantità di dati sulle preferenze e sulle opinioni degli utenti. Gli stessi, infatti, lasciano traccia delle proprie preferenze, dei pareri relativi a personaggi, eventi piuttosto che relativi a prodotti commerciali.

Questo è lo scenario in cui si colloca il focus dell'attività di ricerca che ha come obiettivo lo sviluppo e la validazione di una metodologia per la determinazione del "sentiment" contenuto all'interno di "post" o commenti scritti in linguaggio naturale. Si è investigato circa la possibilità di estrapolare, quanto più possibile in maniera automatica, gli orientamenti degli utenti in genere (piuttosto che specificamente di consumatori) relativamente ad alcuni aspetti proprio a partire da stringhe di testo da loro prodotte.

La classificazione automatica del Sentiment consente, infatti, di rispondere a molteplici esigenze o applicazioni quali l'analisi del gradimento di un soggetto politico (emblematico è il caso delle recenti consultazioni americane che ha visto un largo impiego di queste tecniche), oppure la percezione di un prodotto o servizio e delle sue caratteristiche da parte dei potenziali fruitori. Numerosi, quindi sono gli ambiti applicativi e i possibili utilizzi sia in ambito commerciale, ad esempio con la possibilità di costruire campagne di marketing basate sui risultati di questa analisi preventiva, sia in ambito politico e sociale con il monitoraggio continuo dei coefficienti di gradimento di una determinata corrente politica, una coalizione o un esponente di governo.

Nelle pagine che precedono, dopo la definizione degli obiettivi e l'analisi dello scenario e degli ambiti applicativi possibili in modo da

motivare la scelta del percorso di ricerca, si è prodotta una rassegna di definizioni e una disamina della nomenclatura adottata in questi ambiti.

Successivamente è stata presentata una descrizione delle problematiche relative alla Sentiment Analysis e ai numerosi ambiti applicativi possibili con un'attenta analisi dello stato dell'arte in materia di classificazione del sentiment.

In particolare proprio in considerazione della ingente mole di dati relativi alle preferenze e ai comportamenti degli utenti e soprattutto alla differente tipologia di rappresentazione degli stessi (si tratta di files di testo, di pagine web, di raccolte video, di audio ambientali, di registrazioni del parlato singolo e così via) si è deciso di circoscrivere l'ambito di ricerca e l'identificazione di un metodo efficace solo a documenti in linguaggio naturale.

Tale assunzione iniziale, tuttavia, non risulta limitativa poiché in numerosissimi casi è possibile ricondurre le numerose tipologie di rappresentazione proprio a documenti testuali più o meno complessi. Ad esempio, nel caso dell'audio, è possibile (a patto di rinunciare al patrimonio informativo relativo alle inflessioni tonali e ad altri parametri) trascrivere il parlato producendo (anche grazie a programmi di scrittura assistita automatica) proprio dei testi.

Inoltre, la maggior parte degli studi effettuati in passato si è concentrata sull'affrontare direttamente la stima del sentiment di un documento scritto in linguaggio naturale, mentre pochi altri lavori si sono rivolti anche alla questione del modello costruito attorno all'argomento (e allo specifico dominio nel quale si colloca) di cui si vuole studiare la reputazione sul Web.

A ogni dominio corrisponde un certo vocabolario, e più il dominio è ristretto, più il linguaggio che lo caratterizza, e i relativi indicatori di mood (e quindi di orientamento del sentiment), sarà specifico. Tanto meglio si è in grado di rappresentare il dominio, tanto meglio si può affrontare la determinazione del sentiment.

E questo dell'identificazione a priori del dominio che accomunasse i testi oggetto di una particolare indagine ha rappresentato un elemento di criticità riscontrato in tutta la letteratura studiata. Un valore

aggiunto che, quindi, con la ricerca avviata, si voleva raggiungere era proprio la possibilità di estrapolare in maniera rapida e con buoni livelli di performances una rappresentazione del dominio che avrebbe semplificato i passi successivi per l'analisi del sentiment.

Ciò proprio perché, elemento comunemente accettato in letteratura, un sistema di sentiment analysis è fortemente dipendente dal dominio e dal suo vocabolario.

Il problema del riconoscimento del "sentiment" di un documento richiede lo sviluppo di metodologie e tecniche che permettano di:

1. analizzare il testo attraverso metodologie proprie della linguistica computazionale,
2. costruire il dominio a cui il testo appartiene,

e, infine,

3. inferire automaticamente azioni a valle dell'analisi del testo.

Dei molteplici approcci alla Sentiment Analysis analizzati una metodologia particolarmente promettente e alla quale è stata indirizzata l'attenzione in queste pagine, prevede l'utilizzo di tecniche per la costruzione di domini in grado di rappresentare il contesto in cui il documento si colloca sia per tipologia di argomenti che per "sentiment" espresso.

Nello specifico, la metodologia sviluppata si basa sull'utilizzo, in particolare, della Latent Dirichlet Allocation per rappresentare un documento attraverso una rappresentazione grafica (si sono introdotti i mixed Graph of Terms) contenente i principali termini presenti nel testo, le loro relazioni e il suo "mood".

Inoltre, allo scopo di rendere più appetibile il modello, migliorandone le prestazioni, si è provveduto a completarlo attraverso il supporto di repository di conoscenza strutturata (quali Wordnet e SentiWordnet). In particolare tale apporto consente di arricchire l'analisi con un livello di indagine (e di valorizzazione dei risultati) di tipo semantico.

Non si è fatto solo riferimento alla presenza o meno di particolari termini all'interno di una stringa, ma anche allo specifico peso che

ciascuno di questi termini ha all'interno del discorso. Ed è proprio questo ulteriore livello semantico, oltre alla possibilità di operare sui commenti oggetto di indagine in tempo reale con prestazioni elevate, il significativo valore aggiunto offerto dalla ricerca prodotta.

Per completezza descrittiva e per garantire alle successive fasi sperimentali degli idonei parametri di confronto delle prestazioni del modello proposto, sono comunque state affrontate le principali metodologie adottate sia in ambito puramente probabilistico, sia in ambito semantico oltre alle tecniche presenti in letteratura che consentono di ottenere un approccio misto.

Nella seconda parte della tesi si è quindi sviluppata la specifica metodologia che a partire da testi, precedentemente etichettati in base al loro "sentiment", interagisce con le strutture, denominate "mixed graph of terms", in grado di rappresentare termini, e i loro relativi legami, tipici di un certo mood.

La costruzione degli mGT avviene con metodologie di tipo statistico-probabilistiche quale quella denominata LDA (Latent Dirichlet Allocation) che consente di scomporre i documenti e le pagine web in più cluster, suddividendoli per keywords o argomenti simili.

Si riescono, in questo modo, a classificare documenti in modo completamente automatico sulla base del sentiment, senza bisogno di far riferimento ad altri parametri quali, ad esempio, la posizione dei termini all'interno dei nei testi. Solo successivamente, come anticipato, si è aggiunta anche una componente semantica andando a valorizzare i termini aggreganti (quelli cioè che ci hanno permesso di caratterizzare il dominio) mediante SentiWordNet 3.0. Ciò ha consentito di perfezionare un'apposita funzione di score che, tenendo conto anche di questo aspetto, consentisse di estrarre in maniera efficiente. oltre che efficace, il sentiment ricercato.

Dopo l'introduzione, l'analisi dello stato dell'arte e la presentazione del modello in tutte le sue fasi, nella quarta parte sono stati riportati i dettagli delle differenti attività sperimentali.

Per le sperimentazioni si è fatto riferimento a datasets standard ampiamente adottati dagli autori precedenti. In questo modo è stato

possibile estrapolare i parametri che ci hanno consentito di confrontare le prestazioni del modello proposto con gli altri.

In una prima fase sono stati utilizzati solo dei sottoinsiemi di un dataset comune, al solo scopo di validare le funzionalità generali del modello e i risultati quantitativi e non quelli navigativi.

Nella seconda fase sulla scia primi risultati sono state completate altre campagni sperimentali i cui risoltati sono tutti riportati nelle pagine precedenti.

Riepilogando, è stata progettata una metodologia per la Sentiment Analisys basata su LDA in grado di:

- Contestualizzare i commenti in un preciso dominio di conoscenza attraverso la costruzione di mGT a partire dall'analisi dei testi
- Rilevare in maniera rapida ed efficace i termini in grado di caratterizzare l'orientamento da parte degli utenti
- Integrarsi con strutture di conoscenza presenti in letteratura.

Il metodo quindi, se da un lato consente di far emergere rapidamente il vettore rappresentativo di un dato dominio, così come già possibile con altri metodi presenti in letteratura, dall'altro è in grado di evidenziare quei termini che, all'interno di uno specifico dominio, sono indicativi di un sentiment positivo e di uno negativo. Uno stesso termine, infatti, può caratterizzare un orientamento positivo all'interno di un dominio e un orientamento neutro o negativo all'interno di un altro.

Il modello proposto, poi, si caratterizza per i seguenti punti di forza. In primo luogo riduce sensibilmente i tempi necessari all'estrazione dei vettori rappresentativi di un dominio poiché non va a riferirsi a tutti i termini presenti in un testo, ma, grazie all'utilizzo di LDA, fa riferimento ai soli termini cosiddetti aggregatori, con notevole risparmio di tempo e risorse computazionali.

Ancora, le performance ottenute possono essere incrementate facilmente andando ad arricchire i vettori caratterizzanti con ulteriori termini via via acquisiti nell'ambito di successive sperimentazioni.

Infine, come evidenziato nelle tabelle di riepilogo di ciascuna sperimentazione, i tempi di esecuzione dei singoli script realizzati sono estremamente ridotti.

Tutta la sperimentazione condotta ha consentito di affinare un ambiente web (e quindi fruibile su qualsiasi tipo di dispositivo connesso alla rete) in grado di sfruttare le potenzialità del modello in tempo reale.

Direttamente dall'interfaccia è possibile analizzare un set di commenti e, per lo specifico dominio al quale si riferiscono, far generare i vettori rappresentativi sia del dominio che degli orientamenti positivi e negativi.

Gli ambiti di sviluppo ulteriori sono molteplici e, in particolare, si può evidenziare la possibilità di arricchire ulteriormente il modello costruendo una libreria di vettori rappresentativi grazie ai quali sia possibile, dato un generico commento testuale, in primo luogo e in maniera totalmente automatica, andare a identificare lo specifico dominio al quale esso si riferisce.

In questo modo, ad esempio, potrebbe essere possibile capire, dato un determinato commento, se lo stesso è relativo a un parere politico piuttosto che alla recensione di un film o un prodotto.

Naturalmente, subito dopo l'identificazione del dominio, sarà possibile anche valutare l'eventuale sentiment del commento specifico.

Ulteriore ambito di sviluppo interessante sarebbe quello di completare il modello con ulteriori moduli corrispondenti a più complessi livelli semantici rappresentativi di molteplici livelli emozionali.

In questo modo, particolarmente per le informazioni estratte da un video con parlato, si avrebbe la possibilità di estrarre diversi livelli di gradimento tipici, ad esempio delle espressioni facciali.

E le prospettive più interessanti in questo ambito, come già si è accennato nel corso del secondo capitolo, sono quelli di una possibile applicazione nell'ambito della diagnosi basata su riprese televisive (non solo di patologie, ma anche di risposte a terapie farmacologiche o stimoli esterni).

Infine, l'aver efficacemente prodotto l'intero ambiente sperimentale su piattaforma web based suggerisce l'impiego di questa metodologia per lo sviluppo di applicazioni smart ideate per il mondo mobile e in particolare per ambienti Android, iOS e Windows Mobile, magari finalizzate all'analisi del sentiment direttamente in tempo reale.

Indice delle figure